forests

MDPI

Article

Spectral Weaver: A Study of Forest Image Classification Based on SpectralFormer

Haotian Yu [1,†], Xuyang Li [2,†], Xinggui Xu [1,*], Hong Li [1] and Xiangsuo Fan [2,3]

[1] School of Information, Yunnan University of Finance and Economics, Kunming 650500, China; yuhaotian@yeah.net (H.Y.); zz2108@ynufe.edu.cn (H.L.)
[2] School of Automation, Guangxi University of Science and Technology, Liuzhou 545006, China; 221077062@stdmail.gxust.edu.cn (X.L.); 100002085@gxust.edu.cn (X.F.)
[3] Guangxi Collaborative Innovation Centre for Earthmoving Machinery, Guangxi University of Science and Technology, Liuzhou 545006, China
* Correspondence: zz2146@ynufe.edu.cn
† These authors contributed equally to this work and should be considered co-first authors.

Abstract: In forest ecosystems, the application of hyperspectral (HS) imagery offers unprecedented opportunities for refined identification and classification. The diversity and complexity of forest cover make it challenging for traditional remote-sensing techniques to capture subtle spectral differences. Hyperspectral imagery, however, can reveal the nuanced changes in different tree species, vegetation health status, and soil composition through its nearly continuous spectral information. This detailed spectral information is crucial for the monitoring, management, and conservation of forest resources. While Convolutional Neural Networks (CNNs) have demonstrated excellent local context modeling capabilities in HS image classification, their inherent network architecture limits the exploration and representation of spectral feature sequence properties. To address this issue, we have rethought HS image classification from a sequential perspective and proposed a hybrid model, the Spectral Weaver, which combines CNNs and Transformers. The Spectral Weaver replaces the traditional Multi-Head Attention mechanism with a Channel Attention mechanism (MCA) and introduces Centre-Differential Convolutional Layers (Conv2d-cd) to enhance spatial feature extraction capabilities. Additionally, we designed a cross-layer skip connection that adaptively learns to fuse "soft" residuals, transferring memory-like components from shallow to deep layers. Notably, the proposed model is a highly flexible backbone network, adaptable to both hyperspectral and multispectral image inputs. In comparison to traditional Visual Transformers (ViT), the Spectral Weaver innovates in several ways: (1) It introduces the MCA mechanism to enhance the mining of spectral feature sequence properties; (2) It employs Centre-Differential Convolutional Layers to strengthen spatial feature extraction; (3) It designs cross-layer skip connections to reduce information loss; (4) It supports both multispectral and hyperspectral inputs, increasing the model's flexibility and applicability. By integrating global and local features, our model significantly improves the performance of HS image classification. We have conducted extensive experiments on the Gaofen dataset, multispectral data, and multiple hyperspectral datasets, validating the superiority of the Spectral Weaver model in forest hyperspectral image classification. The experimental results show that our model achieves 98.59% accuracy on multispectral data, surpassing ViT's 96.30%. On the Jilin-1 dataset, our proposed algorithm achieved an accuracy of 98.95%, which is 2.17% higher than ViT. The model significantly outperforms classic ViT and other state-of-the-art backbone networks in classification performance. Not only does it effectively capture the spectral features of forest vegetation, but it also significantly improves the accuracy and robustness of classification, providing strong technical support for the refined management and conservation of forest resources.

check for updates

Academic Editor: Eraldo Aparecido Trondoli Matricardi

Received: 29 November 2024
Revised: 17 December 2024
Accepted: 23 December 2024
Published: 26 December 2024

Citation: Yu, H.; Li, X.; Xu, X.; Li, H.; Fan, X. Spectral Weaver: A Study of Forest Image Classification Based on SpectralFormer. *Forests* **2025**, *16*, 21. https://doi.org/10.3390/f16010021

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Keywords: spectral weaver; center-differential convolutional layer; cross-layer jump connections

1. Introduction

Remote-sensing images [1], with their near-continuous spectral information, are able to finely identify materials by capturing subtle spectral differences [2]. This property makes HS images have great potential for high-level Earth-observation (EO) tasks such as precision agriculture, target detection, urban planning, tree species classification, and mineral exploration [3–5]. Especially in forest ecosystems, the application of hyperspectral imagery provides unprecedented opportunities for fine identification and classification [6]. The diversity and complexity of forest cover make it difficult for traditional remote-sensing techniques to capture subtle spectral differences, while hyperspectral images can reveal subtle changes in different tree species, vegetation health status, and soil composition through their near-continuous spectral information. Such fine spectral information is crucial for monitoring, management, and protection of forest resources [7,8].

Hyperspectral image classification is an important task in the field of remote sensing, and its goal is to identify different feature classes by analyzing the spectral information in the images. Traditional classification methods mainly rely on hand-designed feature extraction and classifiers, such as support vector machine (SVM) and random forest (RF) [9,10]. However, these methods often perform poorly when dealing with complex and diverse HS images, especially when the training samples are limited. In recent years, deep learning (DL) techniques have achieved remarkable success in image classification tasks, especially in the application of Convolutional Neural Networks (CNNs) [11,12]. CNNs are capable of automatically extracting spatial features in an image through multi-layer convolutional and pooling operations, thus significantly improving the classification performance. However, CNNs have some inherent limitations in processing HS images. First, CNN focuses mainly on the spatial structure information of the image and ignores the importance of the spectral sequence information. Second, CNNs perform poorly in capturing spectral sequence attributes (especially medium- and long-term dependencies), which leads to a performance bottleneck in HS image classification tasks. Specifically, CNNs perform poorly in capturing spectral sequence attributes (especially medium- and long-term dependencies), which inevitably leads to performance bottlenecks in HS image classification tasks, especially when the categories to be classified are diverse and the spectral features are extremely similar. In addition, CNNs focus too much on spatial content information, resulting in distorted spectral sequence information in the learned features, which increases the difficulty of mining diagnostic spectral attributes. In recent years, multimodal deep learning (MDL) has demonstrated significant advantages in processing complex and high-dimensional data [13]. By integrating information from different modalities, multimodal data fusion techniques are able to provide more comprehensive and accurate feature representations, thus achieving breakthroughs in various applications. In hyperspectral image classification tasks, single-modal features are often insufficient to capture complex feature information, so combining multimodal data for classification has become an important research direction.

In the realm of hyperspectral image classification, researchers have proposed various deep learning architectures aimed at enhancing performance. Among them, the Multimodal Deep Convolutional Neural Network (M3DDCNN) is designed to capture spatial and spectral features more comprehensively by integrating hyperspectral images with other relevant modalities, therefore improving classification accuracy [14]. However, there are practical challenges associated with this approach. First, the acquisition and preprocessing of multimodal data can be complex, requiring substantial computational resources and

time. Second, noise and errors in different modalities may adversely affect the fusion process, leading to a decline in model performance. Additionally, the high complexity of the M3DDCNN results in lengthy training and inference times, which can limit its applicability in real-world scenarios. In addition to M3DDCNN, Another architecture, the Residual and Spatial Attention Network (RSSAN), effectively captures spatial and spectral features through the introduction of residual connections and a spatial attention mechanism. The objective of this network is to enhance classification performance when dealing with classes that exhibit similar spectral characteristics. However, RSSAN also has its limitations, including high model complexity and extended training and inference times, which may restrict its practical utility. Moreover, processing large-scale hyperspectral data may encounter computational resource limitations, and the similarity of spectral features might lead to overfitting issues [15]. The Attention-Based Long Short-Term Memory network (ABLSTM) aims to improve classification performance by adaptively focusing on the crucial parts of spectral sequences through the incorporation of an attention mechanism [16]. The goal is to enhance the LSTM's ability to capture key information from spectral sequences. Nevertheless, ABLSTM also faces blanks such as high model complexity, long training and inference times, and potential issues with computational resource constraints and overfitting when dealing with large-scale hyperspectral data and classes with similar spectral features. The Deep Feature Fusion Network (DFFN) addresses the issue of traditional Convolutional Neural Networks (CNNs) neglecting spectral sequence information by integrating multi-level feature representations to simultaneously capture spatial and spectral characteristics [17]. Despite its significant achievements in improving classification performance, DFFN also grapples with blanks like high model complexity and extended training and inference times. It may also encounter computational resource limitations and overfitting when processing large-scale hyperspectral data and classes with similar spectral features. To overcome these challenges, researchers are exploring deep learning-based solutions such as the Self-Supervised Feature Transformation Network (SSFTT) [18]. SSFTT aims to enhance the model's feature extraction capabilities through self-supervised learning, with the objective of boosting the performance of hyperspectral image classification tasks. However, the design of SSFTT is not without its blanks; the self-supervised task design is complex and requires considerable computational resources and time for training. Moreover, the performance of the self-supervised task is contingent on the rationality of the task design; poor design may lead to the learning of irrelevant features, thus affecting classification performance. SSFTT may also face computational resource constraints when processing large-scale hyperspectral data.

In the current research on hyperspectral image classification, several critical gaps and research objectives have been identified. First, traditional machine learning methods such as Support Vector Machines and Random Forests, while contributing to classification accuracy, have gaps in feature extraction and dimensionality handling. These methods often require manual feature extraction, which is not only time-consuming but may also fail to fully exploit the multidimensional nature of the data. Moreover, when faced with the high dimensionality of hyperspectral images, these traditional approaches are prone to the curse of dimensionality, limiting their applicability. Therefore, one of the research objectives is to explore more effective feature extraction and dimensionality reduction techniques. Second, the application of Convolutional Neural Networks (CNNs) in hyperspectral image classification has made significant progress, but it has also revealed some gaps. CNNs may overlook the spectral sequential nature of the images when processing spectral information, which can affect classification performance. Additionally, CNN models have a large number of parameters, requiring extensive training data to prevent overfitting, and the training process demands high computational resources. Hence, the research focus is

shifting towards how to better integrate spectral and spatial information and how to optimize network structures for improved efficiency and accuracy. Deep Belief Networks (DBNs), as a form of unsupervised learning, have the research objective of extracting more discriminative features through pre-training in hyperspectral image classification. However, the training process of DBNs is complex and time-consuming, and it may face computational efficiency issues when dealing with large-scale data. Furthermore, optimizing the pre-training phase is a significant challenge, representing gaps that future research needs to address. Finally, ensemble learning methods aim to enhance classification stability and accuracy by combining the predictions of multiple classifiers, but they also face gaps. Ensemble learning requires substantial computational resources, and how to select and combine different classifiers to achieve optimal performance remains an open question. Additionally, the results of ensemble learning may lack interpretability, which is an issue that researchers need to consider. In summary, these gaps and research objectives point the way forward for the development of the hyperspectral image classification field.

In order to overcome these limitations, researchers began to explore the application of the Transformer architecture to HS image classification [19–21]. Transformer was originally designed for Natural Language Processing (NLP) tasks, and its core idea is to capture global dependencies in sequential data through the Self-Attention mechanism. Unlike CNNs and Recurrent Neural Networks (RNNs), Transformers are able to efficiently deal with long-distance dependencies and thus excel in processing sequential data. However, the classical Transformer architecture has some problems when processing HS images. First, Transformer mainly focuses on global sequence information and neglects the capture of local spectral details. Second, Transformer is deficient in cross-layer information transfer, resulting in the loss of valuable information during layer-by-layer propagation. To address these issues, we propose a hybrid model combining CNN and Transformer, which significantly improves the performance of HS image classification by introducing a Channel Attention mechanism (MCA), a Center-Differential Convolutional Layer (Conv2d-cd), and Cross-Layer Jump Connections (CAF). Specifically, MCA better captures the sequential properties of spectral features, Conv2d-cd enhances spatial feature extraction, and CAF passes memory-like components from shallow to deep layers by adaptively learning to fuse "soft" residuals across layers. These improvements enable our model to capture both global and local features when processing HS images, which significantly improves the classification performance.

The main contributions of this paper are as follows:

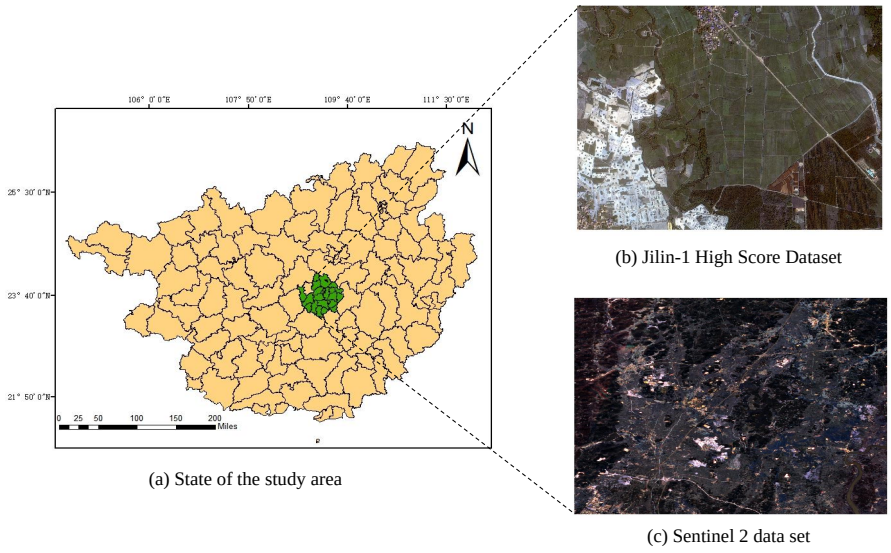
(1) MCA: we introduce the Channel Attention mechanism (MCA) instead of the traditional multiple attention mechanism. The MCA is able to better capture the sequential attributes of the spectral features, thus improving the classification performance.
(2) Center-Differential Convolutional Layer (Conv2d-cd): we introduce a Center-Differential Convolutional Layer (Conv2d-cd) to enhance the spatial feature extraction capability. This convolutional layer is able to add a center difference term to the standard convolution, thus enhancing the spatial feature extraction capability of the model.
(3) Cross-layer Jump Connection (CJC): We design a Cross-Layer Jump Connection that passes memory-like components from shallow to deep layers by adaptively learning to fuse "soft" residuals across layers. This design reduces the possibility of losing valuable information during layer-by-layer propagation.
(4) Flexible backbone network: our model is a highly flexible backbone network that can be applied to hyperspectral images and multispectral image inputs. Integrating global and local features significantly improves the performance of forest hyperspectral image classification.

The paper is structured as follows: Section 2 details the material data as well as the design of the proposed model and its various modules; Section 3 describes the experimental results and provides an in-depth analysis; Section 4 discusses potential research directions; and, finally, Section 5 encapsulates the key contributions of this work and envisions potential avenues for future investigation.

2. Study Area and Data

2.1. Overview of the Research Area

Xingbin District of Laibin City is in the central part of Guangxi Zhuang Autonomous Region and the central part of Laibin City. It borders Wuxuan County in the east, Qintang District of Guigang City in the south, Liujiang District of Liuzhou City in the west, and Xincheng County in the north. The district has a total area of 4364.5 km^2, which is the largest district in Laibin City. Xingbin District has diverse topography and geomorphology, including hills, plains, rivers, lakes, etc. It is rich in feature types, such as farmland, forests, and towns. The rich feature types provide sufficient training samples and validation data, which is conducive to improving the accuracy and generalization ability of the classification model. The forest coverage rate of Xingbin District reaches more than 45%, with rich vegetation types, such as evergreen broad-leaved forests, coniferous forests, shrub forests, grasslands, and so on. The high vegetation cover provides good conditions for vegetation classification research, which can be carried out in forest resources surveys, vegetation cover monitoring, and ecological environment assessment. Xingbin District is an important agricultural area in Guangxi, and the main crops include rice, sugar cane, fruits and vegetables. Agricultural remote sensing can be applied to crop planting area monitoring, growth monitoring, yield prediction, disaster assessment, etc., providing a scientific basis for agricultural production. In summary, Xingbin District, as a study area for remote-sensing image classification, has the advantages of rich feature types, high vegetation coverage, developed agriculture, convenient transportation, and policy support, which makes it an ideal area for remote-sensing image classification research. At the same time, using Jilin-1 L3D data, the study area district Fenghuang Town, Huang'an high-quality double-high sugar cane base area, at 11:30:30 on 4 October 2024, the data JK1KF01C, panchromatic image resolution: 0.5 m, bands: blue, green, red, near-infrared, radiometric accuracy of less than 5%, geometric accuracy of less than 2 m (CE90) to carry out experiments. The geographical location of the study area is shown in Figure 1.



(a) State of the study area

(b) Jilin-1 High Score Dataset

(c) Sentinel 2 data set

Figure 1. Situation of the study area.

2.2. Image Preprocessing of the Study Area

This study employs the Sentinel-2 satellite, equipped with a Multispectral Instrument (MSI), to capture multispectral images covering the research area. The MSI serves as the data source. Initially, the Sen2Cor2.10 software is utilized to geocode the Level 1C data, which involves atmospheric correction to eliminate the impact of the atmosphere on the images, as well as image enhancement and mask extraction. Subsequently, the SNAP 8.0 software is used to resample the band resolution to 10 m. Finally, the ENVI 5.6 software is employed to select all bands, and the Jilin-1 high-resolution data are used to mark the Region of Interest (ROI) to obtain the sample library data, with 60% for training, 20% for validation, and 20% for testing. The categories and numbers of ground objects are shown in Figure 2. In this study, the Sentinel-2 data size is (3312, 2881, 12) for the mirror image shape, while the Jilin-1 data size is (2896, 3955, 4) for the mirror image shape. Sentinel-2 is suitable for large-scale, long-time series analysis but may encounter issues with cloud coverage and insufficient resolution; Jilin-1 is suitable for high-resolution tasks but has a limited coverage area and may be more complex to acquire data. Therefore, this paper opts to use Sentinel-2 for validating the model over extensive regions, while employing Jilin-1 for focused, small-area analysis (Table 1).

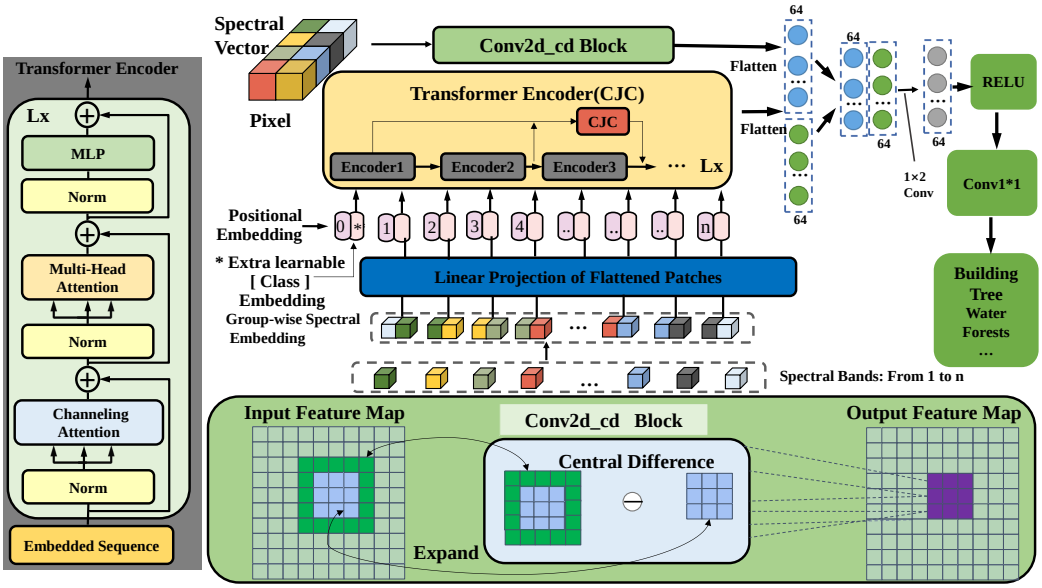


Figure 2. Schematic of the Spectrum Weaver architecture.

Table 1. Jilin 1 High Score and Sentinel 2 Sample Bank Data.

NO.	Class	Jilin-1 High Score Dataset			Sentinel 2 Data Set		
		Train	Val	Test	Train	Val	Test
1	River	2815	939	939	2250	750	750
2	Buildup	2101	700	701	1891	631	631
3	sugarcane	2620	873	874	2007	669	670
4	Forests	1999	666	667	2056	686	686
5	Road	1879	627	627	1765	588	589
6	Pond	3229	1076	1077	1605	535	536
7	Barren	1909	637	637	1771	591	591

2.3. The Architecture Proposed in This Paper Spectral Weaver

The parallel network architecture Spectral Weaver proposed in this paper, which integrates the Multi-Head Attention mechanism with the multi-scale residual cascade, is shown

in Figure 2, which mainly consists of the Center-Differential Convolution, Cross-Layer Hopping Connection, Channel Attention, and Transformer. First, the model architecture Spectral Weaver proposed in this paper combines the advantages of Vision Transformer (ViT) and Convolutional Neural Network (CNN) by segmenting remote-sensing images into small pieces and extracting the spectral information, utilizing linear embedding and positional coding to convert spectral vectors into Then a Cross-Layer Jump Connection is designed to pass memory-like components from shallow to deep layers by adaptively learning to fuse "soft" residuals across layers. After multiple Transformer layers for global feature extraction, and at the same time through the center-differential convolution layer to extract local features, the features extracted from the ViT and center-differential convolution layers are finally spliced and fused to form a high-dimensional feature representation, and the feature expression is enhanced by the introduction of Channel Attention (CA), and finally input into the classifier to make classification decisions, so as to improve the remote-sensing image classification.

2.3.1. Channeling Attention Mechanisms

In the traditional Transformer architecture, the Multi-Head Attention (MHA) mechanism captures different features in the input sequence by computing multiple attention heads in parallel. However, MHA has some limitations in processing hyperspectral (HS) images, mainly in its insufficient ability to capture the attributes of spectral feature sequences. To overcome this problem, we introduce the Channel Attention mechanism (MCA), which replaces the traditional MHA. The MCA is able to better capture the sequence attributes of the spectral features, thus improving the classification performance. Specifically, MCA generates an attention matrix through the interaction of query (Q), key (K), and value (V) and performs a weighted summation of the values. In this way, instead of the traditional multi-attention mechanism, MCA is able to better capture the sequential attributes of spectral features, thus improving the classification performance. The core idea of MCA is to generate the attention matrix through the interaction of Q, K, and V, and perform weighted summation on the values. Specifically, the computational process of MCA can be divided into the following steps:

1. Input embedding: first, the input hyperspectral image sequence is embedded to obtain the representation of Q, K, and V;
2. Attention matrix computation: the attention matrix is generated by computing the dot product between the Q and K. In order to stabilize the gradient, the dot product result is usually scaled, i.e.,

$$s_{i,j} = \frac{q_i \cdot k_j}{\sqrt{d}}_{\omega} \quad (1)$$

where q_i and k_j denote the i-th query and j-th key, respectively, and d is the dimension of the query and key.

3 Attention weight calculation: softmax operation is performed on the attention matrix to obtain the attention weight;

4 Weighted summation: weighted summation of values (V) using attention weights to obtain the final attention representation:

$$z_i = \sum_j \hat{s}_{i,j} v_{j+} \quad (2)$$

MCA has several significant advantages over traditional multi-attention mechanisms:

1. Better Spectral Feature Capture: the MCA is able to better capture the sequence properties of spectral features through the interaction of Q, K, and V. Specifically,

MCA is able to effectively capture medium and long-term dependencies in spectral sequences, thus improving classification performance.

2. Stronger expression ability: MCA is able to express the relationship between spectral features more flexibly by generating an attention matrix and performing weighted summation. This flexibility allows MCA to perform well when dealing with complex and diverse HS images.
3. Higher Classification Accuracy: Experimental results in the next section show that MCA significantly outperforms the traditional multi-attention mechanism in classification accuracy on multiple hyperspectral datasets. Specifically, our model achieves significant improvements in overall accuracy (OA), average accuracy (AA), and Kappa coefficient (K).

2.3.2. Center Difference Convolutional LayerCenter Difference Convolutional Layer (Conv2d-cd)

In traditional Convolutional Neural Networks (CNNs), a standard convolutional layer extracts local spatial features by sliding a convolutional kernel over the input image. However, the standard convolutional layer has limitations in capturing subtle spatial changes in the image. To enhance the spatial feature extraction capability of the model, we introduce the Center-Differential Convolutional Layer (Conv2d-cd). This convolutional layer can add a center difference term to the standard convolution, thus enhancing the spatial feature extraction ability of the model.

The core idea of the Center-Differential Convolutional Layer (Conv2d-cd) is to enhance the model's ability to capture subtle spatial variations in an image by adding a center difference term to the standard convolution. Specifically, the computational process of Conv2d-cd can be divided into the following steps:

1. Standard Convolution: First, the input image is subjected to standard convolution operation to obtain a preliminary feature map. The formula for standard convolution is as follows:

$$\text{Conv2d}(x, w) = \sum_{i,j} w_{i,j} \cdot x_{i,j^{\omega}} \tag{3}$$

where x denotes the input image, w denotes the convolution kernel, and i and j denote the row and column indices of the convolution kernel, respectively.

2. Center Difference Term: Adds a center difference term to the standard convolution. The center difference term captures the subtle spatial changes in the image by calculating the difference between the center positions of the convolution kernels. The formula for calculating the center difference term is as follows:

$$\text{CD}(x, w) = \sum_{i,j} w_{i,j} \cdot (x_{i,j} - x_{c,c})_{\omega} \tag{4}$$

In the actual implementation, the calculation process of Conv2d-cd can be represented by the following equation:

$$Conv2d - cd(x, w) = Conv2\theta(x, w) + \theta \cdot \sum_{i,j} w_{i,j} \cdot (x_{i,j} - x_{c,c})_{\omega} \tag{5}$$

where θ is a learnable parameter that controls the weight of the central difference term. In this way, Conv2d-cd can effectively enhance the spatial feature extraction ability of the model, thus improving the classification performance.

2.3.3. Cross-Layer Jump Connection (CJC)

In deep neural networks, valuable information is often lost during layer-by-layer propagation, especially in deep networks. To address this problem, we design a Cross-Layer Jump Connection (CJC) that passes memory-like components from shallow to deep layers by adaptively learning to fuse "soft" residuals across layers. This design improves the performance of the model by reducing the possibility of losing valuable information during layer-by-layer propagation. The core idea of cross-layer jump connectivity (CJC) is to transfer valuable information from the shallow layer to the deep layer by adaptively learning to fuse "soft" residuals across layers. Specifically, the computational process of CJC can be divided into the following steps:

1. Shallow feature extraction: first, features are extracted from the shallow network. These features contain rich low-level information such as edges, textures, etc.
2. Deep feature extraction: next, features are extracted from the deep network. These features contain high-level information such as semantic information, global structure, etc.
3. Adaptive fusion: shallow features and deep features are fused by means of adaptive learning. Specifically, we introduce a learnable parameter matrix for controlling the fusion ratio of shallow and deep features. The formula for adaptive fusion is as follows:

$$\mathrm{CJC}(x^{(l)}, x^{(l-2)}) = \hat{w} \left[\begin{matrix} x^{(l)} \\ x^{(l-2)} \end{matrix} \right]_{\omega} \quad (6)$$

where $x^{(l)}$ denotes the features in layer l,$x^{(l-2)}$ denotes the features in layers l–2, $\hat{w}$ is a learnable parameter matrix that controls the fusion ratio of the features.

4 Feature passing: the fused features are passed to the next layer to continue feature extraction and classification.

Compared with traditional hopping connections (e.g., short hopping connections in ResNet), CJC has several significant advantages:

1. Reducing information loss: CJC passes valuable information from the shallow layer to the deep layer by means of adaptive learning, thus reducing the possibility of losing valuable information during layer-by-layer propagation.
2. Enhance information transfer: CJC enhances the information transfer between shallow and deep layers through Cross-Layer Jump Connection, which enables the model to better utilize the features of both shallow and deep layers.
3. Improvement of model robustness: the adaptive fusion mechanism of CJC makes the model better adaptable to the features of different layers, thus improving the robustness of the model.

We significantly improve the performance of remote sensing spectral image classification by introducing cross-layer jump connectivity (CJC). CJC passes valuable information from the shallow layer to the deeper layer by means of adaptive learning, thus reducing the possibility of losing valuable information in the layer-by-layer propagation process. In the future, we will further explore the application of CJC in other remote sensing tasks and investigate how to further optimize the performance of CJC.

2.4. Evaluation Indicators and Loss Functions

When training and evaluating deep learning models in pixel classification of remote sensing images, it is crucial to choose appropriate evaluation metrics and loss functions. The evaluation metrics are used to measure the performance of the model, while the loss function is used to guide the training process of the model. In order to comprehensively

evaluate the model's classification performance, the overall accuracy (OA) can intuitively reflect the model's classification performance on all samples and is an important metric for evaluating the model's performance. The average accuracy (AA) can reflect the classification performance of the model on each category, which is suitable for the case of category imbalance. The Kappa coefficient takes into account the effect of random classification, which can more accurately assess the classification consistency of the model. The cross-entropy loss function is widely used in classification tasks, which can effectively measure the difference between the model prediction results and the real labels and guide the learning process of the model.

1. Overall Accuracy (OA): Overall accuracy is the proportion of all samples that are correctly categorized.

$$\mathrm{OA} = \frac{\Sigma_{i=1}^{C} \mathrm{TP}_i}{\sum_{i=1}^{C} (\mathrm{TP}_i + \mathrm{FN}_i)^{\downarrow}} \tag{7}$$

where C denotes the number of categories, TP denotes the number of correctly categorized samples in category i, and FN denotes the number of incorrectly categorized samples in category i.

2 Average Accuracy (AA): Average Accuracy is the average of the classification accuracy of all categories.

$$\mathrm{AA} = \frac{1}{C} \sum_{i=1}^{C} \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i^{\omega}} \tag{8}$$

3 Kappa coefficient (K): The Kappa coefficient is a statistical measure of categorization consistency that takes into account the effects of random categorization.

$$\mathrm{K} = \frac{\mathrm{OA} - \mathrm{p}_e}{1 - \mathrm{p}_e \ _{\omega}} \tag{9}$$

where OA denotes the overall accuracy and pe denotes the expected accuracy of randomized classification, calculated as:

$$\mathrm{p}_e = \frac{\sum_{i=1}^{C} (\mathrm{TP}_i + \mathrm{FP}_i) \cdot (\mathrm{TP}_i + \mathrm{FN}_i)}{\left(\sum_{i=1}^{C} (\mathrm{TP}_i + \mathrm{FN}_i) \right)^2} \tag{10}$$

where FP denotes the number of samples in category i that were misclassified as other categories.

CrossEntropyLoss: The cross-entropy loss function is used to measure the difference between the probability distribution predicted by the model and the true label.

$$\mathrm{CrossEntropyLoss} = -\sum_{i=1}^{C} y_i \log(\widehat{y}_i)_{\omega} \tag{11}$$

where C denotes the number of categories, denotes the true label of category i (0 or 1), and denotes the probability of category i predicted by the model.

3. Results

3.1. Ablation Study

To assess the effectiveness of the algorithm proposed in this paper, ablation experiments were carried out. The results are detailed in Table 2. The OA of the standard ViT model is 96.30%, which is an impressive classification result, indicating that ViT is highly appropriate for the domain of remote sensing image classification. Although the OA of the ViT combined with the MCA module (ViT+MCA) is improved to 96.97%, this also attests to the efficacy of the MCA scheme in achieving accurate image classification. The

introduction of Conv2d-cd raises the OA to 97.79%, with a slight improvement in AA and Kappa coefficient, indicating that on multispectral data, the addition of Conv2d-cd may enhance the classification performance of ViT for specific categories. When all modules are integrated, the Spectral Weaver achieves an accuracy of 98.59%, representing a notable improvement in OA compared to the ViT alone.

Table 2. Results of ablation experiments on the study area dataset using spectral weavers with different combinations of modules.

Different Methods	Different Module			Metric		
	MCA	Conv2d-cd	CJC	OA (%)	AA (%)	Kappa
ViT	×	×	×	96.30	96.20	95.65
ViT	✓	×	×	96.97	96.84	96.27
Spectral Weaver	×	✓	×	97.79	97.86	97.41
Spectral Weaver	×	✓	✓	98.17	98.22	97.85
Spectral Weaver	✓	✓	×	98.33	98.43	98.04
Spectral Weaver	×	✓	✓	98.42	98.47	98.15
Spectral Weaver	✓	✓	✓	**98.59**	**98.67**	**98.34**

3.2. Multi-Methods for Experimentation

3.2.1. Multi-Method Comparison of Gilead 1 and Sentinel 2

The correctness of all categories of Jilin I, as well as the Sentinel 2 study area dataset and the total OA, AA, and kappa results of different methods, are shown in Tables 3 and 4. It can be learned that VIT has the worst results and the lowest of all three evaluation metrics, but it also achieves relatively good results. It can be observed that in Sentinel-2 data, the ViT algorithm has achieved a very commendable result of 96.30%. However, the performance for the Barren category is relatively subpar; the dffn algorithm also does not fare well in classifying this category. The other three algorithms are fairly good, but our algorithm has reached 99.89%, which is an outstanding outcome. In the classification of Jilin-1 data, only the Pond category was not fully recognized by the m3ddcnn algorithm. Additionally, none of the other algorithms achieved above 90% accuracy in classifying the Road category, except for our algorithm. Moreover, for all other categories, our algorithm consistently achieved accuracy rates above 99%, which amply demonstrates the potential of our method.

Table 3. Jilin-1 Land Use and Coverage Results.

C N.	Different Methods						
	ViT	Speformer	m3ddcnn	Rssan	Ablstm	Dffn	Spectral Weaver
River	95.45	92.22	91.54	92.93	96.05	96.09	96.58
Buildup	97.28	97.47	97.19	96.57	97.00	97.04	98.09
sugarcane	99.38	99.61	99.80	99.92	99.88	99.96	100.00
Forests	97.29	98.54	97.54	98.64	98.54	98.59	99.44
Road	94.03	95.95	96.59	91.75	92.86	98.24	98.24
Pond	95.94	95.85	97.86	97.89	94.89	96.03	98.38
Barren	93.97	98.32	96.54	99.42	99.16	93.55	99.89
OA(%)	96.30	96.66	96.68	96.77	96.85	97.07	98.59
AA(%)	96.20	96.86	96.73	96.74	96.92	97.08	98.67
Kappa	95.65	96.07	96.09	96.21	96.30	96.56	98.34

The classification map of the Jilin 1 high score L3D dataset obtained using various methods is shown in Figure 3, and it can be seen that Spectral Weaver performs the best on

Forests, which is almost never misclassified, especially in the species crossover condition it can be clearly seen that Spectral Weaver is clearer than other algorithms.

Table 4. Sentinel 2 Land Use and Coverage Results.

C N.	Different Methods						
	ViT	Speformer	m3ddcnn	Rssan	Ablstm	Dffn	Spectral Weaver
River	99.20	98.82	99.77	99.77	99.77	99.73	99.91
Buildup	95.76	96.19	98.20	99.10	97.51	99.47	99.47
sugarcane	99.95	98.95	99.95	100.00	99.95	99.95	99.85
Forests	99.75	98.75	99.90	99.80	99.70	99.90	99.90
Road	82.71	88.00	88.49	86.96	89.68	89.51	93.65
Pond	100.00	100.00	99.87	100.00	100.00	100.00	100.00
Barren	98.81	99.15	98.64	99.83	99.20	98.98	99.37
OA(%)	96.78	96.96	97.97	98.06	98.09	98.34	98.95
AA(%)	96.60	96.84	97.84	97.93	97.98	98.22	98.88
Kappa	96.23	96.62	97.63	97.73	97.77	98.06	98.77

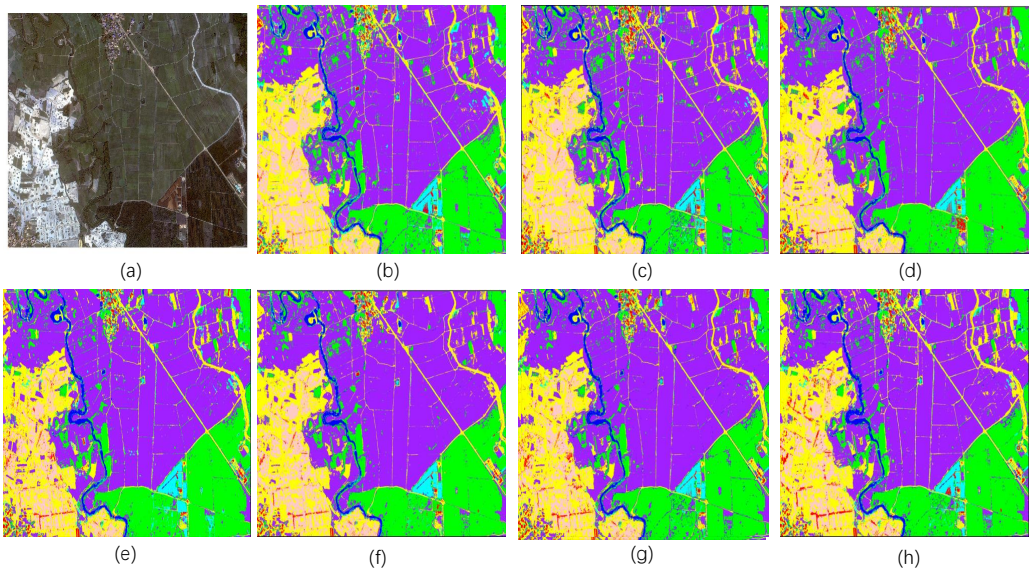


Figure 3. Classification results obtained by different models on the Jilin 1 high-resolution L3D dataset, (**a**) Image. (**b**) ViT. (**c**) speformer. (**d**) m3ddcnn. (**e**) rssan. (**f**) ablstm. (**g**) dffn. (**h**) Spectral Weaver.

This can be seen as shown in Figure 4 as the results are all good and all close. The proposed Spectral Weaver prediction map effect on the Sentinel 2 dataset is also clearer in detail compared to the other algorithms, and it is still Spectral Weaver that does a better job of edge classification for the object class in this case, where there are fewer bands.

3.2.2. Hyperspectral Multi-Method Comparison

The following table shows the details of the hyperspectral HyRANK-Loukia, WHU-Hi-HongHu, and WHU-Hi-LongKou datasets as well as the sample delineation information as shown in Table 5, which will be used for the experiments in this section.

Tabular plots of the classification results for hyperspectral HyRANK-Loukia, WHU-Hi-HongHu, and WHU-Hi-LongKou data are shown in Tables 6–8.

In the hyperspectral classification shown in Tables 6 and 7, our algorithm shows significant improvement, which is due to the fact that each category is well classified. The other two effective learning algorithms might also incur some errors in such complex environments. In Table 8, the improvement of our algorithm is not as pronounced because all three

algorithms are already at 99%. We speculate that the categories in this case are relatively easy to distinguish, which also confirms the superiority of our comparative algorithm.

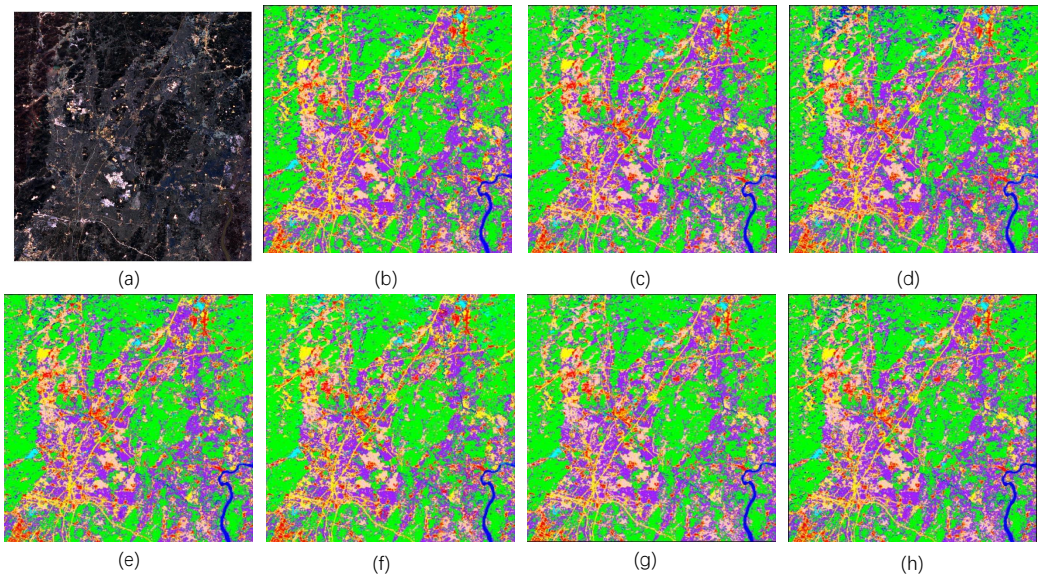


Figure 4. Classification results obtained by different models on Sentinel 2 dataset, (**a**) Image. (**b**) ViT. (**c**) speformer. (**d**) m3ddcnn. (**e**) rssan. (**f**) ablstm. (**g**) dffn. (**h**) Spectral Weaver.

Table 5. Status of hyperspectral datasets.

Dataset	**HyRANK-Loukia**		**WHU-Hi-HongHu**		**WHU-Hi-LongKou**	
Sensor	Hyperspectral Imager		DJI Matrice 600 Pro		DJI M600 Pro	
platform	Small Satellite Platform		unmanned aerial vehicle		unmanned aerial vehicle	
Loc and Time	European, 2018		Honghu City, 2017		Longkou Town, 2018	
GSD	30 m		0.043 m		0.463 m	
Wavelength	400–1000 nm		400–1000 nm		400–1000 nm	
Data size	250 × 1000 × 176		940 × 475 × 270		550 × 400 × 270	
#	Class	Samples	Class	Samples	Class	Samples
1	Dense Urban Fabric	288	Red roof	1404	Corn	34,511
2	Mineral Extraction Sites	67	Road	351	Cotton	31,897
3	Non-Irrigated Arable Land	542	Bare soil	2182	Sesame	3031
4	Fruit Trees	79	Cotton	17,732	Broad-leaf soybean	63,212
5	Olive Groves	1401	Cotton firewood	973	Narrow-leaf soybean	4151
6	Broad-leaved Forest	223	Rape	6638	Rice	11,854
7	Coniferous Forest	500	Chinese cabbage	20,142	Water	67,056
8	Mixed Forest	1072	Pakchoi	1378	Roads and houses	7124
9	Dense Sclerophyllous Vegetation	3793	Cabbage	7720	Mixed weed	5229
10	Sparse Sclerophyllous Vegetation	2803	Tuber mustard	21,381		
11	Sparsely Vegetated Areas	404	*Brassica parachinensis*	2480		
12	Rocks and Sand	487	*Brassica chinensis*	8615		
13	Water	1393	Small *Brassica chinensis*	23,632		
14	Coastal Water	451	*Lactuca sativa*	3216		
			Celtuce	8715		
			Film covered lettuce	24,358		
			Romaine lettuce	3517		
			Carrot	9037		
			White radish	25,229		
			Garlic sprout	3866		
			Broad bean	9170		
			Tree	25,633		

Table 6. HyRANK-Loukia classification results table.

Class		SPEFORMER	M3D-DCNN	Spectral Weaver
1	Dense Urban Fabric	38.92 ± 13.71	55.75 ± 3.91	67.82 ± 9.15
2	Mineral Extraction Sites	100.00 ± 0.00	95.33 ± 5.31	79.51 ± 19.15
3	Non-Irrigated Arable Land	69.84 ± 9.15	86.89 ± 2.73	92.69 ± 2.14
4	Fruit Trees	1.69 ± 2.25	30.70 ± 8.42	35.73 ± 11.62
5	Olive Groves	76.83 ± 4.00	88.44 ± 1.41	93.95 ± 1.87
6	Broad-leaved Forest	12.24 ± 9.66	47.26 ± 13.14	55.92 ± 11.54
7	Coniferous Forest	42.71 ± 2.61	63.82 ± 6.50	70.84 ± 5.46
8	Mixed Forest	66.34 ± 8.98	74.38 ± 3.61	82.65 ± 6.49
9	Dense Sclerophyllous Vegetation	76.14 ± 2.80	84.18 ± 1.95	83.45 ± 2.15
10	Sparse Sclerophyllous Vegetation	72.20 ± 1.60	82.13 ± 1.96	83.54 ± 2.54
11	Sparsely Vegetated Areas	64.68 ± 5.52	69.09 ± 10.28	76.45 ± 11.54
12	Rocks and Sand	92.74 ± 1.97	94.38 ± 1.59	94.41 ± 1.41
13	Water	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
14	Coastal Water	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
	OA(%)	74.47 ± 0.72	83.38 ± 0.52	86.16 ± 0.54
	AA(%)	65.31 ± 1.44	76.60 ± 1.60	79.45 ± 2.54
	Kappa	69.48 ± 0.93	80.17 ± 0.66	84.56 ± 0.41
	inference	76.45 it/s	196.57 it/s	247.41 it/s

Table 7. Table of WHU-Hi-HongHu classification results.

Class	SPEFORMER	M3D-DCNN	Spectral Weaver
Red roof	98.47 ± 0.30	96.51 ± 0.27	99.31 ± 0.51
Road	94.41 ± 1.72	89.95 ± 2.85	98.15 ± 0.87
Bare soil	96.27 ± 0.78	96.10 ± 0.67	99.24 ± 0.38
Cotton	99.59 ± 0.13	99.67 ± 0.09	99.95 ± 0.04
Cotton firewood	90.83 ± 3.67	88.60 ± 1.65	98.68 ± 1.32
Rape	98.78 ± 0.26	98.04 ± 0.48	99.78 ± 0.16
Chinese cabbage	95.32 ± 0.38	92.09 ± 1.62	99.12 ± 0.44
Pakchoi	75.62 ± 2.47	64.40 ± 7.02	95.65 ± 2.12
Cabbage	99.09 ± 0.12	98.54 ± 0.49	99.01 ± 0.99
Tuber mustard	90.40 ± 1.44	90.93 ± 1.01	98.12 ± 1.25
Brassica parachinensis	90.39 ± 0.98	89.41 ± 3.25	97.69 ± 0.79
Brassica chinensis	88.56 ± 2.15	82.95 ± 1.19	98.08 ± 0.81
Small *Brassica chinensis*	93.32 ± 0.70	89.94 ± 1.72	98.98 ± 0.64
Lactuca sativa	94.65 ± 1.56	95.15 ± 0.88	98.79 ± 0.57
Celtuce	90.33 ± 2.17	87.52 ± 4.82	97.77 ± 1.98
Film covered lettuce	98.60 ± 0.35	97.63 ± 1.40	99.73 ± 0.19
Romaine lettuce	92.57 ± 2.53	94.94 ± 4.54	98.85 ± 0.96
Carrot	91.37 ± 1.25	94.25 ± 0.54	98.29 ± 1.52
White radish	94.71 ± 0.65	93.70 ± 1.19	97.81± 1.35
Garlic sprout	94.44 ± 1.88	93.03 ± 1.46	97.81 ± 1.21
Broad bean	82.53 ± 10.69	43.60 ± 6.08	96.84 ± 1.95
Tree	93.80 ± 1.44	86.67 ± 5.11	98.49 ± 0.48
OA(%)	96.84 ± 0.18	95.74 ± 0.22	99.05 ± 0.17
AA(%)	92.91 ± 0.81	89.25 ± 0.53	98.01 ± 0.15
Kappa	96.00 ± 0.23	94.61 ± 0.28	98.87 ± 0.41
inference	38.01 it/s	110.49 it/s	161.19 it/s

Table 8. WHU-Hi-LongKou classification result table.

Class	SPEFORMER	M3D-DCNN	Spectral Weaver
Corn	99.87 ± 0.04	99.73 ± 0.13	99.98 ± 0.01
Cotton	98.77 ± 0.38	98.65 ± 0.40	99.82 ± 0.11
Sesame	99.14 ± 0.58	97.09 ± 1.40	99.87 ± 0.16
Broad-leaf soybean	99.62 ± 0.06	99.60 ± 0.08	99.98 ± 0.01
Narrow-leaf soybean	96.62 ± 1.04	95.69 ± 1.59	99.67 ± 0.48
Rice	99.87 ± 0.13	99.87 ± 0.14	99.99 ± 0.02
Water	99.99 ± 0.00	99.99 ± 0.00	99.98 ± 0.01
Roads and houses	98.01 ± 0.72	98.19 ± 0.49	99.05 ± 0.65
Mixed weed	98.72 ± 0.34	97.14 ± 0.39	99.55 ± 0.35
OA(%)	99.61 ± 0.07	99.50 ± 0.07	99.91 ± 0.06
AA(%)	98.96 ± 0.26	98.44 ± 0.17	99.72 ± 0.06
Kappa	99.49 ± 0.09	99.34 ± 0.09	99.89 ± 0.07
inference	41.30 it/s	129.31 it/s	167.24 it/s

Figures 5–7 show the classification results of hyperspectral HyRANK-Loukia, WHU-Hi-HongHu, and WHU-Hi-LongKou data.

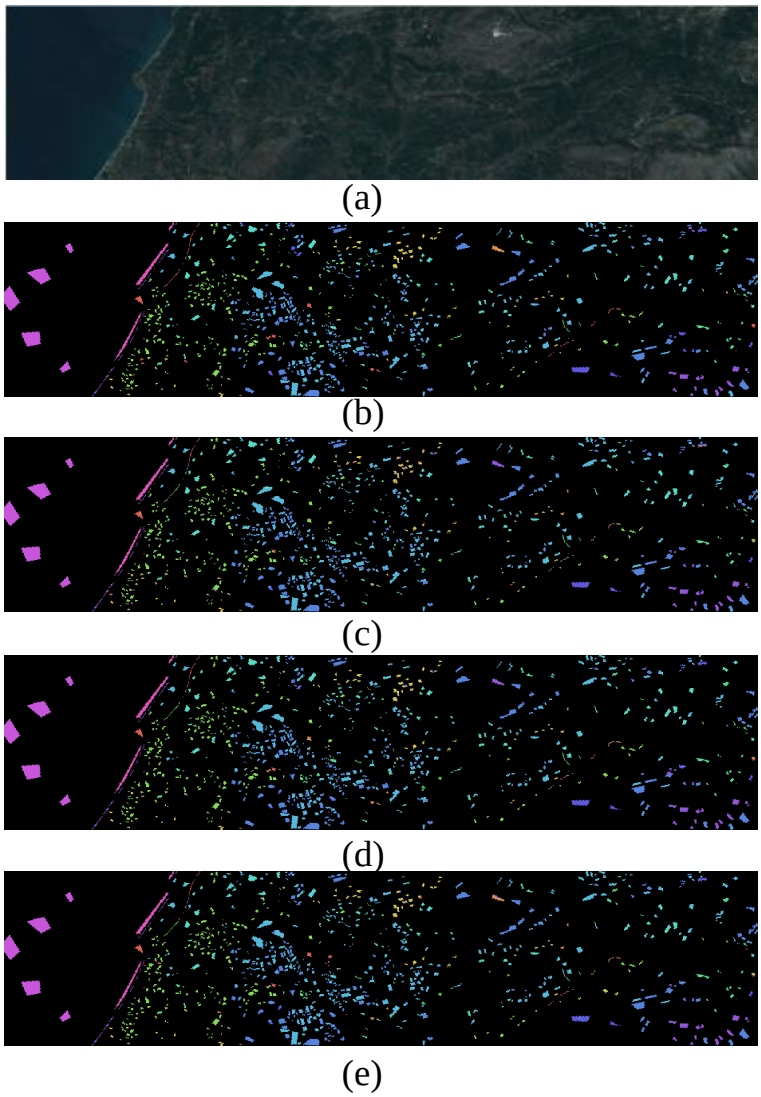


Figure 5. Classification results obtained by different models on the HyRANK-Loukia dataset, (**a**) Image. (**b**) gt. (**c**) speformer. (**d**) m3ddcnn. (**e**) Spectral Weaver.

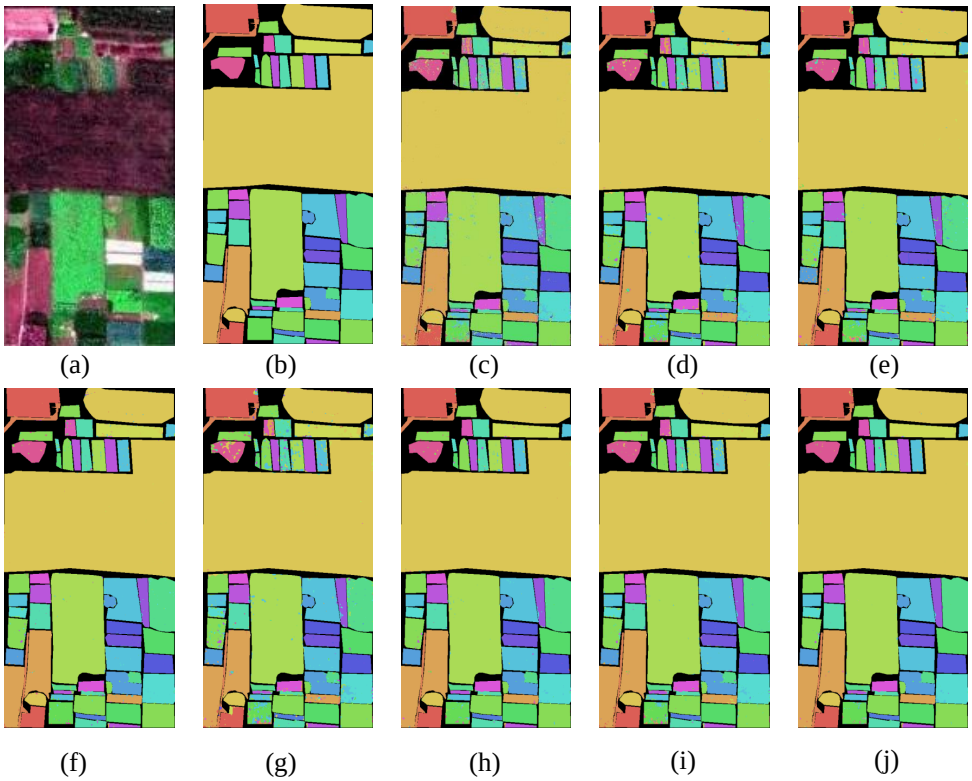


Figure 6. Classification results obtained by different models on the WHU-Hi-HongHu dataset, (**a**) Image.(**b**) gt. (**c**) ViT. (**d**) speformer. (**e**) m3ddcnn. (**f**) rssan. (**g**) ablstm. (**h**) dffn. (**i**) ssftt. (**j**) Spectral Weaver.

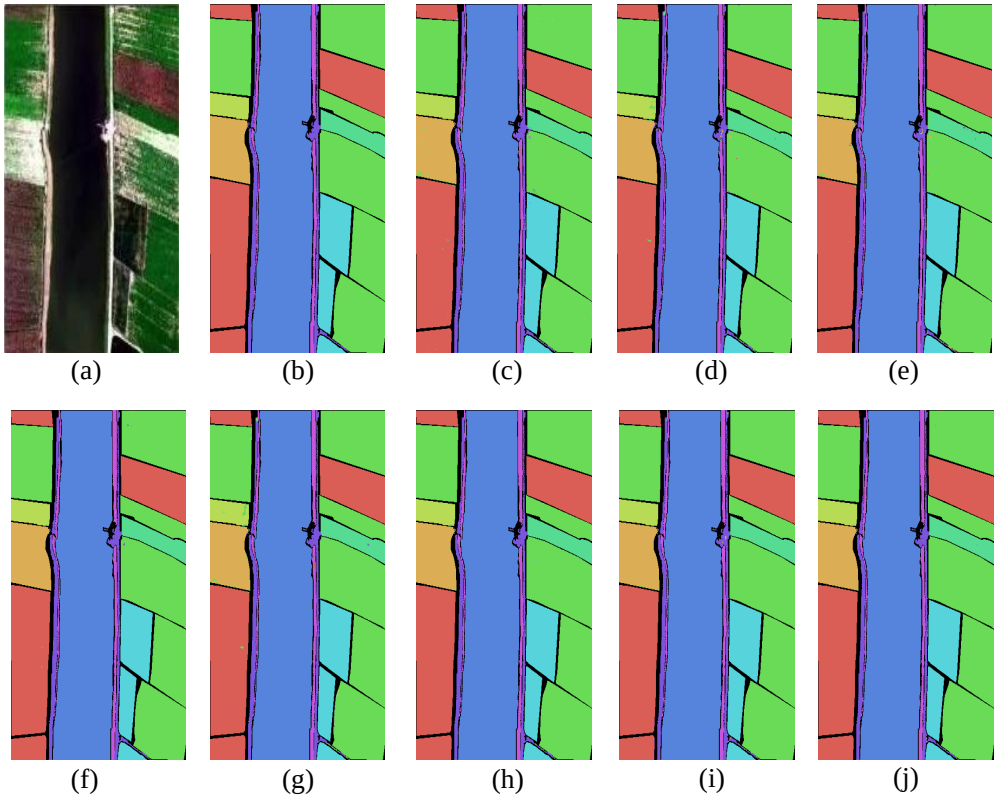


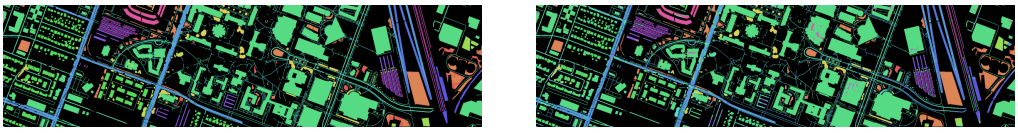
Figure 7. Classification results obtained by different models on the WHU-Hi-LongKou dataset, (**a**) Image.(**b**) gt. (**c**) ViT. (**d**) speformer. (**e**) m3ddcnn. (**f**) rssan. (**g**) ablstm. (**h**) dffn. (**i**) ssftt. (**j**) Spectral Weaver.

In addition to this, we looked for the latest dataset, Houston 2018 Dataset, As shown in Table 9 below.

Table 9. Houston 2018 Dataset Sample Dataset.

NO.	Class	Train 0.05	Val 0.05	Test 0.9
1	Healthy grass	490	490	8819
2	Stressed grass	1625	1625	29,252
3	Artificial turf	34	34	616
4	Evergreen trees	679	680	12,236
5	Deciduous trees	251	251	4519
6	Bare earth	226	226	4064
7	Water	14	13	239
8	Residential buildings	1989	1989	35,795
9	Non-residential buildings	11,187	11,188	201,377
10	Roads	2293	2293	41,280
11	Sidewalks	1702	1702	30,626
12	Crosswalks	76	76	1366
13	Major thoroughfares	2317	2318	41,713
14	Highways	493	493	8879
15	Railways	347	347	6243
16	Paved parking lots	575	575	10,350
17	Unpaved parking lots	8	7	131
18	Cars	328	327	5892
19	Trains	269	268	4832
20	Stadium seats	341	341	6142

The predicted plot of the experimental results is shown in Figure 8 below.



(a) Houston_gt (b) Houston_proposed_out

Figure 8. Houston 2018 Dataset Data Classification Chart.

The training loss as well as correctness graphs are shown below in Figure 9.

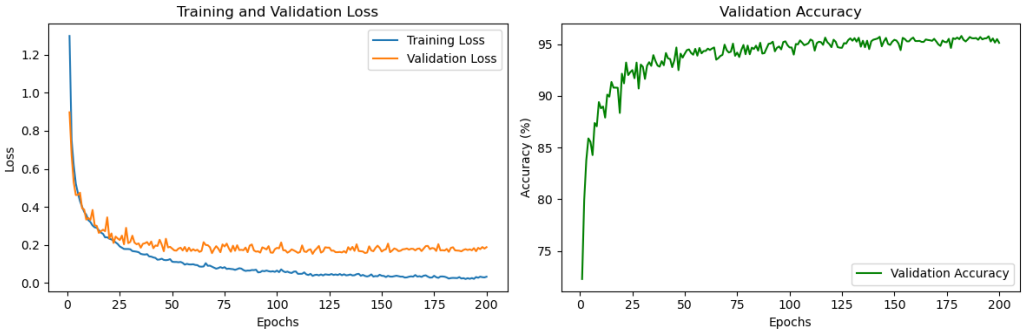


Figure 9. Houston 2018 Dataset LOSS and correctness plot.

It can be seen that our proposed algorithm has also played well in hyperspectral datasets in recent years, which also fully proves the applicability of our algorithm.

4. Discussion

The Spectral Weaver model proposed in this paper innovates in several aspects, significantly enhancing the performance of hyperspectral image classification. First, the

introduction of the Multi-Channel Attention (MCA) mechanism replaces the traditional Multi-Head Attention mechanism, effectively enhancing the mining of sequential attributes in spectral features. MCA, through the interaction of Q, K, and V, captures long-term and medium-term dependencies in spectral sequences more effectively, therefore improving classification performance. Second, the inclusion of the Central Difference Convolution layer (Conv2d-cd) strengthens the model's spatial feature extraction capabilities. Conv2d-cd adds a central difference term to the standard convolution, effectively capturing subtle spatial changes within images and further enhancing the model's classification accuracy. Additionally, the design of Cross-Layer Jump Connections (CJC) reduces information loss during layer-by-layer propagation by adaptively learning to fuse "soft" residuals across layers, therefore enhancing the model's robustness and information transmission capabilities. Experimental results on hyperspectral datasets (such as HyRANK-Loukia, WHU-Hi-HongHu, and WHU-Hi-LongKou) further validate the superiority of the Spectral Weaver. On these datasets, the classification accuracy of the Spectral Weaver significantly surpasses that of other comparative models, especially in classes with similar spectral features, where the Spectral Weaver can classify more accurately. This indicates that the Spectral Weaver is more effective in addressing the challenge of class diversity with similar spectral features in hyperspectral image classification tasks. The Spectral Weaver model has a highly flexible backbone network suitable for both hyperspectral and multispectral image inputs. This flexibility extends the applicability of the Spectral Weaver to various types of remote-sensing image classification tasks. Experimental results show that the Spectral Weaver not only excels in hyperspectral image classification tasks but also has significant advantages in multispectral image classification, further demonstrating the model's flexibility and applicability. Through experimentation, we found that MCA significantly enhances the model's classification performance when processing spectral sequence data, particularly excelling in classes with similar spectral features. The spatial feature extraction capability of Conv2d-cd is outstanding in capturing subtle spatial changes, especially in complex scenes, leading to more accurate target identification. The optimization of information transmission by CJC significantly reduces information loss and enhances the model's robustness, which is particularly evident in multi-layer networks. However, some limitations were encountered; although the Spectral Weaver performs excellently in classification, it has a high computational complexity, especially when dealing with large-scale datasets, where the model's training time and resource consumption are substantial. Therefore, our scope cannot be expanded to the provincial level. Despite the Spectral Weaver's superior performance in most categories, classification errors may still occur in a few categories. These are issues that we will need to consider in our future research. Although the Spectral Weaver has achieved significant improvements in classification performance on multiple hyperspectral datasets, there are still potential research directions worth exploring further. First, the design of MCA and Conv2d-cd could be further optimized to explore more advanced attention mechanisms and convolution operations to enhance the model's performance. Second, the research could focus on how to embed physical properties (such as spectral absorption characteristics) and prior knowledge (such as the distribution of ground object categories) into the model to improve its interpretability and generalization capabilities. Additionally, one could explore the design of lightweight Transformer architectures that maintain high performance while reducing the model's complexity and computational cost. Finally, further research could investigate the application of the Spectral Weaver in other remote-sensing tasks (such as target detection, change detection, etc.) to validate its versatility and effectiveness in different tasks.

5. Conclusions

Within this article, we introduce a hybrid model named Speformer, which integrates CNN and Transformer architectures. Speformer significantly enhances the performance of hyperspectral image classification through an innovative multi-channel attention mechanism (MCA), Central-Differential Convolutional Layer (Conv2d-cd), and cross-layer skip connections (CAF). Experimental results demonstrate that Speformer achieves substantial improvements in OA, AA, and Kappa coefficient on both the Jilin-1 and Sentinel-2 datasets. For instance, on the Jilin-1 dataset, the classification accuracy reaches 98.95%, which is 2.17% higher than the classical ViT model, and the overall accuracy on the multispectral dataset is 98.59%, significantly outperforming ViT's 96.30%. Particularly in the Forests category, Speformer excels with almost no misclassifications, especially in situations with species overlap, where it can achieve clearer classification, indicating its significant advantage in handling complex and diverse HS images. Moreover, Speformer not only effectively captures the spectral features of forest vegetation but also significantly improves the accuracy and robustness of classification, providing strong technical support for the refined management and conservation of forest resources. This study successfully achieves the goals of enhancing the mining of spectral feature sequence attributes, strengthening the extraction of spatial features, and optimizing information transmission and robustness through the introduction of MCA, Conv2d-cd, and CAF. Although Speformer has achieved remarkable improvements in classification performance on multiple hyperspectral datasets, future research can explore more advanced attention mechanisms, self-supervised learning methods, the integration of physical characteristics and prior knowledge, as well as lightweight architecture design to further advance the field of remote-sensing spectral image classification.

Author Contributions: Conceptualization, H.Y., X.L., X.X. and X.F.; methodology, H.Y., X.L., X.X., H.L. and X.F.; software, H.Y., X.L., X.X., H.L. and X.F.; validation, H.Y., X.L., X.X., H.L. and X.F.; formal analysis, H.Y., X.L., X.X., H.L. and X.F.; investigation, H.Y., X.L., X.X., H.L. and X.F.; resources, H.Y., X.L., X.X., H.L. and X.F.; data curation, H.Y., X.L., X.X., H.L. and X.F.; writing—original draft preparation, H.Y., X.L., X.X., H.L. and X.F.; writing—review and editing, H.Y., X.L., X.X., H.L. and X.F. All authors have read and agreed to the published version of the manuscript.

Funding: The work was supported by the National Natural Science Foundation of China under Grant No. 62161051.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sun, X.; Wang, P.; Yan, Z.; Xu, F.; Wang, R.; Diao, W.; Chen, J.; Li, J.; Feng, Y.; Xu, T.; et al. FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *184*, 116–130. [CrossRef]
2. Xu, Q.; Li, Y.; Nie, J.; Liu, Q.; Guo, M. UPanGAN: Unsupervised pansharpening based on the spectral and spatial loss constrained generative adversarial network. *Inf. Fusion* **2023**, *91*, 31–46. [CrossRef]
3. Saranya, T.; Deisy, C.; Sridevi, S.; Anbananthen, K.S.M. A comparative study of deep learning and Internet of Things for precision agriculture. *Eng. Appl. Artif. Intell.* **2023**, *122*, 106034. [CrossRef]
4. Lei, F.; Tang, F.; Li, S. Underwater target detection algorithm based on improved YOLOv5. *J. Mar. Sci. Eng.* **2022**, *10*, 310. [CrossRef]
5. Fan, X.; Li, X.; Yan, C.; Fan, J.; Chen, L.; Wang, N. Converging Channel Attention Mechanisms with Multilayer Perceptron Parallel Networks for Land Cover Classification. *Remote Sens.* **2023**, *15*, 3924. [CrossRef]
6. Li, X.; Fan, X.; Li, Q.; Zhao, X. RS-Net: Hyperspectral Image Land Cover Classification Based on Spectral Imager Combined with Random Forest Algorithm. *Electronics* **2024**, *13*, 4046. [CrossRef]

7. Li, X.; Fan, X.; Fan, J.; Li, Q.; Gao, Y.; Zhao, X. DASR-Net: Land Cover Classification Methods for Hybrid Multiattention Multispectral High Spectral Resolution Remote Sensing Imagery. *Forests* **2024**, *15*, 1826. [CrossRef]
8. Gschwantner, T.; Alberdi, I.; Bauwens, S.; Bender, S.; Borota, D.; Bosela, M.; Bouriaud, O.; Breidenbach, J.; Donis, J.; Fischer, C.; et al. Growing stock monitoring by European National Forest Inventories: Historical origins, current methods and harmonisation. *For. Ecol. Manag.* **2022**, *505*, 119868. [CrossRef]
9. Cervantes, J.; Garcia-Lamont, F.; Rodríguez-Mazahua, L.; Lopez, A. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing* **2020**, *408*, 189–215. [CrossRef]
10. Speiser, J.L.; Miller, M.E.; Tooze, J.; Ip, E. A comparison of random forest variable selection methods for classification prediction modeling. *Expert Syst. Appl.* **2019**, *134*, 93–101. [CrossRef]
11. Amini, S.; Saber, M.; Rabiei-Dastjerdi, H.; Homayouni, S. Urban land use and land cover change analysis using random forest classification of landsat time series. *Remote Sens.* **2022**, *14*, 2654. [CrossRef]
12. Vasan, D.; Alazab, M.; Wassan, S.; Safaei, B.; Zheng, Q. Image-Based malware classification using ensemble of CNN architectures (IMCEC). *Comput. Secur.* **2020**, *92*, 101748. [CrossRef]
13. Choi, J.H.; Lee, J.S. EmbraceNet: A robust deep learning architecture for multimodal classification. *Inf. Fusion* **2019**, *51*, 259–270. [CrossRef]
14. He, M.; Li, B.; Chen, H. Multi-scale 3D deep convolutional neural network for hyperspectral image classification. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 3904–3908.
15. Zhu, M.; Jiao, L.; Liu, F.; Yang, S.; Wang, J. Residual spectral–spatial attention network for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 449–462. [CrossRef]
16. Chen, Z.; Zhang, L.; Jiang, C.; Cao, Z.; Cui, W. WiFi CSI based passive human activity recognition using attention based BLSTM. *IEEE Trans. Mob. Comput.* **2018**, *18*, 2714–2724. [CrossRef]
17. Wei, Q.; Tian, X.; Cui, L.; Zheng, F.; Liu, L. WSAFormer-DFFN: A model for rotating machinery fault diagnosis using 1D window-based multi-head self-attention and deep feature fusion network. *Eng. Appl. Artif. Intell.* **2023**, *124*, 106633. [CrossRef]
18. Sun, L.; Zhao, G.; Zheng, Y.; Wu, Z. Spectral–spatial feature tokenization transformer for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
19. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 87–110. [CrossRef]
20. Vaswani, A. Attention is all you need. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2017.
21. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.

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