



OPEN CMTNet: a hybrid CNN-transformer network for UAV-based hyperspectral crop classification in precision agriculture

Xihong Guo¹, Quan Feng² & Faxu Guo^{2✉}

Hyperspectral imaging acquired from unmanned aerial vehicles (UAVs) offers detailed spectral and spatial data that holds transformative potential for precision agriculture applications, such as crop classification, health monitoring, and yield estimation. However, traditional methods struggle to effectively capture both local and global features, particularly in complex agricultural environments with diverse crop types, varying growth stages, and imbalanced data distributions. To address these challenges, we propose CMTNet, an innovative deep learning framework that integrates convolutional neural networks (CNNs) and Transformers for hyperspectral crop classification. The model combines a spectral-spatial feature extraction module to capture shallow features, a dual-branch architecture that extracts both local and global features simultaneously, and a multi-output constraint module to enhance classification accuracy through cross-constraints among multiple feature levels. Extensive experiments were conducted on three UAV-acquired datasets: WHU-Hi-LongKou, WHU-Hi-HanChuan, and WHU-Hi-HongHu. The experimental results demonstrate that CMTNet achieved overall accuracy (OA) values of 99.58%, 97.29%, and 98.31% on these three datasets, surpassing the current state-of-the-art method (CTMixer) by 0.19% (LongKou), 1.75% (HanChuan), and 2.52% (HongHu) in OA values, respectively. These findings indicate its superior potential for UAV-based agricultural monitoring in complex environments. These results advance the precision and reliability of hyperspectral crop classification, offering a valuable solution for precision agriculture challenges.

Keywords Hyperspectral imaging, Crop classification, Multi-output feature constraints, Convolutional neural networks, Transformer

Accurate identification of crop types is crucial for agricultural monitoring, crop yield estimation, growth analysis, and determining the spatial distribution and area of crops¹. It also provides essential reference information for resource allocation, agricultural structure adjustment, and the formulation of economic development strategies in the agricultural production process. In recent years, with the continuous improvement of spectral imaging technology, hyperspectral imaging (HSI) has become a research hotspot for remote sensing data analysis^{2,3}. HSI images consist of hundreds or even thousands of spectral channels containing abundant spatial and spectral information. The high spatial resolution of HSI provides new opportunities for detecting subtle spectral differences between crops, which is beneficial for the fine classification of crops. In addition, HSI is widely used in areas such as plant disease detection⁴, food inspection⁵, reidentification⁶, and geological exploration⁷.

Traditional methods for HSI classification typically include the designed loss⁸ and the designed model⁹. In addition, scholars have also introduced several methods for HSI spectral dimension reduction and information extraction, including principal component analysis, minimum noise fraction transformation, linear discriminant analysis, independent component analysis, and others. However, these methods only consider the spectral information of HSI, ignoring the spatial correlation between pixels in the spatial dimension. This ignores the spatial features contained in the HSI data and ignored rich spatial contextual information, leading to variability in the spectral features of target objects, thus affecting classification performance. To utilize spatial information in the images, scholars have studied various mathematical morphology operators suitable for HSI to extract spatial features from the images, including morphological profile features, extended morphological profile features, extended multi-attribute profile features (EMAP), and extinction profile features^{10,11}. Although

¹Dingxi Sanniu Agricultural Machinery Manufacturing Co., Ltd., Dingxi 743000, China. ²College of Mechanical and Electrical Engineering, Gansu Agriculture University, Lanzhou 730070, China. ✉email: guofax@gsau.edu.cn

hyperspectral image classification methods based on spatial features can effectively capture the spatial information such as the position, structure, and contours of target objects, they neglect the spectral dimension information of hyperspectral imaging, resulting in less than ideal classification results. The generalization and versatility of traditional HSI classification methods are weak and susceptible to salt and pepper noise, which affects classification accuracy.

In recent years, many deep learning-based methods have been applied to HSI classification^{12,13}, as illustrated in Fig. 1. Initially, deep belief networks (DBN)¹⁴, recurrent neural networks (RNN)¹⁵, and one-dimensional convolutional neural networks (1D-CNN)¹⁶ was introduced into the HSI classification field. However, these methods only utilize spectral information and ignore the neighborhood information in the spatial dimension, leading to lower classification accuracy¹⁷. To address this issue, researchers proposed an architecture based on two-dimensional convolutional neural networks (2D-CNN)¹⁸. Subsequently, Xu et al.¹⁹ combined 1D-CNN and 2D-CNN, constructing a dual-branch network structure to extract spectral and spatial features. However, this approach extracts spectral and spatial features separately and cannot effectively utilize the 3D spectral-spatial features of HSI. In order to better extract spectral-spatial features, researchers developed the 3D-CNN¹⁸ architecture and applied it to HSI classification. To overcome the limitation of CNN in capturing global information, scholars have proposed two approaches to improve CNN. One approach is to improve the perceptual range directly of the convolutional kernel, including the use of dilated convolutions²⁰ and the construction of a multiscale feature pyramid²¹; the other method is to embed an attention module²² that can capture global contextual information into the CNN structure^{23,24}, including spectral attention, spatial attention, and spatiotemporal attention. However, these methods still rely on convolutional operations in the backbone network to encode dense features, thus tending to local semantic information interaction²⁵. Capturing long-range dependencies becomes a pivotal breakthrough in overcoming the CNN performance bottleneck.

Recently, a visual transformer (ViT)²⁶ has been applied to various image processing tasks and has been preliminarily applied to the HSI classification field²⁷. ViT originates from the field of natural language processing (NLP) and is a new type of deep neural network composed of a multi-head attention mechanism and feedforward neural network, which can capture long-range dependency relationships in sequences through the multi-head attention mechanism^{28,29}. Compared to CNN, the self-attention mechanism of the Transformer imitates the saliency detection and selective attention of biological vision, and can establish long-distance dependency relationships, solving the limited receptive field problem of convolutional neural networks³⁰. However, ViT is not good at capturing local features. Given this, some scholars have begun to combine CNN and ViT to jointly capture local information, sequence features, and long-range dependency relationships. Existing HSI classification methods based on CNN-Transformer hybrid architectures²⁵ usually adopt manually specified hybrid strategies, such as using convolution to extract local features in the shallow layers and using a Transformer to extract global features in the deep layers^{29,31}, or directly adding the features extracted by CNN and Transformer³². Currently,

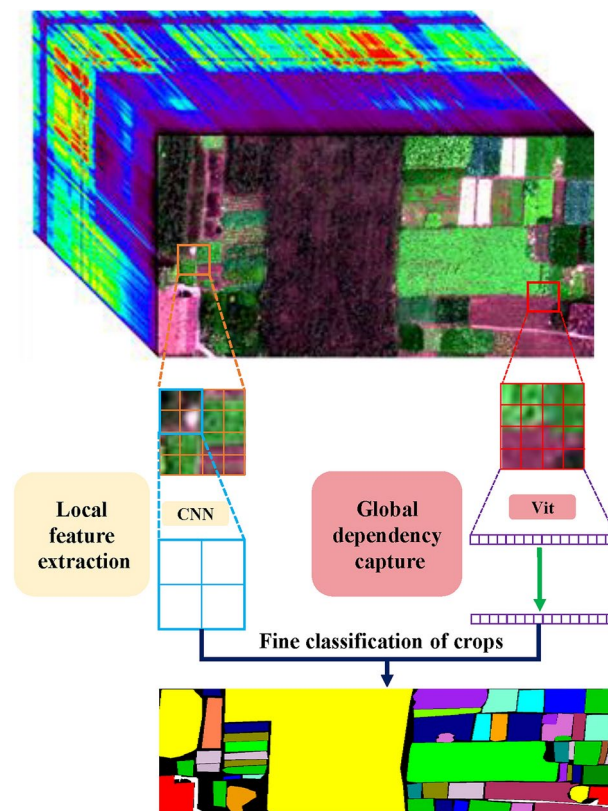


Fig. 1. HSI classification using deep learning approach.

many hyperspectral image classification methods based on hybrid CNN-Transformer architectures employ shallow convolutional layers to extract local features, while introducing Transformers at deeper layers to capture global features. Although this design successfully integrates local and global information, it presents several limitations. First, the direct concatenation or sequential stacking of convolutional layers and Transformers often lacks flexibility in feature fusion, leading to insufficient interaction between local and global information. Second, such methods typically struggle with fine-grained classification in complex scenarios, particularly in agricultural applications where the spectral features of different crops can be highly similar. This similarity makes it challenging to differentiate between crops using global features alone. These limitations highlight the need for a more flexible and integrated approach to feature extraction and fusion. In recent years, hyperspectral imaging has gained attention for crop classification in precision agriculture. However, traditional methods often face limitations due to the use of single-source data. To address this, fusion-based strategies, combining multi-spectral, hyperspectral, and LiDAR data, have been explored to enhance classification accuracy³³. For example, Anandakrishnan et al.³⁴ emphasized the effectiveness of UAV-based hyperspectral imaging in crop monitoring and classification.

In response to these limitations, this study proposes a novel hyperspectral crop classification approach, the Convolutional Meets Transformer Network (CMTNet). CMTNet employs a unique two-branch architecture: a CNN branch extracts local spectral-spatial features, while a Transformer branch captures global spectral-spatial features. This parallel dual-branch design not only mitigates the separation between local and global features seen in traditional methods but also excels in fine-grained classification tasks, particularly in complex agricultural environments. Furthermore, CMTNet enhances the efficiency of feature fusion through a multi-output constraint module, with experimental results demonstrating significant improvements in classification accuracy and generalization capabilities.

The main contributions of this article are given as follows.

- The CMTNet network proposed in this study features a unique dual-branch design that enables parallel extraction and dynamic fusion of local and global features. In contrast to existing hybrid CNN-Transformer methods, which typically stack local and global features sequentially, this design effectively addresses the limitations of traditional feature fusion methods. This innovation not only enhances the model's performance in fine-grained classification tasks but also improves its adaptability in complex agricultural environments.
- A novel multi-constraint module is introduced to enhance classification accuracy by applying cross-constraints on local, global, and combined features. Unlike traditional decision-level fusion, our approach imposes constraints at multiple stages of feature extraction, improving the utilization of spectral-spatial features and enabling better differentiation of fine-grained classes in complex agricultural scenarios.
- The proposed CMTNet employs a dual-branch structure with CNN and Transformer components to extract both local and global spectral-spatial features. Our approach introduces enhancements, such as a multi-output constraint module and optimized feature extraction, leading to significant improvements in classification accuracy. Experimental results on three datasets demonstrate that our method outperforms several state-of-the-art networks, particularly in complex, low-resolution hyperspectral scenarios in agricultural applications.

The rest of the article is organized as follows: Section 2 reviews related work in hyperspectral image classification and UAV-based precision agriculture. Section 3 describes the architecture and key components of the proposed CMTNet. Section 4 presents the experimental setup and datasets used, followed by the results and discussions. Finally, Sect. 5 concludes the article and outlines potential future directions.

Related work

CNN-based methods

CNN is a powerful tool for analyzing HSI images because they can accurately represent the spectral and spatial contextual information contained in the HSI data cube, extracting highly abstract features from the raw data and achieving excellent classification results³⁵. HSI classification tasks are categorized into three based on the distinct features CNN processes. The initial category involves 1D-CNN, focusing on spectral features. The data input for 1D-CNN is typically a single pixel. Li et al.³⁶ proposed a n feature extraction module and feature interaction in the frequency domain to enhance salient features. Chen et al.³⁷ used a multi-layer convolutional network to extract deep features of HSI, improving the classification results with a few training samples. Yue et al.³⁸ utilized principal component analysis for HSI preprocessing before feature extraction. The second category involves 2D-CNN, focusing on spatial features. Li et al.³⁹ used two 2D-CNN networks to extract high spectral and spatial frequency information simultaneously. Zhao et al.⁴⁰ proposed a 2D-CNN model that initially reduces dimensionality using PCA or another method, followed by data input into the model, where the data undergo initial processing by 2D-CNN to extract spatial information, subsequently combined with spectral information. Haut et al.⁴¹ developed a novel classification model guided by active learning, employing a Bayesian approach. The last category is based on spectral-spatial feature methods. In this case, there are two ways of feature processing. One approach involves the use of 3D-CNN. For instance, Li et al.⁴² introduced a 3D-CNN framework for the efficient extraction of deep spectral-spatial combined features from HSI cube data without preprocessing or post-processing. Another approach involves hybrid CNN, with significant research applying this method^{43–45}. Xu et al.¹⁹ integrated multi-source remote sensing data to enhance classification performance, employing 1D-CNN and 2D-CNN for the extraction of spectral and spatial features, respectively. Diakite et al.⁴⁶ suggested a hybrid network combining 3D-CNN and 2D-CNN. However, the current CNN-based methods overlook important differences between spatial pixels and unequal contributions of spectral bands. Convolutional kernels with limited receptive fields are independent of content, resulting in less accurate recognition of ground objects with local contextual similarity and large-scale variations.

Subsequently, the attention mechanism has been widely integrated with CNN frameworks^{43,47–49} due to its capability to assign varying weights to input features, enabling the model to concentrate more on crucial task-related information. Haut et al.⁵⁰ introduced a dual data-path attention module as the basic building block, considering both bottom-up and top-down visual factors to enhance the network's feature extraction capability. Liu et al.⁴³, based on the widely used convolutional block attention module (CBAM) improved accuracy by changing the way the attention module is connected. Tang et al.⁵¹ presented two attention models from spatial and spectral dimensions to emphasize crucial spatial regions and specific spectral bands, offering significant information for the classification task. Additionally, Roy et al.⁵² suggested an attention-based adaptive spectral-spatial kernel to enhance the residual network, capturing discriminative spectral-spatial features through end-to-end training for HSI classification. These attention-based methods are essentially enhanced versions of CNN, yet they are restricted by the inherent constraints of local convolutional kernels. These approaches emphasize local features while neglecting global information, consequently inadequately addressing the remote dependency between spectral sequences and spatial pixels.

Transformer-based methods

The initial design of the Transformer was focused on sequence modeling and transduction tasks. Its remarkable success in natural language processing has prompted researchers to explore its application in the visual domain, where it has demonstrated exceptional performance in tasks such as image classification and joint visual-linguistic modeling. Recent advances in diffusion models^{53,54} have significantly enhanced various image synthesis tasks. For instance, Shen et al.⁵⁵ proposed a progressive conditional diffusion model for story generation, while their later work⁵⁶ introduced a customizable virtual dressing model using diffusion-based approaches, further demonstrating the versatility of diffusion models in dynamic and interactive applications. In their work, Hong et al.⁵⁷ were the first to apply the ViT to HSI classification and achieved impressive results on commonly used hyperspectral image datasets. Additionally, He et al.⁵⁸ utilized a well-trained bidirectional encoder transformer structure for hyperspectral image classification. Furthermore, Qing et al.⁵⁹ introduced the self-attention-based transformer network (SAT-Net) for HSI classification, employing multiple Transformer encoders to extract image features. The encoder modules are directly connected using a multi-level residual structure to address the issues of vanishing gradients and overfitting. Tan et al.⁶⁰ introduced the transformer-in-transformer module for end-to-end classification, building a deep network model that fully utilizes global and local information in the input spectral cube. Sun et al.²⁴ proposed the spatial and spectral attention mechanism fusion network (SSAMF) for HSI classification, which incorporates channel self-attention into the Swin Transformer to better encode the rich spectral-spatial information of HSI, contributing to improved classification by the network. Mei et al.⁶¹ proposed the Group-Aware Hierarchical Transformer (GAHT) for HSI classification, applying multi-head self-attention to local spatial-spectral context and hierarchically constructing the network to improve classifying accuracy. Zhong et al.⁶² developed a spectral-spatial transformer network (SSTN) to overcome the constraints of convolutional kernels. Additionally, stable and efficient network architecture optimization is achieved through fast architecture search. It is evident that these previous studies primarily utilize Transformer to learn strong interactions between comprehensive label information through multiple self-attention modules. However, they are troubled by slow processing speed during inference and high memory usage, and these methods have yet to exploit the rich spatial features of HSI fully.

Hybrid methods

Recently, multiple endeavors have sought to integrate CNN and Transformer to build HSI classification networks that leverage the strengths of both architectures. Zhang et al.⁶³ proposed a dual-branch structure, incorporating both CNN and Transformer branches to capture local-global hyperspectral features. In the multi-head self-attention mechanism, convolutional operations were introduced skillfully to unite CNN and Transformer, further enhancing the classification progress. Liang et al.⁶⁴ integrated multi-head self-attention mechanisms in the spatial and spectral domains, applying them to context through uniform sampling and embedding 1D-CNN and 2D-CNN operations. Yang et al.⁶⁵ integrated CNN and Transformer sequentially and in parallel to fully utilize the features of HSI. Qi et al.³¹ developed the global-local spatial convolution transformer (GACT) to exploit local spatial context features and global interaction between different pixels. Additionally, through the weighted multi-scale spectral-spatial feature interaction (WMSFI) module, trainable adaptive fusion of multi-scale global-local spectral-spatial information is achieved. Song et al.⁶⁶ presented a dual-branch HSI classification framework utilizing 3D-CNN and bottleneck spatial-spectral transformer (B2ST), where both branches use a combination of shallow CNN and deep Transformer. Yang et al.⁶⁷ embedded CNN operations into the Transformer structure to capture subtle spectral differences and convey local spatial context information, then encoded spatial-spectral representation along multiple dimensions through a novel convolution displacer. In our earlier work published on arXiv⁶⁸, we proposed a convolutional and transformer hybrid model for hyperspectral image classification, which serves as the foundation for the improvements presented in this manuscript. However, while these methods have been adapted from natural image processing, challenges remain in effectively integrating CNN's strength in local context exploration with the Transformer's capability in global spectral-spatial modeling, particularly in achieving adaptive fusion of spectral-spatial features across multiple attributes and scales in low spatial resolution HSI.

Proposed method

The proposed method CMTNet's framework is illustrated in Fig. 2. CMTNet comprises a spectral-spatial feature extraction module, a local-global feature extraction module, and a multi-scale output constraint module. The spectral-space feature extraction module initially extracts shallow features from hyperspectral images by solely utilizing the spectral-space information present in the images. Subsequently, a parallel local-global feature

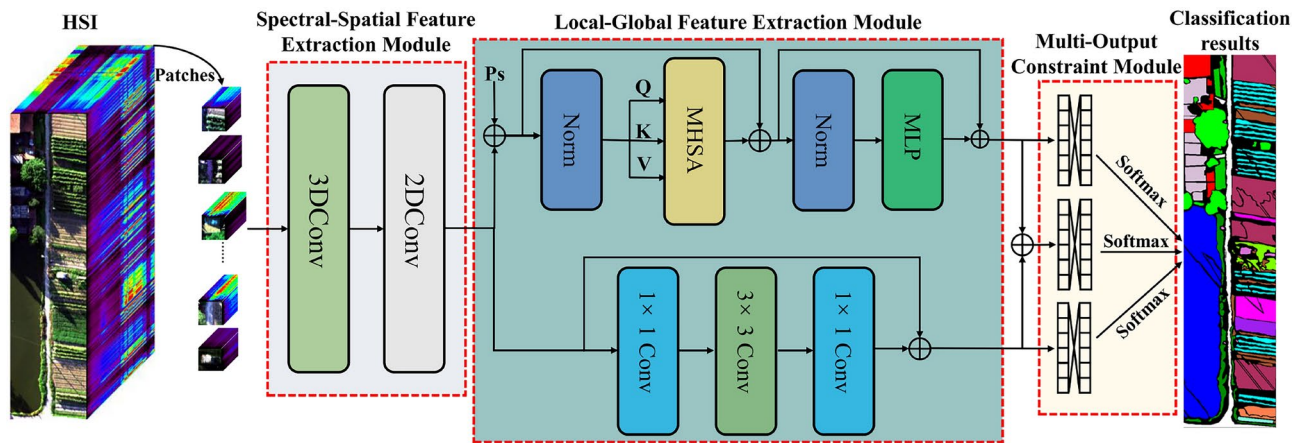


Fig. 2. CMTNet overall network framework. CMTNet consists of three main modules, spectral-spatial feature extraction module, local-global feature extraction module and multi-output constraint module.

extraction module, consisting of a Transformer branch and a CNN branch, is employed to deeply extract local and global features from the hyperspectral images. Finally, the classification results are generated using the multi-output constraint module, which calculates multi-output losses and cross-constraints on local, global, and joint features from various feature perspectives.

Spectral-spatial feature extraction module

The structure of the spectral-spatial feature extraction module outlined in this study is illustrated in Fig. 2. This module primarily utilizes convolutional neural networks to process the segmented hyperspectral image block. It begins by employing a 3D convolutional layer to extract spectral-spatial features, followed by a 2D convolutional layer to capture shallow spatial features. Let the hyperspectral dataset be denoted as $H \in \mathbb{R}^{h \times w \times d}$, with the spatial dimensions' height and width represented as h and w , respectively, and the number of spectral bands as d . Each pixel in H comprises d spectral dimensions, with its corresponding class label vector denoted as $V = (v_1, v_2, \dots, v_n)$, where n signifies the number of land cover categories in the hyperspectral scene. To manage the extensive hyperspectral image data, block division is necessary during model training to accommodate the computer's computational limitations. Following partitioning, each hyperspectral image block is denoted as $X \in \mathbb{R}^{m \times m \times d}$, with its dimensions specified. Each training image block sample is then inputted into the initial 3D convolutional layer. The convolution kernel within the 3D convolution calculates new convolutional feature maps by summing the dot product between the convolution of the entire spatial dimension and the kernel. The calculation formula is presented in Eq. (1):

$$v_{i,j}^{p,q,u} = \sum_{\eta=1}^{\eta} \sum_{h=0}^h \sum_{w=0}^w \sum_{c=0}^c \omega_{i,j,\eta}^{h,w,c} \times v_{i-1,\eta}^{(p+h),(q+w),(u+c)} + b_{i,j} \quad (1)$$

Where η represents the feature related to the j -th convolutional feature cube of the $i-1$ th layer; $v_{i,j}^{p,q,u}$ represents the convolution output value at position (p, q, u) of the j -th convolutional feature cube of the i -th layer, with the convolution kernel size of (h, w, c) ; $\omega_{i,j,\eta}^{h,w,c}$ and $b_{i,j}$ represent the weight parameters and bias at position (h, w, c) related to the η -th convolutional feature cube.

Similar to the 3D convolutional layer, the 2D convolutional layer operates by convolving a two-dimensional kernel to produce new feature maps. The calculation formula for this process is depicted in Eq. (2):

$$v_{i,j}^{p,q} = \sum_{\eta=1}^{\eta} \sum_{h=0}^h \sum_{w=0}^w \omega_{i,j,\eta}^{h,w} \times v_{i-1,\eta}^{(p+h),(q+w)} + b_{i,j} \quad (2)$$

$v_{i,j}^{p,q}$ represents the convolution output value at position (p, q) of the j -th convolutional feature cube of the i -th layer, with the convolution kernel size of (h, w) ; $\omega_{i,j,\eta}^{h,w}$ and $b_{i,j}$ represent the weight parameters and bias at position (h, w) related to the η -th convolutional feature cube.

This module consists of two convolutional layers, two batch normalization layers, and two activation layers using the ReLU activation function. The extraction process and calculation formulas of this module are detailed in Eqs. (3) and (4):

$$v_{i,j}^{p,q} = \Phi \left(g_1 \left(v_{i,j}^{p,q,u} \right) \right) \quad (3)$$

$$y = \Phi \left(g_2 \left(v_{i,j}^{p,q} \right) \right) \quad (4)$$

Where $\Phi(\cdot)$ represents the ReLU activation function, g_1 and g_2 , respectively, represent three-dimensional batch normalization and two-dimensional batch normalization.

Local-global feature extraction module

- (1) Transformer encoder branch: As shown in Fig. 2, the Transformer encoder branch mainly consists of positional encoding embeddings, multi-head self-attention (MHSA) (Fig. 3a), a multilayer perceptron (MLP), and two normalization layers. Residual connections are designed in front of MHSA and MLP. The output features of the spectral-spatial feature extraction module are flattened and linearly mapped to a sequence vector $T \in \mathbb{R}^{n \times z}$ of length s and channel dimension z . Then, a relative positional information vector $P_s \in \mathbb{R}^{n \times z}$ of length s is embedded into N sequence vectors as the input feature T_{in} of the Transformer encoder branch.

$$T_{in} = [T^1; T^2; \dots T^N] + P_s \quad (5)$$

The Transformer encoder’s exceptional performance can be attributed to its MHSA mechanism. MHSA efficiently captures the relationships between feature sequences by utilizing self-attention (SA) (see Fig. 3b). Initially, the Q , K , and V values derived from the convolution mapping are passed to MHSA via SA to extract global features. Within this process, Q and K are used to calculate attention scores, and the softmax function is applied to determine the weights of these attention scores. The formula for SA can be expressed as follows:

$$T_{SA} = \text{Attention}(Q, K, V) = \text{soft max} \left(\frac{QK^T}{\sqrt{d_K}} \right) V \quad (6)$$

Where T_{SA} represents the output of the SA module, and d_K is the dimension of K . MHSA uses multiple sets of weight matrices to generate Q , K , and V , and through a consistent computation process, multiple attention distributions are obtained. These distributions are then aggregated to obtain a comprehensive attention value. Finally, the features obtained by MHSA are passed to the MLP layer.

Traditional 3D-CNNs are constrained by fixed receptive fields, which limit their ability to model long-range spectral dependencies—a critical factor in distinguishing crops with subtle spectral differences. In contrast, the self-attention mechanism in CMTNet’s Transformer branch dynamically captures global spectral relationships across all bands and spatial positions.

- (2) CNN branch: As shown in Fig. 2, the CNN branch mainly consists of a 3×3 convolutional layer, two 1×1 convolutional layers, and residual connections, aiming to extract local features of hyperspectral images.

Multi-output constraint module

When calculating the loss, traditional feature constraints are typically restricted to the highest-level network output features, leading to suboptimal utilization of valuable spatial and spectral information. In hyperspectral

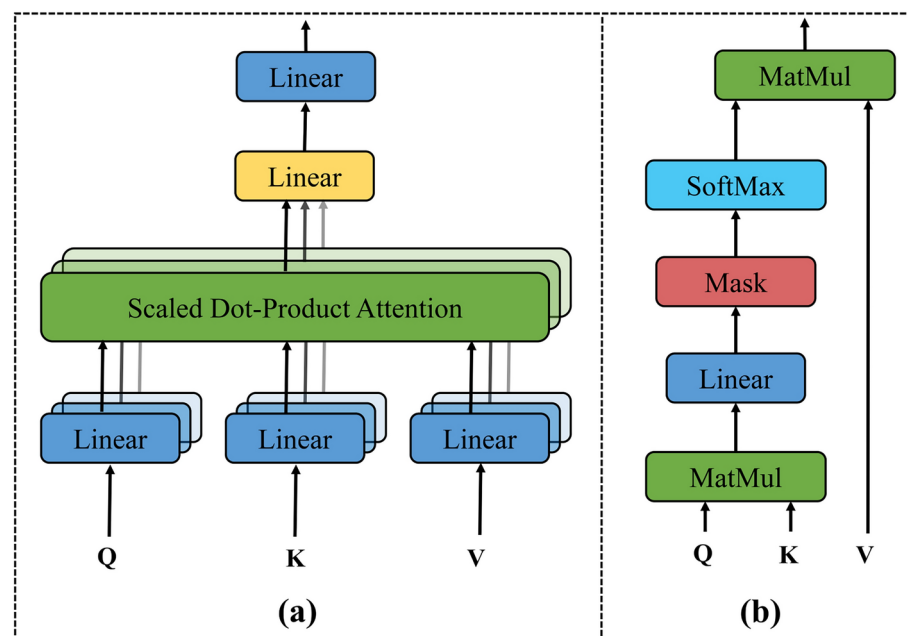


Fig. 3. Attention mechanism of the transformer. (a) Multi-head attention mechanism. (b) Self-attention mechanism.

image classification, high-level spatial and spectral semantic features play a pivotal role, and preserving this valuable information across different scales during multi-scale feature fusion can significantly impact classification performance. To address this limitation, we propose a multi-output constraint module in CMTNet that applies feature constraints at multiple stages (as illustrated in Fig. 4), rather than solely at the final output.

Specifically, the multi-output constraint module independently constrains the local features (CNN branch), global features (Transformer branch), and integration features (fused high-level semantic features). Each feature set F_L, F_G, F_I is sent to the softmax activation function for classification, resulting in the probability outputs P_L, P_G, P_I . The corresponding losses are computed using the categorical cross-entropy loss function:

$$\mathcal{L}_L = - \sum_{i=1}^N y_i \log (P_L(i)) \quad (7)$$

$$\mathcal{L}_G = - \sum_{i=1}^N y_i \log (P_G(i)) \quad (8)$$

$$\mathcal{L}_I = - \sum_{i=1}^N y_i \log (P_I(i)) \quad (9)$$

where $\mathcal{L}_L$ is the loss associated with the local features, $\mathcal{L}_G$ is the loss associated with the global features, $\mathcal{L}_I$ is the loss associated with the integration features, y_i represents the true class label, and N is the number of samples. The total loss is then computed as a weighted sum of the individual losses:

$$\mathcal{L}_T = \lambda_1 \mathcal{L}_L + \lambda_2 \mathcal{L}_G + \lambda_3 \mathcal{L}_I \quad (10)$$

where $\mathcal{L}_T$ is the total loss that combines the contributions from all feature branches, and $\lambda_1, \lambda_2, \lambda_3$ are weighting factors that balance the contribution of each feature type. To determine their optimal values, we conducted a random search over the parameter space $\lambda_1, \lambda_2, \lambda_3 \in [0.1, 1.0]$, constrained by $\lambda_1 + \lambda_2 + \lambda_3 = 1$. During backpropagation, these combined losses guide the optimization of network parameters across all branches, ensuring that local, global, and integration features are simultaneously refined. This enhances the interaction and fusion of local and global information, ultimately leading to a more robust feature representation.

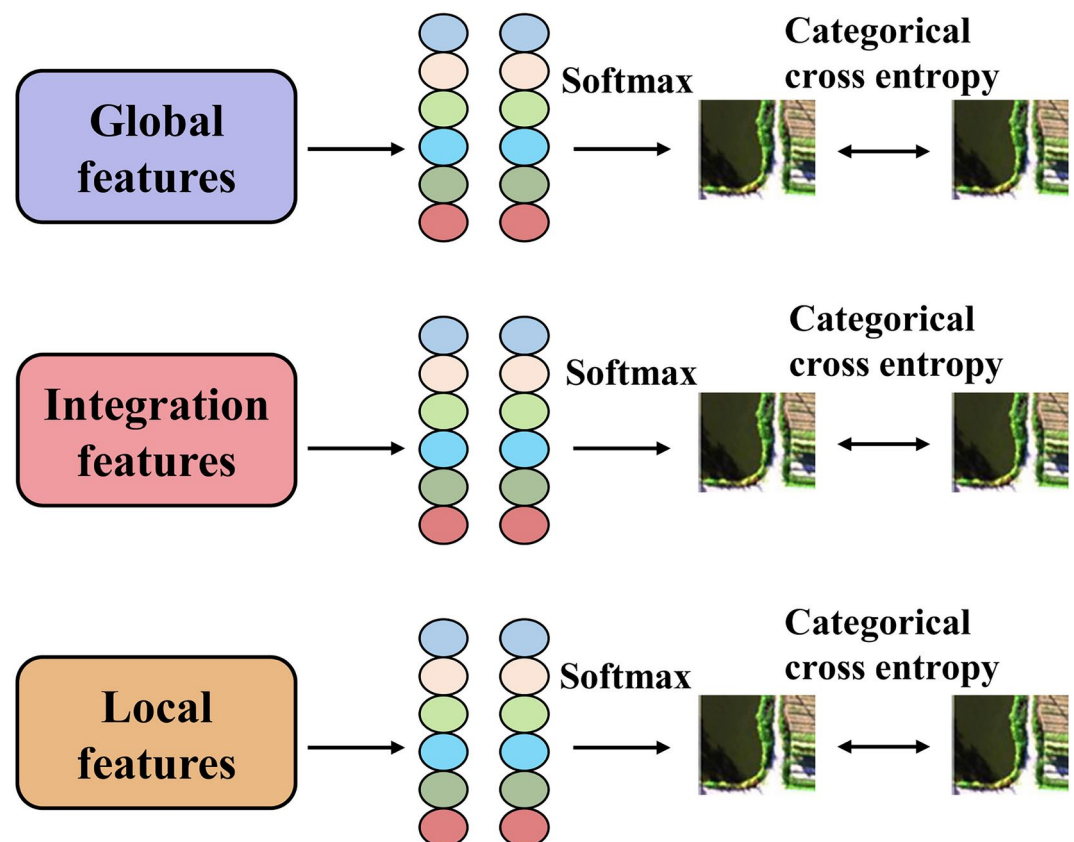


Fig. 4. Multi-output constraint module.

This approach contrasts with traditional methods that constrain only the final output, allowing CMTNet to dynamically adjust the importance of each feature type throughout the training process. The multi-output constraint module thus improves gradient flow and convergence during backpropagation, leading to higher classification accuracy and better performance in fine-grained classification tasks, especially in complex agricultural scenarios where differentiating between spectrally similar crops is challenging.

Experiment and analysis

To validate the proposed CMTNet method's superiority, it is compared with multiple state-of-the-art RF⁶⁹, SVM⁷⁰, 2D-CNN¹⁸, 3D-CNN¹⁸, Resnet⁵², ViT²⁶, SSFTT⁷¹ and CTMixer⁶³ approaches on three large-scale datasets, namely, WHU-Hi-LongKou, WHU-Hi-HanChuan and WHU-Hi-HongHu.

Datasets

This study used the publicly available HSI datasets. The WHU-Hi dataset^{72,73} produced by Wuhan University from a research study located on the Jiangnan plain of Hubei Province, China, with flat topography and abundant crop species (Fig. 5). The WHU-Hi LongKou dataset was acquired using the Headwall Nano-Hyperspec unmanned aerial UAV in LongKou town, Hubei Province, China, on July 17, 2018. The image size is 550×400 pixels, with 270 bands between 400 nm and 1000 nm, and a spatial resolution of approximately 0.463 m. The study area includes 9 land cover types. The image cube and ground-truth image are shown in Fig. 5a. The WHU-Hi HanChuan dataset was acquired using the Headwall Nano-Hyperspec unmanned aerial vehicle hyperspectral imager in Hanchuan City, Hubei Province, China, on June 17, 2016. The image size is 1217×303 pixels, with 274 bands between 400 and 1000 nm and a spatial resolution of approximately 0.109 m. The study area includes 16 land cover types. The image cube and ground-truth image are shown in Fig. 5b. The WHU-Hi HongHu dataset was acquired using the Headwall Nano-Hyperspec unmanned aerial vehicle hyperspectral imager in Honghu City, Hubei Province, China, on November 20, 2017. The image size is 940×475 pixels, with 270 bands between 400 and 1000 nm and a spatial resolution of approximately 0.043 m. The study area includes 22 land cover types. The image cube and ground-truth image are shown in Fig. 5c. Table 1 lists the overall crop category names, number of training samples, and number of test samples for these three datasets. Each dataset is divided into training and sample sets, with 0.5 % randomly selected from the total samples as the training set.

Evaluation metrics

This study uses overall classification accuracy (OA), average classification accuracy (AA), kappa coefficient, and accuracy under individual categories as evaluation metrics. It also visually presents classification diagrams as a visualization of the results.

Experimental setup

The experiment utilized the SITONHOLY IW4202 rack server, equipped with an Intel® Xeon® CPU E5-2620 v4 @ 2.10GHz and four NVIDIA TITAN Xp GPUs, each with 12 GB of memory and a total of 128 GB running memory. The software platform included Ubuntu 16.04.6 LTS 64-bit OS, CUDA Toolkit 10.1, CUDNN v7.6.5,

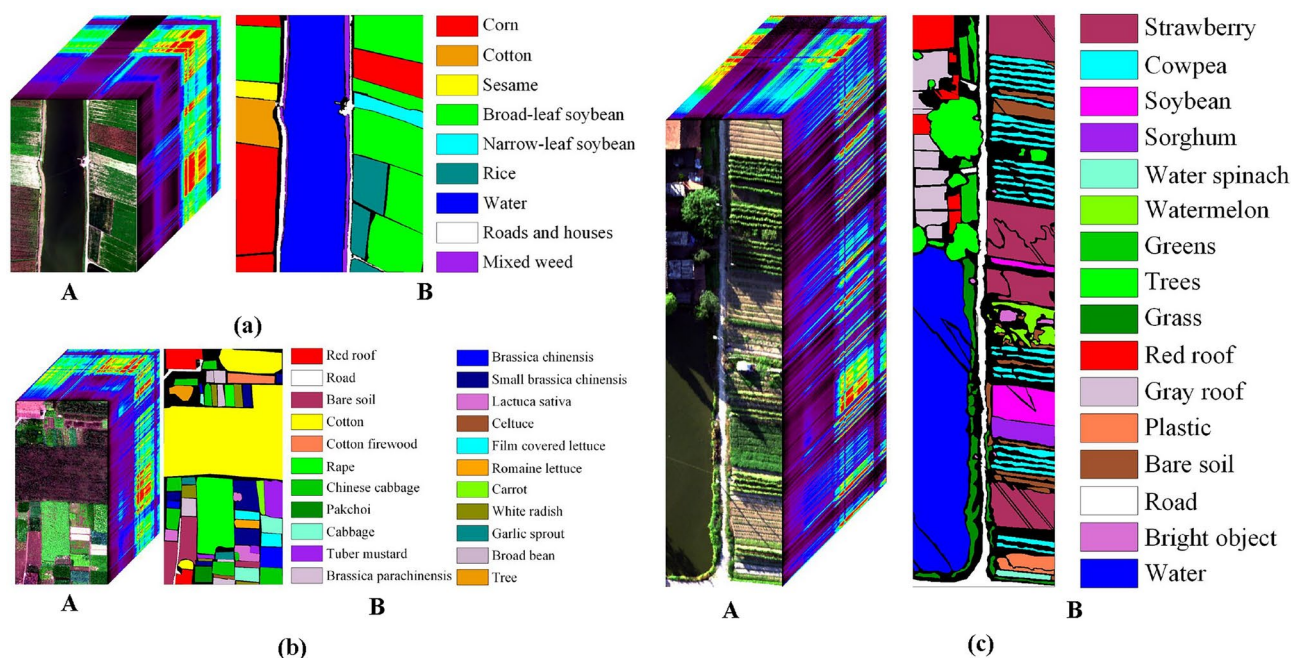


Fig. 5. Wuhan UAV-borne hyperspectral image. A and B denote image cube and ground-truth image respectively. (a) WHU-Hi-LongKou dataset. (b) WHU-Hi-HanChuan dataset. (c) WHU-Hi-HongHu dataset.

NO.	WHU-Hi LongKou			WHU-Hi HanChuan			WHU-Hi HongHu		
	Class	Training	Test	Class	Training	Test	Class	Training	Test
1	Corn	172	34339	Strawberry	223	44521	Red roof	70	13971
2	Cotton	41	8333	Cowpea	113	22640	Road	17	3495
3	Sesame	15	3016	Soybean	51	10236	Bare soil	109	21712
4	Broad-leaf soybean	316	62896	Sorghum	26	5327	Cotton	816	162469
5	Narrow-leaf soybean	20	4131	Water spinach	6	1194	Cotton firewood	31	6187
6	Rice	59	11795	Watermelon	22	4511	Rape	222	44335
7	Water	335	66721	Greens	29	5974	Chinese cabbage	120	23983
8	Roads and houses	35	7089	Trees	89	17889	Pakchoi	20	4034
9	Mixed weed	26	5203	Grass	47	9422	Cabbage	54	10765
10				Red roof	52	10464	Tuber mustard	61	12333
11				Gray roof	84	16827	Brassica parachinensis	55	10960
12				Plastic	18	3661	Brassica chinensis	44	8910
13				Bare soil	45	9071	Small Brassica chinensis	112	22395
14				Road	92	18468	Lactuca sativa	36	7320
14				Bright object	5	1131	Celtuce	5	997
16				Water	377	75024	Film covered lettuce	36	7226
17							Romaine lettuce	15	2995
18							Carrot	16	3201
19							White radish	43	8669
20							Garlic sprout	17	3469
21							Broad bean	6	1322
22							Tree	20	4020
/	Total	1019	203523	Total	1289	256241	Total	1925	384678

Table 1. Training and test sample numbers in the WHU-HiLongKou dataset, theWHU-HiHanChuan dataset, and the WHU-HiHongHu dataset.

Python 3.8, and Pytorch 1.7.0. Each experiment was repeated ten times independently, with the average value taken as the final result to mitigate the impact of random factors. The number of iterations was set to 100, using cross-entropy loss as the loss function and the Adam optimizer for model optimization.

The network, which combines CNN and Transformer, emphasizes global spatial information. To investigate the impact of various input image patch sizes (s) on classification performance, experiments were conducted using image patch sizes ranging from 5 to 15 on three datasets, with adjacent spatial sizes differing by 2. The experimental results are illustrated in Fig. 6. The figure demonstrates that the classification accuracy of the WHU-Hi LongKou dataset increases as the input image s increases. OA initially increases and then stabilizes. The WHU-Hi HanChuan and WHU-Hi HongHu datasets exhibit greater sensitivity to different input image patch sizes, with OA initially increasing and then decreasing with s . When $s = 13$, the OA of all three datasets approaches the maximum value. Consequently, $s = 13$ is chosen as the input image block size for the network proposed in this study.

In order to determine the optimal configuration of the proposed network architecture in terms of learning rate and batch size, a series of targeted experiments were conducted. The experimental results are illustrated in Fig. 7, with Fig. 7a–c representing the WHU-Hi LongKou, WHU-Hi HanChuan, and WHU-Hi HongHu datasets. Different colors in the figures indicate various ranges of OA. It is evident that different learning rates and batch sizes result in different OA values for the same dataset. For the WHU-Hi LongKou dataset (Fig. 7a), the impact of learning rate and batch size on OA is minimal, but there is some interaction between the two. The optimal learning rate and batch size were found to be $1e-3$ and 100, respectively. On the other hand, the WHU-Hi HanChuan and WHU-Hi HongHu datasets show significant sensitivity to learning rate due to variations in crop types used for training. Increasing the learning rate initially boosts OA and then decreases, while increasing the batch size also shows a similar trend. For the WHU-Hi HanChuan dataset Fig. 7b), a batch size of 100 resulted in improved classification performance with the selected learning rate. Similarly, for the WHU-Hi HongHu dataset (Fig. 7c), the best classification performance was achieved with a learning rate of $1e-3$ and a batch size of 100. Consequently, based on the parameter experiments, the optimal learning rate and batch size for the proposed classification network were determined to be $1e-3$ and 100, respectively.

This study investigates how the number of encoder layers and attention heads can impact the model's robustness and stability. Experimental results, as shown in Fig. 8, demonstrate the effects on the WHU-Hi LongKou, WHU-Hi HanChuan, and WHU-Hi HongHu datasets. The histograms in Fig. 8a–c reveal that the differences in OA histograms across different layers and heads are minimal, with OAs remaining stable within specific ranges: LongKou: 99.52–99.68, HanChuan: 97.42–97.59, HongHu: 98.51–98.62. To ensure uniformity in the network structure across all datasets, this study opts for the CMTNet with one transformer layer and four attention heads as the final network configuration.

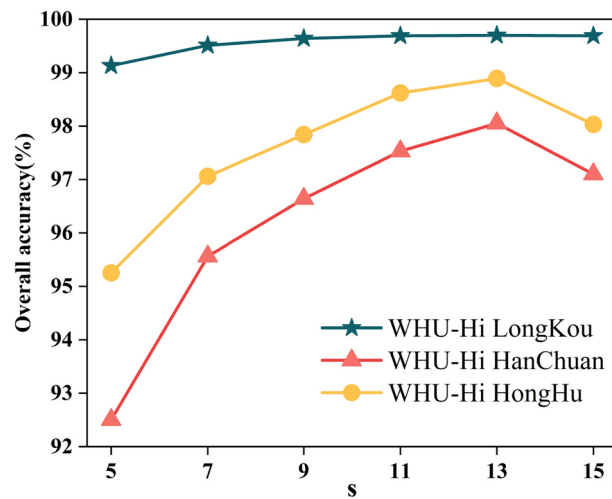


Fig. 6. Impact of different input space sizes on OA.

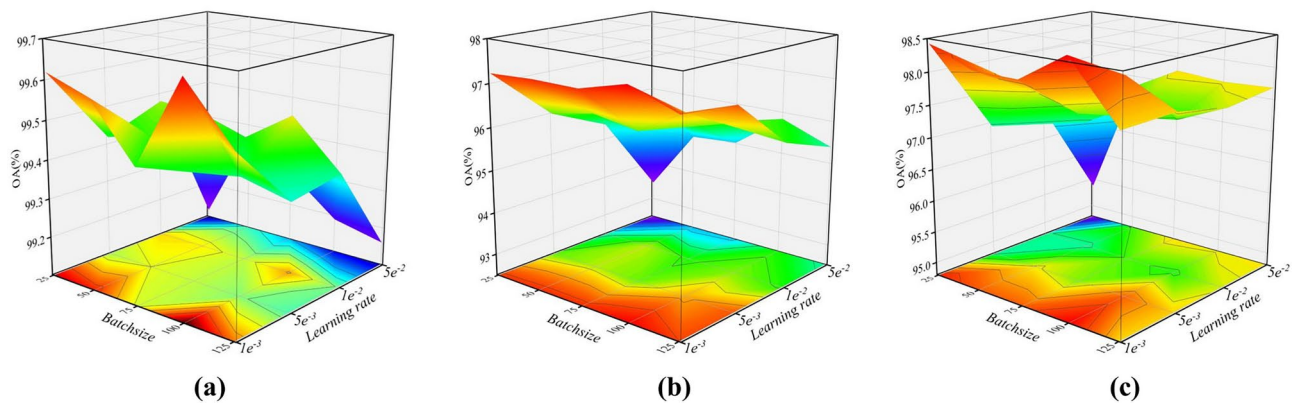


Fig. 7. Effect of different learning rates and batch sizes on performance accuracy OA. (a) Experimental results on WHU-Hi LongKou dataset. (b) Experimental results on WHU-Hi HanChuan dataset. (c) Experimental results on WHU-Hi HongHu dataset.

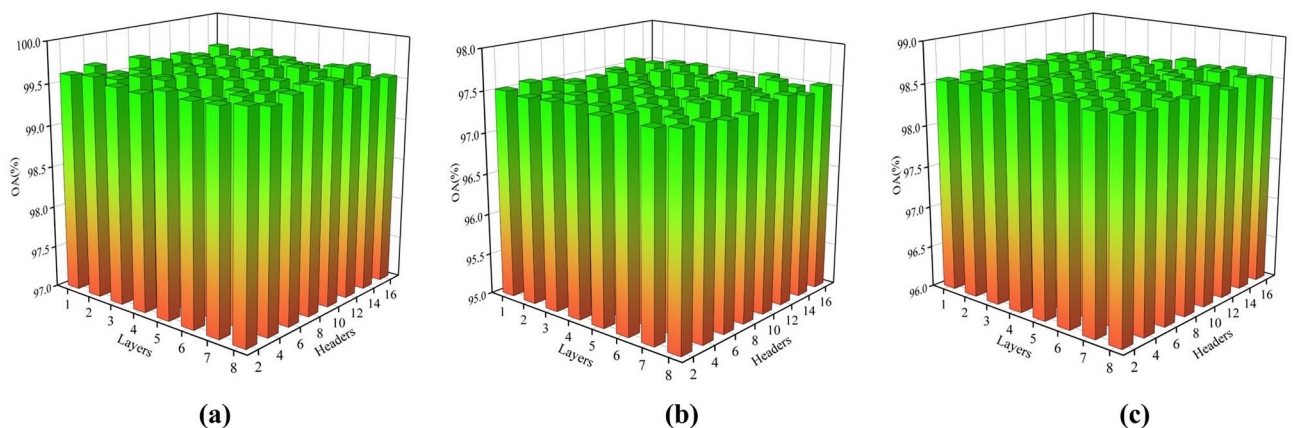


Fig. 8. OA under different numbers of transformer encoder layers and MHSA heads on (a) WHU-Hi LongKou, (b) WHU-Hi HanChuan, and (c) WHU-Hi HongHu, respectively.

NO.	RF	SVM	2D-CNN	3D-CNN	Resnet	ViT	SSFTT	CTMixer	CMTNet
1	89.04	90.99	94.83	98.56	88.09	92.3	99.94	99.62	99.75
2	45.10	46.25	58.63	65.53	89.97	72.22	98.32	99.69	99.53
3	90.23	89.87	95.47	97.73	88.31	97.63	99.93	100.00	99.83
4	87.56	87.83	84	93.91	90.07	90.15	99.36	99.31	99.67
5	32.59	42.57	55.11	75.11	86.88	92.35	98.94	98.38	98.61
6	83.85	85.18	90.79	96.62	85.26	85.53	99.89	99.95	99.94
7	86.86	86.83	97.66	97.61	86.83	86.65	99.90	99.79	99.84
8	64.61	65.55	66.35	81.44	82.83	81.71	95.6	95.51	99.32
9	44.06	65.87	71.83	90.21	84.56	87.58	95.1	97.83	98.55
OA(%)	84.65	85.21	88.75	93.93	90.43	94.05	99.37	99.39	99.58
AA(%)	75.33	75.26	79.41	88.53	90.68	93.84	97.66	97.60	98.62
k×100	86.24	87.56	85.36	92.38	88.67	94.99	99.18	99.20	99.45

Table 2. Classification performance obtained by different methods for WHU-Hi LongKou dataset (optimal results are bolded).

NO.	RF	SVM	2D-CNN	3D-CNN	Resnet	ViT	SSFTT	CTMixer	CMTNet
1	77.94	70.74	94.72	87.16	86.61	90.87	95.97	95.65	96.09
2	72.55	49.21	87.18	93.39	78.74	75.90	97.90	95.60	97.66
3	38.76	71.1	91.33	87.66	87.35	89.31	95.54	92.16	94.30
4	86.13	94.12	94.25	92.53	90.35	92.67	94.72	98.24	98.28
5	14.23	81.35	94.39	64.22	89.68	85.3	84.03	97.41	98.35
6	22.75	47.53	64.42	75.41	85.02	83.37	82.82	89.89	96.11
7	49.02	88.82	85.59	74.36	86.40	85.21	85.84	89.16	96.29
8	35.31	59.46	85.32	90.43	77.78	77.02	85.92	91.18	94.63
9	87.62	61.63	84.18	87.84	87.43	80.21	81.32	89.64	92.70
10	87.94	87.94	89.44	95.32	87.03	88.86	97.62	98.39	98.15
11	47.85	92.05	91.35	89.71	90.06	86.84	92.04	95.59	96.86
12	26.17	61.61	55.8	75.31	85.87	84.86	77.92	93.41	97.30
13	68.66	56.39	71.28	82.21	84.09	80.36	87.63	88.39	94.25
14	95.06	63.48	86.77	89.09	83.73	83.79	89.98	92.84	97.96
15	38.07	70.93	36.68	89.65	90.54	88.46	90.6	97.10	96.01
16	93.25	94	94.72	97.79	89.57	91.07	99.66	99.36	99.79
OA(%)	71.03	76.05	89.47	91.03	87.36	92.86	94.17	95.54	97.29
AA(%)	69.83	71.9	81.71	85.35	81.33	84.67	87.22	89.34	94.01
k×100	70.71	72.58	88.52	89.43	85.04	91.63	93.18	94.77	96.83

Table 3. Classification performance obtained by different methods for WHU-Hi HanChuan dataset (optimal results are bolded).

Comparison of experimental results with SOTA

The OA, AA, and Kappa values of CMTNet and other comparative methods on the WHU-Hi LongKou, WHU-Hi HanChuan, and WHU-Hi HongHu datasets are presented in Tables 2, 3 and 4, accompanied by visual representations in Figs. 9, 10 and 11. The best values are highlighted in bold in the tables, clearly indicating the superior performance of the proposed CMTNet method. Analysis of Table 2 reveals that CMTNet excels in capturing both global and local spectral features of hyperspectral imaging separately, effectively integrating high-dimensional information to achieve outstanding classification results across different land cover targets. When compared to the CNN and Transformer hybrid networks SSFTT and CTMixer, CMTNet outperforms in final classification results, showing an increase in OA of 0.21 and 0.19, respectively. This improvement can be attributed to the multi-output constraint module of CMTNet, which optimally reallocates feature weights. However, the performance of RF and SVM could be enhanced, particularly in the classification of cotton and soybeans with limited training samples, where individual accuracies fall below 47%. Examination of Fig. 9 demonstrates that CMTNet significantly enhances classification performance, reducing misclassifications and ensuring complete classification edges through the fusion of local-global spectral features.

The WHU-Hi HanChuan dataset captured images in the afternoon with a lower sun angle, resulting in numerous shadow patches. The classification results for the RF and SVM methods show many misclassifications. Both 2D-CNN and 3-DCNN models display significant fragmentation, highlighting the necessity for methods to enhance model generalization. The SSFTT synthesizes the use of 3D convolutional layers and attention

NO.	RF	SVM	2D-CNN	3D-CNN	Resnet	ViT	SSFTT	CTMixer	CMTNet
1	92.35	83.93	85.67	94.6	98.25	97.22	96.23	96.90	98.32
2	49.67	97.34	76.96	85.17	80.08	98.62	81.87	90.30	96.14
3	97.97	72.85	98.08	98.62	99.96	99.31	92.19	92.81	97.31
4	96.20	78.96	92.83	97.48	99.75	98.43	98.89	98.96	99.79
5	22.34	77.25	67.59	80.09	51.02	72.64	84.47	91.17	97.60
6	22.91	81.95	82.35	76.71	80.21	87.18	98.51	97.68	98.98
7	46.88	59.25	62.94	92.24	82.01	92.70	88.21	91.71	94.41
8	14.44	41.63	49.68	40.38	39.35	63.25	92.35	88.71	96.23
9	82.68	90.86	86.23	100	99.79	100.00	97.33	97.93	97.92
10	30.46	54.08	84.83	59.51	76.25	86.25	94.62	94.13	98.26
11	14.36	48.31	73.53	85.25	83.84	82.45	89.34	92.65	96.54
12	14.46	61.31	64.91	67.92	38.14	48.18	90.25	87.4	98.36
13	21.21	49.86	43.02	30.89	29.95	38.6	90.8	87.79	96.25
14	57.30	63.78	62.46	59.17	94.37	94.02	98.37	98.19	98.08
15	9.90	85.92	58.00	76.92	100.00	95.24	90.14	99.01	97.70
16	78.29	78.01	99.17	96.25	98.94	98.55	97.76	96.65	99.10
17	57.71	70.65	100	91.51	84.27	99.49	80.68	91.45	99.73
18	18.07	79.24	82.63	67.70	56.69	63.49	95.21	94.07	97.29
19	47.87	68.22	78.91	53.06	71.58	63.06	95.20	93.90	97.99
20	26.86	77.85	17.87	72.76	69.72	81.72	84.9	87.67	96.31
21	16.52	74.67	98.78	48.80	55.00	92.77	84.13	66.16	90.86
22	10.14	81.14	74.64	54.84	47.99	46.55	92.76	90.17	95.91
OA(%)	54.06	73.55	87.81	89.48	85.14	91.53	95.56	95.79	98.31
AA(%)	49.32	71.23	82.14	85.99	83.54	87.26	89.02	89.58	95.26
k×100	52.38	68.05	86.63	88.20	84.43	91.51	94.37	94.68	97.87

Table 4. Classification performance obtained by different methods for WHU-Hi HongHu dataset (optimal results are bolded).

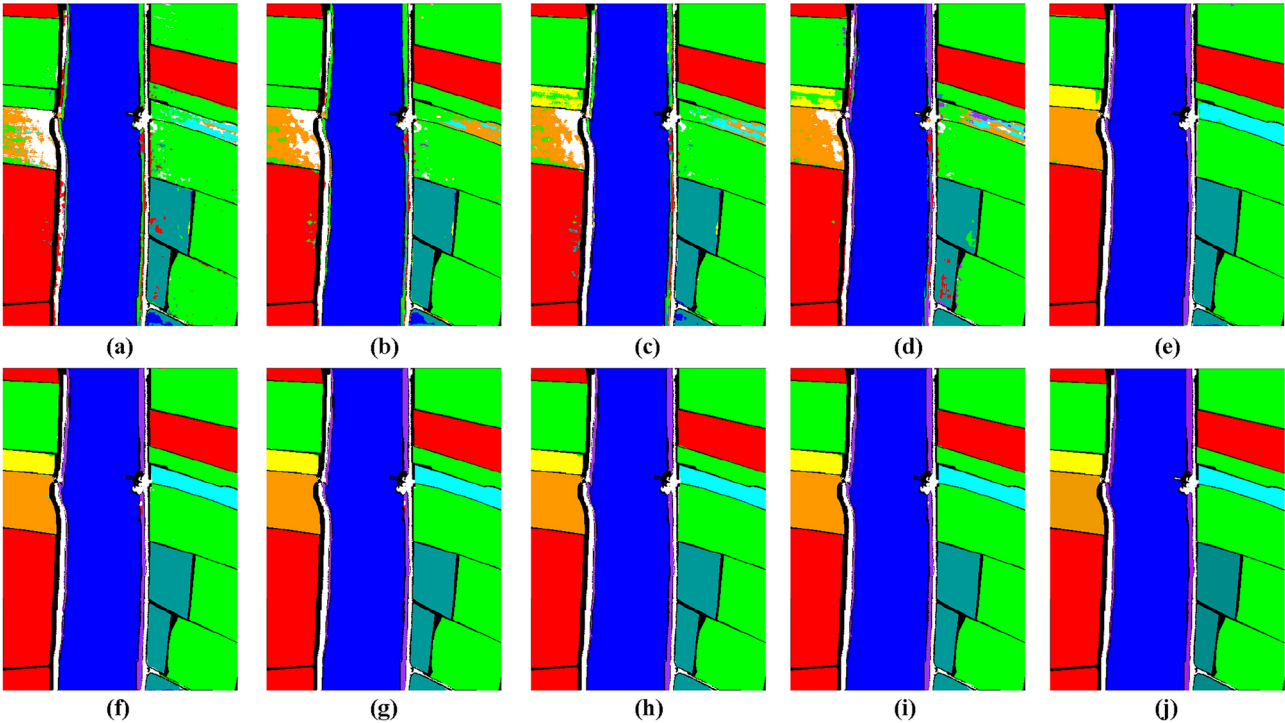


Fig. 9. Classification visualization maps of all methods on the WHU-Hi LongKou dataset. (a)–(i) Classification map of RF, SVM, 2-DCNN, 3-D CNN, Resnet, ViT, SSFTT, SSTN, CTMixer, and CMTNet, respectively. (j) Real ground feature map. CMTNet model classification results (OA = 99.58%, Kappa = 99.45).

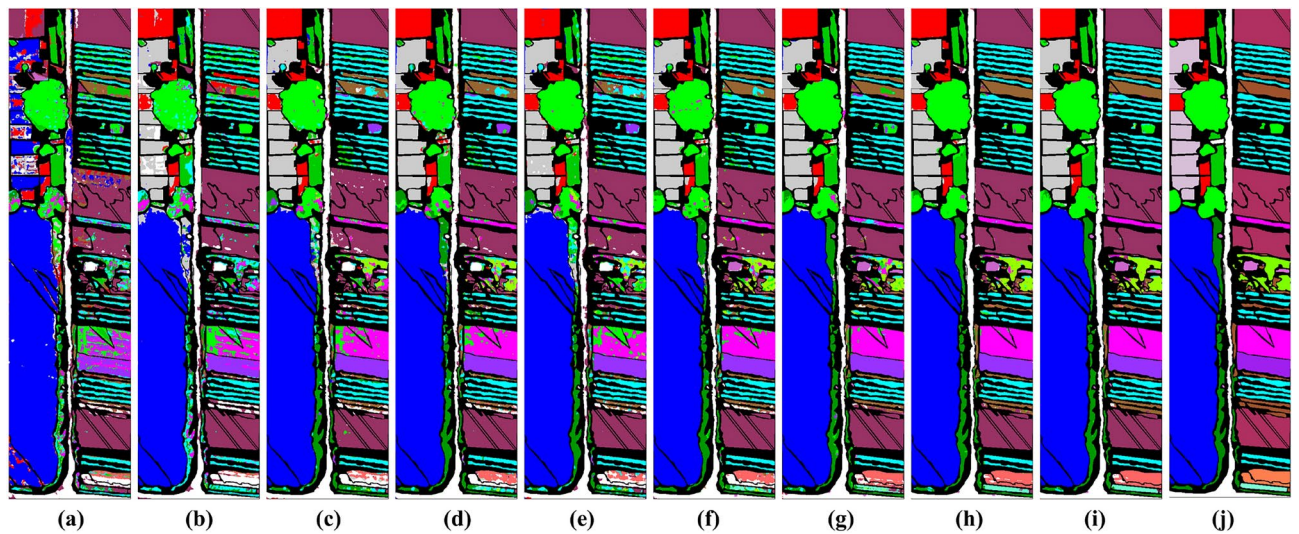


Fig. 10. Classification visualization maps of all methods on the WHU-Hi HanChuan dataset. (a)–(i) Classification map of RF, SVM, 2-DCNN, 3-D CNN, Resnet, ViT, SSFTT, SSTN, CTMixer, and CMTNet, respectively. (j) Real ground feature map. CMTNet model classification results (OA = 97.29%, Kappa = 96.83).

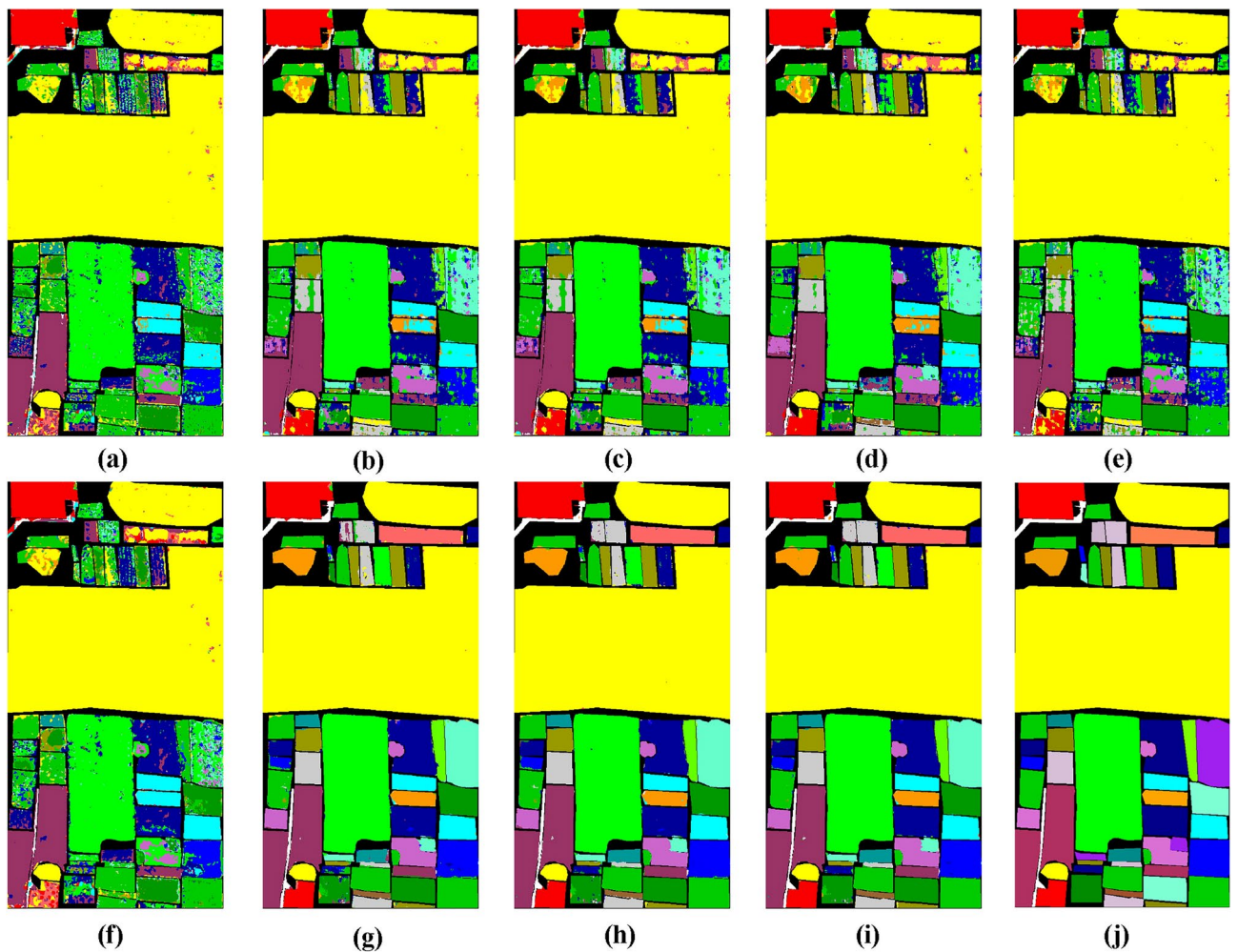


Fig. 11. Classification visualization maps of all methods on the WHU-Hi HanChuan dataset. (a)–(i) Classification map of RF, SVM, 2-DCNN, 3-D CNN, Resnet, ViT, SSFTT, SSTN, CTMixer, and CMTNet, respectively. (j) Real ground feature map. CMTNet model classification results (OA = 98.31%, Kappa = 97.87).

mechanism modules to realize the abstraction extraction of joint spectral-spatial features, effectively mitigating classification errors caused by ‘same material different spectrum, different materials same spectrum’. However, owing to its serial extraction of spectral-spatial features without effective selection, there is still a problem of performance plummeting in the classification of small-sample targets, with the OA for categories such as Watermelon and Plastic (NO.6 and 12, respectively) being only 82.42% and 77.92%. ResNet exhibits clear misclassifications of soybeans and gray rooftops. On the other hand, ViT and CTMixer methods achieve high-precision classifications overall, but errors persist in shadow-covered areas. Despite this, CMTNet outperforms in identifying similar spectral features through multi-feature fusion extraction, leading to reduced fragmentation compared to other methods.

In the WHU-Hi HongHu dataset, traditional classification algorithms struggle with misclassifications due to slight spectral differences among crops of the same type. Specifically, *Brassica parachinensis*, *Brassica chinensis*, and *Small Brassica chinensis* exhibit low classification accuracy. Deep learning methods have notably enhanced hyperspectral classification over traditional approaches. However, 2D-CNN and 3D-CNN tend to only capture local features in hyperspectral images, resulting in fragmented classification outcomes. The ViT model, on the other hand, leverages global perceptual spectral features to mitigate this fragmentation. While models like SSFTT and CTMixer combine CNN and Transformer architectures to effectively utilize spectral-spatial information for improved classification, they still struggle with misclassifications in land cover categories with limited samples. CMTNet demonstrates the best effectiveness in categorizing various terrestrial objects due to its capability to capture spatial and spectral characteristics separately, and efficiently filter and integrate high-dimensional information. It delivers exceptionally good results for different categories, with OA for Red roof, Cotton, Rape, Tuber mustard, and *Lactuca sativa* reaching 98.32%, 99.7%, 98.98%, 98.26%, and 98.08% respectively, and the overall OA and Kappa coefficient being 98.31% and 97.87%. CTMixer focuses on the effective use of global and local multi-scale features, achieving better outcomes in mixed terrestrial feature regions, yet its OA and Kappa coefficients are reduced by 2.52% and 3.19% compared to CMTNet. Visual and quantitative analyses reveal that CMTNet achieves the highest accuracy and excels at classifying land cover categories with limited samples. This suggests that incorporating the multi-output constraint module can enhance the model’s robustness and stability.

Ablation experiments

To thoroughly verify the effectiveness of the proposed method, ablation experiments were conducted on three datasets using different components of the network model. The baseline network was Transformer, with modules from CMTNet sequentially added to assess their contributions. Five combined models were analyzed, and the impact of each component on the OA was measured. The results of all ablation experiments are presented in Table 5. The checkmark symbol “✓” indicates module usage, while the cross symbol “×” indicates non-usage. Analysis revealed that using only the Transformer module resulted in relatively low OAs across the datasets, suggesting its limitations in extracting local features for hyperspectral image classification. Addition of the spectral-spatial feature extraction module in Case 2 and Case 3 led to an increase in OA. Case 4 introduced a CNN branch in parallel with the Transformer branch to enhance local feature extraction, resulting in a significant OA improvement. Case 5 further improved the integration of features from each branch by incorporating the multi-output constraint module. Experimental findings demonstrated that Case 5 consistently outperformed Case 4 on all three datasets, highlighting the effectiveness of the Multi-Output Constraint Module (MOCM).

Model efficiency analysis

To evaluate the computational efficiency of the proposed method, we conducted efficiency tests on all approaches, with Table 6 presenting the experimental results. As shown in Table 6, traditional machine learning models (such as RF and SVM) exhibited the fastest running speeds, with RF requiring 36.82, 62.66, and 74.54 s on the three datasets, respectively. In contrast, deep learning models like 3D-CNN, Resnet, and ViT demand significantly more computational resources. For instance, ViT required 857.20, 1458.26, and 1735.34 s on the Longkou, Hanchuan, and Honghu datasets, respectively, reflecting its inherent complexity. Compared with the SSFTT method, the training and testing times for CMTNet are slightly longer, which can be attributed to its dual-branch Transformer architecture that enhances feature representation while introducing additional computational overhead during optimization. In contrast to the CTMixer method, CMTNet requires less running time. Overall, Transformer-based methods demonstrate significantly higher efficiency than CNN-based methods. While CMTNet achieves state-of-the-art classification accuracy, its computational cost is higher than traditional machine learning methods. This trade-off is critical for precision agriculture applications where accuracy is prioritized. Future work will focus on model compression and edge deployment frameworks to further bridge the efficiency gap.

Case	CNN Branch	Conv3D	Conv2D	MOCM	LongKou	HanChuan	HongHu
1	×	×	×	×	94.97	92.86	91.53
2	✓	×	×	×	96.25	95.21	93.89
3	✓	✓	×	×	98.77	95.76	95.34
4	✓	✓	✓	×	99.12	96.05	97.29
5	✓	✓	✓	✓	99.58	97.29	98.31

Table 5. Impact of different modules on network OA value (%).

Datasets	RF	SVM	2D-CNN	3D-CNN	Resnet	ViT	SSFTT	CTMixer	CMTNet
WHU-Hi LongKou	36.82	163.31	591.22	658.82	726.10	857.20	701.21	1149.38	931.63
WHU-Hi HanChuan	62.66	277.65	1005.48	1120.93	1235.35	1458.26	1192.85	1955.42	1584.60
WHU-Hi HongHu	74.54	330.32	1196.46	1333.98	1469.82	1735.34	1419.31	2327.04	1885.32

Table 6. Running time of different methods on three datasets (s).

Discussion

While CMTNet achieves state-of-the-art performance on diverse datasets, its accuracy on shadow-affected regions (e.g., WHU-Hi HanChuan) reveals a dependency on consistent illumination conditions. Shadows introduce spectral ambiguities that challenge the current feature extraction modules. To address this, future iterations could integrate shadow-invariant feature learning techniques, such as normalization based on illumination-invariant indices⁷⁴, or leverage multi-temporal data to disentangle shadow effects from intrinsic spectral signatures.

To evaluate the risk of overfitting in classes with limited training samples, we specifically analyze the performance of Cotton (NO.2) and Narrow-leaf soybean (NO.5) in the WHU-Hi LongKou dataset. As shown in Table 2, these classes have only 41 and 20 training samples, respectively. Despite the small sample size, CMTNet achieves OA values of 99.53% for Cotton and 98.61% for Narrow-leaf soybean. However, compared to classes with abundant samples (e.g., Broad-leaf soybean (NO.4) with 316 training samples and 99.67% OA), the accuracy gaps (0.14% and 1.06%) indicate potential overfitting risks. Although MOCM partially alleviates overfitting, the performance of extremely small sample classes (e.g., Narrow-leaf soybean) has not yet reached its optimal level. To address this issue, in future work, we will initialize the feature extractor using a pre-trained model on large-scale hyperspectral datasets (such as WHU-Hi HongHu) and then fine-tune it on the target small-sample classes⁷⁵. Additionally, semi-supervised learning techniques, such as consistency regularization or pseudo-labeling⁷⁶, will be employed to incorporate unlabeled data, which can enhance the model’s generalization without requiring additional labeled samples. These methods can synergistically improve the robustness of CMTNet’s dual-branch architecture in imbalanced agricultural scenarios.

Conclusions

In order to enhance the precision and efficiency of crop classification in areas with imbalanced samples and diverse land cover types, this study introduces a novel method called CMTNet. This method incorporates a dual-branch structure featuring parallel CNN and Transformer components, enabling the extraction of local-global features from hyperspectral images. A convolutional layer combination spectral-spatial feature extraction module is employed to capture low-level spectral-spatial features, while a multi-output constraint module effectively addresses information loss post multi-scale feature fusion. Experimental results demonstrate the method’s effectiveness in enhancing classification performance. Although CMTNet demonstrates excellent performance in complex agricultural scenarios, the current study still exhibits several limitations. Specifically, the model’s robustness against extreme shadows and occlusions requires improvement, and its computational cost restricts deployment on resource-constrained devices. To address these issues, future research will focus on the following directions. First, by integrating multimodal data (e.g., LiDAR elevation information), we aim to enhance classification robustness in complex environments characterized by shadows and occlusions. Second, we plan to optimize real-time performance and explore model compression techniques (such as pruning and quantization) to reduce computational costs. Finally, by incorporating edge computing frameworks, we seek to achieve efficient deployment on drone platforms, thereby advancing real-time monitoring applications in precision agriculture.

Data availability

The datasets analyzed during the current study are publicly available in the WHU-Hi repository at http://rsidea.wvu.edu.cn/resource_WHUHi_sharing.htm.

Received: 21 January 2025; Accepted: 2 April 2025

Published online: 11 April 2025

References

1. Weiss, M., Jacob, F. & Duveiller, G. Remote sensing for agricultural applications: A meta-review. *Remote Sens. Environ.* **236**, 111402 (2020).
2. Khan, M. J., Khan, H. S., Yousaf, A., Khurshid, K. & Abbas, A. Modern trends in hyperspectral image analysis: A review. *IEEE Access* **6**, 14118–14129 (2018).
3. Hu, J., Huang, Z., Shen, F., He, D. & Xian, Q. A robust method for roof extraction and height estimation. In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium* (IEEE, 2023).
4. Qiao, C. et al. A novel multi-frequency coordinated module for sar ship detection. In *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)* 804–811 (IEEE, 2022).
5. Soni, A., Dixit, Y., Reis, M. M. & Brightwell, G. Hyperspectral imaging and machine learning in food microbiology: Developments and challenges in detection of bacterial, fungal, and viral contaminants. *Compr. Rev. Food Sci. Food Saf.* **21**, 3717–3745 (2022).
6. Shen, F. et al. An efficient multiresolution network for vehicle reidentification. *IEEE Internet Things J.* **9**, 9049–9059 (2021).
7. Hu, J., Huang, Z., Shen, F., He, D. & Xian, Q. A bag of tricks for fine-grained roof extraction. In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium* (IEEE, 2023).

8. Wu, H. et al. A sample-proxy dual triplet loss function for object re-identification. *IET Image Proc.* **16**, 3781–3789 (2022).
9. Xu, R., Shen, F., Wu, H., Zhu, J. & Zeng, H. Dual modal meta metric learning for attribute-image person re-identification. In *2021 IEEE International Conference on Networking, Sensing and Control (ICNSC)*, vol. 1, 1–6 (IEEE, 2021).
10. Dalla Mura, M., Benediktsson, J. A., Waske, B. & Bruzzone, L. Morphological attribute profiles for the analysis of very high resolution images. *IEEE Trans. Geosci. Remote Sens.* **48**, 3747–3762 (2010).
11. Ghamisi, P. et al. Extinction profiles for the classification of remote sensing data. *IEEE Trans. Geosci. Remote Sens.* **54**, 5631–5645 (2016).
12. Hennessy, A., Clarke, K. & Lewis, M. Hyperspectral classification of plants: A review of waveband selection generalisability. *Remote Sens.* **12**, 113 (2020).
13. Ranjan, P. & Girdhar, A. A comprehensive systematic review of deep learning methods for hyperspectral images classification. *Int. J. Remote Sens.* **43**, 6221–6306 (2022).
14. Chen, Y., Zhao, X. & Jia, X. Spectral-spatial classification of hyperspectral data based on deep belief network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **8**, 2381–2392 (2015).
15. Wu, H. & Prasad, S. Convolutional recurrent neural networks for hyperspectral data classification. *Remote Sens.* **9**, 298 (2017).
16. Hsieh, T.-H. & Kiang, J.-F. Comparison of CNN algorithms on hyperspectral image classification in agricultural lands. *Sensors* **20**, 1734 (2020).
17. Paoletti, M., Haut, J., Plaza, J. & Plaza, A. Deep learning classifiers for hyperspectral imaging: A review. *ISPRS J. Photogramm. Remote Sens.* **158**, 279–317 (2019).
18. Bera, S., Shrivastava, V. K. & Satapathy, S. C. Advances in hyperspectral image classification based on convolutional neural networks: A review. *CMES-Comput. Model. Eng. Sci.* **133**, 219–250 (2022).
19. Xu, X. et al. Multisource remote sensing data classification based on convolutional neural network. *IEEE Trans. Geosci. Remote Sens.* **56**, 937–949 (2017).
20. Shi, C., Liao, D., Zhang, T. & Wang, L. Hyperspectral image classification based on expansion convolution network. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–16 (2022).
21. Xu, Q., Yuan, X., Ouyang, C. & Zeng, Y. Attention-based pyramid network for segmentation and classification of high-resolution and hyperspectral remote sensing images. *Remote Sens.* **12**, 3501 (2020).
22. Li, M., Wei, M., He, X. & Shen, F. Enhancing part features via contrastive attention module for vehicle re-identification. In *2022 IEEE International Conference on Image Processing (ICIP)* 1816–1820 (IEEE, 2022).
23. Shen, F., Zhu, J., Zhu, X., Xie, Y. & Huang, J. Exploring spatial significance via hybrid pyramidal graph network for vehicle re-identification. *IEEE Trans. Intell. Transp. Syst.* **23**, 8793–8804 (2021).
24. Sun, J. et al. Fusing spatial attention with spectral-channel attention mechanism for hyperspectral image classification via encoder-decoder networks. *Remote Sens.* **14**, 1968 (2022).
25. Shen, F., Xie, Y., Zhu, J., Zhu, X. & Zeng, H. Git: Graph interactive transformer for vehicle re-identification. *IEEE Trans. Image Process.* **32**, 1039–1051 (2023).
26. Ming, Y. et al. Visuals to text: A comprehensive review on automatic image captioning. *IEEE/CAA J. Autom. Sin.* **9**, 1339–1365 (2022).
27. Aleissae, A. A. et al. Transformers in remote sensing: A survey. *Remote Sens.* **15**, 1860 (2023).
28. Mauricio, J., Domingues, I. & Bernardino, J. Comparing vision transformers and convolutional neural networks for image classification: A literature review. *Appl. Sci.* **13**, 5521 (2023).
29. Weng, W., Wei, M., Ren, J. & Shen, F. Enhancing aerial object detection with selective frequency interaction network. *IEEE Trans. Artif. Intell.* **1**, 1–12 (2024).
30. Xu, Y. et al. Transformers in computational visual media: A survey. *Comput. Vis. Media* **8**, 33–62 (2022).
31. Qi, W. et al. Global-local three-dimensional convolutional transformer network for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **61**, 1–20 (2023).
32. Shen, F., Shu, X., Du, X. & Tang, J. Pedestrian-specific bipartite-aware similarity learning for text-based person retrieval. In *Proceedings of the 31th ACM International Conference on Multimedia* (2023).
33. Sangaiah, A. K. et al. Edge-IoT-UAV adaptation toward precision agriculture using 3d-lidar point clouds. *IEEE Internet Things Mag.* **8**, 19–25 (2024).
34. Anandakrishnan, J., Sundaram, V. M. & Paneer, P. STA-AgriNet: A spatio-temporal attention framework for crop type mapping from fused multi-sensor multi-temporal sits. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **18**, 1817–1826 (2024).
35. Mohan, A. & Venkatesan, M. HybridCNN based hyperspectral image classification using multiscale spatio-spectral features. *Infrared Phys. Technol.* **108**, 103326 (2020).
36. Li, H., Zhang, R., Pan, Y., Ren, J. & Shen, F. Lr-fpn: Enhancing remote sensing object detection with location refined feature pyramid network. Preprint at [arXiv:2404.01614](https://arxiv.org/abs/2404.01614) (2024).
37. Chen, Y., Jiang, H., Li, C., Jia, X. & Ghamisi, P. Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **54**, 6232–6251 (2016).
38. Yue, J., Zhao, W., Mao, S. & Liu, H. Spectral-spatial classification of hyperspectral images using deep convolutional neural networks. *Remote Sens. Lett.* **6**, 468–477 (2015).
39. Li, X., Ding, M. & Pižurica, A. Deep feature fusion via two-stream convolutional neural network for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **58**, 2615–2629 (2019).
40. Zhao, W. & Du, S. Learning multiscale and deep representations for classifying remotely sensed imagery. *ISPRS J. Photogramm. Remote Sens.* **113**, 155–165 (2016).
41. Haut, J. M., Paoletti, M. E., Plaza, J., Li, J. & Plaza, A. Active learning with convolutional neural networks for hyperspectral image classification using a new Bayesian approach. *IEEE Trans. Geosci. Remote Sens.* **56**, 6440–6461 (2018).
42. Li, Y., Zhang, H. & Shen, Q. Spectral-spatial classification of hyperspectral imagery with 3D convolutional neural network. *Remote Sens.* **9**, 67 (2017).
43. Liu, J. et al. An investigation of a multidimensional CNN combined with an attention mechanism model to resolve small-sample problems in hyperspectral image classification. *Remote Sens.* **14**, 785 (2022).
44. Yang, J., Zhao, Y.-Q. & Chan, J.C.-W. Learning and transferring deep joint spectral-spatial features for hyperspectral classification. *IEEE Trans. Geosci. Remote Sens.* **55**, 4729–4742 (2017).
45. Yu, S., Jia, S. & Xu, C. Convolutional neural networks for hyperspectral image classification. *Neurocomputing* **219**, 88–98 (2017).
46. Diakite, A., Jiangsheng, G. & Xiaping, F. Hyperspectral image classification using 3D 2D CNN. *IET Image Process.* **15** (2020).
47. Shen, F., Wei, M. & Ren, J. HSGNet: Object re-identification with hierarchical similarity graph network. Preprint at [arXiv:2211.05486](https://arxiv.org/abs/2211.05486) (2022).
48. Hu, X., Wang, X., Zhong, Y. & Zhang, L. S3ANet: Spectral-spatial-scale attention network for end-to-end precise crop classification based on UAV-borne H2 imagery. *ISPRS J. Photogramm. Remote Sens.* **183**, 147–163 (2022).
49. Shen, F., Du, X., Zhang, L. & Tang, J. Triplet contrastive learning for unsupervised vehicle re-identification. Preprint at [arXiv:2301.09498](https://arxiv.org/abs/2301.09498) (2023).
50. Haut, J. M., Paoletti, M. E., Plaza, J., Plaza, A. & Li, J. Visual attention-driven hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **57**, 8065–8080 (2019).
51. Tang, X. et al. Hyperspectral image classification based on 3-d octave convolution with spatial-spectral attention network. *IEEE Trans. Geosci. Remote Sens.* **59**, 2430–2447 (2020).

52. Roy, S. K., Manna, S., Song, T. & Bruzzone, L. Attention-based adaptive spectral-spatial kernel ResNet for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **59**, 7831–7843 (2020).
53. Shen, F. & Tang, J. Imagpose: A unified conditional framework for pose-guided person generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems* (2024).
54. Shen, F. et al. Long-term talkingface generation via motion-prior conditional diffusion model. Preprint at [arXiv:2502.09533](https://arxiv.org/abs/2502.09533) (2025).
55. Shen, F. et al. Boosting consistency in story visualization with rich-contextual conditional diffusion models. Preprint at [arXiv:2407.02482](https://arxiv.org/abs/2407.02482) (2024).
56. Shen, F. et al. Imagdressing-v1: Customizable virtual dressing. Preprint at [arXiv:2407.12705](https://arxiv.org/abs/2407.12705) (2024).
57. Hong, D. et al. SpectralFormer: Rethinking hyperspectral image classification with transformers. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–15 (2021).
58. He, J., Zhao, L., Yang, H., Zhang, M. & Li, W. HSI-BERT: Hyperspectral image classification using the bidirectional encoder representation from transformers. *IEEE Trans. Geosci. Remote Sens.* **58**, 165–178 (2019).
59. Qing, Y., Liu, W., Feng, L. & Gao, W. Improved transformer net for hyperspectral image classification. *Remote Sens.* **13**, 2216 (2021).
60. Tan, X., Gao, K., Liu, B., Fu, Y. & Kang, L. Deep global-local transformer network combined with extended morphological profiles for hyperspectral image classification. *J. Appl. Remote Sens.* **15**, 038509–038509 (2021).
61. Mei, S., Song, C., Ma, M. & Xu, F. Hyperspectral image classification using group-aware hierarchical transformer. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–14 (2022).
62. Zhong, Z., Li, Y., Ma, L., Li, J. & Zheng, W.-S. Spectral-spatial transformer network for hyperspectral image classification: A factorized architecture search framework. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–15 (2021).
63. Zhang, J., Meng, Z., Zhao, F., Liu, H. & Chang, Z. Convolution transformer mixer for hyperspectral image classification. *IEEE Geosci. Remote Sens. Lett.* **19**, 1–5 (2022).
64. Liang, M. et al. A dual multi-head contextual attention network for hyperspectral image classification. *Remote Sens.* **14**, 3091 (2022).
65. Yang, L. et al. FusionNet: A convolution-transformer fusion network for hyperspectral image classification. *Remote Sens.* **14**, 4066 (2022).
66. Song, R., Feng, Y., Cheng, W., Mu, Z. & Wang, X. BS2T: Bottleneck spatial-spectral transformer for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–17 (2022).
67. Yang, X., Cao, W., Lu, Y. & Zhou, Y. Hyperspectral image transformer classification networks. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–15 (2022).
68. Guo, F., Feng, Q., Yang, S. & Yang, W. CMTNet: Convolutional meets transformer network for hyperspectral images classification. Preprint at [arXiv:2406.14080](https://arxiv.org/abs/2406.14080) (2024).
69. Ballanti, L., Blesius, L., Hines, E. & Kruse, B. Tree species classification using hyperspectral imagery: A comparison of two classifiers. *Remote Sens.* **8**, 445 (2016).
70. Chen, Y., Zhao, X. & Lin, Z. Optimizing subspace SVM ensemble for hyperspectral imagery classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **7**, 1295–1305 (2014).
71. Sun, L., Zhao, G., Zheng, Y. & Wu, Z. Spectral-spatial feature tokenization transformer for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–14 (2022).
72. Zhong, Y. et al. Mini-UAV-borne hyperspectral remote sensing: From observation and processing to applications. *IEEE Geosci. Remote Sens. Mag.* **6**, 46–62 (2018).
73. Zhong, Y. et al. WHU-Hi: UAV-borne hyperspectral with high spatial resolution (H2) benchmark datasets and classifier for precise crop identification based on deep convolutional neural network with CRF. *Remote Sens. Environ.* **250**, 112012 (2020).
74. Guo, D. et al. Face illumination normalization based on generative adversarial network. *Nat. Comput.* **22**, 105–117 (2023).
75. Xie, F., Gao, Q., Jin, C. & Zhao, F. Hyperspectral image classification based on superpixel pooling convolutional neural network with transfer learning. *Remote Sens.* **13**, 930 (2021).
76. Li, Z. et al. Pseudo-labelling contrastive learning for semi-supervised hyperspectral and LiDAR data classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **17**, 17099–17116 (2024).

Acknowledgements

This research was funded by the National Natural Science Foundation of China (32160421) and Major Science and Technology Project of Gansu Province (24ZDNJ001).

Author contributions

X.G. and Q.F. conceptualized the study. X.G. developed the methodology. F.G. wrote the software. X.G. drafted the original manuscript. Q.F. and F.G. reviewed and edited the manuscript. Q.F. and F.G. provided supervision. X.G. and Q.F. curated the data. X.G. and Q.F. conducted the investigation. Q.F. provided resources. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to F.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025