PS2former: Parallel Spatial-Spectral Transformer Network for Hyperspectral Image Classification

Guanliang Wan[1], Danqing Liu[2], Yunxin Liu[1], Tengyue Yang[1], Yanhui Guo[3]
{202333331060@stu.qhnu.edu.cn, liudanqing@cdut.edu.cn, 202333331047@stu.qhnu.edu.cn
344744841@qq.com, guoyanhui@sdwu.edu.cn}
[1]The College of Computer, Qinghai Normal University, Xining, CO 810008, China
[2]The College of Computer Science and Cyber Security, Chengdu University of Technology, Chengdu, CO 610059, China
[3]The School of Artificial Intelligence, Shandong Women's University, Ji'nan, CO 250300, China
Corresponding author: Yanhui Guo. These authors contributed equally.

Abstract. Recent advances in hyperspectral image (HSI) classification demonstrate the potential of hybrid architectures that combine convolutional neural networks (CNNs) with Transformer modules. Yet, existing approaches fail to fully exploit the joint spatial–spectral characteristics of HSIs, as they often focus on either local or global features extracted separately by CNNs or Transformers. To address this limitation, we propose PS2former, a parallel spatial–spectral Transformer network. Specifically, we design a Parallel Extraction Module (PEM) that simultaneously captures primary spatial and spectral information and integrates them through feature fusion. Furthermore, we introduce a Parallel Hybrid Transformer (PHT) to jointly model local details and global context. The PHT incorporates a hollow CNN for parallel extraction of local features and a convolution Transformer for long-range global dependency modeling. Extensive experiments on two benchmark datasets demonstrate that PS2former consistently outperforms several recent state-of-the-art methods in HSI classification.

Keywords: Convolutional Neural Networks (CNNs); Transformer; Feature fusion; hyperspectral image (HSI) classification.

1 Introduction

Hyperspectral image (HSI) is acquired through hyperspectral remote sensing technology and typically consists of dozens to hundreds of contiguous spectral bands, spanning a broad range from the visible to the infrared spectrum. Hyperspectral image classification involves assigning each pixel in the image to a predefined land-cover category by leveraging both spectral and spatial information[1]. Traditional HSI classification methods are primarily statistical, relying on spectral information alone. Representative examples include K-Nearest Neighbor (KNN)[2], Support Vector Machine (SVM)[3], and Linear Discriminant Analysis (LDA)[4]. In recent years, deep learning–based approaches have gained considerable attention for HSI classification, as they can

IIKI 2025, December 18-21, Chengdu, People's Republic of China

DOI 10.4108/eai.18-12-2025.2365300

effectively capture complex spectral patterns and integrate spatial context. Convolutional Neural Networks (CNNs), in particular, have been widely applied in HSI processing. 2-D CNN employ two-dimensional convolution operations to compress and extract HSI features[5],[6], making them effective for spatial feature extraction. In contrast, 3-D CNN utilize more complex 3-D convolution kernels to simultaneously exploit spectral and spatial information[7],[8], thereby enabling comprehensive modeling of the spatial–spectral correlations inherent in HSIs. More recently, Transformer-based models[9],[10],[11] have been introduced into HSI classification, owing to their strong capability for modeling long-range dependencies in spectral–spatial data. State-of-the-art approaches often adopt hybrid architectures that integrate the strengths of both CNN and Transformer[12],[13], thereby capturing fine-grained local details while simultaneously modeling the global context of HSI.

Some recent methods fail to fully exploit joint spatial–spectral representations at either the local or global scale, resulting in incomplete modeling of spatial–spectral band correlations for ground objects. To address the aforementioned challenges, we propose a Parallel Spatial–Spectral Transformer Network (PS2former) for HSI classification. The framework consists of a Parallel Extraction Module (PEM) and a Parallel Hybrid Transformer (PHT) connected in series. In the PEM, spectral extraction and spatial residual feature extraction are performed in parallel to effectively capture complementary spectral and spatial representations. The PHT integrates hollow convolutional neural network (hollow CNN) with convolution Transformers to simultaneously model local details and global contextual information. Finally, a dedicated feature fusion strategy and classification head are employed to produce the final classification results. The main contributions of this work are as follows:

1) We design a PEM consisting of two components: Spectral Extraction (SE) and Spatial Residual Extraction (SRE). SE employs three-dimensional convolution with varying kernel sizes to capture multi-scale spectral information, thereby reducing the impact of redundant spectral bands. SRE incorporates multiple residual connections to preserve spatial information continuity. By performing SE and SRE in parallel, the module can extract spatial and spectral features simultaneously while maintaining computational efficiency.

2) We construct a PHT that integrates a hollow CNN with a convolution Transformer to jointly model local and global contextual information. The hollow CNN applies convolutions with different dilation rates to enhance the network's ability to capture fine-grained local details. In the convolution Transformer, convolution is first used to generate the QKV matrices, after which two-dimensional convolution is applied to the key and value matrices to suppress noisy spectral bands and reduce computational overhead.

2 Methodology

2.1 Overall framework

The architecture of PS2former, illustrated in Fig. 1, consists of two main components: the PEM and the PHT. Within PEM, the Spectral Extraction (SE) integrates spectral information, whereas the Spatial Residual Extraction (SRE) captures spatial features. In PHT, a hollow convolutional

neural network and a convolution Transformer are employed to model local details and long-range dependencies in HSI. Finally, the extracted features are fused through a straightforward aggregation strategy.

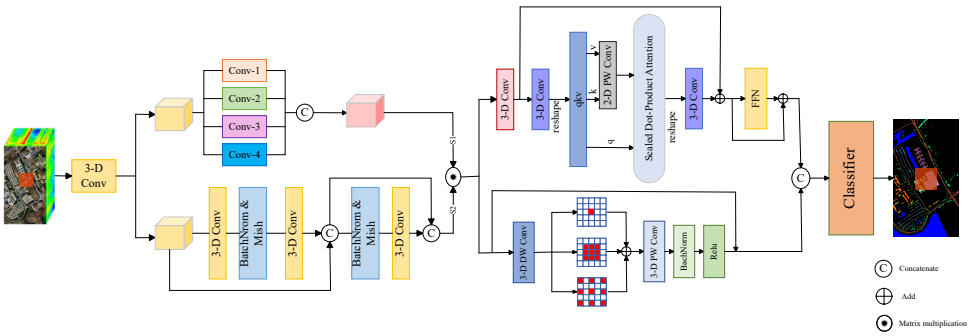

Fig. 1. Framework of the proposed PS2former.

2.2 Parallel Extraction Module (PEM)

The input patch of the original HSI is denoted as $X_{hsi} \in \mathbb{R}^{B\times 1\times H\times W\times S}$, where B is the batch size, 1 is the number of channels, H and W are the height and width of the image, and S is the number of spectral bands. A three-dimensional convolution is first applied to reduce the number of bands, resulting in $X'_{hsi} \in \mathbb{R}^{B\times C\times H\times W\times n}$, where $C > 1$ and $n < S$. This operation aims to remove redundant bands, preserve informative features, and balance computational efficiency. The processed tensor X'_{hsi} is then fed into two parallel branches. In the SE, the channels are uniformly partitioned into four groups, each processed with a convolution kernel of a distinct size to extract spectral features at multiple scales. The grouped features are denoted as $X_i \in \mathbb{R}^{B\times c_i\times H\times W\times n}$, and the convolution kernel size is defined as $K = (1,1,k_i)$, where $\{i \in \mathbb{N} \mid 1 \le i \le 4\}$. Batch normalization and ReLU activation are subsequently applied to stabilize the feature distribution, followed by concatenation along the channel dimension to produce the final multi-scale spectral representation. This process can be mathematically formulated as

$$X_{SE} = \text{Concat}(\sigma(\text{Conv}(X_i))). \tag{1}$$

where σ denotes the activation function. In the SRE, X'_{hsi} is first further compressed along the spectral dimension to emphasize spatial regions using a three-dimensional convolution. A second convolution is then applied within the module, after which the resulting feature maps are merged through a residual connection. Following a final convolution, another residual connection is employed to obtain the output feature maps. These two residual connections help preserve the integrity of the original features and mitigate the loss of valuable information. The output is denoted as

$X_{SRE} \in \mathbb{R}^{B\times C\times H\times W\times 1}$. Finally, the features from the two branches are fused to generate the final output, expressed as

$$X_{out} = X_{spa} \otimes X_{spe}. \tag{2}$$

where $\otimes$ denotes matrix multiplication.

2.3 Parallel Hybrid Transformer (PHT)

We design a Parallel Hybrid Transformer (PHT) to integrate convolution and Transformer-based representations. The features extracted by the PEM are fed into both a hollow CNN and a convolution Transformer, whose outputs are unified across dimensions via depth-wise three-dimensional convolution. These unified features are concatenated along the channel dimension and subsequently processed with point-wise three-dimensional convolution to reduce channel dimensionality, enabling cross-channel feature fusion. Finally, the classification header applies global pooling followed by linear mapping to generate the classification results.

In the hollow convolutional neural network (CNN), three distinct convolution operations are employed to capture local detail representations. Specifically, we use a 1×1 convolution with a dilation rate of 0, a 3×3 convolution with a dilation rate of 0, and a 3×3 convolution with a dilation rate of 1. This design extracts local features at multiple scales, expands the spatial receptive field while preserving spectral continuity, and enables the fusion of local spatial–spectral features. Finally, a residual connection is added to preserve feature integrity and produce the final output.

First, the QKV matrices are generated using a $3 \times 3 \times 1$ three-dimensional convolutional layer applied within each head channel. This design ensures that each convolution operates only across a single spectral band, meaning that each band forms an independent token. Replacing the conventional linear projection with this convolutional mapping provides two advantages. The sliding-window mechanism of convolution naturally encodes local positional relationships, removing the need for additional positional encoding. Convolution can also adaptively extract features at different spatial–spectral scales, whereas linear projection processes all spatial regions uniformly and struggles to capture scale variations present in HSI scenes.

Next, spatial and channel self-attention modules interact with the query matrix Q. Spatial self-attention dynamically emphasizes informative spatial positions and strengthens the focus on key regions. Channel self-attention prevents all channels from being treated uniformly, reweights their contributions based on task relevance, and explicitly models cross-channel dependencies. A two-dimensional convolution with an $m \times 1$ kernel and an $n \times 1$ stride is then applied to the key matrix K and the value matrix V, which reduces spectral redundancy and introduces additional spatial–spectral contextual cues.

Scaled Dot-Product Attention is used to obtain the attended feature map. After reshaping, the features are refined through layer normalization and a residual connection. The refined representation is subsequently fed into a feedforward network consisting of a three-dimensional convolutional layer, a ReLU activation function, and a dropout layer for dimensional transformation and nonlinear enhancement. The output of the convolutional Transformer is finally obtained through a residual connection, leading to feature representations that are jointly strengthened in the spatial domain and

the spectral domain. If the output of the hollow CNN is denoted as X_{CNN} and the output of the convolution Transformer as X_{TR}, the entire process can be expressed as

$$X_{mid} = \text{Concat}(X_{CNN}, X_{TR}) \tag{3}$$

$$X_{out} = \text{ReLU}(\text{BN}(\text{Conv}(X_{mid}))). \tag{4}$$

The complete procedure is summarized in Algorithm 1.

Algorithm 1 The workflow of the PS2former framework

Input: Feature $X \in \mathbb{R}^{B\times 1\times H\times W\times S}$, the i-th spectral group c_i, and kernel size of each group $K = (1,1,k_i)$, where $\{i \in \mathbb{N} \mid 1 \le i \le 4\}$

Output: The final classification result produced by the classification head

$X' \leftarrow \text{Conv}(X)$
The input X is processed by SRE to produce X_{spa}
The input X is processed by SE and grouped to obtain X_i
for $i \leftarrow 1$ to 4 **do**
    $X_i \leftarrow \text{Conv}(X_i)$
    $X_{spe} \leftarrow \text{Concat}(\text{ReLU}(\text{BN}(\text{Conv}(X_i))))$
end for
$X_{out} \leftarrow X_{spa} \otimes X_{spe}$
The output X_{out} is passed through the hollow CNN to obtain X_{CNN}
The output X_{out} is passed through the convolution Transformer to obtain X_{TR}
X_{CNN} and X_{TR} are processed by the classifier and integrated to produce the final output X_{output}

3 Experiments

3.1 Datasets and Experimental Setting

We conducted experiments on two publicly available datasets, including Pavia University (PU) and WHU-Hi LongKou (LK). For the PU dataset, 0.7%, 0.7%, and 98.6% of the samples were randomly selected for training, validation, and testing, respectively. For the LK dataset, 0.1%, 0.1%, and 99.8% of the samples were used for training, validation, and testing.

All experiments were conducted on a vGPU with 32GB memory. To ensure fair comparison with other models, the training batch size was uniformly set to 128, and the model was trained for 300 epochs on both datasets. The focal loss was adopted as the training objective. Optimization was performed using the Adam optimizer with a weight decay of 0.001, and the learning rate for both datasets was set to 0.001. The patch size was set to 9. Model performance was evaluated using overall accuracy (OA), average accuracy (AA), and the Kappa coefficient.

Table 1: Classification Results of Different Classification Methods on the **PU** Dataset of 0.7% Training Samples

Class	A2S2KResNet	SpectralFormer	SSFTT	Morphformer	CTMixer	GSC-ViT	DBCTNet	Ours
1	95.28	36.83	91.93	82.05	98.89	**96.88**	96.99	99.07
2	99.96	89.18	99.71	95.23	96.65	99.14	99.5	**99.54**
3	**86.52**	45.7	78.74	42.27	88.01	86.9	87.56	91.53
4	98.54	71.56	**87.17**	87.62	94.24	94.44	93.35	97.23
5	100	43.51	100	81.67	100	99.67	99.90	**99.87**
6	97.98	7.70	92.97	45.69	100	96.85	**98.71**	99.36
7	98.85	8.23	96.97	54.10	87.03	**94.36**	97.32	98.93
8	95.18	26.74	84.54	57.22	93.19	**93.57**	95.01	96.38
9	**99.89**	77.73	95.85	96.57	99.03	97.38	97.26	96.99
OA	97.73	58.51	94.32	79.18	96.36	96.93	97.50	**98.55**
AA	96.67	45.24	91.99	71.05	95.23	95.47	96.18	**97.65**
Kappa	97.00	44.75	92.44	71.95	95.20	95.93	96.68	**98.08**

3.2 Comparison With State-of-the-Art Methods

We selected eight state-of-the-art (SOTA) models for comparative experiments, including DBCTNet[14], GSC-ViT, CTMixer[15], MorphFormer[16], SSFTT[17], SpectralFormer[9], and A2S2KResNet. Among them, DBCTNet adopts a dual-branch network structure to jointly extract spatial and spectral features. GSC-ViT mainly uses depthwise separable convolution to capture spectral characteristics in HSI. CTMixer models local information through CNNs and global information through transformers, followed by an innovative combination of the two. MorphFormer incorporates spectral and spatial morphological convolution operations and couples them with attention mechanisms to strengthen spatial–spectral interactions. SSFTT employs a feature tokenization transformer to learn spectral–spatial representations and high-level semantics. SpectralFormer integrates the transformer architecture with HSI to capture sequential dependencies among adjacent spectral bands. A2S2KResNet progressively extracts information using spatial–spectral feature extraction and residual connections.

To evaluate the robustness of the proposed model, all experiments were repeated ten times, and the average results were reported. The comparison results are shown in Tables 1 and 2. For qualitative experimental analysis, the model yielding the highest OA among the ten runs was selected to visualize the predicted samples for each SOTA method. The results are presented in Figs. 2 and 3.

Table 2: Classification Results of Different Classification Methods on the **LK** Dataset of 0.1% Training Samples

Class	A2S2KResNet	SpectralFormer	SSFTT	Morphformer	CTMixer	GSC-ViT	DBCTNet	Ours
1	97.21	85.59	99.60	94.30	97.64	**99.72**	99.38	99.55
2	97.69	24.45	95.31	57.78	97.58	92.62	95.23	**96.91**
3	**75.60**	6.87	89.36	2.21	82.05	85.58	88.61	93.93
4	99.08	75.36	**98.86**	94.06	99.52	98.90	97.33	97.3
5	61.86	28.12	57.51	36.88	50.39	84.61	80.71	**95.67**
6	94.81	27.23	98.83	83.86	93.75	97.47	**99.33**	97.86
7	99.77	97.85	99.84	98.66	99.96	**99.36**	99.55	99.41
8	90.05	59.52	82.92	76.58	95.86	**89.41**	85.24	86.70
9	**93.98**	67.03	70.50	65.12	70.81	80.89	80.57	85.49
OA	97.14	76.85	96.90	89.46	96.81	97.54	97.12	**97.64**
AA	90.01	52.45	88.08	67.16	87.51	92.25	91.77	**94.76**
Kappa	96.23	69.85	95.90	86.39	95.80	96.76	96.22	**96.91**

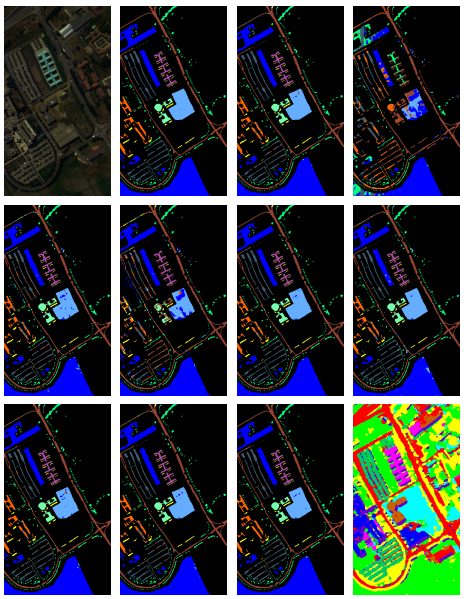

Fig. 2. Classification maps of the PU dataset with 0.7% training samples. (a) False-color image. (b) Ground truth. (c) A2S2KResNet. (d) SpectralFormer. (e) SSFTT. (f) MorphFormer. (g) HybridFormer. (h) CTMixer. (i) GSC-ViT. (j) DBCTNet. (k) PS2former. (l) Entire image classification result.

On the PU dataset, PS2former achieved 98.55%, 97.65%, and 98.08% in OA, AA, and Kappa, respectively—at least 1.05%, 1.48%, and 1.41% higher than all compared models. The highest accuracy was obtained in five of the nine categories, and even the lowest accuracy reached 96.39%. For the three classes with the fewest training samples—Painted metal sheets, Bitumen, and Shadows—our method still produced nearly perfect results, demonstrating its strong ability to enhance the classification performance of rare categories[18]. In contrast, MorphFormer and SpectralFormer performed poorly, likely due to the limited number of samples, which prevents them from fully learning discriminative features.Their ViT-based[19],[20] architectures also lack effective mechanisms for capturing local spatial–spectral representations. The remaining models delivered acceptable but clearly inferior performance.

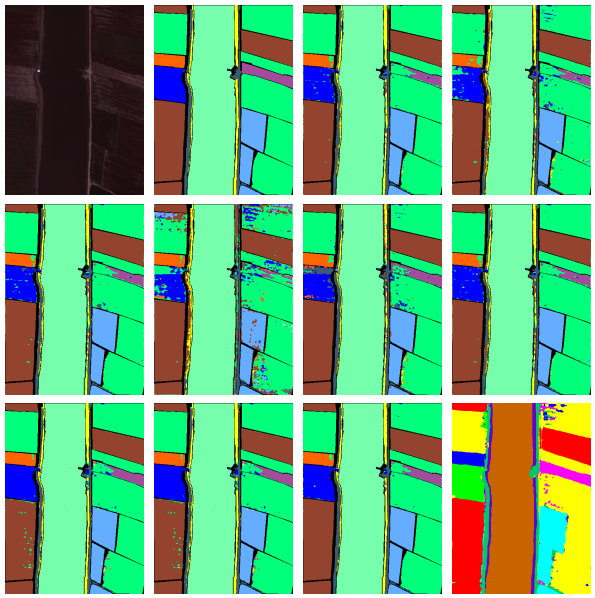

Fig. 3. Classification maps of the LK dataset with 0.1% training samples. (a) False-color image. (b) Ground truth. (c) A2S2KResNet. (d) SpectralFormer. (e) SSFTT. (f) MorphFormer. (g) HybridFormer. (h) CTMixer. (i) GSC-ViT. (j) DBCTNet. (k) PS2former. (l) Entire image classification result.

For the LK dataset, PS2former achieved 97.64%, 94.76%, and 96.91% in OA, AA, and Kappa, respectively—at least 0.11%, 1.07%, and 0.16% higher than all compared models. The LK dataset was trained using only 0.1% of the available samples. Although all models maintained high OA and Kappa values under the challenges of severe class imbalance and limited training data, AA dropped considerably. By comparing the gap between OA and AA, PS2former exhibited the smallest decline of only 2.88%, while the second-best method, GSC-ViT, showed a gap of 5.29%. For the Narrow-leaf soybean class, PS2former outperformed the second-best method by 7.08%. The LK dataset contains extremely uneven sample distributions, high inter-class similarity, and complex spa-

tial structures. Even under these conditions, PS2former demonstrates strong robustness. The model not only captures the characteristics of small-sample classes but also maintains reliable performance for classes with abundant samples.

The superior performance of our method arises from the parallel use of two dedicated branches that extract critical spectral and spatial features simultaneously. For spectral information, we adopt a more sophisticated grouped multi-scale extraction strategy that captures informative spectral patterns while suppressing redundant bands. Many existing methods do not explicitly decouple spatial and spectral mining, leading to incomplete feature representation. In addition, we design a shallow CNN and a convolution-based Transformer to model local features and global context, respectively. The use of three-dimensional convolution to generate the QKV matrices in the attention mechanism further enhances the ability to preserve spatial–spectral structure compared with traditional linear projection. These components work jointly to enable the performance improvement of our method.

4 Conclusion

Hyperspectral image classification has become an important research topic in remote sensing because it provides rich spectral information for accurate land-cover recognition and environmental monitoring. However, the high dimensionality of hyperspectral data, strong spectral redundancy, and limited labeled samples still make robust classification challenging. Against this background, this paper introduces an HSI classification network, termed PS2former, which incorporates a Parallel Extraction Module (PEM) consisting of Spectral Extraction (SE) and Spatial Residual Extraction (SRE). In addition, a Parallel Hybrid Transformer (PHT) is developed, leveraging hollow CNNs and convolution Transformers to jointly capture local and global information. By combining complementary spatial and spectral representations in a parallel manner, the proposed framework enhances feature diversity while maintaining an efficient architecture. Extensive experiments on two benchmark datasets demonstrate that PS2former consistently outperforms state-of-the-art methods. Our method achieved improvements in average accuracy of at least 0.52%, 1.48%, 0.96%, and 1.07% over these recent approaches on the corresponding datasets. These results confirm the effectiveness and robustness of the proposed design for HSI classification under limited-sample conditions. For future work, we plan to extend the evaluation to a broader range of benchmark datasets and further investigate lightweight model designs for practical applications.

Acknowledgments

This work was supported in part by Shandong Provincial Natural Science Foundation (grant No. ZR2023MF110) and Jinan City-University Integration Project (JNSX2025052).

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] Deng S, Xu Y, He Y, Yin J, Wu Z. A hyperspectral image classification framework and its application. Information Sciences. 2015;299:379-93.

[2] Huang K, Li S, Kang X, Fang L. Spectral–spatial hyperspectral image classification based on KNN. Sensing and Imaging. 2016;17(1):1.

[3] Okwuashi O, Ndehedehe CE. Deep support vector machine for hyperspectral image classification. Pattern Recognition. 2020;103:107298.

[4] Xia C, Yang S, Huang M, Zhu Q, Guo Y, Qin J. Maize seed classification using hyperspectral image coupled with multi-linear discriminant analysis. Infrared Physics & Technology. 2019;103:103077.

[5] Fırat H, Asker ME, Hanbay D. Classification of hyperspectral remote sensing images using different dimension reduction methods with 3D/2D CNN. Remote Sensing Applications: Society and Environment. 2022;25:100694.

[6] Yu C, Han R, Song M, Liu C, Chang CI. A simplified 2D-3D CNN architecture for hyperspectral image classification based on spatial–spectral fusion. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2020;13:2485-501.

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

[8] He M, Li B, Chen H. Multi-scale 3D deep convolutional neural network for hyperspectral image classification. In: 2017 IEEE International Conference on Image Processing (ICIP). IEEE; 2017. p. 3904-8.

[9] Hong D, Han Z, Yao J, Gao L, Zhang B, Plaza A, et al. SpectralFormer: Rethinking hyperspectral image classification with transformers. IEEE Transactions on Geoscience and Remote Sensing. 2021;60:1-15.

[10] He X, Chen Y, Lin Z. Spatial-spectral transformer for hyperspectral image classification. Remote Sensing. 2021;13(3):498.

[11] Zhang WT, Bai Y, Zheng SD, Cui J, Huang Zz. Tensor transformer for hyperspectral image classification. Pattern Recognition. 2025;163:111470.

[12] Arshad T, Zhang J, Ullah I. A hybrid convolution transformer for hyperspectral image classification. European Journal of Remote Sensing. 2024;57(1):2330979.

[13] Li Z, Huang W, Wang L, Xin Z, Meng Q. CNN and Transformer interaction network for hyperspectral image classification. International journal of remote sensing. 2023;44(18):5548-73.

[14] Wang Q, Jin X, Jiang Q, Wu L, Zhang Y, Zhou W. DBCT-Net: A dual branch hybrid CNN-transformer network for remote sensing image fusion. Expert Systems with Applications. 2023;233:120829.

[15] Zhang J, Meng Z, Zhao F, Liu H, Chang Z. Convolution transformer mixer for hyperspectral image classification. IEEE Geoscience and Remote Sensing Letters. 2022;19:1-5.

[16] Roy SK, Deria A, Shah C, Haut JM, Du Q, Plaza A. Spectral–spatial morphological attention transformer for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing. 2023;61:1-15.

[17] Xu Z, Hu H, Wang T, Zhao Y, Zhou C, Xu H, et al. Identification of growth years of Kudzu root by hyperspectral imaging combined with spectral–spatial feature tokenization transformer. Computers and Electronics in Agriculture. 2023;214:108332.

[18] Zhou D, He J. Rare Category Analysis for Complex Data: A Review. ACM Computing Surveys. 2023;56(5):1-35.

[19] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision; 2021. p. 10012-22.

[20] Pinasthika K, Laksono BSP, Irsal RBP, Shabiyya S, Yudistira N. SparseSwin: Swin transformer with sparse transformer block. Neurocomputing. 2024;580:127433.