Scientific Reports

https://doi.org/10.1038/s41598-025-34756-z

Article in Press

A multiscale transformer with spatial attention for hyperspectral image classification

Received: 4 June 2025

Accepted: 31 December 2025

Published online: 13 January 2026

Cite this article as: Ahmad I., Farooque G., Hadi F. *et al.* A multiscale transformer with spatial attention for hyperspectral image classification. *Sci Rep* (2026). https://doi.org/10.1038/s41598-025-34756-z

Irfan Ahmad, Ghulam Farooque, Fazal Hadi, Abdolraheem Khader, Sara Abdelwahab Ghorashi, Ali Ahmed & Eatedal Alabdulkreem

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

A Multiscale Transformer with Spatial Attention for Hyperspectral Image Classification

Irfan Ahmad[1], Ghulam Farooque[2], Fazal Hadi[3],
Abdolraheem Khader[1*], Sara Abdelwahab Ghorashi[4],
Ali Ahmed[5], Eatedal Alabdulkreem[4]

[1*]School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, 210094, Jiangsu, China.
[2]Department of Computer Science & IT, University of Lahore, Lahore, Punjab, Pakistan.
[3]Taizhou Key Laboratory of Minimally Invasive Interventional Therapy Artificial Intelligence, Taizhou Campus of Zhejiang Cancer Hospital, Taizhou, 210094, Zhejiang, China.
[4]Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia.
[5]Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, 21589, Saudi Arabia.

*Corresponding author(s). E-mail(s): abdolraheem@njust.edu.cn;
Contributing authors: irfan_yummy@njust.edu.cn;
ghulam.farooque@cs.uol.edu.pk; fazalhadi@zjcc.org.cn;
saabdelghani@pnu.edu.sa; aabdelrahim@kau.edu.sa;
eaalabdulkareem@pnu.edu.sa;

Abstract

Hyperspectral images (HSIs) are renowned for their rich spatial and spectral information, which is crucial for accurate classification. The acquisition of discriminative spectral-spatial features plays a pivotal role in determining classification results. While convolutional neural networks (CNNs) have demonstrated remarkable performance in HSI classification, increasing network depth can lead to performance degradation. Furthermore, their fixed scale and limited receptive field restrict the ability to capture long-range dependencies, hindering effective feature learning and, consequently, affecting the generalization capability of the framework. This paper presents a novel HSIs classification framework, MTSA-Net, which integrates a multiscale transformer with a spatial attention mechanism, resulting in a more robust, flexible, and high-performing approach. Initially, the proposed framework utilizes 3-D and 2-D convolution layers, followed by spatial attention to prioritize and focus on the most critical spatial features. These enhanced features are then passed through multiscale transformer encoders to capture local and global representations, effectively modeling long-range dependencies. Finally, a feature fusion module combines features extracted at varying scales, leading to a more robust and comprehensive feature representation for final classification. Extensive experiments on five widely used benchmark HSIs datasets demonstrate that the proposed MTSA-Net method outperforms state-of-the-art approaches, particularly with limited training samples. The overall accuracies of 98.84%, 98.77%, 99.80%, 97.84%, and 95.87% are achieved on the Indian Pines, Pavia University, Salinas Valley, Houston-13, and Houston-18 datasets, respectively. The source code for this work will be accessible at https://github.com/irfan01000 for reproducibility.

Keywords: Hyperspectral image classification, convolutional neural networks (CNNs), spatial attention, multiscale transformers, spectral-spatial features.

1 Introduction

Recent advancements in hyperspectral imaging technology have enabled the feasible capture and analysis of the light spectrum reflected or emitted by objects and scenes. This technique entails the acquisition of numerous closely spaced wavelength bands, encompassing a wide electromagnetic radiation spectrum from ultraviolet to long-wave infrared [1]. The resultant information, termed a hyperspectral image (HSI), holds the potential for identifying and quantifying materials within a scene and detecting nuanced alterations in the environment. Hyperspectral images are represented as three-dimensional (3D) data, combining one-dimensional spectral features with two-dimensional (2D) spatial information [2]. These spectral attributes signify narrow spectral bands replete with informative content about land cover [3], while the spatial component adeptly captures the arrangement of land covers. In the realm of HSI classification, both spectral signatures and spatial data emerge as dominant features, furnishing valuable cues that contribute to precise outcomes [4]. Hyperspectral imaging finds diverse applications, including mineral exploration [5], agriculture [6], military purposes [7], urban planning [8], environmental monitoring [9] and forestry management [10]. In order to harness the full potential of hyperspectral imaging data, researchers have investigated various data processing methods, including denoising [11], unmixing [12, 13], image fusion [14], target detection [15], and classification [16, 17]. Among these methods, the classification of land-cover information has garnered considerable interest. In hyperspectral imaging, the abundance of spectral information at hand has prompted researchers to delve into a spectrum of conventional machine-learning techniques to harness its potential for classification purposes. These methods encompass a diverse range of approaches, each tailored to extract valuable insights from the multi-dimensional HSI data. Among these approaches, the k-nearest neighbour [18] technique leverages proximity-based relationships within the data space to make accurate predictions. Support vector machine (SVM) [19] stands out for its ability to construct effective decision boundaries by mapping the data into higher-dimensional spaces. Logistic regression [20] seeks to model the probabilities of class memberships, while extreme learning machine [21] employs a single hidden layer neural network to discern intricate patterns within the data. Meanwhile, random forest [22] capitalises on an ensemble of decision trees to enhance the predictive accuracy. Principal component analysis (PCA) [23] excels in dimensionality reduction, offering a compressed representation of the original data by capturing the most informative spectral signatures. Decision trees [24] guide classification based on a hierarchical structure of conditions, enabling a comprehensive exploration of data space.

Recently, there has been a surge in the popularity of deep learning approaches due to their ability to autonomously acquire adaptable and resilient features from training data, surpassing conventional techniques [25, 26]. This impressive achievement has been evident in diverse research fields, such as natural language processing [27] and computer vision [28], where deep learning methods have showcased remarkable achievements. Numerous deep learning techniques, which effectively capture both spectral and spatial details, have been utilized for HSI classification tasks [29]. In [30], a Convolutional Neural Network (CNN) was harnessed to extract spatial characteristics. These spatial features were then combined with spectral features acquired through balanced local discriminant embedding, forming a comprehensive framework for HSI classification. This approach highlights the effectiveness of CNNs in capturing essential information from both spectral and spatial domains, thereby enhancing the accuracy of HSI classification. In [31], an innovative feature extraction technique based on CNNs was introduced. This method effectively learned discriminative representations from pairs of pixels and incorporated a voting mechanism to enhance the smoothness of final classification maps.

To strengthen the acquisition of spectral-spatial features, several CNN-based approaches have been introduced, encompassing 1D, 2D, and 3D CNNs [32]. In [33], the author presented a CNN-based approach for HSI classification. This approach involved extracting spatial features using a 2D-CNN technique, which leveraged the initial principal component channels of the original HSI. Utilizing 2D-CNN for HSI analysis offers notable benefits, including the capability to efficiently extract features from the raw input images. This method has confirmed its promising performance across diverse fields such as image processing and computer vision, including tasks like object detection and image classification. In this context, [34] introduced an advanced contextual CNN method for the prediction of pixel labels. This method harnessed localized spectral-spatial features, with spectral and spatial attributes being captured from multi-scale filters through the application of a 2D CNN. These extracted features were then integrated to generate a unified feature map. Nevertheless, the use of 2D CNN-based approaches faced challenges in effectively leveraging both spectral and spatial data simultaneously. This led to the potential loss of information during the process of feature learning [35]. As an alternative, another technique [36] proposed a novel hybrid architecture comprising layers of both 1D and 2D CNN. This architecture was specifically designed to separately acquire spectral and spatial information. Through this innovative approach, the aim was to prevail over the limitations associated with the concurrent utilization of spectral and spatial attributes.

To overcome the constraints posed by 2D CNN-based networks, the adoption of 3D CNN approaches was embraced to directly obtain spatial-spectral features from HSI. In this regard, investigations carried out in [37–39] proposed techniques based on 3D CNNs for HSI classification. These methodologies aimed to extract comprehensive spectral and spatial features from HSI data, yielding notable enhancements in classification performance. The significance of the spatial properties of ground objects lies in their capacity to offer insights into the structure and contextual position. Objects close to each other often share the same class. The integration of spatial features in the classification process enhances the capacity to capture these associations, ultimately leading to improved classification accuracy. Furthermore, the potential of 3D CNNs was utilized in [40], where they directly extracted deep spectral–spatial features from raw hyperspectral images and exhibited prominent results. Similarly, in [41], an extended investigation was conducted on 3D CNNs for spectral–spatial classification, utilizing input cubes with reduced spatial dimensions from HSIs. These models were designed to generate thematic maps by directly processing raw HSIs. Despite the achievements of CNN-based methods in extracting spatial and spectral information, they possess specific constraints. One of these constraints relates to their challenge in extracting sequential features, particularly middle and long-range spectral correlations. In addition, their ability to extract local attributes is hampered by the fixed dimensions of the receptive field, which may not be sufficient to comprehensively capture complex details and localized fluctuations within the data.

In recent years, the Transformer architecture has quickly emerged as a formidable foundation for image-related tasks in the domain of computer vision, primarily due to its exceptional capabilities in modeling and processing visual data [42–44]. In [45], the author presented the diverse applications of Vision Transformers in medical computer vision. These applications span a wide range, including disease classification from medical images, segmentation of anatomical structures, image registration, detection of lesions in specific regions, image captioning, report generation, and image reconstruction. These versatile applications significantly contribute to medical diagnosis and enhance the overall treatment process. Furthermore, in [46], the author proposed an innovative backbone network called SpectralFormer to effectively capture local spectral sequence information from neighbouring bands of HSIs, resulting in group-wise spectral embeddings. To avoid the loss of vital information during the propagation of data across layers, they proposed a cross-layer skip connection. This connection adaptively merges memory-like components from surface-level to comprehensive layers, learning to combine subtle residuals across the layers. A hybrid CNN and vision transformer approach for anomaly detection is suggested [47]. This approach combined spatial and temporal information in two steps: an efficient CNN extracted spatial features, which were then processed by a transformer-based model to capture long-range temporal relationships among complicated events. The model incorporated temporal self-attention to effectively learn spatial-temporal features and identify anomalies. To utilize spectral and spatial features a novel approach was presented in [48] named Spectral-Spatial Feature Tokenization Transformer, comprising of 3D and 2D convolution layers, to obtain low-level features from the data. A Gaussian-weighted feature tokenizer was deployed to transform these extracted features and fed them into a transformer encoder module for further feature representation and learning. Finally, the linear layer was used for classification. A similar approach, morphological transformer (morphFormer) [49], has been adopted to incorporate a trainable spectral and spatial morphological network, employing spectral and spatial morphological convolution operations along with the attention mechanism. In [50], a novel framework was proposed for HSI classification. Initially, HSIs were transformed into sequences. Simultaneously, spatial information was incorporated by adding a learned positional embedding. Subsequently, a conventional transformer encoder was utilized to acquire feature representations. Finally, these multilevel features were processed by decoders to produce classification results. Another similar approach was presented in [51], where the author introduced two branches for extracting pixel-wise multiscale features. Subsequently, a multiscale Spectral Embedding Module was developed to boost the portrayal of local details among adjacent spectral bands. Furthermore, leveraging the cross-attention operation, a single token in each branch acts as a query, facilitating the swapping of information with other modules. In [52], the author proposed a Multi-Attention and Transformer Network (MATNet) that captured spatial-spectral features by utilizing spatial attention (SA) and channel attention (CA), followed by tokenizer and transformer module to perform deep semantic feature extraction. Finally, the Lpoly loss function is employed.

The current literature indicates that CNNs are proficient at capturing local spectral-spatial information but encounter difficulties when dealing with comprehensive spectral-spatial features. In contrast, transformers have demonstrated exceptional proficiency in understanding complex relationships among long-range features. Consequently, in HSI classification the integration of these two architectural approaches has the capability to elevate spectral-spatial feature learning by effectively addressing both local and global relationships. On the other hand, multiscale HSI classification models have made significant progress, but several issues persist. Conventional CNN-based multiscale approaches, while adept at extracting contextual information across various scales, are constrained by fixed receptive fields that hinder their ability to capture long-range spectral relationships. Additionally, these methods often rely on deeper networks to expand the receptive field, which can result in overfitting, loss of detailed information, and higher computational demands.

To tackle these challenges, we propose MTSA-Net, a novel multiscale transformer framework with spatial attention for HSI classification. The model begins with a 3D convolution layer to extract shallow spectral–spatial features, followed by a 2D convolution layer and a spatial attention module. This design reduces feature redundancy, mitigates inaccuracies that often arise in deeper networks, and emphasizes the most discriminative spatial features, thereby alleviating the limited spatial resolution of HSIs. The refined feature vectors are then processed by multiple parallel transformer encoder branches with varying hidden dimensions, enabling simultaneous modeling of fine-grained local details, intermediate relationships, and global representations. Finally, a multiscale feature fusion module integrates the outputs from different branches to balance feature representation across scales and enhance robustness. By combining spatial attention with hidden-dimension diversity, MTSA-Net addresses the shortcomings of existing multiscale models and achieves superior classification performance. The primary contributions of this paper can be outlined as follows:

1. This study introduces a straightforward CNN architecture augmented with a spatial attention mechanism to extract spectral and spatial features by focusing on crucial areas and eliminating redundant information. The spatial attention module generates a spatial attention map by exploiting the spatial interconnections among features.
2. A multiscale transformer encoder is proposed to capture local and global representations, effectively modeling long-range dependencies. This is followed by the feature fusion module that enriches the feature representation by leveraging multiscale information and effectively mitigating the issue of imbalanced feature representation.
3. The generalization capability and effectiveness of the proposed MTSA-Net model have been validated through extensive experiments conducted on five benchmark HSI datasets, demonstrating superior performance compared to state-of-the-art approaches.

The remainder of this work is summarized as follows. Related works to HSI are reviewed in Section 2. The section 3 elaborates on the proposed methodology. Section 4 illustrates the HSI datasets, experimental settings, and offers an in-depth analysis of classification results and ablation studies. Finally, conclusions are provided in Section 5.

2 Related Work

2.1 CNN-based HSI classification frameworks

The task of HSI classification involves assigning a land-cover label to each individual pixel, which has garnered substantial interest in recent times [53]. CNN has gained widespread popularity for its ability to extract comprehensive spectral-spatial features while retaining important spatial structure information. Initially, in [54], the author proposed CNN in HSI classification, using CNN layers to extract spectral features. To learn spectral-spatial features, numerous variations of CNNs, including 1D-CNN, 2D-CNN, and 3D-CNN, have been developed [32]. In [33], a 2D CNN is presented for HSI classification, providing significant advantages, such as the rapid extraction of features from the original input images. This methodology entails the extraction of spatial characteristics. Based on 2D CNN, a spectral-spatial feature extraction architecture was proposed in [30]. However, these models faced limitations in adequately capturing spectral features due to their relatively simple architectural design.

To enhance the exploitation of the spatial and spectral interaction in HSI classification, 3D CNNs were employed. However, 3D CNNs require greater computational resources compared to their 2D counterparts; their capability to learn spatial-spectral features preserves the inherent relationship between spatial and spectral information without degradation. In [40], a 3D CNN was employed to directly and efficiently learn spectral-spatial features from the original HSI, showcasing promising outcomes in terms of classification performance. Similarly, a 3D CNN model has been introduced that leverages both spatial and spectral features to enhance the performance of HSI classification [55]. The HSI cube is initially subdivided into slightly overlapping 3D patches. These patches are then subjected to processing, resulting in the generation of 3D feature maps. A novel approach [4], referred to as HybridSN, has been introduced, which combines both 3D and 2D CNNs. HybridSN employs larger spatial dimensions while working with smaller spectral bands. It employs three 3D convolution layers initially to collect spatial-spectral attributes, followed by immediate enhancement using 2D convolution to focus on spatial features. Impressively, the integration of hybrid CNNs leads to a model with reduced complexity when compared to using 3D CNN as the sole component. Furthermore, [56] presented an approach to HSI classification utilizing a multiscale self-looping CNN. These networks incorporate self-looping blocks, where each layer serves as both the input and output for every other layer, effectively establishing a looping structure within the network. The network's loopy connections, which maximizes information flow, result in extracting high-level features. For instance, in [57], the contextual feature was extracted from the HSI at different scales using multiscale convolution. Octave 3D CNN was then utilized to minimize spatial redundancy and expand the receptive field. To investigate and enhance the discerning features, the approach included the utilization of a channel attention module and a spatial attention module. These modules were incorporated to improve the feature maps and ultimately enhance the classification results. Similarly, a composite neighbour-aware convolutional metric network (CNCMN) was proposed in [58], which intends to learn the representation of each target batch-wise from its composite neighbours (Euclidean and non-Euclidean neighbours). A composite convolution (CoConv) combines traditional image convolution with graph convolution to perform versatile operations on these composite neighbors, allowing for the extraction of adaptively fused features from them. CNN-based techniques have demonstrated their capacity to efficiently extract both spatial and contextual information from HSI. However, CNNs have limitations, including fixed receptive fields, loss of fine-grained details through downsampling, high data requirements, computational demands, and limited contextual understanding. In addition, CNNs face challenges in capturing long-term dependencies and handling sequential attributes. Conversely, transformer architecture based on self-attention mechanisms grasps intricate dependencies in sequential data and has achieved great success in the field of NLP [59].

2.2 Attention-based methods

The attention mechanism, initially motivated by the human visual system's ability to discern salient regions within images to facilitate classification, has gained significant interest in the realm of remote sensing [60]. Numerous attention-based mechanisms have demonstrated significant effectiveness in HSI classification. In [61], a dual attention mechanism was proposed in a two-stage process. In the initial stage, it learns features from the overall region and condenses them into a compact set using second-order attention pooling. In the subsequent stage, it intelligently chooses and disperses

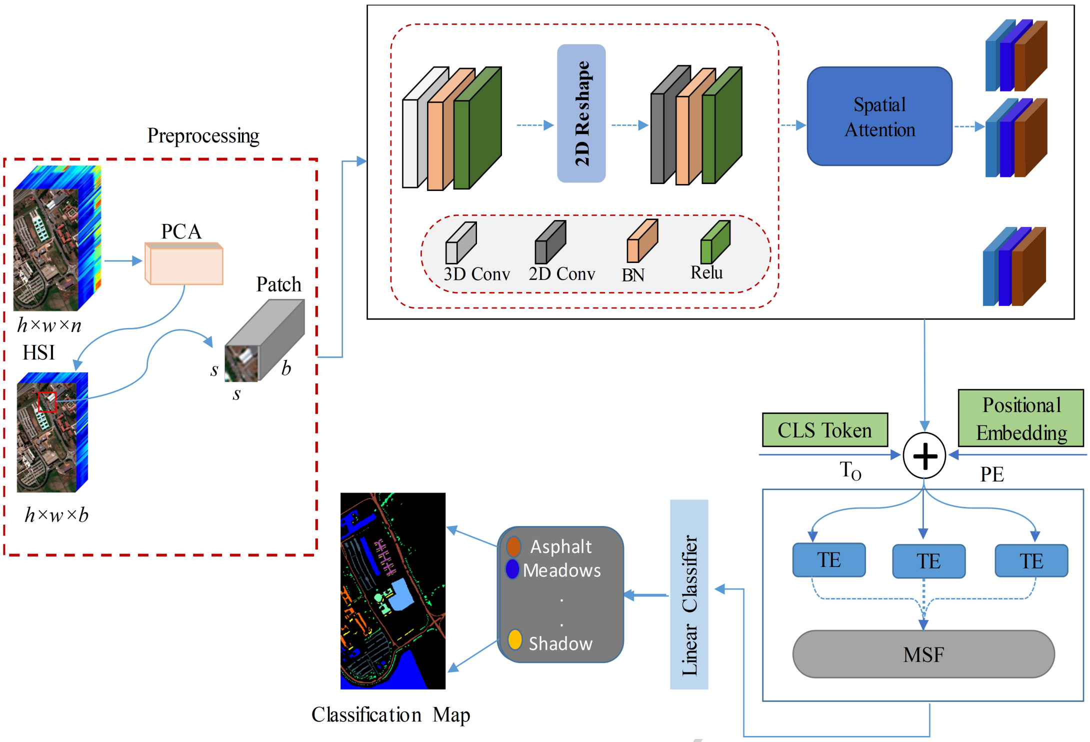


Fig. 1: Illustration of the proposed MTSA-Net framework. The upper left part presents preprocessing. The uppermost centre part shows the spectral-spatial feature extraction with spatial attention. The bottom row shows the multiscale transformer at the right and the classification map at the bottom left

.

features to each specific location through another attention mechanism. In addition, the authors in [62] sequentially calculate attention maps along with two different dimensions: channels and spatial. These attention maps are subsequently multiplied by the input feature map to achieve adaptive feature refinement. Similarly, an efficient channel attention (ECA) model was suggested in [63], demonstrating that maintaining suitable cross-channel interaction can both enhance performance and substantially reduce model complexity. In [64], a gating mechanism is used to dynamically recalibrate spectral bands. It accomplishes this by selectively amplifying informative bands while downplaying less important ones.

Recently, the self-attention mechanism has emerged as a pivotal tool for modelling extensive, long-range dependencies in various machine learning and natural language processing domains. Unlike CNN models with rigid, local receptive fields, self-attention empowers each element or token within a sequence to dynamically weigh its relevance to all other elements in the sequence. This dynamic weighting enables the model to effectively capture intricate long-distance relationships, transcending the limitations of fixed positional dependencies. For instance, modern state-of-the-art vision transformer framework [65], employs the self-attention mechanism to comprehensively model sequences and adeptly learn extensive long-range dependencies.

Furthermore, a hybrid approach incorporating both CNN and Transformer networks [48] was proposed to effectively calculate spectral-spatial and semantic features for HSI classification. In [66], Swin Transformer was presented, which calculates representations through a shifted window strategy. This approach allows for efficient processing of visual data with diverse scales and higher resolutions. Adopting this strategy effectively addresses the challenges presented by scale and resolution disparities, rendering it well-suited for HSI classification.

3 Proposed Methodology

The main components encompass a 3D convolution layer, a 2D convolution layer and spatial attention (SA). In addition, a multiscale transformer encoder is introduced, incorporating a feature fusion module followed by a softmax function. The proposed structure is presented in Figure 1, and its detailed explanation can be found in the following sections.

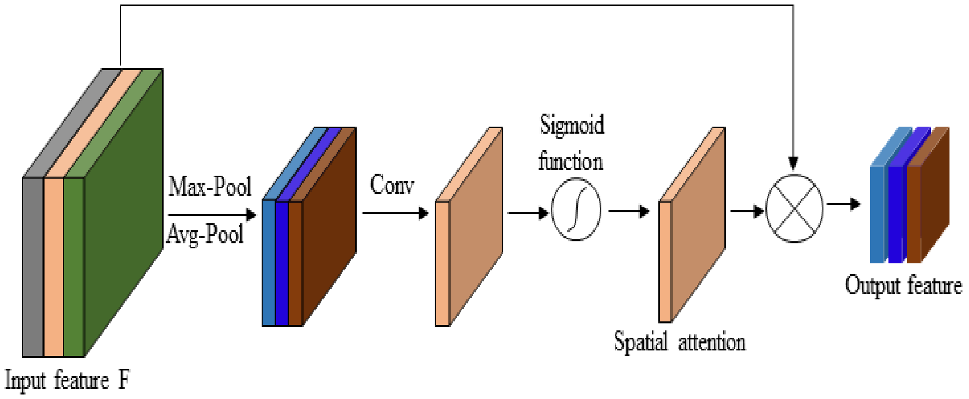


Fig. 2: The structure of the spatial attention mechanism

3.1 Preprocessing

The original HSI data $\mathbf{I} \in \mathbb{R}^{h\times w\times n}$, where $h \times w$ denotes the spatial dimension and n represents useful spectral information. The HSI n bands contain valuable spectral data but also introduce substantial computational demands due to their high dimensionality. To address this, we apply Principal Component Analysis (PCA) to the HSI data $\mathbf{I}$. This operation decreases the spectral dimension from n to b while preserving the spatial dimension. By doing so, we mitigate spectral band redundancy and alleviate the computational load. After PCA, the transformed hyperspectral data is denoted as $\mathbf{P} \in \mathbb{R}^{h\times w\times b}$, with b representing reduced spectral dimension. The size of each patch is $\mathbf{Q} \in \mathbb{R}^{s\times s\times b}$, where $s \times s$ shows the patch size and the label assigned to each patch corresponds to the label of its center pixel. While extracting the patch around a unique pixel, the boundary pixels are not accessible. To address this issue, a padding operation of $\frac{s-1}{2}$ is applied to these pixels. After removing unlabeled pixels, the remaining data are partitioned into training sets and test sets. Each 3D patch of size $s \times s \times b$ is input to the 3D convolution layer for the extraction of spectral-spatial features.

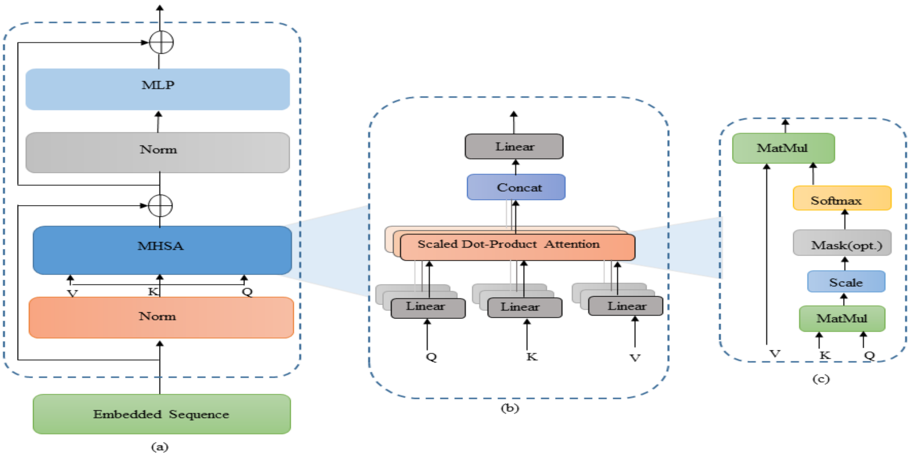


Fig. 3: The structure of (a) Transformer Encoder, (b) Multihead self-attention and (c) self-attention.

3.2 CNN for Feature Learning

CNNs demonstrate superior achievements in HSI classification owing to their capability to autonomously extract contextual features. CNNs have been previously validated for their effectiveness in obtaining high-level features, irrespective of the data source modality. We have developed a straightforward CNN-based feature extractor with the specific aim of efficiently capturing local semantic details from HSI. In the proposed network architecture, after each convolutional layer, there is a sequence of operations that includes batch normalization (BN) and rectified linear unit (ReLU) activation.

Our proposed MTSA-Net leverages the sequential layer of Conv3D and Conv2D to extract resilient and distinctive features from HSIs. In the 3D convolutional layer, the computed output value at a given spatial location (x, y, z) for the jth feature map within the ith layer is depicted as follows:

$$\boldsymbol{v}_{i,j}^{(x,y,z)} = \Phi\left(\sum_{k}\sum_{l=0}^{L_i-1}\sum_{m=0}^{M_i-1}\sum_{n=0}^{N_i-1}\boldsymbol{w}_{i,j,k}^{(l',m',n')}v_{i-1,k}^{(x+l',y+m',z+n')}+b_{i,j}\right) \tag{1}$$

where k is labeled as a feature map in the $(i-1)$th layer. The variables L_i, M_i, and N_i denote the height, width, and channel of a 3-D convolution kernel. $\boldsymbol{w}_{i,j,k}^{l',m',n'}$ is the parameter weight of (l', m', n') and $b_{i,j}$ is the bias.

Similarly, for the 2D convolution layer, its mathematical representation can be stated as follows

$$\boldsymbol{v}_{i,j}^{(x,y)} = \Phi\left(\sum_{k}\sum_{l=0}^{L_i-1}\sum_{m=0}^{M_i-1}\boldsymbol{w}_{i,j,k}^{(l',m')}v_{i-1,k}^{(x+l',y+m')}+b_{i,j}\right) \tag{2}$$

The initial HSI data is organized into subcubes, each having dimensions of $(13 \times 13 \times b)$. These subcubes are then transformed into a format of $(1 \times 13 \times 13 \times b)$ and employed as input for a Conv3D layer with a kernel size of $(3 \times 3 \times 3)$. Following the application of 3D convolution, the resulting output possesses dimensions of $(8 \times 11 \times 11 \times (b-2))$, where 8 represents the number of channels generated by the convolution operation and $(b-2)$ is the spectral bands. The output shape resulting from the 3D convolution undergoes rearrangement and is then fed into the 2D convolution layer, resulting in dimensions of $(64 \times 9 \times 9)$.

3.3 Spatial Attention

Recently, attention mechanisms have been extensively used in HSI classification. The attention mechanism pertains to the selective emphasis on certain information while ignoring irrelevant information. In HSI classification, it is consistently observed that spatial information holds greater importance than spectral information [67].

Similarly, in [68] only spatial attention was utilized to focus on spatial features and maximize the diversity of features. HSI encompasses rich spectral bands that are readily captured. However, they often suffer from limited spatial resolution. Consequently, our proposed model utilizes spatial attention mechanisms to enhance distinctive features, addressing this inherent spatial limitation. The structure of the spatial attention mechanism is visualized in Figure 2.

Initially, this stage processes the input features through individual average pooling and maximum pooling operations. Subsequently, it combines the resulting feature sets and ultimately generates spatial attention feature maps through a convolutional layer. The spatial attention is expressed as follows:

$$\mathbf{SA(F)} = \sigma\left(f^{7x7}\left([\text{Avg-Pool}(\mathbf{F}); \text{Max-Pool}(\mathbf{F})]\right)\right) \tag{3}$$

$$\mathbf{SA(F)} = \sigma\left(f^{7x7}\left([\mathbf{F}_{avg};\mathbf{F}_{max}]\right)\right) * \mathbf{F} \tag{4}$$

where σ depicts the sigmoid function, $[F_{avg};F_{max}]$ concatenates two feature maps and f^{7x7} denotes the convolution operation. * represents the element-wise product.

3.4 Transformer Encoder (TE)

The structure of the transformer encoder is depicted in Figure 3. The transformer encoder is employed to acquire global information and comprises two normalization layers (LN), a multihead self-attention module (MSA), and a multilayer perceptron (MLP). Residual skip connections are incorporated prior to the MSA block and the MLP layer. Layer Normalization is incorporated to address the challenge of gradient vanishing and to strengthen the model's capacity for feature representation. The effectiveness of the transformer architecture stems primarily from its core MSA block. Within this block, the utilization of a self-attention mechanism, as depicted in Figure 3(c), adeptly captures the interrelationships among feature sequences. The matrices Q (query), K (key), and V (value) are employed during the computation procedure. The attention score is obtained by taking the dot product between Q and K, and the weight of this score is determined by applying the softmax function.

In summary, self-attention is expressed as follows:

$$\mathbf{SA} = \text{Atten}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}\left(\frac{\mathbf{QK}^T}{\sqrt{d_K}}\right)\mathbf{V} \tag{5}$$

where d_k is the dimension of $\mathbf{K}$.

Furthermore, MSA concatenates the output obtained from multiple SAs, as shown in Figure 3 (b).

$$\text{MHSA}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{Concat}(\mathbf{SA}_1,\mathbf{SA}_2,\ldots,\mathbf{SA}_\text{n})\mathbf{W} \tag{6}$$

where n denotes the head number and $\mathbf{W}$ represents the learned parameter.

Subsequently, the weight matrix acquired from the preceding step is fed into the MLP layer, which comprises of two fully connected layers. A nonlinear activation function called Gaussian Error Linear Unit (GELU) is placed between these two layers. GELU introduces non-linearity into the model expressed as follows,

$$\text{GELU}(x) = x\Phi x = \frac{1}{2}x\left(1+\tanh\left(\sqrt{\frac{2}{\pi}}\left(x+0.044715\cdot x^3\right)\right)\right) \tag{7}$$

where $\Phi(x)$ is the cumulative distribution of a Gaussian distribution.

The proposed multiscale representation learning is realized through multiple parallel transformer encoder branches, each processing the same tokenized input but with distinct MLP hidden dimensions in their feed-forward networks. This design enables each branch to emphasize features at different representational scales: smaller hidden dimensions capture fine-grained local patterns, while larger hidden dimensions model broader contextual dependencies. Unlike conventional multiscale strategies that rely on varying patch sizes or spatial resolutions, our approach achieves multiscale diversity solely through hidden-dimension variation, thereby avoiding additional preprocessing or resolution changes. In the proposed design, we employ three parallel branches with hidden dimensions of 128, 256, and 512, respectively. These values were selected based on both design intuition and empirical validation: the smaller hidden size (128) emphasizes fine-grained local cues, the intermediate size (256) balances local and global features, and the larger size (512) captures rich contextual dependencies. This hierarchical setup ensures complementary feature extraction across scales, leading to improved classification accuracy and generalization.

3.5 Multiscale Feature Fusion

In the final stage, multiple feature matrices are produced through transformer encoders operating at different scales. Multiscale feature fusion is achieved using multiple parallel Transformer branches, each configured with distinct MLP hidden dimensions while sharing the same input resolution and tokenization. Each branch captures features at a different representational scale, enabling the model to extract both local and global contextual information. Feature matrices obtained from different scales enhance the proposed model's capability to learn diverse patterns and details in the data, yielding improved performance and generalization. These feature matrices are then integrated and passed through a linear layer in order to compute the probabilities associated with each class, as determined by the softmax function. The category of the sample corresponds to the label with the highest probability value. Since cross-entropy (CE) is more effective in handling multi-categorization tasks, it has been deployed as the loss function in the proposed MTSA-Net. This loss function evaluates the discrepancies between the predicted class and the target class for each pixel; the mathematical formula is given below:

$$Loss = -\frac{1}{S}\sum_{s=1}^{S}\sum_{l=1}^{L}\mathbf{Y}_l^s \; log(\hat{\mathbf{Y}})_l^s) \tag{8}$$

where L represnts the number of classes and $\mathbf{Y}_l^s$ and $\hat{\mathbf{Y}}_l^s$ are true and corresponding predicted labels, respectively. S indicates the number of samples.

3.6 Implementation

Algorithm 1 outlines the procedural steps of the proposed model structure.

4 Experiment and analysis

In this section, five HSI datasets are evaluated using overall accuracy (OA), average accuracy (AA), and Kappa coefficient (K). Subsequently, the detailed experimental configuration is presented, where we quantitatively and visually compare the proposed method with alternative approaches based on CNNs and Transformers. The impact of varying patch sizes, as well as the influence of diverse training sample ratios on model performance, has been investigated. Additionally, the optimal choice for the

```
Algorithm 1 MTSA-Net
Input: HSI data I is input.
Output: Classification results.
Preprocessing: Set batch size to 128, patchsize, $s = 13$; PCAbands, $b = 30$; training sample ratio %, optimizer Adam (lr: $\eta = 0.001$), epochs number $e$ to 100
Iterate:
1. Apply a 3D convolutional layer to generate 3D feature maps.
2. Utilize a 2D convolutional layer along with spatial attention to produce 2D feature maps.
3. Concatenate the class tokens.
4. Embed position information.
5. Perform TE module.
6. Perform the multiscale feature fusion operation.
7. Input the initial classification token into the final linear layer.
8. Apply the softmax function to predict the class.
   end for: Apply the trained model to the test dataset for prediction.
```

Table 1: Detail of IP dataset with numbers of training, validation, and test samples.

Class	Name	Training	Validation	Testing	Total
1	Alfalfa	2	2	42	46
2	Corn-n	71	71	1286	1428
3	Corn-m	42	42	746	830
4	Corn	12	12	213	237
5	Grass-p	24	24	435	483
6	Grass-t	37	37	656	730
7	Grass-p-m	2	2	24	28
8	Hay-wd	18	18	430	478
9	Oats	1	1	18	20
10	Soybean-n	49	49	874	972
11	Soybean-m	123	123	2209	2455
12	Soybean-c	30	30	533	593
13	Wheat	10	10	185	205
14	Wood	63	63	1139	1265
15	Building-g-t-d	19	19	348	386
16	Stones-s-t	5	5	83	93
Total	-	509	509	9231	10,249

spectral dimension is determined. Finally, a discussion of the results from the ablation experiments is provided.

4.1 HSI datasets

Five HSI datasets were employed to analyze the effectiveness of the proposed MTSA-Net.

The Indian Pines (IP) dataset was gathered over agricultural and forest regions using the AVIRIS sensor. The data exhibits a comparatively low spatial resolution

Table 2: Detail of SA dataset with numbers of training, validation, and test samples.

Class	Name	Training	Validation	Testing	Total
1	Brocoli-gn-wd-1	60	60	1889	2009
2	Brocoli-gn-wd-2	112	112	3502	3726
3	Fallow	59	59	1858	1976
4	Fallow-rh-pw	42	42	1310	1394
5	Fallow-smoth	80	80	2518	2678
6	Stubble	119	119	3721	3959
7	Celery	107	107	3365	3579
8	Grapes-u	338	338	10595	11,271
9	Soil-vd-dp	186	186	5831	6203
10	Corn-sd-gn-ws	99	99	3080	3278
11	Lettuce-ro-4w	32	32	1004	1068
12	Lettuce-ro-5w	58	58	1811	1927
13	Lettuce-ro-6w	28	28	860	916
14	Lettuce-ro-7w	32	32	1006	1070
15	Vinyard-ud	218	218	6832	7268
16	Vinyard-vl-ts	54	54	1699	1807
Total	-	1624	1624	50,881	54,129

of almost 20 m/pixel, a size of 145 × 145 with 10,249 labelled pixels, spanning the wavelength range $0.4 - 2.5 \times 10^{-6}$ m. It has 224 spectral bands; after eliminating 24 regions of water absorption, the remaining 200 are used in the experiment. This dataset is divided into sixteen vegetation classes, where each class has a distinct sample size of between 20 and 2455. As indicated in Table 1, the total samples for each category are partitioned into training, validation, and testing sets.

Table 3: Detail of UP dataset with numbers of training, validation, and test samples.

Class	Name	Training	Validation	Testing	Total
1	Asphalt	199	199	6233	6631
2	Medows	560	560	17,529	18,649
3	Gravel	63	63	1973	2099
4	Trees	92	92	2880	3064
5	Paintd-m-s	40	40	1265	1345
6	Bare-s	151	151	4727	5029
7	Bitumen	40	40	1250	1330
8	Self-b-b	111	111	3460	3682
9	Shadow	29	29	889	947
Total	-	1285	1285	40,206	42,776

The Salinas (SA) dataset was captured by the AVIRIS sensor in the Salinas Valley and comprises 224 bands. However, 20 bands were discarded from experiments being noisy, and 204 bands were adopted for assessment. The dataset offers a spatial resolution of 3.7 m/pixel and a spatial dimension of 512 × 217. It encompasses 16 distinct land cover classes, including areas of bare soil, vineyard fields, and vegetables. Table 2 presents the distribution of total samples for each class, demonstrating splitting the data into train, validation, and test sets.

Table 4: Detail of Houston 2013 dataset with numbers of training, validation, and test samples.

Class	Name	Training	Validation	Testing	Total
1	Healthy-Grass	63	63	1125	1251
2	Stressed-Grass	63	63	1128	1254
3	Synthetic-Grass	35	35	627	697
4	Tree	62	62	1120	1244
5	Soil	62	62	1118	1242
6	Water	16	16	293	325
7	Residential	64	64	1140	1268
8	Commercial	62	62	1120	1244
9	Roads	63	63	1126	1252
10	Highway	61	61	1105	1227
11	Railway	62	62	1111	1235
12	Parking-1	62	62	1109	1233
13	Parking-2	24	24	421	469
14	Tennis-Court	22	22	384	428
15	Running-Track	33	33	594	660
Total	-	754	754	13521	15029

The University of Pavia (UP) dataset was collected over the city of Pavia using the ROSIS sensor. It offers a higher spatial resolution with 1.3 m/pixel and a spatial size of 610 × 340. The image consists of 115 channels, of which 12 are discarded from experimentation for being noisy; the dataset contains 103 spectral bands available for analysis. It comprises nine distinct land-cover classes, each having a varying number of samples. Table 3 illustrates the distribution of samples across the different land-cover categories, and the dataset is split up into training, validation, and test sets accordingly.

The Houston 2013 (H-13) dataset was gathered using the ITRES CASI-1500 sensor and covers the University of Houston premises and the neighbouring rural regions. This dataset has been publicly employed for assessing the effectiveness of HSI classification methods. This dataset has an image size of 349×1905 pixels and includes 144 spectral bands. It encompasses 15 challenging distinct classes, making it a valuable resource for land cover classification studies. The dataset samples have been divided into training, validation, and test sets, with the allocation details provided in Table 4.

The Houston 2018 (H-18) dataset was initially employed in 2018 IEEE GRSS Data Fusion competition, it was collected and published by the University of Houston campus and its neighboring areas. This dataset comprises HSI, multispectral LiDAR, and very high-resolution RGB images. The HSI dataset was collected by an ITRES CASI 1500 instrument, acquiring data in 48 bands within the spectral range of 380-1050 nm, with a resolution of 1 meter. The multispectral data was collected using an Optech Titan MW (14SEN/CON340), and the RGB data was captured using a VHR RGB imager (DiMAC ULTRALIGHT) equipped with a 70 mm focal length lens. The Houston 2018 dataset comprises 601 × 2384 pixels categorized into twenty distinct classes. Table 5 provides details regarding the number of training, validation, and test samples for this dataset.

Table 5: Detail of the Houston 2018 dataset with numbers of training, validation, and test samples.

Class	Name	Training	Validation	Testing	Total
1	Healthy-Grass	294	294	9211	9799
2	Stressed-Grass	975	975	30552	32502
3	Synthetic-Grass	21	21	642	684
4	Evergreen-Trees	408	408	12772	13588
5	Deciduous-Trees	152	152	4744	5048
6	Soil	136	136	4244	4516
7	Water	8	8	250	266
8	Residential	1193	1193	37376	39762
9	Commercial	6711	6711	210262	223684
10	Road	1374	1374	43062	45810
11	Side walk	1020	1020	31962	34002
12	Cross walk	46	46	1424	1516
13	Major T-fares	1391	1391	43576	46358
14	Highway	296	296	9257	9849
15	Railway	208	208	6521	6937
16	Paved P-L	344	344	10787	11475
17	Gravel P-L	5	5	139	149
18	Cars	197	197	6184	6578
19	Trains	161	161	5043	5365
20	Seats	205	205	6414	6824
Total	-	15141	15141	474422	504712

Table 6: Classification results (%) of various models on the IP dataset.

class	2DCNN	3DCNN	HybridSN	SSCRN	DATN	MSDCA	SSFTT	MACLSTM	MTSA-Net
1	21.12	41.46	87.8	**94.85**	94.80	75.49	86.36	59.09	82.61
2	81.15	90.51	94.39	95.15	96.81	71.23	94.69	94.47	**96.54**
3	90.98	79.36	96.52	95.87	97.22	80.45	**98.85**	98.35	98.57
4	70.43	46.01	83.89	98.01	97.51	93.52	88.88	86.67	**98.85**
5	96.83	95.17	98.16	97.81	98.83	95.53	**99.78**	97.6	98.23
6	98.19	**99.7**	99.54	96.77	98.54	86.11	99.42	96.39	98.86
7	92.59	88	92.97	88.89	79.59	**99**	92.53	88.89	74.20
8	99.99	99.98	99.89	99.98	97.99	98.32	98.25	98.77	**100**
9	88.21	48.89	86.27	88.85	67.23	87.45	84.21	**89.47**	77.85
10	86.69	86.06	97.94	98.01	95.45	79.41	96.64	96.75	**98.45**
11	89.5	97.51	**99.5**	97.99	98.01	77.50	98.58	96.87	98.85
12	65.53	74.91	94.57	98.15	97.86	75.46	92.53	92.36	**98.54**
13	99.95	99.46	94.59	**99.85**	97.15	97.51	100	98.46	97.96
14	99.93	99.74	99.29	98.95	**100**	94.24	99.91	**100**	**100**
15	82.65	84.1	92.35	**97.95**	98.96	79.61	92.91	92.1	98.53
16	74.45	93.95	97.98	98.55	95.66	**100**	94.31	92.45	89.55
OA	84.47	91.03	95.62	94.17	94.17	82.54	96.35	94.31	**98.84**
AA	83.63	82.80	94.72	**96.60**	94.40	87.60	94.86	92.41	94.22
K	83.55	88.68	95.29	96.45	94.75	81.43	95.98	93.55	**97.24**

4.2 Experimental Setting

In this article, three evaluation metrics are utilized to examine the effectiveness of the proposed MTSA-Net: overall accuracy (OA), average accuracy (AA), and the Kappa coefficient. The proposed approach is implemented within the Python programming language, employing the PyTorch framework. All experiments described in this paper have been executed on a single system equipped with an NVIDIA RTX 3050 GPU and an 11th Gen Intel(R) Core i7-11800 CPU, which had 16 GB of memory.

Table 7: Classification results (%) of various models on the SA dataset.

class	2D CNN	3D CNN	Hybrid SN	SSCRN	DATN	MSDCA	SSFTT	MACLST	MTSA-Net
1	98.87	99.73	99.45	99.65	98.43	84.19	99.56	98.46	**99.82**
2	99.49	96.94	98.15	99.95	99.43	98.32	98.21	96.31	**100**
3	99.54	88.48	99.08	99.86	99.80	95.95	98.78	97.6	**100**
4	99.22	97.45	99.53	98.88	99.44	96.32	99.52	99.43	**99.85**
5	97.51	95.91	98.85	**99.85**	99.11	99.03	99.79	92.28	99.38
6	**100**	**100**	98.99	99.25	99.81	95.16	**100**	95.54	99.76
7	99.86	98.28	99.68	99.85	**99.98**	88.30	99.94	96.35	99.94
8	91.57	85.64	97.82	**99.99**	95.23	89.11	99.23	96.11	99.88
9	99.91	98.57	**100**	99.93	99.91	87.85	**100**	95.64	**100**
10	99.23	93.87	98.86	**100**	95.54	90.45	99.96	**100**	99.9
11	98.18	90.01	99.23	99.85	98.57	98.23	**100**	98.49	**100**
12	99.34	96.61	99.91	99.97	99.81	89.15	98.52	91.32	**100**
13	99.28	98.56	**99.85**	99.84	99.35	74.47	99.84	91.22	99.10
14	98.74	94.29	98.56	98.89	99.26	95.89	99.12	97.91	**99.88**
15	91.82	79.02	95.99	99.07	84.71	**99.98**	99.95	90.23	99.68
16	99.27	96.69	98.78	98.85	**100**	99.85	**100**	99.78	98.99
OA	96.76	96.65	98.84	99.67	98.80	93.55	99.51	96.69	**99.80**
AA	98.28	94.37	98.92	99.60	98.40	92.64	99.52	96.04	**99.64**
K	96.39	92.44	97.23	98.71	98.87	89.47	98.55	96.45	**99.44**

In the experimental configuration, the model operates with a dataset batch size set at 128, employs a learning rate of 0.001, and optimizes model parameters using the Adam optimizer. The maximum number of epochs is limited to 100. The dataset division is as follows: the IP and H-13 datasets are divided with a split ratio of 5% for training, 5% for validation, and 90% for testing purposes, while the UP, SA, and H-18 datasets are divided with split ratios of 3%, 3%, and 94%, respectively. To ensure a reliable evaluation, each experiment was conducted ten times with different random initializations for each model. The average Overall Accuracy (OA) across these runs was reported to assess performance stability. Classification maps were subsequently generated based on the aggregated results

4.3 Classification results

The robustness of the MTSA-Net is confirmed by comparing it with eight deep learning-based methods, including 2D-CNN [69], 3D-CNN [70], HybridSN [4], SSCRN

Table 8: Classification results (%) of various models on the UP dataset.

class	2D	3D	Hybrid	SSCRN	DATN	MSDCA	SSFTT	MACLST	MTSA-
	CNN	CNN	SN						Net
1	93.28	93.31	95.51	99.88	95.21	81.54	**99.92**	97.23	97.31
2	94.93	93.99	99.49	97.55	97.54	93.52	96.33	96.28	**99.91**
3	75.55	90.19	94.18	**98.09**	83.63	73.66	97.44	97.18	96.44
4	93.87	91.29	99.55	99.52	95.28	92.22	98.08	97.34	92.32
5	95.51	95.47	96.71	96.07	97.46	97.54	95.87	95.74	**97.12**
6	70.05	93.85	99.43	96.58	92.22	89.21	95.69	96.38	**99.61**
7	70.92	81.45	90.11	96.78	90.51	88.20	97.22	95.55	**100**
8	90.3	92.73	93.11	95.61	92.23	95.54	95.21	**99.22**	97.44
9	97.89	95.46	95.22	94.37	100	95.35	98.35	98.05	97.41
OA	90.19	92.01	98.16	98.11	96.54	94.88	96.33	98.45	**98.77**
AA	86.91	91.97	95.92	97.16	94.11	89.22	97.12	96.99	**97.55**
K	87.52	90.87	96.36	96.35	97.11	96.27	97.89	97.57	**98.91**

Table 9: Classification results (%) of various models on the H-13 dataset.

class	2D	3D	Hybrid	SSCRN	DATN	MSDCA	SSFTT	MACLST	MTSA-
	CNN	CNN	SN						Net
1	91.68	95.13	95.78	96.22	93.12	88.54	**98.54**	97.81	97.56
2	90.11	92.54	94.51	96.01	96.46	92.79	96.31	95.89	**96.76**
3	80.23	98.21	96.52	98.52	97.53	94.11	97.54	98.11	**98.95**
4	90.21	91.35	92.01	92.01	90.33	90.36	**95.25**	94.51	93.38
5	96.41	97.15	97.89	96.21	98.35	98.85	95.14	93.33	**98.92**
6	89.22	97.63	97.89	**98.77**	97.75	96.41	96.32	94.15	85.56
7	80.28	80.51	95.14	**98.99**	95.61	90.32	97.12	95.66	93.56
8	93.89	89.51	96.21	95.23	95.55	88.44	**98.51**	96.11	91.42
9	99.33	92.32	94.21	98.77	97.41	89.74	97.43	94.82	96.21
10	99.78	95.69	98.86	96.84	96.14	85.26	92.31	97.21	97.22
11	98.18	98.08	93.56	94.68	98.67	98.80	95.65	98.74	**99.88**
12	96.54	96.39	94.51	95.21	95.36	88.55	95.64	92.37	**98.66**
13	70.25	85.54	85.14	80.13	84.99	71.88	87.21	90.39	**92.01**
14	92.36	96.29	94.75	98.51	**98.96**	94.78	92.32	97.48	98.08
15	94.33	93.97	97.15	89.37	83.55	**99.53**	94.51	93.97	99.02
OA	93.55	95.21	96.54	96.56	97.11	93.01	97.12	95.55	**97.84**
AA	90.85	93.35	94.94	94.99	95.12	92.44	95.32	95.37	**96.44**
K	92.33	94.98	95.28	96.11	96.55	93.89	96.99	95.11	**98.47**

[3], DATN [71], MSDCA [72], SSFTT [48], MACLST [73]. Additionally, in Table 6 to 10, a comparative evaluation of the proposed model against modern HSI classification methods is provided, using five well-known datasets.

As evident from the tables, the proposed MTSA-Net model consistently demonstrates exceptional classification performance across all five datasets while utilizing a small number of training samples.

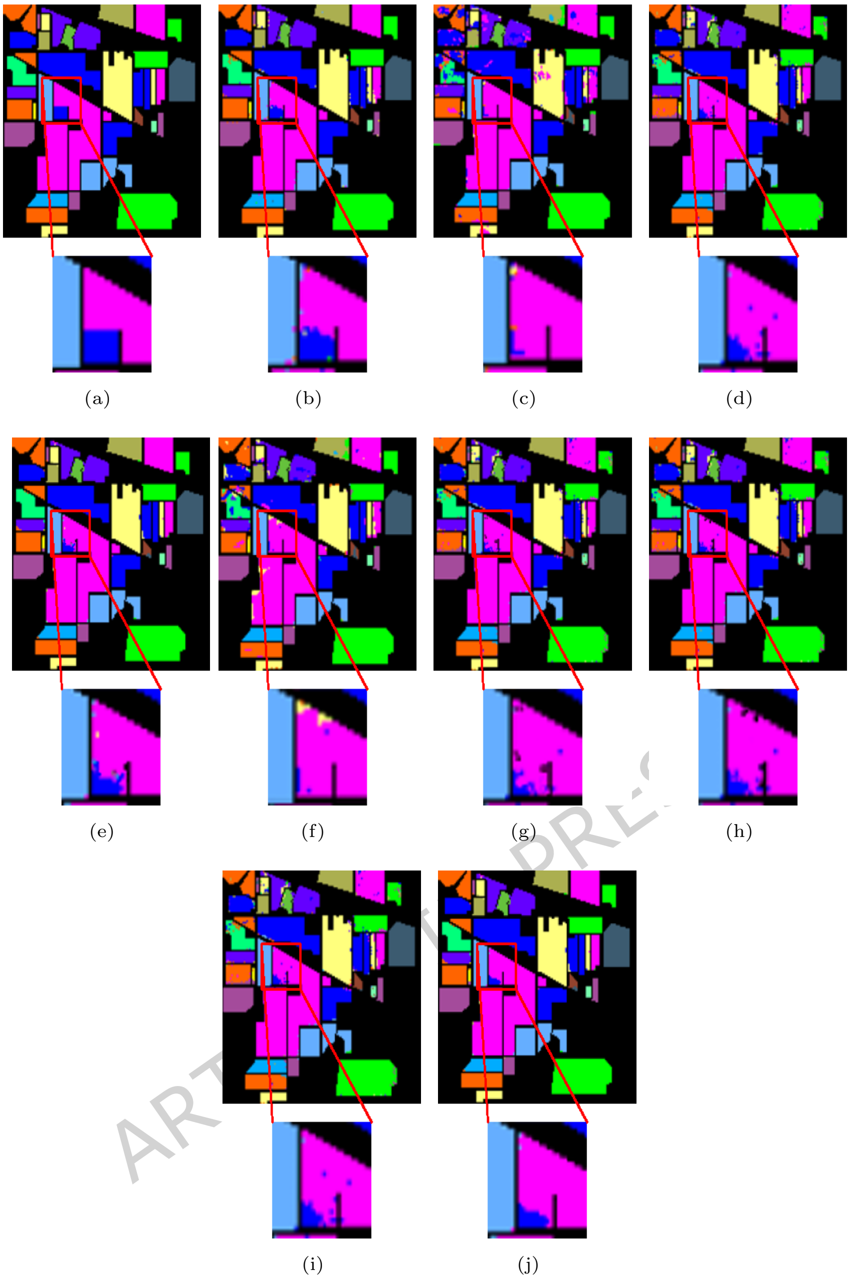
Tables 6 to 10 present quantitative classification results, encompassing OA, AA, and K for each class, acquired by various models on the IP, SA, UP, H-13, and H-18

Table 10: Classification results (%) of various models on the H-18 dataset.

class	2D CNN	3D CNN	Hybrid SN	SSCRN	DATN	MSDCA	SSFTT	MACLST	MTSA-Net
1	81.25	82.58	86.88	87.12	82.64	86.88	70.11	87.22	**89.55**
2	90.54	90.04	93.22	94.05	89.45	95.28	93.55	**97.11**	93.88
3	90.08	88.48	99.08	**99.86**	98.22	95.23	95.11	97.56	98.88
4	96.33	97.45	99.53	98.88	98.22	95.21	90.19	98.77	**99.66**
5	98.11	98.76	98.85	99.24	99.35	99.88	88.78	91.28	**99.96**
6	93.87	99.71	98.99	99.25	98.84	93.44	99.32	88.51	**99.55**
7	88.11	98.28	99.17	99.53	99.33	88.56	99.13	93.33	**99.43**
8	90.22	91.22	91.851	94.55	88.64	88.11	92.06	91.11	**93.44**
9	96.36	98.57	98.01	97.85	97.88	90.52	**99.25**	94.74	98.11
10	98.77	94.93	98.86	96.54	97.43	76.87	**99.96**	98.81	86.97
11	90.11	95.85	**99.23**	95.65	98.02	76.23	80.12	97.45	89.88
12	48.84	49.92	51.69	50.11	50.23	66.22	54.33	56.85	**76.82**
13	91.71	98.56	99.85	**99.84**	99.11	74.93	65.01	84.75	90.73
14	98.74	94.29	98.56	98.89	99.11	96.32	**99.85**	95.45	83.55
15	91.82	79.02	95.99	94.54	84.85	95.87	95.14	80.22	**96.68**
16	99.27	96.69	98.78	98.85	**99.91**	99.23	88.98	98.78	89.56
17	85.21	87.77	76.07	74.55	77.56	42.66	70.18	71.25	**86.33**
18	90.22	83.98	79.94	78.45	75.91	78.02	90.21	**90.25**	88.90
19	89.78	89.83	87.32	88.98	90.51	89.76	89.92	88.87	**93.57**
20	88.32	90.89	85.45	85.45	84.43	94.44	88.21	**97.25**	96.88
OA	92.11	92.06	92.32	93.66	93.32	88.61	92.58	93.55	**95.87**
AA	89.83	90.34	91.86	91.60	91.46	86.44	87.45	89.98	**92.87**
K	93.66	92.44	89.58	92.58	92.98	87.55	89.14	92.85	**94.83**

datasets. The highest notable OA values are achieved on the IP, SA, UP, H-13, and H-18 datasets, with accuracy rates of 98.84%, 99.80%, 98.77%, 97.84%, and 95.87%, respectively. It is observed from the results that CNN-based methods such as 2D CNN, 3D CNN, HybridSN, SSCRN and DATN exhibit exceptional performance. Sub-optimal results in certain classes can be attributed to the limited number of training samples available for these specific classes. In addition, attention-based methods with convolution, such as SSFTT, and MACLST, obtained superior results. However, the MTSA-Net outperformed all the compared methods, achieving outstanding results in terms of OA, AA and k. Additionally, the model consistently demonstrates high accuracy across all classes, confirming its efficiency and stability, especially in scenarios with few samples for specific classes.

The classification results of all compared approaches using the IP dataset are presented in Table 6. Certain classes in the IP dataset comprise a few samples, making it challenging to extract easily distinguishable features. The performance of existing methods is compromised due to the inherent difficulty of addressing imbalanced class challenges. For instance, 2D CNN and 3D CNN produce less than 50% accuracy for classes 1, 4, and 9. The proposed method surpasses all the compared models, achieving a remarkable accuracy of more than 80% on classes with limited samples. Specifically, the proposed method obtained OA of 98.84%, AA of 94.22%, and 97.24% of k.



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j)

Fig. 4: Classification maps generated by various models on the IP dataset. (a) Ground truth, (b) 2DCNN, (c) 3DCNN, (d) HybridSN, (e) SSCRN, (f) DATN, (g) MSDCA, (h) SSFTT, (i) MACLST,(j) MTSA-Net.

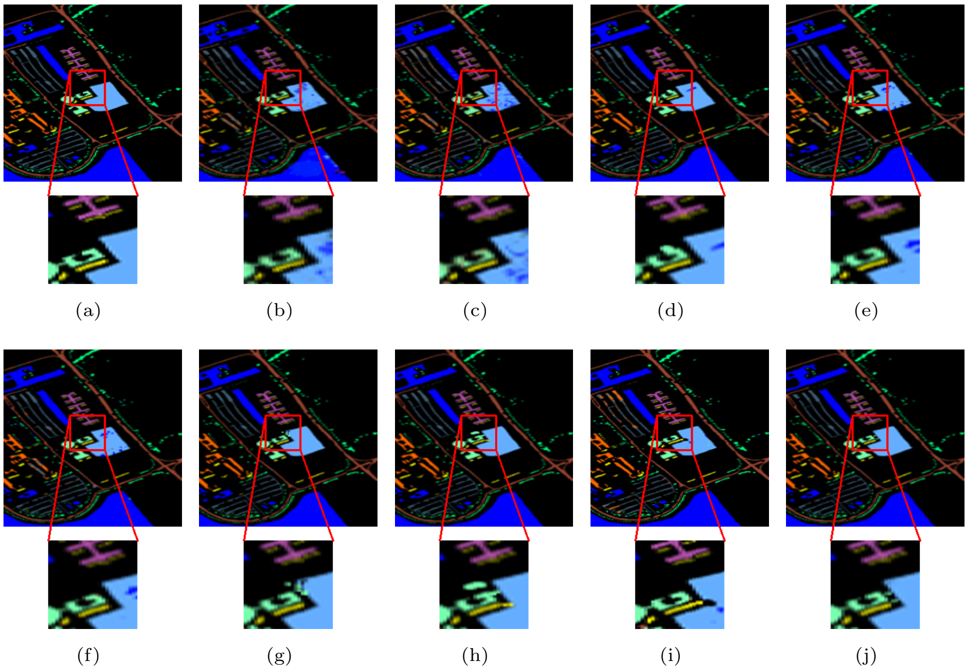


Fig. 5: Classification maps generated by various models on the UP dataset. (a) Ground truth. (b) 2DCNN. (c) 3DCNN. (d) HybridSN. (e) SSCRN. (f) DATN. (g) MSDCA. (h) SSFTT. (i) MACLST. (j) MTSA-Net.

The classification outcomes for the SA dataset can be observed from Table 7. Notably, MSDCA and MACLST exhibited subpar results for classes 12 and 13, while 3D CNN, DATN, and MACLST also delivered less satisfactory results for class 15 in this dataset. SSFTT and SSCRN demonstrated comparable results, outperforming other methods except the MTSA-Net. Our proposed method achieved an accuracy of over 99% in each individual class, boasting impressive overall metrics of 99.80% for OA, 99.64% for AA, and 99.44% for k.

Table 8 presents the OA, AA, k, and class-based accuracies for all models on the UP dataset. Several modern methods have demonstrated superior performance on this dataset, primarily owing to their maximum number of training samples. Nevertheless, the challenge in this dataset lies in learning discriminative features, as the presence of interfering pixels makes this a challenging task. SSCRN and SSFTT outperformed all other methods in class 3 and class 1, respectively. In addition, the proposed method demonstrated outstanding performance in class-based accuracies and attained an impressive overall accuracy of 98.77%.

The classification results for the Houston 2013 dataset can be seen in Table 9. On this dataset, the 2D CNN and MSDCA methods exhibit poor performance, while DATN and SSFTT surpass other techniques in terms of overall accuracy. To be more specific, our proposed method yields favourable classification results in classes 13, 14,

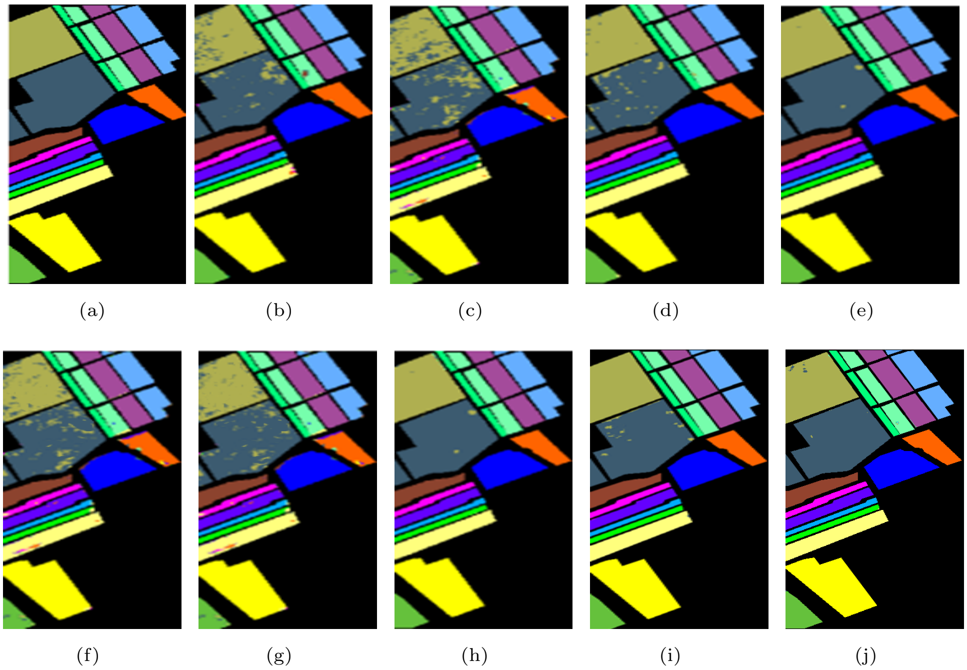


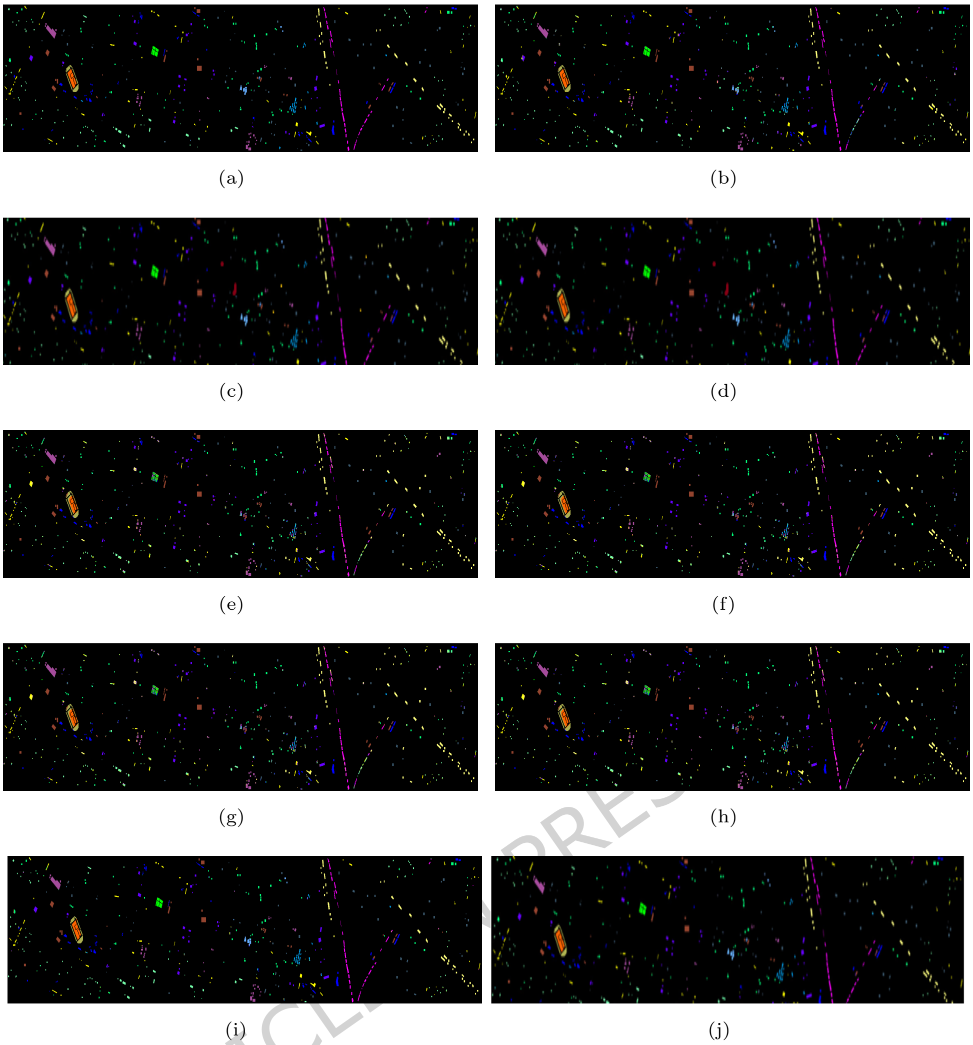
Fig. 6: Classification maps generated by various models on the SA dataset (a) Ground truth. (b) 2DCNN. (c) 3DCNN. (d) HybridSN. (e) SSCRN. (f) DATN. (g) MSDCA, (h) SSFTT. (i) MACLST. (j) MTSA-Net.

and 15, despite the few training samples available for these classes. Table 9 reported that the proposed MTSA-Net achieved exceptional results of 97.84%, 96.44%, and 98.47% in terms of OA, AA, and k, respectively.

In Table 10, the classification result of various methods on Houston 2018 dataset is presented, which is relatively low as compared to other datasets. This discrepancy lies due to the complex nature and identical classes of H-18 dataset. In terms of overall accuracy, 3D CNN, hybridSN, and SSFTT demonstrate comparable results, while SSCRN and MACLST exhibit relatively improved performance. However, the proposed method surpasses its counterparts by achieving the highest overall accuracy of 95.87%.

4.4 Classification maps

To visualize the classification results more effectively, we display the corresponding visual classification maps in Figures 4 to 8. These visual representations display false-color images along with their associated ground-truth maps for the provided five datasets. From the classification maps, it is clear that our proposed model produces less noise and is closest to the ground truth. For instance, the proposed approach effectively discriminates between ground features, preserves the quality of boundary


(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

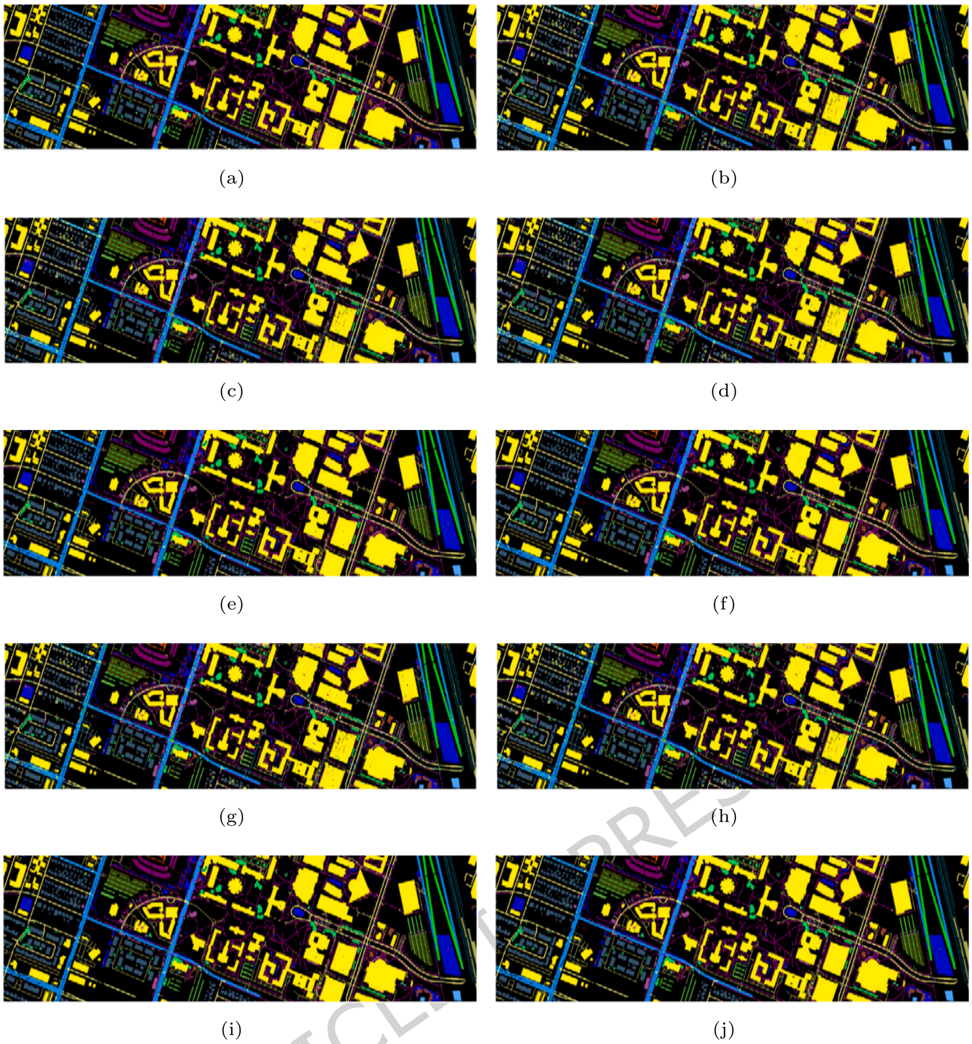
(i)

(j)

Fig. 7: Classification maps generated by various models on H-13 dataset. (a) Ground truth. (b) 2DCNN. (c) 3DCNN. (d) HybridSN. (e) SSCRN. (f) DATN. (g) MSDCA (h) SSFTT. (i) MACLST. (j) MTSA-Net.

regions, and generates more genuine classification maps. This is achieved by taking into account both local and global representations.

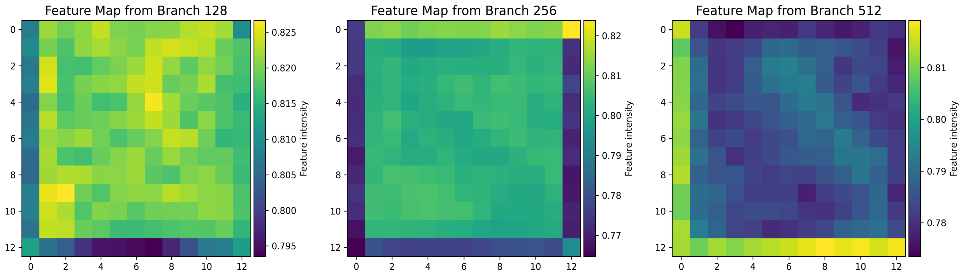
To complement the classification outcomes, we provide feature intensity heatmaps generated from MLP blocks with hidden dimensions of 128, 256, and 512 to empirically demonstrate the multiscale representational strength of the proposed model. As shown in Figure 9, the 128-dimensional branch primarily focuses on detailed spatial textures and localized structural patterns, while the 512-dimensional branch encodes



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Fig. 8: Classification maps generated by various models on the H-18 dataset. (a) Ground truth. (b) 2DCNN. (c) 3DCNN. (d) HybridSN. (e) SSCRN. (f) DATN. (g) MSDCA. (h) SSFTT. (i) MACLST. (j) MTSA-Net.

broader, global representations across larger areas. The 256-dimensional branch strikes a balance, capturing both fine local details and overarching global information. These distinct patterns across different scales confirm that the multiscale MLP modules successfully extract hierarchical feature representations, providing clear evidence of their role in boosting discriminative capability and enhancing classification accuracy.



(a) (b) (c)

Fig. 9: Feature intensity heatmaps obtained from the three transformer branches with hidden dimensions 128, 256, and 512 on the IP dataset. Brighter regions highlight areas with stronger feature activations, demonstrating the multiscale representation capability of the proposed model.

4.5 Ablation studies

In this section, we evaluate the effectiveness of the proposed model by conducting three crucial experiments: exploring the impact of varying patch sizes, evaluating the influence of different training sample sizes, and assessing the significance of altering the number of principal components.

4.5.1 Ablation study of MTSA-Net modules

The performance of the proposed model has been experimentally evaluated to assess the impact of various component parts of the module. Table 11 and 12 presents the results from various implementations of the proposed model and comparison of FLOPs and parameter counts between MTSA-Net and the other baseline methods respectively. To evaluate the efficiency and complexity of the proposed MTSA-Net method, we assessed several key metrics using an input image size of 1×30×15×15 with a batch size of 64. These metrics include the number of parameters, Floating Point Operations (FLOPs), memory size, and inference time on the IP dataset.

The proposed method achieves an overall accuracy (OA) of 98.24%, accompanied by a marginal increase in FLOPs, while other performance metrics remain largely consistent with minor variations. In contrast, module lacking multiscale attention demonstrate a significant decline in OA.

Table 11: Analysis of the Different Module Components in MTSA-Net

Model	Parameters (M)	FLOPS (M)	Size in Memory (MB)	Inference Time (ms)	OA (%)
CNN+SA	0.53	750	2.32	0.64	80.56
CNN+SA+TE	0.55	845	2.54	2.09	94.54
MTSA-Net	0.59	960	2.78	2.26	98.24

Table 12: Comparison of FLOPs and parameters of different methods. OA values are averaged over 10 runs.

Method	Params (M)	FLOPs (M)	OA (%)
2D-CNN	0.42	610	84.47
3D-CNN	0.48	720	91.03
HybridSN	0.51	815	95.62
SSCRN	0.55	850	97.02
DATN	0.58	930	97.45
MSDCA	0.56	860	97.80
SSFTT	0.56	880	96.35
MACLST	0.58	920	98.00
MTSA-Net	**0.59**	**960**	**98.24**

4.5.2 Impact of input patch sizes

The complexity of feature extraction is significantly influenced by the spatial size. A very small patch size provides limited spatial information, while a larger size encompasses numerous pixels with diverse categories and intricate spatial details, potentially hindering the classification process. Therefore, experiments have been conducted with various spatial input sizes ranging from 7 to 15, aiming to analyze their impact on the performance of the proposed MTSA-Net. As depicted in Figure 10, a patch size of 13 persistently outperforms other patch sizes by achieving the utmost overall accuracy among all datasets. Consequently, based on these results, a patch size of 13 is selected for further experiments, guaranteeing optimal classification performance.

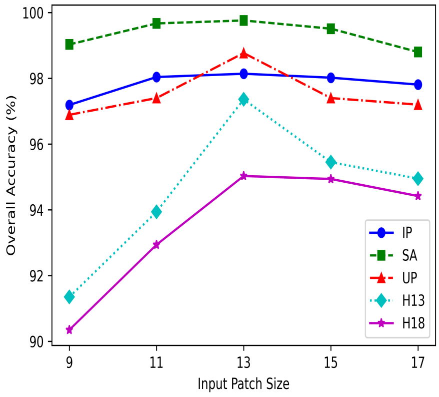


Fig. 10: Impact of various patch sizes on OA(%) for the IP, UP, SA, H-13, and H-18 datasets.

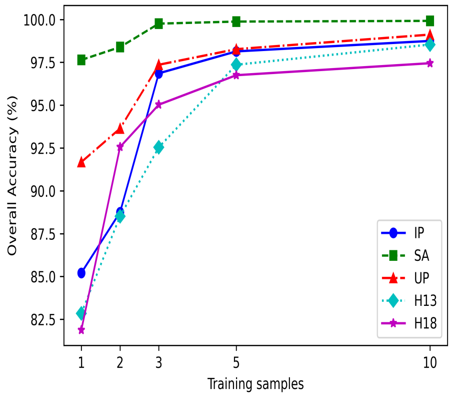


Fig. 11: Impact of randomly selected training samples on OA(%) for the IP, UP, SA, H-13, and H-18 datasets.

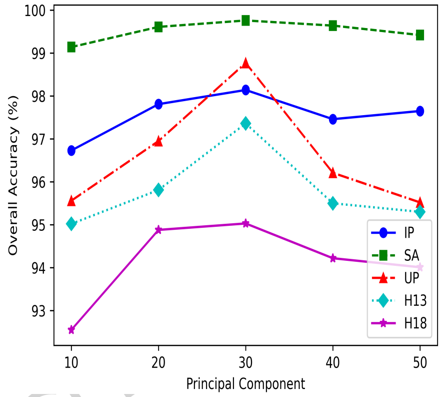
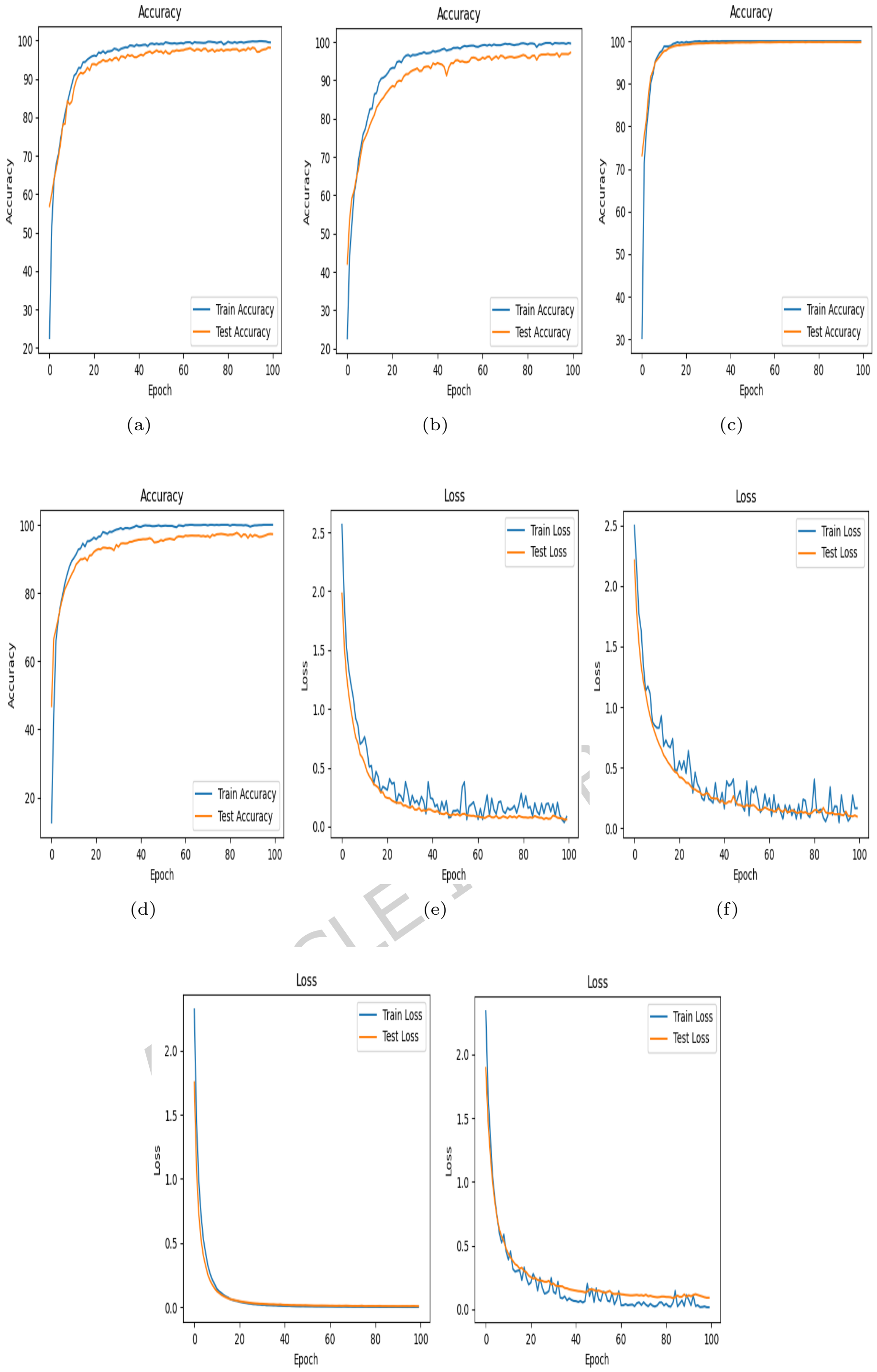


Fig. 12: Comparison of OA(%) with different principal components for the IP, UP, SA, H-13, and H-18 Datasets.

4.5.3 Impact of diverse training data size

To confirm the robustness of the proposed model, various training sample ratios have been utilized on five datasets, as shown in Figure 11. The training samples with the ratios of 10%, 5%, 3%, 2%, and 1% of the overall samples are randomly selected. The proposed model consistently achieves outstanding classification performance across all sample sizes on the five datasets.

As depicted in Figure 11, an increase in the proportion of training samples contributes to moderate improvement in overall accuracy for each dataset. However, the



(a) (b) (c)

(d) (e) (f)

(g) (h)

Fig. 13: Accuracy curves (a-d) and loss curves (e-h) for the training and testing sets of the IP, H-13, SA and UP datasets.

proposed method demonstrates stability and robustness even with a small proportion of training samples.

4.5.4 Impact of Principal components

In the proposed model, PCA is employed to minimize the parameters by decreasing the dimensionality of the original HSI. The experimental analysis demonstrates that the different numbers of principal components significantly influence the extraction of spectral-spatial features. The total number of spectral bands preserved by PCA and its impact on overall accuracy is investigated. Varying numbers of PCA components (10, 20, 30, 40, and 50) are selected for analysis. The results, illustrated in Figure 12, indicate that the best classification performance is attained when selecting 30 principal components in our proposed model. Consequently, in light of these findings, we utilised 30 principal components for our experimentations.

4.5.5 Convergence curve

Figure 13 displays the accuracy and loss curves obtained for each epoch during the experiments conducted on the five datasets. With only 5% of trained samples from the IP and H-13 datasets, and 3% of trained samples from the SA, UP, and H-18 datasets, the accuracy of the proposed model during training and testing shows a consistent increase as the number of epochs increases. Additionally, the loss curve exhibits a steady decrease throughout the training and testing process. This indicates that the model is effectively learning and improving its performance on both training and test data.

4.5.6 Time cost comparison

To evaluate the inference efficiency of the proposed MTSA-Net, we measured both the training and testing durations across the benchmark datasets. As observed from the data in table 13, MTSA-Net attains the fastest training time on the IP dataset (6.8 minutes), outperforming both CNN-based and transformer-based baselines. This confirms that the proposed multiscale design achieves superior accuracy without incurring additional training overhead. Among all the evaluated datasets, Houston 2018 (H-18) exhibits the longest training time due to its high spatial and spectral complexity. Compared to the contrast methods, MTSA-Net offers significant advantages in terms of both efficiency and overall classification accuracy.

5 Conclusion

In this work, a novel framework, MTSA-Net, is introduced for HSI classification, harnessing the combined strengths of spatial attention and multiscale transformers to effectively utilize the spatial-spectral information in hyperspectral data. The spatial attention mechanism enhances selective spatial features by considering the relationships between adjacent pixels, thereby increasing the discriminative power of the learned representations while suppressing irrelevant information. Followed by a multiscale transformer module that captures long-range dependencies, enabling the

Table 13: Time comparison of the proposed MTSA-Net with contrast methods across benchmark datasets.

	IP		UP		SA		H-13		H-18	
Methods	Train (m)	Test (s)	Train (m)	Test (s)	Train (m)	Test (s)	Train (m)	Test (s)	Train (m)	Test (s)
2D CNN	7.5	4.32	12.5	5.05	15.7	10.21	17.4	11.1	22.59	24.98
3D CNN	9.85	13.4	14.1	22.5	16.5	26.6	18.4	24.52	25.4	21.01
Hybrid SN	14	4.8	20.52	6.6	25.7	9.15	23.11	18.56	25.4	20.55
SSCRN	16.3	14.4	18.45	16.82	23.45	28.02	25.4	20.1	32.5	28.51
DATN	10.12	8.5	14.6	9.5	18.4	12.21	19.4	11.3	25	19.5
MSDCA	11.54	8.9	15.31	18.32	18.01	20.5	20.58	12.8	26.5	9.8
SSFTT	8.9	3.6	13.6	5.9	14.6	9.2	16.8	9.1	20.5	10.5
MACLST	12.4	8.9	16.6	14.2	18.9	16.2	19.5	8.98	22.5	11.5
MTSA-Net	6.8	2.6	16.8	9.0	20.3	11.4	16.5	6.3	21.5	16.2

framework to emphasize more distinguishing features. By adequately leveraging spatial and spectral information at different scales, the MTSA-Net method achieved excellent discriminative feature representations. Finally, a multiscale feature fusion approach is employed to integrate features from various levels, maximizing their contribution to robust and discriminative feature learning. Experimental results validate that MTSA-Net achieves the highest classification accuracy compared to state-of-the-art methods across five challenging benchmark datasets. Moreover, it demonstrates strong robustness even with limited training samples, highlighting its superiority and effectiveness in HSI classification. The generalization capability of MTSA-Net also makes it suitable for extension to other hyperspectral remote sensing datasets. In future work, we plan to explore more efficient multiscale designs and lightweight transformer modules to further reduce computational costs without compromising overall performance, making the framework even more suitable for HSI classification tasks.

Acknowledgements

This paper was funded by the Princess Nourah bint Abdulrahman University Researchers Supporting Project, under grant No. (PNURSP2025R755), Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. The authors, therefore, gratefully acknowledge and thank Nourah bint Abdulrahman University for its technical and financial support.

Competing interests

The authors declare no competing interests.

Data availability statement

The datasets analyzed during this research are available at: https://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes https://machinelearning.ee.uh.edu/2013-ieee-grss-data-fusion-contest/ https://machinelearning.ee.uh.edu/2018-ieee-grss-data-fusion-challenge-fusion-of-multispectral-lidar-and-hyperspectral-data/

References

[1] Li, S., Song, W., Fang, L., Chen, Y., Ghamisi, P., Benediktsson, J.A.: Deep learning for hyperspectral image classification: An overview. IEEE Transactions on Geoscience and Remote Sensing **57**(9), 6690–6709 (2019) https://doi.org/10.1109/TGRS.2019.2907932

[2] Vaddi, R., Manoharan, P.: Hyperspectral image classification using cnn with spectral and spatial features integration. Infrared Physics & Technology **107**, 103296 (2020)

[3] Farooque, G., Xiao, L., Yang, J., Sargano, A.B.: Hyperspectral image classification via a novel spectral–spatial 3d convlstm-cnn. Remote Sensing **13**(21), 4348 (2021)

[4] Roy, S.K., Krishna, G., Dubey, S.R., Chaudhuri, B.B.: Hybridsn: Exploring 3-d–2-d cnn feature hierarchy for hyperspectral image classification. IEEE Geoscience and Remote Sensing Letters **17**(2), 277–281 (2019)

[5] Wang, C., Zhang, P., Zhang, Y., Zhang, L., Wei, W.: A multi-label hyperspectral image classification method with deep learning features. In: Proceedings of the International Conference on Internet Multimedia Computing and Service, pp. 127–131 (2016)

[6] He, L., Li, J., Liu, C., Li, S.: Recent advances on spectral–spatial hyperspectral image classification: An overview and new guidelines. IEEE Transactions on Geoscience and Remote Sensing **56**(3), 1579–1597 (2017)

[7] Hong, D., Han, Z., Yao, J., Gao, L., Zhang, B., Plaza, A., Chanussot, J.: Spectralformer: Rethinking hyperspectral image classification with transformers. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–15 (2021)

[8] Zhou, P., Han, J., Cheng, G., Zhang, B.: Learning compact and discriminative stacked autoencoder for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing **57**(7), 4823–4833 (2019)

[9] Zheng, X., Yuan, Y., Lu, X.: Dimensionality reduction by spatial–spectral preservation in selected bands. IEEE Transactions on Geoscience and Remote Sensing **55**(9), 5185–5197 (2017)

[10] Liang, J., Li, P., Zhao, H., Han, L., Qu, M.: Forest species classification of uav hyperspectral image using deep learning. In: 2020 Chinese Automation Congress (CAC), pp. 7126–7130 (2020). https://doi.org/10.1109/CAC51589.2020.9327690

[11] Zhang, C., Zhou, L., Zhao, Y., Zhu, S., Liu, F., He, Y.: Noise reduction in the spectral domain of hyperspectral images using denoising autoencoder methods. Chemometrics and Intelligent Laboratory Systems **203**, 104063 (2020)

[12] Hadi, F., Yang, J., Ullah, M., Ahmad, I., Farooque, G., Xiao, L.: Dhcae: Deep hybrid convolutional autoencoder approach for robust supervised hyperspectral unmixing. Remote Sensing **14**(18), 4433 (2022)

[13] Hadi, F., Yang, J., Farooque, G., Xiao, L.: Deep convolutional transformer network for hyperspectral unmixing. European Journal of Remote Sensing **56**(1), 2268820 (2023)

[14] Khader, A., Xiao, L., Yang, J.: A model-guided deep convolutional sparse coding network for hyperspectral and multispectral image fusion. International Journal of Remote Sensing **43**(6), 2268–2295 (2022)

[15] Zhang, G., Zhao, S., Li, W., Du, Q., Ran, Q., Tao, R.: Htd-net: A deep convolutional neural network for target detection in hyperspectral imagery. Remote Sensing **12**(9), 1489 (2020)

[16] Farooque, G., Xiao, L., Sargano, A.B., Abid, F., Hadi, F.: A dual attention driven multiscale-multilevel feature fusion approach for hyperspectral image classification. International Journal of Remote Sensing **44**(4), 1151–1178 (2023)

[17] Shenming, Q., Xiang, L., Zhihua, G.: A new hyperspectral image classification method based on spatial-spectral features. Scientific Reports **12**(1), 1541 (2022)

[18] Miranda-Vega, J.E., Rivas-López, M., Fuentes, W.F.: k-nearest neighbor classification for pattern recognition of a reference source light for machine vision system. IEEE Sensors Journal **21**(10), 11514–11521 (2021) https://doi.org/10.1109/JSEN.2020.3024094

[19] Shi, X., Sun, L.: Hyperspectral image classification with support vector machines based on the maximum noise fraction. In: 2022 IEEE 5th International Conference on Electronics Technology (ICET), pp. 1193–1197 (2022). https://doi.org/10.1109/ICET55676.2022.9824122

[20] Bajpai, S., Singh, H.V., Kidwai, N.R.: Feature extraction & classification of hyperspectral images using singular spectrum analysis & multinomial logistic regression classifiers. In: 2017 International Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT), pp. 97–100 (2017). IEEE

[21] Duan, W., Li, S., Fang, L.: Spectral-spatial hyperspectral image classification using superpixel and extreme learning machines. In: Pattern Recognition: 6th Chinese Conference, CCPR 2014, Changsha, China, November 17-19, 2014. Proceedings, Part I 6, pp. 159–167 (2014). Springer

[22] Zhang, Y., Cao, G., Li, X., Wang, B.: Cascaded random forest for hyperspectral image classification. IEEE journal of selected topics in applied earth observations and remote sensing **11**(4), 1082–1094 (2018)

[23] Islam, M.R., Ahmed, B., Hossain, M.A.: Feature reduction based on segmented principal component analysis for hyperspectral images classification. In: 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), pp. 1–6 (2019). https://doi.org/10.1109/ECACE.2019.8679394

[24] Champa, A.I., Rabbi, M.F., Mahedy Hasan, S.M., Zaman, A., Kabir, M.H.: Tree-based classifier for hyperspectral image classification via hybrid technique of feature reduction. In: 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), pp. 115–119 (2021). https://doi.org/10.1109/ICICT4SD50815.2021.9396809

[25] Li, E., Du, P., Samat, A., Meng, Y., Che, M.: Mid-level feature representation via sparse autoencoder for remotely sensed scene classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **10**(3), 1068–1081 (2016)

[26] Chen, Y., Lin, Z., Zhao, X., Wang, G., Gu, Y.: Deep learning-based classification of hyperspectral data. IEEE Journal of Selected topics in applied earth observations and remote sensing **7**(6), 2094–2107 (2014)

[27] Otter, D.W., Medina, J.R., Kalita, J.K.: A survey of the usages of deep learning for natural language processing. IEEE Transactions on Neural Networks and Learning Systems **32**(2), 604–624 (2021) https://doi.org/10.1109/TNNLS.2020.2979670

[28] Chai, J., Zeng, H., Li, A., Ngai, E.W.: Deep learning in computer vision: A critical review of emerging techniques and application scenarios. Machine Learning with Applications **6**, 100134 (2021)

[29] Zhang, Q., Yang, L.T., Chen, Z., Li, P.: A survey on deep learning for big data. Information Fusion **42**, 146–157 (2018)

[30] Zhao, W., Du, S.: Spectral–spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach. IEEE Transactions on Geoscience and Remote Sensing **54**(8), 4544–4554 (2016)

[31] Li, W., Wu, G., Zhang, F., Du, Q.: Hyperspectral image classification using deep pixel-pair features. IEEE Transactions on Geoscience and Remote Sensing **55**(2), 844–853 (2016)

[32] Liu, Q., Xiao, L., Yang, J., Wei, Z.: Cnn-enhanced graph convolutional network with pixel-and superpixel-level feature fusion for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing **59**(10), 8657–8671 (2020)

[33] Zhang, H., Li, Y., Zhang, Y., Shen, Q.: Spectral-spatial classification of hyperspectral imagery using a dual-channel convolutional neural network. Remote sensing

letters **8**(5), 438–447 (2017)

[34] Lee, H., Kwon, H.: Going deeper with contextual cnn for hyperspectral image classification. IEEE Transactions on Image Processing **26**(10), 4843–4855 (2017)

[35] Feng, F., Wang, S., Wang, C., Zhang, J.: Learning deep hierarchical spatial–spectral features for hyperspectral image classification based on residual 3d-2d cnn. Sensors **19**(23), 5276 (2019)

[36] Jia, S., Zhao, B., Tang, L., Feng, F., Wang, W.: Spectral–spatial classification of hyperspectral remote sensing image based on capsule network. The Journal of Engineering **2019**(21), 7352–7355 (2019)

[37] Chen, C., Zhang, J.-J., Zheng, C.-H., Yan, Q., Xun, L.-N.: Classification of hyperspectral data using a multi-channel convolutional neural network. In: Intelligent Computing Methodologies: 14th International Conference, ICIC 2018, Wuhan, China, August 15-18, 2018, Proceedings, Part III 14, pp. 81–92 (2018). Springer

[38] Hao, S., Wang, W., Ye, Y., Nie, T., Bruzzone, L.: Two-stream deep architecture for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing **56**(4), 2349–2361 (2017)

[39] Gong, H., Farooque, G., Khader, A., Xiao, L.: Multiscale semantic alignment graph convolution network for single-shot learning based hyperspectral image classification. In: Fourteenth International Conference on Graphics and Image Processing (ICGIP 2022), vol. 12705, pp. 462–473 (2023). SPIE

[40] Chen, Y., Jiang, H., Li, C., Jia, X., Ghamisi, P.: Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. IEEE transactions on geoscience and remote sensing **54**(10), 6232–6251 (2016)

[41] Li, Y., Zhang, H., Shen, Q.: Spectral–spatial classification of hyperspectral imagery with 3d convolutional neural network. Remote Sensing **9**(1), 67 (2017)

[42] Zhao, F., Zhang, J., Meng, Z., Liu, H., Chang, Z., Fan, J.: Multiple vision architectures-based hybrid network for hyperspectral image classification. Expert Systems with Applications **234**, 121032 (2023)

[43] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision, pp. 213–229 (2020). Springer

[44] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning, pp. 10347–10357 (2021). PMLR

[45] Parvaiz, A., Khalid, M.A., Zafar, R., Ameer, H., Ali, M., Fraz, M.M.: Vision transformers in medical computer vision—a contemplative retrospection. Engineering Applications of Artificial Intelligence **122**, 106126 (2023)

[46] Hong, D., Han, Z., Yao, J., Gao, L., Zhang, B., Plaza, A., Chanussot, J.: Spectralformer: Rethinking hyperspectral image classification with transformers. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–15 (2022) https://doi.org/10.1109/TGRS.2021.3130716

[47] Ullah, W., Hussain, T., Ullah, F.U.M., Lee, M.Y., Baik, S.W.: Transcnn: Hybrid cnn and transformer mechanism for surveillance anomaly detection. Engineering Applications of Artificial Intelligence **123**, 106173 (2023)

[48] Sun, L., Zhao, G., Zheng, Y., Wu, Z.: Spectral–spatial feature tokenization transformer for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–14 (2022)

[49] Roy, S.K., Deria, A., Shah, C., Haut, J.M., Du, Q., Plaza, A.: Spectral–spatial morphological attention transformer for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–15 (2023) https://doi.org/10.1109/TGRS.2023.3242346

[50] Yu, H., Xu, Z., Zheng, K., Hong, D., Yang, H., Song, M.: Mstnet: A multilevel spectral–spatial transformer network for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–13 (2022) https://doi.org/10.1109/TGRS.2022.3186400

[51] He, W., Huang, W., Liao, S., Xu, Z., Yan, J.: Csit: A multiscale vision transformer for hyperspectral image classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **15**, 9266–9277 (2022) https://doi.org/10.1109/JSTARS.2022.3216335

[52] Zhang, B., Chen, Y., Rong, Y., Xiong, S., Lu, X.: Matnet: A combining multi-attention and transformer network for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–15 (2023)

[53] Luo, F., Huang, H., Duan, Y., Liu, J., Liao, Y.: Local geometric structure feature for dimensionality reduction of hyperspectral imagery. Remote Sensing **9**(8), 790 (2017)

[54] Hu, W., Huang, Y., Wei, L., Zhang, F., Li, H.: Deep convolutional neural networks for hyperspectral image classification. Journal of Sensors **2015**, 1–12 (2015)

[55] Ahmad, M., Khan, A.M., Mazzara, M., Distefano, S., Ali, M., Sarfraz, M.S.: A fast and compact 3-d cnn for hyperspectral image classification. IEEE Geoscience and Remote Sensing Letters **19**, 1–5 (2020)

[56] Pande, S., Banerjee, B.: Hyperloopnet: Hyperspectral image classification using multiscale self-looping convolutional networks. ISPRS Journal of Photogrammetry and Remote Sensing **183**, 422–438 (2022)

[57] Xu, Q., Xiao, Y., Wang, D., Luo, B.: Csa-mso3dcnn: Multiscale octave 3d cnn with channel and spatial attention for hyperspectral image classification. Remote Sensing **12**(1), 188 (2020)

[58] Liu, Q., Xiao, L., Huang, N., Tang, J.: Composite neighbor-aware convolutional metric networks for hyperspectral image classification. IEEE Transactions on Neural Networks and Learning Systems (2022)

[59] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[60] Haut, J.M., Paoletti, M.E., Plaza, J., Plaza, A., Li, J.: Visual attention-driven hyperspectral image classification. IEEE transactions on geoscience and remote sensing **57**(10), 8065–8080 (2019)

[61] Chen, Y., Kalantidis, Y., Li, J., Yan, S., Feng, J.: Aˆ 2-nets: Double attention networks. Advances in neural information processing systems **31** (2018)

[62] Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)

[63] Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: Eca-net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11534–11542 (2020)

[64] Mou, L., Zhu, X.X.: Learning to pay attention on spectral domain: A spectral attention module-based convolutional network for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing **58**(1), 110–122 (2019)

[65] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

[66] 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, pp. 10012–10022 (2021)

[67] Xu, Y., Du, B., Zhang, L.: Beyond the patchwise classification: Spectral-spatial fully convolutional networks for hyperspectral image classification. IEEE Transactions on Big Data **6**(3), 492–506 (2020) https://doi.org/10.1109/TBDATA.2019.2923243

[68] Yin, J., Qi, C., Huang, W., Chen, Q., Qu, J.: Multibranch 3d-dense attention network for hyperspectral image classification. IEEE Access **10**, 71886–71898 (2022) https://doi.org/10.1109/ACCESS.2022.3188853

[69] Makantasis, K., Karantzalos, K., Doulamis, A., Doulamis, N.: Deep supervised learning for hyperspectral data classification through convolutional neural networks. In: 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 4959–4962 (2015). IEEE

[70] Kanthi, M., Sarma, T.H., Bindu, C.S.: A 3d-deep cnn based feature extraction and hyperspectral image classification. In: 2020 IEEE India Geoscience and Remote Sensing Symposium (InGARSS), pp. 229–232 (2020). IEEE

[71] Shu, Z., Wang, Y., Yu, Z.: Dual attention transformer network for hyperspectral image classification. Engineering Applications of Artificial Intelligence **127**, 107351 (2024)

[72] Jiang, N., Geng, S., Zheng, Y., Sun, L.: Msdca: A multi-scale dual-branch network with enhanced cross-attention for hyperspectral image classification. Remote Sensing **17**(13) (2025) https://doi.org/10.3390/rs17132198

[73] Farooque, G., Liu, Q., Sargano, A.B., Xiao, L.: Swin transformer with multiscale 3d atrous convolution for hyperspectral image classification. Engineering Applications of Artificial Intelligence **126**, 107070 (2023)