frontiers | Frontiers in Remote Sensing

TYPE Original Research
PUBLISHED 25 June 2025
DOI 10.3389/frsen.2025.1583560

OPEN ACCESS

EDITED BY
Zenghui Zhang,
Shanghai Jiao Tong University, China

REVIEWED BY
Zhang Ying,
Hunan University of Technology, China
Xiaojie Gao,
Harvard University, United States

*CORRESPONDENCE
Yan Li,
834623022@qq.com

RECEIVED 26 February 2025
ACCEPTED 09 June 2025
PUBLISHED 25 June 2025

CITATION
Li Y, Li Y, Chen G, Li L, Jin S and Zhou L (2025) RCTNet: Residual conv-attention transformer network for corn hyperspectral image classification.
Front. Remote Sens. 6:1583560.
doi: 10.3389/frsen.2025.1583560

RCTNet: Residual conv-attention transformer network for corn hyperspectral image classification

Yihan Li[1,2], Yan Li[3]*, Gongchao Chen[1], Linfang Li[1], Songlin Jin[1] and Ling Zhou[1]

[1]School of Information Engineering, Henan Institute of Science and Technology, Xinxiang, China, [2]Faculty of Science, Asia Pacific University of Technology and Innovation, Kuala Lumpur, Malaysia, [3]School of Software, Henan Institute of Science and Technology, Xinxiang, China

Classifying corn varieties presents a significant challenge due to the high-dimensional characteristics of hyperspectral images and the complexity of feature extraction, which hinder progress in developing intelligent agriculture systems. To cope with these challenges, we introduce the Residual Convolution-Attention Transformer Network (RCTNet), an innovative framework designed to optimize hyperspectral image classification. RCTNet integrates Conv2D with Channel Attention (2DWCA) and Conv3D with Spatial Attention (3DWSA) modules for efficient local spatial-spectral feature extraction, ensuring meaningful feature selection across multiple dimensions. Additionally, a residual transformer module is incorporated to enhance global feature learning by capturing long-range dependencies and improving classification performance. By effectively fusing local and global representations, RCTNet maximizes feature utilization, leading to superior accuracy and robustness in classification tasks. Extensive experimental results on a corn seed hyperspectral image dataset and two widely used remote sensing datasets validate the effectiveness, efficiency, and generalizability of RCTNet in hyperspectral image classification applications.

KEYWORDS

hyperspectral image classification, corn seed, intelligent agriculture, corn identification, deep learning

1 Introduction

As a staple crop of global significance, corn plays a vital role in ensuring food security and sustaining agricultural economies. The quality and diversity of corn seeds directly impact crop yields, which in turn influence food production and the broader agricultural market. Given its importance, accurately classifying and identifying corn varieties is essential for optimizing agricultural practices and improving overall efficiency. However, traditional seed identification methods suffer from notable limitations, including susceptibility to human subjectivity, inefficiency in processing large sample volumes, and reduced accuracy under varying lighting conditions (Xu et al., 2022; Yuan et al., 2023; Zhang et al., 2023a; Zhao et al., 2025; Zhang et al., 2025b). These challenges undermine the scalability and reliability of conventional techniques. In response to these challenges, hyperspectral imaging technology has emerged as a powerful non-destructive testing method, offering significant advantages in seed identification. This advanced imaging technique enhances classification accuracy by capturing rich spectral and

spatial information and has become an indispensable tool in intelligent agriculture (Zhang L. et al., 2024; Barbedo, 2023; Zhang W. et al., 2024; Zhang et al., 2025a).

In the early stage, various machine learning techniques were employed for seed recognition tasks (Su et al., 2022; Jin et al., 2023). These early methods typically relied on manual or semi-automatic feature extraction processes, which were labor-intensive and prone to overlooking crucial information (Liang et al., 2024). Furthermore, these methods exhibited inconsistent performance across different datasets and under varying environmental conditions. As a result, the effectiveness of early seed identification approaches was limited. However, with rapid advancements in deep learning, convolutional neural networks (CNNs) have become increasingly popular due to their robust ability to learn hierarchical feature representations from raw image data automatically.

CNN-based methods have demonstrated strong capabilities in hyperspectral image classification due to their proficiency in extracting local spatial features. For instance, Zhang et al. (2021) proposed a rapid, non-destructive corn seed purity assessment method by leveraging hyperspectral imaging with CNNs, significantly improving accuracy and efficiency. Later, Zhang et al. (2022) enhanced this framework by integrating 2D/3D convolutions and attention mechanisms, allowing for the effective extraction of spatial, spectral, and textural features. Zheng et al. (2022) further addressed spectral interference using CSpeA and RSpaA models, which focused on reducing redundant spectral bands and capturing rotation-invariant features. Diao et al. (2023) introduced residual 3D frequency convolutions and spectral-spatial attention modules to enhance spatial-spectral learning, while Tian et al. (2024) incorporated hierarchical representations to improve classification performance. Despite these advances, CNN-based models often fall short in capturing global dependencies across the high-dimensional hyperspectral data. In contrast, transformer-based methods have garnered attention since the introduction of Vision Transformer (ViT) (Dosovitskiy et al., 2020), which effectively models long-range dependencies (Dai et al., 2021). CoAtNet (2021) exemplifies this trend by integrating convolution with self-attention while optimizing computational efficiency. SpectralFormer (Hong et al., 2021) reframes HSI classification as a sequence modeling task, and other frameworks (Zhang et al., 2023b; Li Y. et al., 2024; Shi et al., 2024; Song et al., 2024) propose various attention-enhanced transformers capable of better spectral-spatial feature representation. These models demonstrate strong global modeling but often lack precision in capturing fine-grained local information.

Recognizing these complementary strengths and limitations, we propose RCTNet, a hybrid architecture that combines convolutional networks for local spatial-spectral feature extraction with transformers for global representation learning. Specifically, we design 2DWCA and 3DWSA modules to refine channel and spatial information, respectively, and introduce a Residual Transformer Module (RTM) to enhance the modeling of long-range dependencies. Our RCTNet achieves an accuracy of 99.32% on the CSHID dataset, surpassing state-of-the-art models by up to 0.49%. This performance, along with consistent results on Salinas-A and Botswana datasets, confirms the generalizability and robustness of RCTNet. The major contributions are as follows.

- We propose a residual convolution-attention transformer network designed for classifying corn hyperspectral images, which integrates the complementary strengths of convolution operations and transformer structures from local and global perspectives.
- We design the 2DWCA and 3DWSA modules, refining traditional Conv2D to focus on critical channel features while ignoring irrelevant or redundant information and adjusting traditional Conv3D to prioritize the spatial dimension, enhancing the representation of features in crucial regions.
- We develop the Residual Transformer Module (RTM), which enhances the network's global feature extraction by integrating max and average pooling into the traditional transformer structure and incorporating residual connections between transformer structures.

The structure of this paper is as follows: Section 2 shows the related work of hyperspectral image classification, Section 3 outlines the proposed method, Section 4 presents the experiments and analyses, and Section 5 concludes the study.

2 Related work

This section provides a concise overview of existing deep learning methods for hyperspectral image classification, focusing on how they attempt to address the challenges of modeling spatial and spectral information. We group prior works into three categories: CNN-based, Transformer-based, and CNN-Transformer collaborative network-based methods, and highlight their key limitations that motivate the design of our RCTNet.

CNN-based methods have been widely adopted in hyperspectral image classification due to their ability to extract local spatial-spectral features through convolutional operations. For instance, Sellami and Tabbone (2022) introduced a method using multi-view deep neural networks to fuse spectral and spatial features with a limited number of labeled samples. Sun et al. (2023) introduced a large kernel spectral-spatial attention network to solve two issues, which neglect spatial properties and do not fully consider the dependence between spectral and spatial information. Paoletti et al. (2023) developed a method that automatically designs and optimizes convolutional neural networks for hyperspectral image classification by using channel-based attention mechanisms. Although these models achieve good performance, they often require deep architectures to expand receptive fields, making them computationally intensive and prone to overfitting when training data is scarce.

Transformer-based methods, inspired by the Vision Transformer (ViT), provide powerful capabilities in modeling long-range dependencies and global contextual information. Yu et al. (2022) designed a multilevel spatial-spectral transformer method for hyperspectral image classification, addressing the limitations of CNNs, such as limited receptive fields, information loss, and high computational costs. Wang et al. (2023) introduced a tri-spectral image generation pipeline that converts hyperspectral images into high-quality tri-spectral images, enabling the use of ImageNet pre-trained networks for feature extraction. Ahmad et al.

(2024a) introduced a pyramid-based hierarchical spatial-spectral transformer that organizes input data hierarchically into pyramid segments to improve processing efficiency. Moreover, Ahmad et al. (2024b) presented a novel transformer-based method that employs wavelet transforms for invertible downsampling, avoiding information loss from average pooling. While they have shown promise in hyperspectral image classification tasks, pure transformer architectures often lack precision in modeling fine-grained local spatial details, especially under conditions of limited training data. Their reliance on large-scale data and high computational cost further restricts their application in certain remote sensing scenarios.

CNN-Transformer collaborative network-based methods integrate the advantages of both CNNs and transformers, using CNNs for local feature extraction and transformers for modeling global context. These hybrid architectures have shown great promise in hyperspectral image classification by combining the fine-grained feature extraction capabilities of CNNs with the long-range dependency modeling of transformers. For example, Sun et al. (2022) introduced an SSFTT method, which extracts low-level and high-level features through a spectral-spatial feature extraction module, integrates a Gaussian-weighted feature tokenizer, and employs a transformer encoder module for feature representation and learning. Yang et al. (2022) proposed a fusion network for hyperspectral image classification that uses both serial and parallel mechanisms to fully leverage spectral-spatial features, with CNNs capturing local spatial features and Transformers capturing global features. Yan et al. (2023) developed a hybrid convolution and vision transformer network, addressing the issue of obtaining actual global features in existing transformer-based methods.

Liang et al. (2023) developed an HSI classification method that integrates transformers with SimAM-based CNNs to address limitations in spatial and global feature extraction. The CNN module employs a hierarchical 2-D dense structure to enhance spatial feature representation, guided by a dual attention unit (DAU) that focuses on discriminative pixel and channel features. Spectral features are modeled using a squeeze-enhanced axial transformer to capture global dependencies and local details.

Xu et al. (2024) designed a double branch convolution-transformer network to address the high computational complexity and underutilization of spectral information in existing hybrid models. These collaborative approaches benefit from the complementary strengths of CNNs and transformers, offering improved feature representation by jointly modeling local details and global dependencies. However, they often suffer from ineffective integration strategies or increased architectural complexity, which may limit their scalability and generalization.

From the above discussion, it is evident that our proposed RCTNet also follows the CNN-Transformer collaborative paradigm, which itself is not a novel concept. However, RCTNet introduces architectural innovations that effectively enhance the synergy between convolutional and transformer components. These improvements result in a more balanced trade-off between local feature preservation and global context modeling, enabling more efficient and accurate hyperspectral image classification, particularly under limited training data conditions.

3 Proposed methods

Figure 1 illustrates the workflow of RCTNet. The framework begins with data preprocessing, where principal component analysis is applied to reduce the dimensionality of the hyperspectral images, thereby retaining the most informative components while eliminating redundant or noisy features. Next, the model extracts and enhances local spatial-spectral features through a dual-branch structure, which consists of two key modules: the 2D Convolution with Channel Attention (2DWCA) and the 3D Convolution with Spatial Attention (3DWSA). Following the local feature extraction, the model transitions to the global feature extraction stage using a residual transformer module. The transformer is responsible for capturing global spatial, spectral, and texture features by modeling long-range dependencies between different spectral bands and spatial regions. The residual connections within the transformer help preserve important feature information across layers, alleviating the potential issue of vanishing gradients and ensuring the robustness of the model during training. Finally, the classifier processes the extracted local and global features and outputs the specific corn seed variety. Table 1 presents the output size and the number of parameters for each layer of the proposed RCTNet, which helps readers better understand the overall network structure.

3.1 Mixed convolution module

Convolutional neural networks have demonstrated remarkable performance in various image processing and computer vision tasks, including image classification, segmentation, and object detection. Traditionally, 2D convolutional operations capture local features by sliding a fixed-size filter over the image's spatial dimensions. However, for more complex datasets like hyperspectral images, which include both spatial and spectral information, extending this capability to 3D convolutions can be highly beneficial. 3D convolutional operations enable the model to capture both spatial and spectral features simultaneously, providing richer representations of image data. Despite these advantages, directly using traditional 2D or 3D convolutions for hyperspectral data often results in suboptimal performance due to the increased complexity of feature extraction across multiple dimensions. To alleviate this limitation, we propose the use of a dual-branch architecture that combines 2DWCA and 3DWSA. This design enables us to efficiently process the corn hyperspectral image data by integrating spectral, spatial, and texture information. By fusing the features extracted from both branches, the overall performance of the model is enhanced.

In the 2DWCA branch, we utilize a 2D convolution operation integrated with a channel attention mechanism. This integration enhances the network's ability to focus on the most relevant features from critical channels, which is particularly important in hyperspectral image processing. The goal is to fully capture the spectral-spatial feature information from the corn hyperspectral images. Let us denote I_{CHS} as the input image data and first perform the correlated convolution operation, which can be expressed as Equation 1

$$I^{2}_{\mathrm{CHS}} = \mathrm{Relu}\left(\mathrm{BN}\left(\mathrm{Conv2D}\left(I_{\mathrm{CHS}}\right)\right)\right), \tag{1}$$

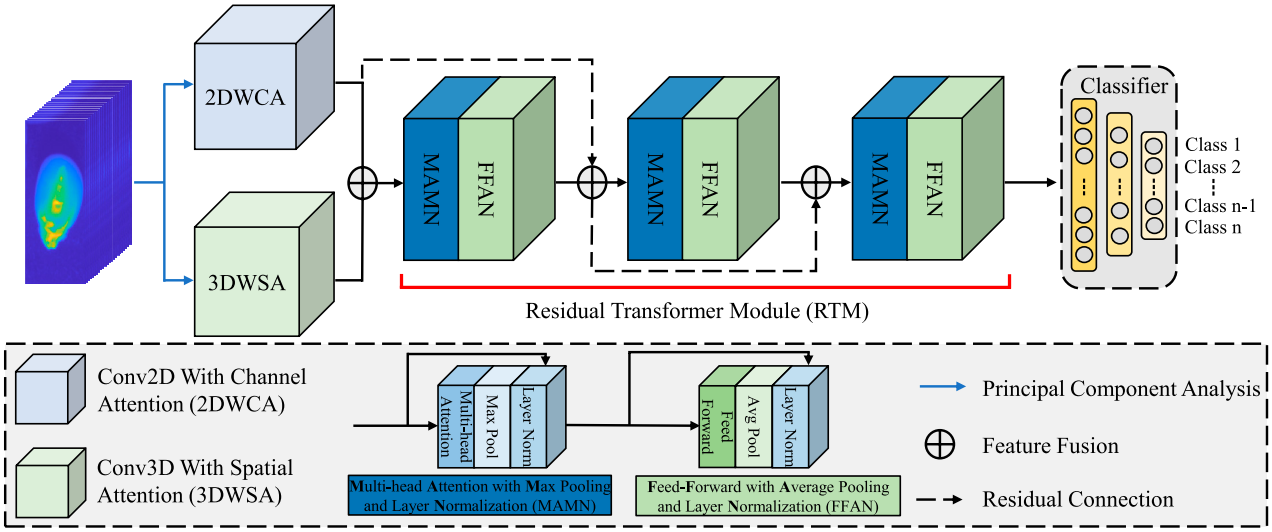

FIGURE 1
The workflow of the residual conv-attention transformer network. Begins with acquiring corn hyperspectral image data, which undergoes initial processing through region extraction and principal component analysis. Subsequently, local spatial-spectral features are extracted and fused using a dual-branch pattern incorporating 2DWCA and 3DWSA. The global feature information is then processed via a transformer module with residual connections introduced between the transformer structures. Finally, the classifier produces the classification and recognition results.

TABLE 1 Output size and parameters of each layer of the RCTNet.

Layer (type)	Output size	#Param
2DWCA_1	(64 × 1 × 1, 32)	1154
3DWSA_1	(304 × 1 × 1, 32)	15616
MAMN_1	(32 × 368, 1)	460
FFAN_1	(32 × 368, 1)	629660
MAMN_2	(16 × 368, 1)	736
FFAN_2	(16 × 368, 1)	1085968
MAMN_3	(8 × 368, 1)	1012
FFAN_3	(8 × 368, 1)	1542246
Linear_1	10/6	48522
Total params: 3,325,404		

where Conv2D$(\cdot)$ represents a 2D convolution with 64 output channels, a stride of 1×1, and a convolution kernel size of 3×3. BN$(\cdot)$ refers to batch normalization, which helps in accelerating network convergence by reducing internal covariate shift. Following this, the convolutional features I^2_{CHS} are passed through a channel attention mechanism to enhance the most informative channels. The attention mechanism is implemented as

$$I^{2'}_{\mathrm{CHS}} = \sigma\left(\mathrm{fc}_2\left(\mathrm{Relu}\left(\mathrm{fc}_1\left(\mathrm{AP}\left(I^2_{\mathrm{CHS}}\right) \oplus \mathrm{MP}\left(I^2_{\mathrm{CHS}}\right)\right)\right)\right)\right) \times I^2_{\mathrm{CHS}}, \quad (2)$$

where AP$(\cdot)$ and MP$(\cdot)$ represent the average pooling and max pooling operations, respectively. The operations $\mathrm{fc}_1(\cdot)$ and $\mathrm{fc}_2(\cdot)$ perform dimensionality reduction and restoration of the feature vector, respectively. $\sigma(\cdot)$ is the Sigmoid activation function, which normalizes the channel attention map to a range between 0 and 1.

In the 3DWSA branch, we leverage 3D convolutions in combination with a spatial attention mechanism. The spatial attention mechanism dynamically adjusts the focus of the convolution operation on important regions in the image, effectively capturing spectral, spatial, and texture information. This enhances the overall feature extraction and improves the robustness of the learned representations. The process begins with a 3D convolution operation, which is applied to the input image I_{CHS} as Equation 3

$$I^3_{\mathrm{CHS}} = \mathrm{Relu}\left(\mathrm{BN}\left(\mathrm{Conv3D}\left(I_{\mathrm{CHS}}\right)\right)\right), \quad (3)$$

where Conv3D$(\cdot)$ indicates a 3D convolution operation with a convolution kernel size of $7 \times 3 \times 3$ and 16 output channels. Following the convolution, the extracted features I^3_{CHS} are processed by a spatial attention mechanism, which can be expressed as Equation 4

$$I^{3'}_{\mathrm{CHS}} = \sigma\left(\mathrm{Conv2}D\left(\mathrm{AP}\left(I^3_{\mathrm{CHS}}\right) \oplus \mathrm{MP}\left(I^3_{\mathrm{CHS}}\right)\right)\right) \times I^3_{\mathrm{CHS}}, \quad (4)$$

where AP$(\cdot)$, MP$(\cdot)$, and σ are defined in Equation 2. Additionally, Conv2D$(\cdot)$ represents a 2D convolution operation with a kernel size of 7×7 and a stride of 3×3. After the features are extracted and enhanced through the 2DWCA and 3DWSA branches, the next step is to fuse the extracted features for further processing.

The features $I^{2'}_{\mathrm{CHS}}$ and $I^{3'}_{\mathrm{CHS}}$ from both branches are combined, and the resulting fused features are passed to a residual transformer module for further refinement. The feature fusion process is expressed as Equation 5

$$I'_{\mathrm{CHS}} = I^{2'}_{\mathrm{CHS}} \oplus I^{3'}_{\mathrm{CHS}}. \quad (5)$$

The fusion strategy effectively combines the complementary information captured by the 2D and 3D convolutions, enhancing both the spatial and spectral representation of the image data.

3.2 Residual transformer module

The novel residual transformer module integrates residual connectivity with an enhanced transformer structure, combining the strengths of both residual learning and attention mechanisms. This module is designed to improve the performance and efficiency of hyperspectral image processing by facilitating the learning of both local and global feature representations. The residual transformer module consists of two primary sub-modules: Multi-head Attention with Max Pooling and Layer Normalization (MAMN) and Feed-Forward with Average Pooling and Layer Normalization (FFAN). These sub-modules are connected through residual concatenation, which allows the network to preserve vital information while enhancing its learning capacity.

The MAMN leverages the multi-head attention mechanism to capture global relationships between the input feature sequences. This mechanism allows the model to learn from different subspaces, providing a richer feature representation. In hyperspectral images, the relationships between spatial and spectral dimensions are complex, and multi-head attention helps the network focus on different aspects of the image simultaneously. After the attention mechanism, a max pooling operation is applied to extract the most salient features. Max pooling helps in retaining the highest activation values while discarding less important information, allowing the model to concentrate on the key features that are essential for downstream tasks. The incorporation of layer normalization further accelerates convergence by stabilizing the learning process and ensuring consistent training behavior. The computational process of the MAMN is given by Equation 6

$$I_{\mathrm{MAMN}} = \mathrm{LN}\left(I'_{\mathrm{CHS}}\right) \oplus \mathrm{LN}\left(\mathrm{Dropout}\left(\mathrm{MP}\left(\mathrm{MH}\left(I'_{\mathrm{CHS}}, I'_{\mathrm{CHS}}, I'_{\mathrm{CHS}}\right)\right)\right)\right), \tag{6}$$

where LN denotes layer normalization, which normalizes the output across the feature dimensions. MH refers to the multi-head attention operation, which enables the model to process multiple attention heads in parallel, enhancing its ability to model complex dependencies. The Dropout is set to 0.3 to prevent overfitting and promote generalization during training. The max pooling operation MP helps retain the most significant features while downsampling the feature maps.

The FFAN starts with two fully connected layers designed to enhance the feature representation through nonlinear mapping. These layers serve to increase the expressive power of the model by transforming the features into a higher-dimensional space. Afterward, an average pooling operation is applied, which aggregates local features across the feature maps. Average pooling helps reduce noise and smoothens the output, ensuring that the model focuses on more stable and prominent features that are less sensitive to small variations in the input data. This process results in a more robust representation, which is essential for effective image analysis in complex scenarios such as hyperspectral image classification. The final step in the FFAN is the application of layer normalization, which accelerates network convergence and ensures stability. This normalization technique has been shown to improve the training efficiency and overall performance of deep learning models. The computational process of the FFAN is expressed as Equation 7

$$I_{\mathrm{FFAN}} = \mathrm{LN}\left(I_{\mathrm{MAMN}}\right) \oplus \mathrm{LN}\left(\mathrm{Dropout}\left(\mathrm{AP}\left(\mathrm{FFN}\left(I_{\mathrm{MAMN}}\right)\right)\right)\right), \tag{7}$$

where *FFN* refers to the operation performed by the two fully connected layers, which introduce nonlinear transformations to enhance feature representations. The average pooling operation AP aggregates local features, while Dropout helps prevent overfitting. The features extracted after processing through the residual transformer module can be expressed as Equation 8

$$I_{\mathrm{RTM}} = I'_{\mathrm{CHS}} \oplus I_{\mathrm{FFANi}}, i \in [0, 3]. \tag{8}$$

This residual fusion strategy ensures that both the initial and enhanced features contribute to the final output, preserving critical information while refining the learned representations. The resulting features, enriched by the combined capabilities of the multi-head attention, max pooling, average pooling, and fully connected layers, are then ready for subsequent processing in the model's classification tasks.

3.3 Loss function

The objective of RCTNet in corn hyperspectral images variety classification is to forecast the probability distribution of an input image corresponding to a particular corn variety. The network produces a set of logits $z = [z_{1,j}, z_{2,j}, \ldots, z_{m,j}]$, which are then transformed into a probability distribution $\hat{y} = [\hat{y}_{1,j}, \hat{y}_{2,j}, \ldots, \hat{y}_{m,j}]$ using the Softmax function Equation 9

$$\hat{y}_{i,j} = \frac{\exp\left(z_{i,j}\right)}{\sum_{k=1}^{m} \exp\left(z_{k,j}\right)}, \tag{9}$$

where $z_{i,j}$ and $\hat{y}_{i,j}$ represent the logit and predicted probability for the j – th sample in the i – th class, respectively. m denotes the total number of classes. The CrossEntropyLoss function measures the discrepancy between the predicted probability $\hat{y}_{i,j}$ and true label $y_{i,j}$, which is well-suited for multi-categorical problems, and the loss is calculated as Equation 10

$$L\left(y, \hat{y}\right) = -\sum_{i=1}^{n} \sum_{j=1}^{m} y_{i,j} \log\left(\hat{y}_{i,j}\right), \tag{10}$$

where $y_{i,j}$ denotes the one-hot encoded true label, and n is the total number of sample.

4 Experiments and analysis

In the experiments and analysis section, we examine how specific parameter settings influence the performance of the proposed RCTNet for classifying corn seeds. We also perform several comparative experiments to showcase RCTNet's effectiveness compared to other image classification methods. Furthermore, experiments conducted on two general hyperspectral remote sensing image datasets confirm the robustness of RCTNet. Lastly, ablation studies were carried out to verify the contribution of each component.

4.1 Experimental configuration

To validate the effectiveness and generalization of RCTNet, we conducted experiments and analyses on a system running Windows

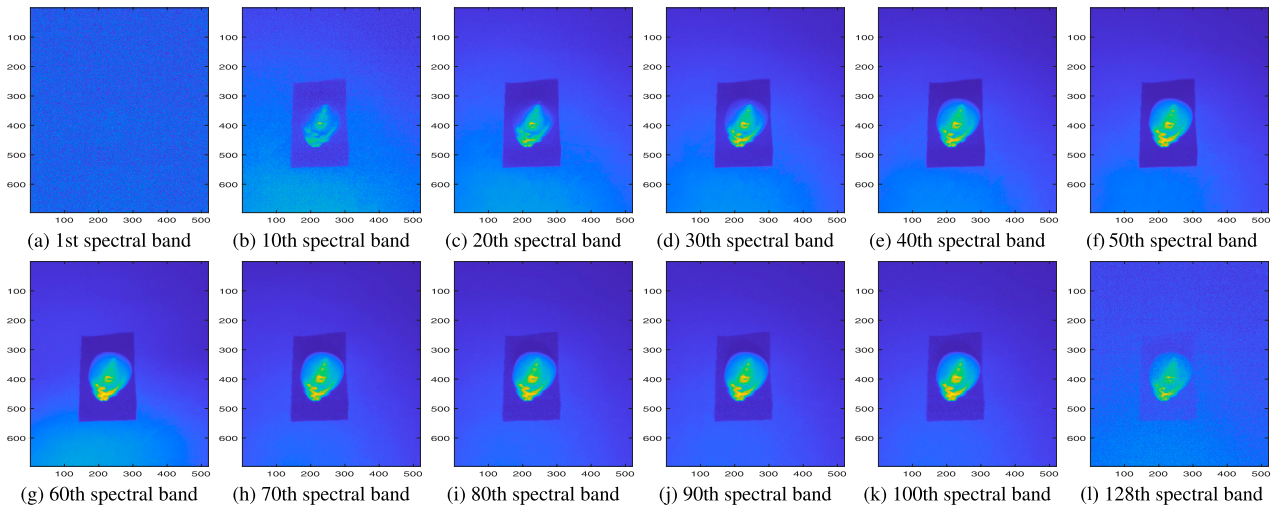

FIGURE 2
Selected spectral bands of BaiYu818 corn seeds. **(a)** first spectral band **(b)** 10th spectral band **(c)** 20th spectral band **(d)** 30th spectral band **(e)** 40th spectral band **(f)** 50th spectral band **(g)** 60th spectral band **(h)** 70th spectral band **(i)** 80th spectral band **(j)** 90th spectral band **(k)** 100th spectral band **(l)** 128th spectral band.

10 Professional, equipped with an Intel i9-14900 KF CPU at 3.20 GHz, an NVIDIA RTX 4060 GPU, 64 GB of RAM, Python 3.11, and PyTorch-GPU 2.1.0. During the experiments, RCTNet was optimized using the Adam optimizer with an initial learning rate of 0.0001, and the total number of iterations was set to 200. We first conducted experiments on the CSHID (Zhang et al., 2022) dataset, setting the test batch size to 8 and designing multiple experimental setups with training batch sizes of 16, 32, and 48. To evaluate the impact of parameter settings on network performance, we used training-to-test sample ratios of 7:3, 8:2, and 9:1. Additionally, to assess the generalization capability of the network, we tested it on the Salinas-A scene (Li et al., 2018) dataset and Botswana (Xu et al., 2019) dataset. For these datasets, we increased the test batch size to 16, maintained the training batch size at 32, and reduced the number of training samples to 10%, while still achieving robust classification results.

4.2 Experimental data

In this work, we utilize the CSHID dataset introduced in SSTNet (Zhang et al., 2022), which comprises ten types of corn seed hyperspectral images from the Henan region. It including varieties such as FengDa601, BaiYu9284, BaiYu8317, BaiYu918, BaiYu897, BaiYu879, BaiYu833, BaiYu818, BaiYu808, and BaiYu607. Each variety contains 120 samples, and a 400–1000 nm range in 128 spectral bands. Figure 2 displays 12 representative bands for the BaiYu818 variety. The raw images have a resolution of 696 × 520 pixels and are reduced to 210 × 200 pixels after extracting the region of interest. Specific details regarding the hyperspectral bands of corn images can be found in SSTNet (Zhang et al., 2022). To further evaluate the performance and generalization capability of RCTNet, we incorporated two publicly available hyperspectral datasets: Salinas-A scene (Li et al., 2018) and Botswana (Xu et al., 2019). The Salinas-A scene dataset, captured by the AVIRIS sensor, represents agricultural terrain in Salinas Valley, California. It consists of 224 spectral bands spanning 380–2500 nm, with a spatial resolution of 83 × 86 pixels, and primarily focuses on six crop categories. Meanwhile, the Botswana dataset, acquired by the Hyperion sensor onboard the EO-1 satellite, covers a broad spectral range of 400–2500 nm with 145 bands and a spatial resolution of 30 m per pixel. By leveraging these two datasets, we aim to comprehensively assess RCTNet across different imaging conditions, ensuring its robustness and effectiveness in real-world applications. Table 2 presents details of specific categories along with the number of training samples and testing samples.

4.3 Classification results

It is worth noting that the proposed RCTNet is primarily designed to classify corn seed varieties in agricultural production. To validate its performance, we conducted a series of experiments on the CSHID dataset (Zhang et al., 2022), compared with ten classification methods, including G-MDRF (Zhang et al., 2025c), SGD (Lei and Tang, 2021), RFA (Chen et al., 2021), SSTNet (Zhang et al., 2022), SSFTT (Sun et al., 2022), SCTNet (Chen et al., 2025), GACNet (Zhang W. et al., 2023), PolSARFormer (Jamali et al., 2023), ERNet (Li X. et al., 2024), and RDTN (Li Y. et al., 2024). All methods were executed with identical configurations on the same device. We employ the common metrics Accuracy, Recall, F1-score, and Precision to measure the classification performance of these methods. Additionally, the robustness of RCTNet was validated on the general Botswana (Xu et al., 2019) and Salinas-A scene (Li et al., 2018) datasets. Next, we will separately analyze the experimental results on the CSHID dataset (Zhang et al., 2022), the Botswana dataset (Xu et al., 2019) and the Salinas-A scene dataset (Li et al., 2018).

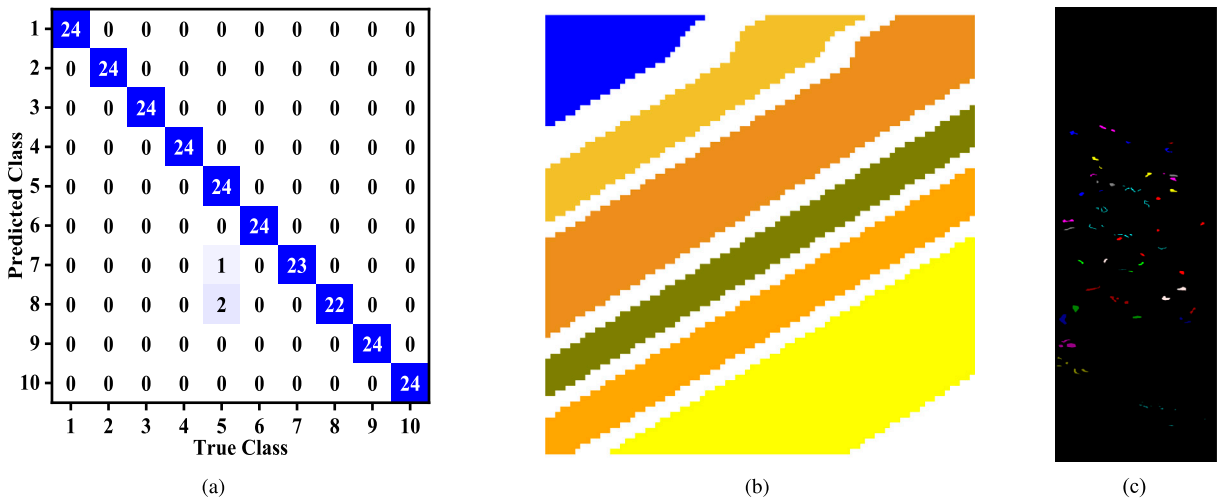

FIGURE 3
(a) Classification confusion matrix of RCTNet experiments on the CSHID (Zhang et al., 2022) dataset. **(b)** Classification map of RCTNet experiments on the Salinas-A scene (Li et al., 2018) dataset. **(c)** Classification map of RCTNet experiments on the Botswana (Xu et al., 2019) dataset.

TABLE 2 Specific categories and corresponding training and testing samples in the Botswana (Xu et al., 2019) and Salinas-A scene (Li et al., 2018) datasets.

No.	Botswana (Xu et al., 2019)			Salinas-A scene (Li et al., 2018)		
	Name	Training Samples	Testing Samples	Name	Training Samples	Testing Samples
1	Water	27	243	Brocoli green weeds 1	39	352
2	Hippograss	10	91	Lettuce romaine 7 weeks	80	719
3	Floodplain grasses 1	25	226	Corn senesced green weeds	134	1209
4	Floodplain grasses 2	22	193	Lettuce romaine 6 weeks	67	607
5	Reeds	27	242	Lettuce romaine 4 weeks	62	554
6	Riparian	27	242	Lettuce romaine 5 weeks	152	1373
7	Firescar	26	233			
8	Island interior	20	183			
9	Acacia woodlands	31	283			
10	Acacia shrublands	25	223			
11	Acacia grasslands	31	274			
12	Short mopane	18	163			
13	Mixed mopane	27	241			
14	Exposed soils	10	85			
	Total	325	2923	Total	534	4814

Quantitative evaluation on the CSHID dataset (Zhang et al., 2022). By comparing with other classification methods, the effectiveness of the proposed RCTNet has been validated. Additionally, we explored the impact of different training batch sizes and testing sample quantities on the model performance. Tables 3–5 show the results for training batch sizes of 16, 32, and 48, and testing sample proportions of 30%, 20%, and 10% of the total samples, further validating the effectiveness of RCTNet. Across all three tables, the results indicate that the training batch size significantly impacts the model's performance. When the batch size is set to 32, RCTNet consistently achieves the best performance in terms of Precision, Recall, F1-score, and Accuracy. With a smaller batch size of 16, the network's performance slightly decreases due to the limited number of samples processed per iteration, and an excessively large batch size of 48 also leads to suboptimal performance. Overall, the results demonstrate that a moderate

TABLE 3 Classification results of RCTNet experiments on the CSHID dataset (Zhang et al., 2022) using the training batch size 16, with test sample ratios of 30%, 20%, and 10%. (Best results are highlighted in red, and second-best results in blue).

Method	30%				20%				10%			
	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
G-MDRF (Zhang et al., 2025c)	95.82	94.90	95.08	95.51	96.04	95.73	95.75	96.10	96.15	95.95	96.00	96.13
SGD (Lei and Tang, 2021)	96.13	96.10	96.12	96.38	96.80	96.50	96.55	96.85	96.55	96.37	96.35	96.83
RFA (Chen et al., 2021)	93.68	93.48	93.50	93.72	94.13	93.89	93.86	94.35	94.90	94.71	94.77	94.89
SSFTT (Sun et al., 2022)	96.39	96.67	96.55	96.96	95.97	95.82	95.73	95.49	95.35	95.08	95.29	95.16
SSTNet (Zhang et al., 2022)	97.55	97.50	97.55	97.68	97.90	97.70	97.81	97.85	98.15	98.05	97.94	98.10
GACNet (Zhang et al., 2023c)	97.21	97.05	96.99	97.18	97.13	96.95	96.82	97.04	96.57	96.38	96.42	96.50
SCTNet (Chen et al., 2025)	97.60	97.45	97.50	97.50	97.85	97.60	97.70	97.80	98.10	97.90	98.05	98.05
PSARF (Jamali et al., 2023)	96.10	96.18	96.17	96.20	96.50	96.40	96.45	96.55	96.70	96.60	96.65	96.75
ERNet (Li et al., 2024a)	97.25	97.58	97.65	97.95	97.89	97.89	97.79	97.95	97.55	97.45	97.46	97.80
RDTN (Li et al., 2024b)	98.10	97.63	97.60	97.93	97.76	98.05	98.06	97.88	97.60	97.38	97.40	97.72
RCTNet	97.76	97.55	97.69	98.02	97.55	97.80	97.86	98.33	97.62	97.93	97.97	98.57

10.3389/frsen.2025.1583560

TABLE 4 Classification results of RCTNet experiments on the CSHID dataset (Zhang et al., 2022) using the training batch size 32, with test sample ratios of 30%, 20%, and 10%.

Method	30%				20%				10%			
	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
G-MDRF (Zhang et al., 2025c)	96.02	95.00	95.50	95.75	96.36	95.85	96.18	96.25	96.25	96.20	96.15	96.23
SGD Lei and Tang, 2021)	96.50	96.10	96.20	96.58	97.02	96.72	96.75	97.08	97.15	96.87	96.88	97.16
RFA (Chen et al., 2021)	93.80	93.49	93.62	93.80	94.30	94.10	94.28	94.58	94.98	94.82	94.83	95.04
SSFTT (Sun et al., 2022)	96.53	96.59	96.67	96.85	96.12	95.97	95.92	95.76	95.63	95.51	95.38	95.29
SSTNet (Zhang et al., 2022)	97.70	97.59	97.60	97.75	98.06	97.92	97.91	97.92	98.32	98.13	98.13	98.20
GACNet (Zhang et al., 2023c)	97.59	97.38	97.51	97.67	97.59	97.30	97.08	97.28	97.32	97.18	97.31	97.50
SCTNet (Chen et al., 2025)	97.59	97.48	97.50	97.13	98.12	97.97	97.95	97.60	98.10	98.02	97.98	97.73
PSARF (Jamali et al., 2023)	96.30	96.00	96.10	96.40	96.58	96.60	96.69	96.67	96.63	96.58	96.63	96.71
ERNet (Li et al., 2024a)	97.80	98.10	98.15	98.05	98.48	98.33	98.14	98.75	98.57	98.46	98.43	98.80
RDTN (Li et al., 2024b)	98.20	97.99	98.12	98.15	98.09	97.93	97.91	98.33	99.05	98.42	98.50	98.83
RCTNet	98.14	98.06	98.05	98.75	98.89	98.75	98.76	99.17	99.23	99.17	99.17	99.32

10.3389/frsen.2025.1583560

TABLE 5 Classification results of RCTNet experiments on the CSHID dataset (Zhang et al., 2022) using the training batch size 48, with test sample ratios of 30%, 20%, and 10%.

Method	30%				20%				10%			
	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
G-MDRF (Zhang et al., 2025c)	95.85	95.05	95.08	95.63	95.97	95.82	95.80	96.31	96.36	96.12	96.15	96.38
SGD (Lei and Tang, 2021)	96.28	96.19	96.21	96.50	96.82	96.58	96.60	96.90	96.47	96.35	96.39	96.86
RFA (Chen et al., 2021)	93.55	93.50	93.53	93.75	94.28	93.98	94.01	94.53	94.98	94.85	94.87	95.03
SSFTT (Sun et al., 2022)	96.38	96.47	96.53	96.79	96.20	96.11	95.86	95.83	95.37	95.62	95.47	95.33
SSTNet (Zhang et al., 2022)	97.82	97.66	97.60	97.77	97.98	97.85	97.86	97.92	98.46	98.25	98.19	98.35
GACNet (Zhang et al., 2023c)	97.53	97.47	97.39	97.34	97.42	97.48	97.27	97.37	97.07	97.31	97.28	97.31
SCTNet (Chen et al., 2025)	97.58	97.42	97.45	97.53	97.93	97.66	97.67	97.89	98.23	97.98	98.15	98.37
PSARF (Jamali et al., 2023)	96.38	96.24	96.27	96.33	96.61	96.52	96.53	96.58	96.85	96.73	96.74	96.82
ERNet (Li et al., 2024a)	97.15	97.51	97.65	97.81	97.93	97.93	97.89	98.02	97.90	97.88	97.89	98.29
RDTN (Li et al., 2024b)	98.15	97.63	97.67	97.99	97.83	98.33	98.30	97.93	97.76	97.58	97.52	97.98
RCTNet	98.17	97.92	97.93	98.28	98.40	98.25	98.28	98.52	98.19	98.33	98.29	98.73

10.3389/frsen.2025.1583560

TABLE 6 Classification results of different methods on the Salinas-A scene (Li et al., 2018) and Botswana (Xu et al., 2019) datasets. (Best results are highlighted in red, and second-best results in blue).

Method	Salinas-A scene (Li et al., 2018)				Botswana (Xu et al., 2019)			
	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
G-MDRF (Zhang et al., 2025c)	96.28	96.33	96.30	96.24	96.37	96.20	96.44	96.39
SGD (Lei and Tang, 2021)	96.21	96.02	96.12	95.98	95.83	96.12	95.94	95.88
RFA (Chen et al., 2021)	94.46	94.50	94.38	95.02	93.87	93.61	93.76	93.92
SSFTT (Sun et al., 2022)	97.58	97.23	97.35	97.40	97.26	97.43	97.45	97.59
SSTNet (Zhang et al., 2022)	96.35	96.50	96.47	96.66	97.34	97.31	97.12	97.53
GACNet (Zhang et al., 2023c)	96.83	96.91	96.76	96.93	95.86	96.05	95.79	96.08
SCTNet (Chen et al., 2025)	98.11	98.12	98.10	98.47	97.15	97.37	97.25	97.29
PSARF (Jamali et al., 2023)	97.15	97.02	96.98	97.23	96.18	96.14	96.13	96.35
ERNet (Li et al., 2024a)	96.66	96.81	96.73	96.58	96.11	96.08	96.24	96.15
RDTN (Li et al., 2024b)	98.24	98.05	98.12	98.58	97.32	97.20	97.09	97.18
RCTNet	99.17	99.07	99.11	99.19	98.38	97.76	98.06	98.11

batch size (e.g., 32) strikes a balance between learning efficiency and generalization ability, yielding the best results for this dataset. The performance of RCTNet improves as the proportion of training samples increases, showcasing the network's ability to learn effectively from more data. This trend is observed across all evaluation metrics and further emphasizes the importance of sufficient training data for achieving optimal results.

Quantitative evaluation on the Salinas-A scene dataset (Li et al., 2018). After validating the excellent performance of RCTNet in the corn seed classification task, we further conducted experiments on the general hyperspectral remote sensing dataset to verify its robustness. In the experiment, we fixed the training batch size at 32, the testing batch size at 16, and progressively reduced the proportion of training samples to evaluate the model's adaptability to limited training data. Even under the extreme condition where the training sample proportion was reduced to 10%, RCTNet maintained satisfactory classification performance, demonstrating its strong learning capacity and adaptability. Table 6 provides detailed comparative results against various traditional and deep learning-based methods. In comparison, RCTNet consistently outperformed all competing methods across all metrics. It achieved a precision of 99.17%, a recall of 99.07%, an F1-score of 99.11%, and an accuracy of 99.19%, clearly establishing its superiority. These results not only validate the model's robustness but also demonstrate its capacity to handle diverse datasets, further solidifying its potential for general remote sensing and classification applications.

Quantitative evaluation on the Botswana dataset (Xu et al., 2019). To further validate the robustness of RCTNet in hyperspectral image classification, we conducted additional experiments on the Botswana dataset (Xu et al., 2019). The experimental setup was identical to that used for the Salinas-A scene dataset (Li et al., 2018), with a training batch size of 32 and a testing batch size of 16. When the training sample proportion was reduced to 10%, RCTNet maintained high classification performance, demonstrating its strong learning capacity and adaptability. Table 6 provides a detailed comparison against various traditional and deep learning-based methods. These results further confirm the model's robustness across diverse datasets, solidifying its potential for general remote sensing applications.

Furthermore, Figure 3a exhibitions the classification results of RCTNet on the CSHID (Zhang et al., 2022) dataset through a confusion matrix. The matrix shows that 100% accuracy was achieved in eight corn varieties. However, two samples of BaiYu8317 were misclassified as BaiYu879, leading to an overall accuracy of 91.67%. The BaiYu918 variety reached a classification accuracy of 95.83%. Additionally, Figure 3b displays the classification map generated by RCTNet on the Salinas-A scene (Li et al., 2018) dataset Figure 3c presents the classification map produced by RCTNet for the Botswana dataset (Xu et al., 2019).

4.4 Ablation study

Based on the comprehensive experiments, we further conducted ablation studies to validate the effectiveness of each component in RCTNet. The experiments were meticulously designed to analyze the impact of individual components on the network's performance. Specifically, we evaluated the following variations: 1) The proposed RCTNet without 2DWCA (-w/o 2DWCA); 2) The proposed RCTNet without 3DWSA (-w/o 3DWSA); 3) The proposed RCTNet without Residual connection (-w/o Res-Conn); 4) The proposed RCTNet cuts down one layer of the transformer structure (-c/d OneTS). In each ablation experiment, only one component of RCTNet is modified while the others remain constant. Table 7 presents the results of the ablation experiments, showing that each component positively influences RCTNet and optimal performance is achieved only when the network is fully intact. The fully intact RCTNet demonstrates superior performance

10.3389/frsen.2025.1583560

TABLE 7 Results of the ablation experiments in each module.

Method	CSHID (Zhang et al., 2022)				Salinas-A scene (Li et al., 2018)				Botswana (Xu et al., 2019)			
	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
-w/o 2DWCA	97.49	97.33	97.33	97.50	97.62	97.31	97.46	97.67	96.73	96.39	96.59	96.63
-w/o 3DWSA	96.61	96.50	96.48	96.67	97.23	96.95	97.08	97.38	96.42	96.44	96.36	96.38
-w/o Res-Conn	98.01	97.92	97.90	98.33	98.41	98.42	98.45	98.50	97.62	97.35	97.19	97.33
-c/d OneTS	97.32	97.08	97.05	97.92	98.36	98.23	98.29	98.26	97.45	97.04	97.23	97.25
RCTNet (full model)	98.89	98.75	98.76	99.17	99.17	99.07	99.11	99.19	98.38	97.76	98.06	98.11

across all metrics, achieving the highest precision, recall, F1-score, and accuracy. This confirms that each component contributes positively to the network's effectiveness, and optimal performance is attained only when RCTNet is fully integrated. These findings validate the design choices and underscore the importance of each architectural element in RCTNet.

5 Conclusion

In this work, we propose a deep learning architecture RCTNet for corn hyperspectral image classification. RCTNet is designed to effectively capture and utilize both local and global feature information, making it well-suited for hyperspectral image analysis. The network leverages a combination of Conv2D with Channel Attention, Conv3D with Spatial Attention, and residual transformer modules to enhance feature extraction and improve classification accuracy. The effectiveness and generalization are validated through experiments on the CSHID and Salinas-A scene datasets. Furthermore, RCTNet introduces a non-destructive, highly efficient approach to seed classification, making it particularly suitable for real-world applications in precision agriculture. By leveraging hyperspectral imaging for accurate seed identification, RCTNet holds significant promise in advancing the field of intelligent agriculture, contributing to more precise and efficient agricultural practices. However, despite its strong performance, RCTNet has a relatively large number of parameters, which affects computational efficiency and limits its deployment on resource-constrained devices. In future research, we aim to explore lightweight model designs to reduce computational overhead while maintaining classification accuracy.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

YiL: Conceptualization, Methodology, Software, Writing – original draft. YaL: Conceptualization, Formal Analysis, Funding acquisition, Methodology, Writing – original draft. GC: Data curation, Writing – original draft. LL: Funding acquisition, Visualization, Writing – original draft. SJ: Formal Analysis, Funding acquisition, Writing – original draft. LZ: Formal Analysis, Funding acquisition, Writing – original draft.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported in part by the China Postdoctoral Science Foundation Project under Grant 2024M750747, in part by the Henan Provincial Science and Technology Research and Development Joint Foundation Project

under Grant 235200810066, in part by the Key Research and Development Project of Henan Province under Grant 241111211800, and in part by the Key Specialized Research and Development Program of Science and Technology of Henan Province under Grants 242102210075, 252102211003, 242102211048, 242102211030, 242102211059.

Conflict of interest

The authors declare that the research was conducted without any commercial or financial relationships that could be perceived as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ahmad, M., Butt, M. H. F., Mazzara, M., Distefano, S., Khan, A. M., and Altuwaijri, H. A. (2024a). Pyramid hierarchical spatial-spectral transformer for hyperspectral image classification. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* 17, 17681–17689. doi:10.1109/jstars.2024.3461851

Ahmad, M., Ghous, U., Usama, M., and Mazzara, M. (2024b). Waveformer: spectral–spatial wavelet transformer for hyperspectral image classification. *IEEE Geoscience Remote Sens. Lett.* 21, 1–5. doi:10.1109/lgrs.2024.3353909

Barbedo, J. G. A. (2023). A review on the combination of deep learning techniques with proximal hyperspectral images in agriculture. *Comput. Electron. Agric.* 210, 107920. doi:10.1016/j.compag.2023.107920

Chen, P., He, W., Qian, F., Shi, G., and Yan, J. (2025). A synergistic cnn-transformer network with pooling attention fusion for hyperspectral image classification. *Digit. Signal Process.* 160, 105070. doi:10.1016/j.dsp.2025.105070

Chen, Y., Zheng, W., Li, W., and Huang, Y. (2021). Large group activity security risk assessment and risk early warning based on random forest algorithm. *Pattern Recognit. Lett.* 144, 1–5. doi:10.1016/j.patrec.2021.01.008

Dai, Z., Liu, H., Le, Q. V., and Tan, M. (2021). "Coatnet: marrying convolution and attention for all data sizes," *Advances in neural information processing systems*. 34, 3965–3977.

Diao, Z., Guo, P., Zhang, B., Yan, J., He, Z., Zhao, S., et al. (2023). Spatial-spectral attention-enhanced res-3d-octconv for corn and weed identification utilizing hyperspectral imaging and deep learning. *Comput. Electron. Agric.* 212, 108092. doi:10.1016/j.compag.2023.108092

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*

Hong, D., Han, Z., Yao, J., Gao, L., Zhang, B., Plaza, A., et al. (2021). Spectralformer: rethinking hyperspectral image classification with transformers. *IEEE Trans. Geoscience Remote Sens.* 60, 1–15. doi:10.1109/tgrs.2021.3130716

Jamali, A., Roy, S. K., Bhattacharya, A., and Ghamisi, P. (2023). Local window attention transformer for polarimetric sar image classification. *IEEE Geoscience Remote Sens. Lett.* 20, 1–5. doi:10.1109/lgrs.2023.3239263

Jin, S., Zhang, F., Zheng, Y., Zhou, L., Zuo, X., Zhang, Z., et al. (2023). Csknn: cost-sensitive k-nearest neighbor using hyperspectral imaging for identification of wheat varieties. *Comput. Electr. Eng.* 111, 108896. doi:10.1016/j.compeleceng.2023.108896

Lei, Y., and Tang, K. (2021). Learning rates for stochastic gradient descent with nonconvex objectives. *IEEE Trans. Pattern Analysis Mach. Intell.* 43, 4505–4511. doi:10.1109/tpami.2021.3068154

Li, F., Zhang, P., and Huchuan, L. (2018). Unsupervised band selection of hyperspectral images via multi-dictionary sparse representation. *IEEE Access* 6, 71632–71643. doi:10.1109/access.2018.2879963

Li, X., Zhai, M., Zheng, L., Zhou, L., Xie, X., Zhao, W., et al. (2024a). Efficient residual network using hyperspectral images for corn variety identification. *Front. Plant Sci.* 15, 1376915. doi:10.3389/fpls.2024.1376915

Li, Y., Yang, X., Tang, D., and Zhou, Z. (2024b). Rdtn: residual densely transformer network for hyperspectral image classification. *Expert Syst. Appl.* 250, 123939. doi:10.1016/j.eswa.2024.123939

Liang, J., Yang, Z., Bi, Y., Qu, B., Liu, M., Xue, B., et al. (2024). A multi-tree genetic programming-based feature construction approach to crop classification using hyperspectral images. *IEEE Trans. Geoscience Remote Sens.* 62, 1–17. doi:10.1109/tgrs.2024.3415773

Liang, L., Zhang, Y., Zhang, S., Li, J., Plaza, A., and Kang, X. (2023). Fast hyperspectral image classification combining transformers and simam-based cnns. *IEEE Trans. Geoscience Remote Sens.* 61, 1–19. doi:10.1109/tgrs.2023.3309245

Paoletti, M. E., Moreno-Álvarez, S., Xue, Y., Haut, J. M., and Plaza, A. (2023). Aatt-cnn: automatic attention-based convolutional neural networks for hyperspectral image classification. *IEEE Trans. Geoscience Remote Sens.* 61, 1–18. doi:10.1109/tgrs.2023.3272639

Sellami, A., and Tabbone, S. (2022). Deep neural networks-based relevant latent representation learning for hyperspectral image classification. *Pattern Recognit.* 121, 108224. doi:10.1016/j.patcog.2021.108224

Shi, C., Yue, S., and Wang, L. (2024). A dual branch multiscale transformer network for hyperspectral image classification. *IEEE Trans. Geoscience Remote Sens.* 62, 1–20. doi:10.1109/tgrs.2024.3351486

Song, L., Feng, Z., Yang, S., Zhang, X., and Jiao, L. (2024). Interactive spectral-spatial transformer for hyperspectral image classification. *IEEE Trans. Circuits Syst. Video Technol.* 34, 8589–8601. doi:10.1109/tcsvt.2024.3386578

Su, Y., Gao, L., Jiang, M., Plaza, A., Sun, X., and Zhang, B. (2022). Nsckl: normalized spectral clustering with kernel-based learning for semisupervised hyperspectral image classification. *IEEE Trans. Cybern.* 53, 6649–6662. doi:10.1109/tcyb.2022.3219855

Sun, G., Pan, Z., Zhang, A., Jia, X., Ren, J., Fu, H., et al. (2023). Large kernel spectral and spatial attention networks for hyperspectral image classification. *IEEE Trans. Geoscience Remote Sens.* 61, 1–15. doi:10.1109/tgrs.2023.3292065

Sun, L., Zhao, G., Zheng, Y., and Wu, Z. (2022). Spectral–spatial feature tokenization transformer for hyperspectral image classification. *IEEE Trans. Geoscience Remote Sens.* 60, 1–14. doi:10.1109/tgrs.2022.3144158

Tian, F., Lei, S., Zhou, Y., Cheng, J., Liang, G., Zou, Z., et al. (2024). Hirenet: hierarchical-relation network for few-shot remote sensing image scene classification. *IEEE Trans. Geoscience Remote Sens.* 62, 1–10. doi:10.1109/tgrs.2023.3348464

Wang, D., Zhang, J., Du, B., Zhang, L., and Tao, D. (2023). Dcn-t: dual context network with transformer for hyperspectral image classification. *IEEE Trans. Image Process.* 32, 2536–2551. doi:10.1109/tip.2023.3270104

Xu, H., Zhang, X., Li, H., Xie, L., Dai, W., Xiong, H., et al. (2022). Seed the views: hierarchical semantic alignment for contrastive representation learning. *IEEE Trans. Pattern Analysis Mach. Intell.* 45, 3753–3767. doi:10.1109/TPAMI.2022.3176690

Xu, R., Dong, X.-M., Li, W., Peng, J., Sun, W., and Xu, Y. (2024). Dbctnet: double branch convolution-transformer network for hyperspectral image classification. *IEEE Trans. Geoscience Remote Sens.* 62, 1–15. doi:10.1109/tgrs.2024.3368141

Xu, Y., Du, B., Zhang, L., Cerra, D., Pato, M., Carmona, E., et al. (2019). Advanced multi-sensor optical remote sensing for urban land use and land cover classification: outcome of the 2018 ieee grss data fusion contest. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* 12, 1709–1724. doi:10.1109/jstars.2019.2911113

Yan, H., Zhang, E., Wang, J., Leng, C., Basu, A., and Peng, J. (2023). Hybrid conv-vit network for hyperspectral image classification. *IEEE Geoscience Remote Sens. Lett.* 20, 1–5. doi:10.1109/lgrs.2023.3287277

Yang, L., Yang, Y., Yang, J., Zhao, N., Wu, L., Wang, L., et al. (2022). Fusionnet: a convolution–transformer fusion network for hyperspectral image classification. *Remote Sens.* 14, 4066. doi:10.3390/rs14164066

Yu, H., Xu, Z., Zheng, K., Hong, D., Yang, H., and Song, M. (2022). Mstnet: a multilevel spectral–spatial transformer network for hyperspectral image classification. *IEEE Trans. Geoscience Remote Sens.* 60, 1–13. doi:10.1109/tgrs.2022.3186400

Yuan, Y., Lin, L., Zhou, Z.-G., Jiang, H., and Liu, Q. (2023). Bridging optical and sar satellite image time series via contrastive feature extraction for crop classification. *ISPRS J. Photogrammetry Remote Sens.* 195, 222–232. doi:10.1016/j.isprsjprs.2022.11.020

Zhang, B., Chen, Y., Li, Z., Xiong, S., and Lu, X. (2023a). Sanet: a self-attention network for agricultural hyperspectral image classification. *IEEE Trans. Geoscience Remote Sens.* 62, 1–15. doi:10.1109/tgrs.2023.3341473

Zhang, B., Chen, Y., Rong, Y., Xiong, S., and Lu, X. (2023b). Matnet: a combining multi-attention and transformer network for hyperspectral image classification. *IEEE Trans. Geoscience Remote Sens.* 61, 1–15. doi:10.1109/tgrs.2023.3254523

Zhang, J., Dai, L., and Cheng, F. (2021). Corn seed variety classification based on hyperspectral reflectance imaging and deep convolutional neural network. *J. Food Meas. Charact.* 15, 484–494. doi:10.1007/s11694-020-00646-3

Zhang, L., Zhang, S., Liu, J., Wei, Y., An, D., and Wu, J. (2024a). Maize seed variety identification using hyperspectral imaging and self-supervised learning: a two-stage training approach without spectral preprocessing. *Expert Syst. Appl.* 238, 122113. doi:10.1016/j.eswa.2023.122113

Zhang, W., Li, Z., Li, G., Zhou, L., Zhao, W., and Pan, X. (2024b). Aganet: attention-guided generative adversarial network for corn hyperspectral images augmentation. *IEEE Trans. Consumer Electron.*, 1. doi:10.1109/tce.2024.3470846

Zhang, W., Li, Z., Li, G., Zhuang, P., Hou, G., Zhang, Q., et al. (2023c). Gacnet: generate adversarial-driven cross-aware network for hyperspectral wheat variety identification. *IEEE Trans. Geoscience Remote Sens.* 62, 1–14. doi:10.1109/tgrs.2023.3347745

Zhang, W., Li, Z., Sun, H.-H., Zhang, Q., Zhuang, P., and Li, C. (2022). Sstnet: spatial, spectral, and texture aware attention network using hyperspectral image for corn variety identification. *IEEE Geoscience Remote Sens. Lett.* 19, 1–5. doi:10.1109/lgrs.2022.3225215

Zhang, Y., Duan, P., Liang, L., Kang, X., Li, J., and Plaza, A. (2025a). Pfs3f: probabilistic fusion of superpixel-wise and semantic-aware structural features for hyperspectral image classification. *IEEE Trans. Circuits Syst. Video Technol.*, 1. doi:10.1109/tcsvt.2025.3556548

Zhang, Y., Liang, L., Mao, J., Wang, Y., and Jia, L. (2025b). From global to local: a dual-branch structural feature extraction method for hyperspectral image classification. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* 18, 1778–1791. doi:10.1109/jstars.2024.3509538

Zhang, Y., Liu, L., and Yang, X. (2025c). Hyperspectral image classification using spectral-spatial dual random fields with Gaussian and markov processes. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* 18, 4199–4212. doi:10.1109/jstars.2025.3528115

Zhao, W., Li, W., Li, Y., Yang, L., Liang, Z., Hu, E., et al. (2025). Constructing balanced training samples: a new perspective on long-tailed classification. *IEEE Trans. Multimedia*, 1–14. doi:10.1109/tmm.2025.3543084

Zheng, X., Sun, H., Lu, X., and Xie, W. (2022). Rotation-invariant attention network for hyperspectral image classification. *IEEE Trans. Image Process.* 31, 4251–4265. doi:10.1109/tip.2022.3177322