Spatial Spectral Transformer with Conditional Position Encoding for Hyperspectral Image Classification

Muhammad Ahmad, Muhammad Usama, Adil Mehmood Khan, Salvatore Distefano, Hamad Ahmed Altuwaijri, and Manuel Mazzara

***Abstract*—In Transformer-based Hyperspectral Image Classification (HSIC), predefined positional encodings (PEs) are crucial for capturing the order of each input token. However, their typical representation as fixed-dimension learnable vectors makes it challenging to adapt to variable-length input sequences, thereby limiting the broader application of Transformers for HSIC. To address this issue, this study introduces an implicit conditional PEs (CPEs) scheme in a Transformer for HSIC, conditioned on the input token's local neighborhood. The proposed SSFormer integrates spatial-spectral information and enhances classification performance by incorporating a CPE mechanism, thereby increasing the Transformer layers' capacity to preserve contextual relationships within the HSI data. Moreover, SSFormer ensembles the cross-attention between patches and proposed learnable embeddings. This enables the model to capture global and local features simultaneously while addressing the constraint of limited training samples in a computationally efficient manner. Extensive experiments on publicly available HSI benchmarking datasets were conducted to validate the effectiveness of the proposed SSFormer model. The results demonstrated remarkable performance, achieving classification accuracies of 97.7% on the Indian Pines dataset and 96.08% on the University of Houston dataset.**

***Index Terms*—Spatial Spectral Transformer (SSFormer); Hyperspectral Image Classification (HSIC).**

I. Introduction

HYPERSPECTRAL IMAGING (HSI) has gained significant attention due to its ability to capture fine-grained spectral information, providing a wealth of data across multiple domains [1]–[7]. However, accurately classifying HSI data poses challenges due to its high dimensionality, spectral variability, complex spatial patterns, and Hughes phenomenon [8]–[10]. Several efforts have focused on developing accurate and computationally efficient algorithms to extract rich spectral information [11]. Recent works have developed effective classification frameworks based on jointly exploiting spectral-spatial features [12]–[14].

M. Ahmad and M. Usama are with the Department of Computer Science, National University of Computer and Emerging Sciences, Chiniot 35400, Pakistan. e-mail: mahmad00@gmail.com

A. M. Khan is with the School of Computer Science, University of Hull, Hull HU6 7RX, UK. e-mail: a.m.khan@hull.ac.uk

S. Distefano and M. Ahmad are with Dipartimento di Matematica e Informatica—MIFT, University of Messina, Messina 98121, Italy. e-mail: sdistefano@unime.it

H.A. Altuwaijri is with the Department of Geography, College of Humanities and Social Sciences, King Saud University, Riyadh, 11451 Saudi Arabia. e-mail: Haaltuwaijri@ksu.edu.sa

M. Mazzara is with the Institute of Software Development and Engineering, Innopolis University, 420500 Innopolis, Russia. e-mail: m.mazzara@innopolis.ru

Recent advances in Transformer-based models have demonstrated remarkable success in HSIC by employing self-attention mechanisms to process image patches directly [15]–[22]. For instance, combining spectral-spatial kernels with improved SST enables the joint extraction of spectral-spatial features [23]. Lifan et al. [24] introduced a spatial-spectral hierarchical Vision Transformer (ViT), while X. He et al. [25] proposed an SST that jointly exploits spectral-spatial information through a dedicated module. Hyper-ES2T and CSiT achieved state-of-the-art performance on several datasets using Transformer-based models [26], [27], and a spatial-spectral Transformer has been utilized for HSI denoising [28]. Despite their effectiveness, these models often face computational constraints and challenges in handling variable-length input sequences. Transformers rely on predetermined positional encodings (PEs) to capture the sequential order of input tokens, typically represented as learnable fixed-dimensional vectors. However, they struggle to adapt to variable-length input sequences, which is crucial for generating meaningful outputs from HSI data. This limitation restricts the broader applicability of Transformers.

The Spatial-Spectral Transformer (SSFormer) proposed in this study addresses the aforementioned challenges by effectively integrating spatial-spectral patterns through Conditional Positional Encoding (CPE) and cross-attention mechanisms. The key contributions can be summarized as:

1) Unlike traditional fixed-dimension encodings, the CPE is conditioned on the input token's local neighborhood, enabling the model to process varying input sequence lengths. This provides the Transformer with crucial pixel arrangement and sequence knowledge, allowing for better differentiation of pixels based on their relative positions and preserving the underlying spatial associations.
2) Cross-attention mechanisms enhance the ability to capture spatial context and relationships between pixels by combining positional information with spectral data. This integration improves the overall accuracy and efficiency of the model.
3) SSFormer leverages a backbone CNN to extract joint spatial-spectral features from the HSI cube, which are then encoded in the Transformer alongside CPEs. This combination enhances the model's ability to exploit spatial-spectral patterns, leading to improved generalization performance.

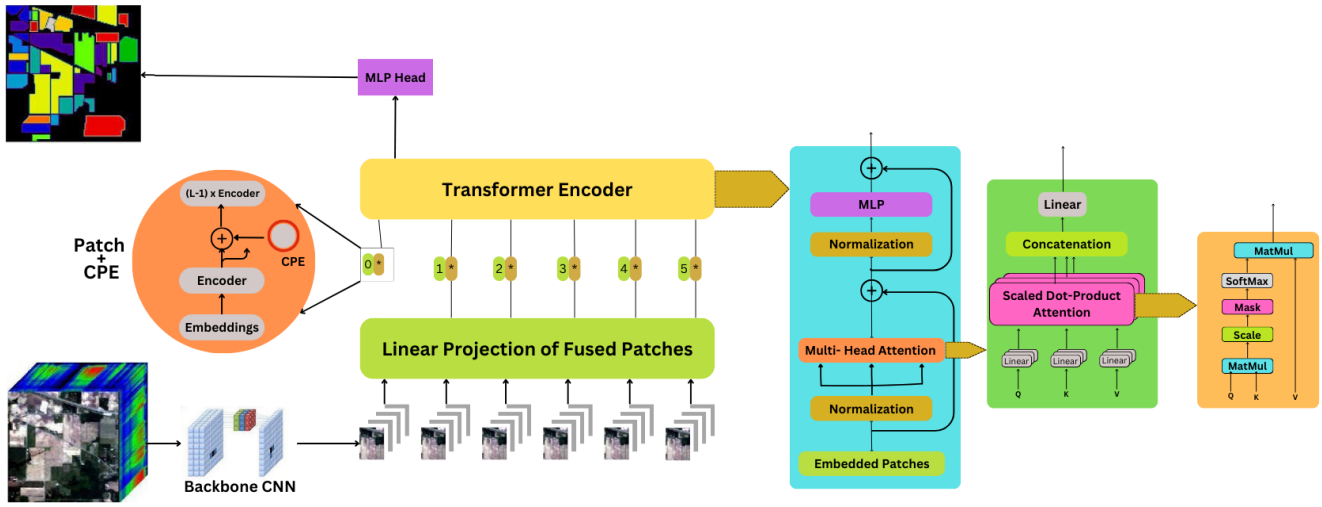


Fig. 1: SSFormer: A backbone CNN extracts joint spatial-spectral features from the HSI cube. These are encoded in the Transformer alongside CPEs, which selectively represent image information better than standard fixed-dimension encodings. The Transformer output then feeds into an MLP head for ground truth generation.

II. Proposed Methodology

To transform a 3D HSI cube $X \in \mathcal{R}^{M\times N\times B}$ into small 3D cubes suitable for the Transformer, the model first divides HSI into overlapping small patches. So, the model reproduces X into a sequence of patches which are then linearly projected into lower-dimensional embedding using a learnable weight matrix $W \in \mathcal{R}^{(D\times N\times S\times S)}$, where D, H, and S are embedding dimension, liner projection, and patch size, respectively:

$$H(n,d,i,j) = \sum_{m}\sum_{k}\sum_{l} X_{patches} * W(d,n,k,l) \quad (1)$$

$$X_{patches} = \phi\Bigg(b_{i,j} + \sum_{\tau=1}^{d_{l-1}}\sum_{\sigma=-\delta}^{\delta}\sum_{\lambda=-v}^{v}\sum_{\rho=-\gamma}^{\gamma} w_{i,j,\tau}^{\sigma,\rho,\lambda} \times v_{i-1,\tau}^{x+\sigma,y+\rho,z+\lambda}\Bigg) \quad (2)$$

where d_{l-1}, $b_{i,j}$, and $w_{i,j}$ correspond to the number of feature maps, bias parameter, and depth of the kernel for the j^{th} feature map at the $(l-1)^{th}$ layer, respectively. Additionally, $2v+1$, $2\gamma+1$, and $2\sigma+1$ represent the depth, width, and height of the kernel of 3D CNN as the backbone model [29]. To better incorporate spatial information, CPEs are added, i.e., for each position (i,j) in the patch grid, we generate a CPE $\mathcal{F} \in \mathcal{R}^{D}$. The CPE is added element-wise (where deemed necessary) to the embeddings, resulting in $H_{pos} \in \mathcal{R}^{(M\times D\times (P//S)\times (Q//S))}$:

$$H_{pos}(m,d,i,j) = H(m,d,i,j) + \mathcal{F}(pos,i) \quad (3)$$

where pos is the position of the element in the sequence and i is the index of the dimension in the embedding vector. The $\mathcal{F}$ provides a unique CPE value for each position and dimension within the input sequence. For instance, to create the reference points, features undergo boundary padding, as an illustrative example, when conducting convolutions on feature maps, the use of zero padding serves to denote the position of the boundary point. The combination of the above CPEs with the patch data allows the SSFormer to learn and capture positional information, ensuring that it can effectively process sequences with different orders or positions. $\mathcal{F}$ can be viewed as a local regularization applied to $X_{patches}$. As the class token does not incorporate position information, it remains unaltered. Consequently, the ultimate output is created by concatenating the unmodified class token and the regularized $X_{patches}$ along the last dimension. The SSFormer employs multi-head attention to capture contextual relationships within the patches. To capture both spatial-spectral information, cross-attention is performed between the output (O_{att}) and the original embeddings (H_{pos}).

$$Attention^{c}(Q_c, K_c, V_c) = \left(\frac{Q_c \times K_c^{T}}{\sqrt{D}}\right) \times V_c \quad (4)$$

where K_c, V_c, and Q_c are the query, key, and value matrices for cross-attention, respectively. The output of the cross-attention is represented as:

$$O_{cross-att} = Attention^{c}(O_{att}, H_{pos}, H_{pos}) \quad (5)$$

The cross-attention output is fed into a feed-forward neural network with two fully connected layers to further process the features. Global average pooling is applied along the spatial dimensions to obtain a fixed-size representation for each band. The output is reshaped back into patches and concatenated to form the final output Y. The softmax function generates class probabilities to produce ground truth maps. The SSFormer effectively handles the HSI cube with reduced complexity by integrating spatial-spectral information through attention mechanisms and projections, making it suitable for resource-constrained environments. The complete model is demonstrated in Figure 1.

$$Y(m,d,i,j) = FFN(O_{cross-att}(m,d,i,j)) \quad (6)$$

III. Experimental Results and Discussion

SSFormer's weights were initially randomized and then optimized over 50 epochs using the Adam optimizer (learning rate 0.0001) and softmax loss. Training occurred in mini-batches of 256 samples per epoch, enabling the SSFormer to learn patterns from repeated exposure to training data and adjust its parameters to reduce loss over multiple iterations.

A. Patch Size and Train-Validation-Test Samples

Two key parameters, train/test split, and patch size, were meticulously examined. Various combinations of these parameters were tested and evaluated to determine the optimal configuration for achieving the best results. The model's performance with different training and test percentages is tested as is shown in Table I, with a fixed patch size of 9×9. Furthermore, Table II details the results for using various patch sizes.

TABLE I: We assess the performance of the proposed model using a 9×9 patch size with various Train/Validation/Test (Tr/Val/Te) splits.

Tr/Val/Te	AA	OA	κ	Tr Time	Te Time
Indian Pines Dataset					
5/5/90	64.44	74.60	70.82	54.49	5.89
10/10/80	77.12	81.92	79.24	66.75	5.92
15/15/70	82.42	86.41	84.42	80.31	3.10
20/20/60	87.97	89.78	88.33	149.64	3.62
Pavia University Dataset					
5/5/90	93.71	96.05	94.74	153.73	21.89
10/10/80	96.46	97.75	97.02	210.88	21.53
15/15/70	97.28	98.26	97.70	204.77	11.22
20/20/60	97.89	98.64	98.20	270.83	8.50
University of Houston Dataset					
5/5/90	94.93	96.08	95.77	46.87	6.2470
10/10/80	96.68	97.38	97.17	65.45	11.2144
15/15/70	97.63	98.09	97.94	154.74	6.2593
20/20/60	98.09	98.68	98.57	109.03	3.7363

TABLE II: Proposed model's performance using a fixed Training(40%)/Validation(40%)/Test(20%) split and different patch sizes.

Patch Size	AA	OA	κ	Tr Time	Te Time	Parameters
Indian Pines Dataset						
3×3	85.35	88.29	86.68	184.39	1.53	784,016
5×5	92.98	93.31	92.37	210.57	2.04	799,376
7×7	94.90	94.09	93.26	211.32	2.03	822,416
9×9	96.53	96.14	95.60	175.90	1.53	853,136
Pavia University Dataset						
3×3	95.52	96.76	95.71	213.72	21.21	782,217
5×5	96.04	97.33	96.46	154.42	11.82	797,577
7×7	96.41	97.51	96.71	149.17	11.69	820,617
9×9	98.39	99.06	98.76	574.58	3.26	851,337
University of Houston Dataset						
3×3	97.46	97.70	97.52	197.73	1.57	779,217
5×5	99.06	99.26	99.20	214.27	1.58	799,119
7×7	99.34	99.40	99.35	212.12	2.04	822,159
9×9	99.13	99.36	99.31	214.47	2.30	852,879

The results indicate that SSFormer excels at capturing minute details and local patterns with smaller patch sizes. However, sensitivity to noise and fluctuations may lead to overfitting, especially with smaller sample sizes. In contrast, larger patch sizes enable SSFormer to encapsulate global features and contextual information, exhibiting enhanced robustness against noise perturbations. As sample sizes increase, SSFormer's robustness and generalization capabilities improve. The optimal patch size depends on the specific sample size, with smaller sample sizes benefiting from larger patch sizes to capture essential features, and larger sample sizes leveraging smaller patch sizes for precise pattern extraction. Increasing the training set size improved performance, particularly with larger patch sizes, by enabling the model to learn complex patterns and reduce overfitting. A moderate validation sample size is crucial for proper hyperparameter tuning. Too small a validation set might lead to suboptimal parameter choices, while too large a set might increase computation time without significant performance gains. Additionally, extremely small test sets might not provide a reliable estimation of SSFormer's generalization ability.

B. Experimental Results with CNN-based Models

A consistent experimental methodology is crucial for fair model comparison. In this paper, we utilized unique sample distributions, geographical locations, and a controlled 9×9 pixel patch size across models. This controlled experimental setup allows for a fair comparison of competing CNN architectures. Figure 2 shows loss and accuracy trends on the University of Houston dataset. The ground truth maps for comparative and SSFormer are presented in Figures 3, 4, and 5.

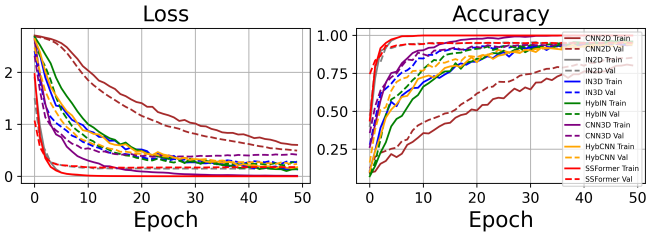
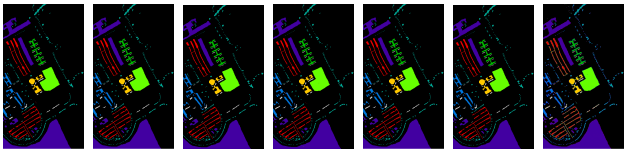


Fig. 2: The comparison above shows the accuracy and loss trends of CNN-based methods on the Houston dataset. The SSFormer (red line) demonstrates quicker convergence compared to traditional CNN models.

(a) [30] (b) [31] (c) [32] (d) [33] (e) [29] (f) [34] (g) SSF

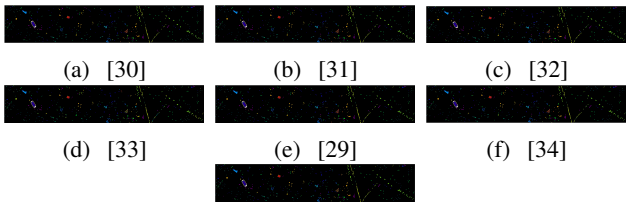
Fig. 3: **Indian Pines Dataset:** The proposed SSFormer achieves OA=97.07% showing competitive performance.



(a) [30] (b) [31] (c) [32] (d) [33] (e) [29] (f) [34] (g) SSF

Fig. 4: **Pavia University Dataset:** The proposed SSFormer achieves OA=99.39% showing competitive performance.

Models evaluated include 3D CNN [29], Hybrid Inception Net [32], 3D Inception Net [31], 2D Inception Net [30], 2D CNN [33], and Hybrid CNN [34]. The findings highlight the advantage of decoupling spatial and spectral information over



(a) [30] (b) [31] (c) [32]

(d) [33] (e) [29] (f) [34]

(g) SSF

Fig. 5: **University of Houston Dataset:** The proposed SSFormer achieves OA=95.99% showing competitive performance.

alternative strategies. SSFormer achieves impressive scores of 97% and performs on par with 3D models. Models employing 2D convolutions exhibit varying performance results on the Indian Pines dataset, with 2D CNN achieving scores of 89% and 2D Inception Net attaining scores of 96%. The marginal difference in accuracy between the SSFormer and comparative CNN-based methods can be attributed to the computational efficiency, overfitting mitigation, and enhanced spatial, and spectral dependencies modeling of the SSFormer. Similar observations can be made for other experimental datasets.

C. Comparative Results with Transformer-Based Networks

Several recently proposed Transformer-based networks for HSIC were selected for comparison with the SSFormer, including Attention is all you need (ViT) [35], Rethinking Hyperspectral Image Classification with Transformers (Spectralformer – SF) [22], Hyperspectral Image Transformer Classification Networks (HiT) [36], A Multiscale Vision Transformer for HSIC (CSiT) [27], Spectral–Spatial Feature Tokenization Transformer for HSIC (SSFTT) [37], Spectrum Transformer for Self-Supervised Learning in HSIC (S3L) [38], Rotation-Invariant Attention Network for HSIC (RIAN) [39], Center Attention Transformer With Stratified Spatial–Spectral Token for HSIC (CAT) [40], Local Transformer With Spatial Partition Restore for HSIC (SPRLT) [41], A Center-Masked Transformer for HSIC (CMT) [42].

A classical SST architecture was adopted and redesigned to accommodate pixel-wise HSI input. The Spectralformer network retained the original architecture, while the HiT architecture incorporated two depth-wise convolution layers for spatial information processing and a point-wise convolution layer for spectral information processing. HiT relies solely on patch-wise input. The CSiT architecture employed a similar backbone Transformer architecture in each branch, with a stack of multiscale Transformer encoders. The CSiT was evaluated both with and without Cross-spectral Attention Fusion (CSAF) modules for insightful comparisons.

The detailed comparison of the aforementioned models is presented in Tables III, IV and V. For the IP dataset, the proposed SSFormer demonstrates superior performance and outperforms all comparative models by wide margins. It surpasses the baseline SST by more than 19% and the SpectralFormer by approximately 18%. For the University of Houston dataset, the proposed SSFormer outperforms other compared models for several classes in the dataset, as shown in Table V. Additionally, the SSFormer achieves better AA, OA, and κ values, emphasizing the potential efficacy of spatial-spectral feature extraction.

TABLE III: Classification accuracy (%) and Kappa (κ) measure for **Indian Pines Dataset**.

Class	ViT	SF	HiT	CSiT	SSFTT	RIAN	CAT	S3L	SPRLT	CMT	SSFormer
1	76.47%	81.82%	80.95%	81.82%	**100.00%**	76.92%	94.87%	88.47%	98.78%	97.32%	**100.00%**
2	77.18%	76.70%	75.11%	81.41%	87.54%	88.58%	84.75%	94.82%	45.56%	61.49%	**95.80%**
3	58.63%	76.11%	57.05%	77.79%	86.22%	79.97%	92.47%	95.27%	51.99%	46.96%	**87.95%**
4	69.16%	58.60%	81.95%	81.44%	98.94%	77.72%	99.46%	84.13%	75.09%	85.69%	**91.66%**
5	84.70%	75.34%	79.39%	86.08%	92.08%	86.13%	93.29%	96.70%	66.05%	68.20%	**100.00%**
6	83.60%	76.69%	93.85%	90.20%	97.11%	82.21%	97.85%	**100.00%**	94.08%	98.34%	**100.00%**
7	62.50%	57.14%	87.50%	40.91%	**100.00%**	**100.00%**	90.91%	94.40%	**100.00%**	**100.00%**	**100.00%**
8	92.83%	93.54%	100%	97.48%	99.07%	99.54%	99.09%	93.16%	99.75%	**100.00%**	**100.00%**
9	19.23%	16.67%	12.50%	57.14%	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
10	80.85%	76.71%	75.95%	77.02%	88.70%	89.11%	90.52%	85.74%	63.31%	81.44%	**90.72%**
11	80.30%	76.06%	76.53%	81.48%	84.86%	88.17%	95.24%	88.69%	52.71%	71.44%	**99.59%**
12	62.45%	75.71%	66.60%	72.45%	76.18%	80.85%	90.25%	96.77%	42.60%	77.35%	**98.30%**
13	86.96%	79.36%	74.77%	98.36%	98.90%	92.59%	92.59%	92.53%	99.65%	99.15%	**100.00%**
14	84.56%	89.44%	91.87%	92.57%	97.81%	98.87%	9895%	**100.00%**	84.23%	92.44%	**99.20%**
15	71.81%	61.36%	78.65%	75.60%	94.51%	95.45%	98.79%	94.47%	89.37%	97.35%	**100.00%**
16	89.87%	93.51%	90.12%	90.24%	**100.00%**	97.78%	97.78%	88.73%	**100.00%**	**100.00%**	**100.00%**
OA	77.84%	78.07%	75.01%	83.40%	91.14%	88.74%	93.69%	93.45%	65.32%	77.34%	**97.07%**
AA	74.73%	75.82%	65.94%	82.27%	91.68%	89.62%	94.80%	92.20%	78.95%	86.07%	**97.70%**
κ	74.70%	74.86%	71.73%	81.02%	89.24%	87.11%	92.75%	92.78%	61.13%	74.56%	**96.66%**

TABLE IV: Classification accuracy (%) and Kappa (κ) measure for **Pavia University**.

Class	ViT	SF	HiT	CSiT	SSFTT	RIAN	CAT	S3L	SPRLT	CMT	SSFormer
1	95.67%	92.67%	95.21%	95.25%	88.58%	90.44%	94.75%	98.76%	78.09%	76.94%	**96.34%**
2	88.37%	92.79%	92.54%	96.64%	96.79%	96.63%	95.54%	95.46%	84.67%	93.70%	**99.61%**
3	73.71%	90.60%	91.18%	81.74%	96.60%	93.50%	**99.83%**	87.75%	53.47%	81.10%	82.74%
4	98.03%	**98.15%**	97.21%	95.50%	87.37%	90.94%	93.99%	91.37%	75.09%	78.42%	96.33%
5	99.01%	98.28%	100.0%	99.59%	99.62%	98.71%	95.82%	92.13%	99.83%	100.00%	**100.00%**
6	89.26%	93.29%	**99.93%**	93.43%	94.94%	98.82%	97.32%	97.10%	52.40%	64.24%	92.31%
7	79.40%	83.01%	**95.75%**	89.24%	92.63%	98.13%	96.58%	96.47%	78.74%	96.27%	92.06%
8	85.54%	84.50%	**97.59%**	88.81%	87.68%	93.95%	95.37%	86.25%	84.35%	58.58%	90.67%
9	99.65%	**99.77%**	99.47%	99.53%	94.04%	95.24%	92.18%	89.27%	99.60%	98.16%	93.30%
OA	89.32%	92.30%	91.35%	94.48%	90.65%	95.25%	95.05%	94.64%	78.23%	83.28%	**96.05%**
AA	87.39%	88.86%	85.07%	93.15%	93.44%	95.15%	93.01%	92.78%	78.47%	83.06%	**93.71%**
κ	86.60%	89.66%	88.94%	92.67%	91.73%	93.55%	94.70%	93.54%	71.55%	77.74%	**94.74%**

TABLE V: Classification accuracy (%) and Kappa (κ) measure for **University of Houston**.

Class	ViT	SF	HiT	CSiT	SSFTT	RIAN	CAT	S3L	SPRLT	CMT	SSFormer
1	90.28%	93.31%	**97.26%**	93.39%	91.78%	83.00%	88.13%	81.23%	80.51%	79.47%	**97.24%**
2	98.21%	97.81%	97.29%	**99.54%**	78.42%	83.74%	90.04%	87.59%	67.45%	79.08%	97.69%
3	96.90%	**100.00%**	98.74%	**100.00%**	80.57%	89.70%	**100.00%**	88.19%	99.12%	100.00%	97.76%
4	**100.00%**	**100.00%**	95.78%	98.09%	90.98%	92.57%	89.49%	77.05%	97.39%	98.51%	97.85%
5	96.89%	98.16%	98.41%	97.70%	87.08%	99.81%	98.20%	90.09%	96.64%	**100.00%**	99.91%
6	98.21%	**100.00%**	91.36%	**100.00%**	90.92%	**100.00%**	99.30%	93.94%	84.66%	87.53%	92.46%
7	82.82%	87.83%	**94.60%**	90.96%	93.02%	87.59%	98.97%	87.53%	78.78%	77.00%	90.70%
8	79.83%	85.91%	91.82%	89.18%	92.36%	73.22%	**98.01%**	93.02%	45.49%	49.77%	96.78%
9	76.88%	75.33%	92.39%	90.62%	93.02%	81.21%	**98.94%**	86.12%	76.05%	72.19%	94.85%
10	80.31%	82.52%	90.61%	93.22%	82.60%	68.15%	98.65%	78.68%	36.93%	94.51%	**99.27%**
11	83.17%	79.19%	89.09%	87.91%	77.68%	89.85%	87.38%	81.65%	64.05%	85.59%	**96.13%**
12	69.42%	72.76%	94.18%	83.15%	70.36%	89.15%	94.33%	89.71%	70.01%	59.63%	**99.18%**
13	63.08%	79.49%	82.51%	84.11%	91.82%	92.28%	98.95%	85.49%	**98.81%**	80.06%	65.16%
14	90.36%	93.90%	91.55%	97.21%	91.71%	**100.00%**	98.79%	93.02%	99.81%	**100.00%**	98.96%
15	94.40%	97.50%	96.72%	**100.00%**	89.87%	**100.00%**	62.16%	88.33%	96.41%	99.57%	**100.00%**
OA	86.45%	88.45	93.06%	93.09%	83.05%	86.30%	93.21%	86.50%	75.65%	82.14%	**96.08%**
AA	86.60%	87.81	86.61%	92.06%	84.40%	88.68%	93.22%	85.25%	79.34%	84.19%	**94.93%**
κ	85.35%	87.50	92.50%	92.53%	81.79%	85.14%	92.63%	85.85%	73.73%	80.70%	**95.76%**

IV. Conclusion

Hyperspectral Image Classification (HSIC) faces challenges due to high-dimensional data and spectral variability. To address these, this work proposed a Spatial Spectral Transformer (SSFormer) for HSIC. The SSFormer effectively combines spatial and spectral information, resulting in improved classification accuracy. By incorporating CPEs, the SSFormer better captures spatial information and enhances contextual relationships within the Transformer layers. Additionally, the integration of cross-attention between patch and learnable embeddings enables the simultaneous capture of global and local features, further enhancing performance. Experimental evaluation on publicly available HSI datasets demonstrated that the SSFormer outperforms SOTA methods, validating its effectiveness for HSIC tasks. Future research may explore how SSFormer and other Transformer-based models perform on different HSI datasets, expanding applicability across real-world applications. Additionally, it opens opportunities to investigate combining SSFormer with transfer learning and domain adaptation techniques.

Acknowledgment

The authors extend their appreciation to the Researchers Supporting Project number (RSPD2024R848), King Saud University, Riyadh, Saudi Arabia.

REFERENCES

[1] M. H. F. Butt, H. Ayaz, M. Ahmad, J. P. Li, and R. Kuleev, "A fast and compact hybrid cnn for hyperspectral imaging-based bloodstain classification," in *2022 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 2022, pp. 1–8.

[2] M. H. Khan, Z. Saleem, M. Ahmad, A. Sohaib, H. Ayaz, M. Mazzara, and R. A. Raza, "Hyperspectral imaging-based unsupervised adulterated red chili content transformation for classification: Identification of red chili adulterants," *Neural Computing and Applications*, vol. 33, no. 21, pp. 14 507–14 521, 2021.

[3] M. Zulfiqar, M. Ahmad, A. Sohaib, M. Mazzara, and S. Distefano, "Hyperspectral imaging for bloodstain identification," *Sensors*, vol. 21, no. 9, 2021.

[4] M. H. Khan, Z. Saleem, M. Ahmad, A. Sohaib, H. Ayaz, and M. Mazzara, "Hyperspectral imaging for color adulteration detection in red chili," *Applied Sciences*, vol. 10, no. 17, 2020.

[5] Z. Saleem, M. H. Khan, M. Ahmad, A. Sohaib, H. Ayaz, and M. Mazzara, "Prediction of microbial spoilage and shelf-life of bakery products through hyperspectral imaging," *IEEE Access*, vol. 8, pp. 176 986–176 996, 2020.

[6] H. Ayaz, M. Ahmad, M. Mazzara, and A. Sohaib, "Hyperspectral imaging for minced meat classification using nonlinear deep features," *Applied Sciences*, vol. 10, no. 21, 2020.

[7] H. Ayaz, M. Ahmad, A. Sohaib, M. N. Yasir, M. A. Zaidan, M. Ali, M. H. Khan, and Z. Saleem, "Myoglobin-based classification of minced meat using hyperspectral imaging," *Applied Sciences*, vol. 10, no. 19, 2020.

[8] M. Ahmad, S. Shabbir, S. K. Roy, D. Hong, X. Wu, J. Yao, A. M. Khan, M. Mazzara, S. Distefano, and J. Chanussot, "Hyperspectral image classification—traditional to deep models: A survey for future prospects," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 968–999, 2021.

[9] R. Hanachi, A. Sellami, I. Farah, and M. Dalla Mura, "Multi-view graph representation learning for hyperspectral image classification with spectral–spatial graph neural networks," *Neural Computing and Applications*, vol. 36, 12 2023.

[10] X. Liao, B. Tu, J. Li, and A. Plaza, "Class-wise graph embedding-based active learning for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–13, 2023.

[11] C. Pan, X. Jia, J. Li, and X. Gao, "Adaptive edge preserving maps in markov random fields for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 10, pp. 8568–8583, 2020.

[12] N. Wambugu, Y. Chen, Z. Xiao, K. Tan, M. Wei, X. Liu, and J. Li, "Hyperspectral image classification on insufficient-sample and feature learning using deep neural networks: A review," *International Journal of Applied Earth Observation and Geoinformation*, vol. 105, p. 102603, 2021.

[13] U. Ghous, M. S. Sarfraz, M. Ahmad, C. Li, and D. Hong, "Exnet: (2+1)d extreme xception net for hyperspectral image classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 5159–5172, 2024.

[14] M. Ahmad and M. Mazzara, "Scsnet: Sharpened cosine similarity-based neural network for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1–4, 2024.

[15] M. H. F. Butt, J. P. Li, M. Ahmad, and M. A. F. Butt, "Graph-infused hybrid vision transformer: Advancing geoai for enhanced land cover classification," *International Journal of Applied Earth Observation and Geoinformation*, vol. 129, p. 103773, 2024.

[16] X. Chen, S.-I. Kamata, and W. Zhou, "Hyperspectral image classification based on multi-stage vision transformer with stacked samples," in *TENCON 2021-2021 IEEE Region 10 Conference (TENCON)*. IEEE, 2021, pp. 441–446.

[17] J. Zou, W. He, and H. Zhang, "Lessformer: Local-enhanced spectral-spatial transformer for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[18] M. Li, Y. Fu, and Y. Zhang, "Spatial-spectral transformer for hyperspectral image denoising," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, 2023, pp. 1368–1376.

[19] P. Tang, M. Zhang, Z. Liu, and R. Song, "Double attention transformer for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[20] H. Yan, E. Zhang, J. Wang, C. Leng, A. Basu, and J. Peng, "Hybrid conv-vit network for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, 2023.

[21] M. Ahmad, U. Ghous, M. Usama, and M. Mazzara, "Waveformer: Spectral–spatial wavelet transformer for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1–5, 2024.

[22] D. Hong, Z. Han, J. Yao, L. Gao, B. Zhang, A. Plaza, and J. Chanussot, "Spectralformer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022, dOI: 10.1109/TGRS.2021.3130716.

[23] A. Wang, S. Xing, Y. Zhao, H. Wu, and Y. Iwahori, "A hyperspectral image classification method based on adaptive spectral spatial kernel combined with improved vision transformer," *Remote Sensing*, vol. 14, no. 15, 2022.

[24] L. Ji, Y. Shao, J. Liu, and L. Xiao, "Spatial-spectral hierarchical vision permutator for hyperspectral image classification," *European Journal of Remote Sensing*, vol. 56, no. 1, p. 2153747, 2023.

[25] X. He, Y. Chen, and Z. Lin, "Spatial-spectral transformer for hyperspectral image classification," *Remote Sensing*, vol. 13, no. 3, 2021.

[26] W. Wang, L. Liu, T. Zhang, J. Shen, J. Wang, and J. Li, "Hyper-es2t: Efficient spatial–spectral transformer for the classification of hyperspectral remote sensing images," *International Journal of Applied Earth Observation and Geoinformation*, vol. 113, p. 103005, 2022.

[27] W. He, W. Huang, S. Liao, Z. Xu, and J. Yan, "Csit: A multiscale vision transformer for hyperspectral image classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 9266–9277, 2022.

[28] M. Li, Y. Fu, and Y. Zhang, "Spatial-spectral transformer for hyperspectral image denoising," 2022.

[29] M. Ahmad, A. M. Khan, M. Mazzara, S. Distefano, M. Ali, and M. S. Sarfraz, "A fast and compact 3-d cnn for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, 2020.

[30] Z. Xiong, Y. Yuan, and Q. Wang, "Ai-net: Attention inception neural networks for hyperspectral image classification," in *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2018, pp. 2647–2650.

[31] X. Zhang, "Improved three-dimensional inception networks for hyperspectral remote sensing image classification," *IEEE Access*, vol. 11, pp. 32 648–32 658, 2023.

[32] H. Fırat, M. E. Asker, M. İ. Bayındır, and D. Hanbay, "Hybrid 3d/2d complete inception module and convolutional neural network for hyperspectral remote sensing image classification," *Neural Processing Letters*, vol. 55, no. 2, pp. 1087–1130, 2023.

[33] X. Yang, Y. Ye, X. Li, R. Y. Lau, X. Zhang, and X. Huang, "Hyperspectral image classification with deep learning models," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 9, pp. 5408–5423, 2018.

[34] S. Ghaderizadeh, D. Abbasi-Moghadam, A. Sharifi, N. Zhao, and A. Tariq, "Hyperspectral image classification using a hybrid 3d-2d convolutional neural networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 7570–7588, 2021.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023.

[36] X. Yang, W. Cao, Y. Lu, and Y. Zhou, "Hyperspectral image transformer classification networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

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

[38] H. Guo and W. Liu, "S3l: Spectrum transformer for self-supervised learning in hyperspectral image classification," *Remote Sensing*, vol. 16, no. 6, 2024.

[39] X. Zheng, H. Sun, X. Lu, and W. Xie, "Rotation-invariant attention network for hyperspectral image classification," *IEEE Transactions on Image Processing*, vol. 31, pp. 4251–4265, 2022.

[40] J. Feng, Q. Wang, G. Zhang, X. Jia, and J. Yin, "Cat: Center attention transformer with stratified spatial–spectral token for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024.

[41] Z. Xue, Q. Xu, and M. Zhang, "Local transformer with spatial partition restore for hyperspectral image classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 4307–4325, 2022.

[42] S. Jia, Y. Wang, S. Jiang, and R. He, "A center-masked transformer for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–16, 2024.