SANet: A Self-Attention Network for Agricultural Hyperspectral Image Classification

Bo Zhang, Yaxiong Chen, Zhiheng Li, Shengwu Xiong, and Xiaoqiang Lu, *Senior Member, IEEE*

***Abstract*— Unlike conventional hyperspectral image (HSI) classification in general scenes, agricultural HSI classification poses greater challenges due to the increased occurrence of "same spectrum different object" and "different spectrum same object" phenomena caused by class similarities. Furthermore, the dense spatial distribution of land cover categories in agricultural scenes and the mixing of spatial–spectral features at crop boundaries add to the complexity of agricultural HSIs. To tackle these issues, we propose SANet, a network designed to enhance crop classification. SANet integrates spectral and contextual information while emphasizing self-correlation within the HSIs. It combines the spatial–spectral nonlocal block structure and the multiscale spectral self-attention (SSA) structure, allocating more attention resources to spatial and spectral dimensions and modeling the existing correlations within the spectral–spatial domain. Additionally, we introduce a two-branch spatial–spectral semantic extraction and fusion structure that can adaptively learn results from both branches. Experimental results demonstrate the promising performance of SANet in agricultural HSI classification by effectively utilizing spectral data, contextual information, and self-attention mechanisms.**

Manuscript received 12 August 2023; revised 8 November 2023 and 21 November 2023; accepted 24 November 2023. Date of publication 12 December 2023; date of current version 21 December 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2022ZD0160604; in part by the Project of Sanya Yazhou Bay Science and Technology City under Grant SCKJ-JYRC-2022-76 and Grant SKJC-2022-PTDX-031; in part by the Key Research and Development Program of Hubei Province under Grant 2023BAB083; in part by NSFC under Grant 62176194 and Grant 62101393; in part by the Hainan Province "Nanhai New Star" Technology Innovation Talent Platform Project under Grant NHXXRCXM202361; in part by the Youth Fund Project of Hainan Natural Science Foundation under Grant 6220N344; in part by CAAI-Huawei MindSpore Open Fund under Grant CAAIXSJLJJ-2022-001A; in part by the Knowledge Innovation Program of Wuhan-Basic Research; in part by the Sanya Science and Education Innovation Park of Wuhan University of Technology under Grant 2022KF0020; in part by the High-Performance Computing Platform of YZBSTCACC; in part by the Natural Science Foundation of Chongqing under Grant cstc2021jcyj-msxmX1148; and in part by MindSpore, which is a new deep learning computing framework. *(Corresponding authors: Yaxiong Chen; Shengwu Xiong.)*

Bo Zhang is with the Sanya Science and Education Innovation Park, Wuhan University of Technology, Sanya 572000, China, also with the School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan 430070, China, also with the Wuhan University of Technology Chongqing Research Institute, Chongqing 401122, China, and also with the Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China.

Yaxiong Chen, Zhiheng Li, and Shengwu Xiong are with the School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan 430070, China, also with with the Sanya Science and Education Innovation Park, Wuhan University of Technology, Sanya 572000, China, also with the Wuhan University of Technology Chongqing Research Institute, Chongqing 401122, China, and also with the Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China (e-mail: chenyaxiong@whut.edu.cn; xiongsw@whut.edu.cn).

Xiaoqiang Lu is with the College of Physics and Information Engineering, Fuzhou University, Fuzhou 350108, China.

Digital Object Identifier 10.1109/TGRS.2023.3341473

***Index Terms*— Agriculture hyperspectral image (HSI) classification, deep learning (DL), nonlocal self-attention, transformer.**

I. Introduction

HYPERSPECTRAL imaging captures detailed spectral information by collecting data from multiple adjacent wavebands, allowing for a more accurate representation of object characteristics. With the rapid development of spectral imaging technology, hyperspectral image (HSI) classification has always been a concern of researchers in the field of remote sensing but also derived a series of problems and challenges that need to be solved. Because the features of HSIs usually contain a lot of redundant or irrelevant information, feature selection or dimensionality reduction is required before classification. Correct selection of relevant features and removing irrelevant features is crucial to improve classification performance. The principal component analysis (PCA) [1] is a classical method for dimension reduction. In order to retain spatial–spectral information useful for classification, Xia et al. [2] a rotating random forest (RoRF-KPCA) for nuclear PCA. In addition, HSIs may be affected by a variety of noise, such as sensor noise, atmospheric influence, and illumination changes. These noises can interfere with the real information in the image, increasing the complexity of the classification, and may lead to false classification results. Therefore, it is necessary to estimate the noise to analyze its impact on the spectral features of ground objects and guide model design. Qian et al. [3] proposed a 3-D nonlocal means filter for denoising, while Mahmood et al. [4] introduced a modified residual method for noise estimation.

Due to limitations in spatial resolution, mixed pixels are frequently encountered in HSI, resulting in significant spectral feature mixing. In agricultural hyperspectral scenes, characterized by dense crop spatial distribution, object boundaries often appear as gradient or mixed regions, posing a challenge for hyperspectral classification. To ensure accurate extraction of ground object information, Hong et al. [8] proposed an innovative spectral mixture model called the augmented LMM. This model addresses spectral variability by employing a data-driven learning strategy in the inverse problems of hyperspectral unmixing. Luo et al. [9], considering hybrid image element decomposition, introduced a novel algorithm for bilinear spectral unmixing of HSIs based on particle swarm optimization. These data preprocessing methods effectively remove redundant information to a certain extent and facilitate subsequent input into the network model for feature learning and extraction.

1558-0644 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Jaypee Insituite of Information Technology-Noida Sec 128 (L3). Downloaded on January 15,2025 at 10:36:02 UTC from IEEE Xplore. Restrictions apply.

During the early stages of deep learning (DL) methods, machine learning techniques were also employed to fulfill various visual tasks. Khazai et al. [10] proposed a support vector data description (SVDD) approach to enhance HSI classification. Yu et al. [11] introduced a novel supervised classification method for HSI that incorporates both spectral and spatial information. This method improved the accuracy and stability of nonlinear classification in small samples, reducing the impact of similar foreign objects. Building upon traditional machine learning approaches, Yu et al. [12] presented locality sensitive discriminant analysis for group sparse representation-based hyperspectral imagery classification (LSDA-GSRC). The method addressed the influence of limited samples through feature dimension reduction and incorporated global spatial and local spectral similarity information to strengthen collaborative constraints. However, machine learning-based methods sometimes necessitate manual design of artificial features and parameter initialization. This limitation poses challenges in terms of generalization ability, robustness of the model, and distinguishing subtle differences between similar categories.

With the maturation of DL methods, Chen et al. [13], [14], Chen and Lu [15], and Chen et al. [16] researchers have proposed various techniques and approaches to enhance HSI classification, particularly in agricultural scenes. These methods aim to leverage both spatial and spectral information to improve classification accuracy and enhance robustness. In this article, we provide an overview of the challenges encountered in hyperspectral classification and discuss different approaches that have been developed to address these issues.

Specifically, we focus on pixel-based classification, object-based classification, and DL-based methods including convolutional neural networks (CNNs) and Transformer Networks. For instance, Guo et al. [17] introduced a deep collaborative attention network model that combines 2-D and 3-D CNNs for efficient classification of HSIs. Xu et al. [18] designed a multiscale spectral–spatial convolution module to extract feature representations at multiple scales, capturing both spectral and spatial associations in the images. Yu et al. [19] proposed an intensive CNN based on feedback attention for HSI classification. Tu et al. [20] proposed a transformer network based on local semantic feature aggregation. The method extracts the global representation of an image by aggregating local semantic features and utilizes the transformer network for HSI classification. Additionally, Zhang et al. [21] proposed a MATNet network that integrates multiattention and Transformer networks for HSI classification. Also, Zheng et al. [22] proposed a rotation-invariant attention network method aimed at addressing the rotational invariance issue in HSI classification. In order to alleviate the differences in feature representation during feature extraction, many researchers have proposed feature fusion across scales. Guo et al. [23] proposed a Dual-View network, which utilizes the spectral and global spatial feature fusion to improve classification performance. And Luo et al. [24] proposed a method for HSI change detection, which utilizes multiscale diff-changed feature fusion to balance variability between features at different levels. To address the issue of rich spectral bands, Ma et al. [25] suggested a method for selecting the most representative and diverse bands based on spectral correlation measurements. To further enhance spatial–spectral feature representation, Yang et al. [26] designed multiple spectral–spatial Transformers to handle features at different levels. Zhou et al. [27] proposed a 3-D multihead self-attention mechanism that effectively combines spectral and spatial information during feature extraction. Liu et al. [28] proposed a central attention network for HSI classification. The central attention prioritizes the information of central pixels by adaptively learning the weight of each pixel. In addition, Wang et al. [29] proposed an HSI classification method based on hyper kernel neural architecture search. This method improves classification performance by automatically searching for the structure and hyperkernel functions of neural networks. Zhang et al. [30] also proposed a method that fuses hyperspectral and LiDAR data features by optimizing the transmission matrix, which enables classification using both types of data.

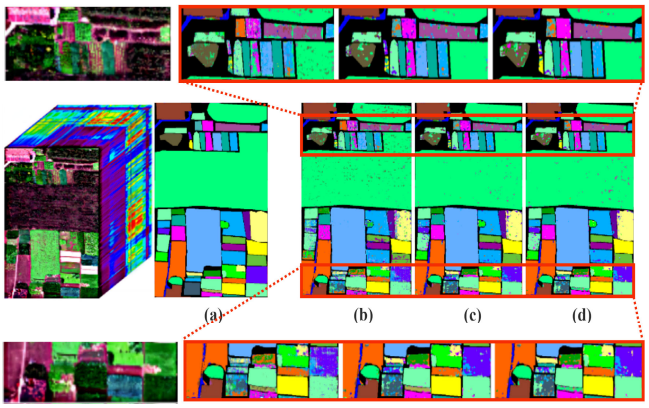


Fig. 1. Agricultural HSI classification visualization of different DL methods. Poor classification in spatially spectrally intermingled crop boundary regions. (a) Ground truth. (b) PyResNet [5]. (c) SSRN [6]. (d) A2S2KResNet [7].

Due to the significant differences between HSIs generated by different source sensors and in different scenes, cross-scenes learning among spectral images has become one of the research hotspots for HSI classification in recent years. Zhang et al. [31] proposed a technique that combines graph information aggregation and cross-domain few-shot learning to improve the accuracy of HSI classification. They further introduced a language-aware domain generalization network [32] and a single-source domain expansion network [33] for cross-scene HSI classification. Additionally, Zhang et al. [34] also proposed a topological structure and semantic information transfer network for cross-scene HSI classification. This method enhances classification performance by modeling the topological structure and utilizing semantic information transfer mechanisms. While DL-based methods have demonstrated effectiveness in classifying HSIs in general scenes, some popular approaches [5], [6], [7] lack robustness in agricultural hyperspectral scenes, as illustrated in Fig. 1.

However, existing HSI classification models designed for general scenarios face challenges in achieving robust classification performance due to the high spectral similarity and significant spatial correlation present in agricultural scenes. Specifically, in crop classification in agricultural hyperspectral scenarios, the following challenges may indeed exist.

1) Similar spectral features among different land cover categories in agricultural HSIs result in ambiguous boundary delineation.
2) High-dimensional HSIs pose computational and storage complexities due to the large volume of data.

Authorized licensed use limited to: Jaypee Insituite of Information Technology-Noida Sec 128 (L3). Downloaded on January 15,2025 at 10:36:02 UTC from IEEE Xplore. Restrictions apply.

3) Challenges in classifying adjacent crop boundaries in agricultural HSI classification are as follows.
   a) Spectral confounding: Adjacent crops may exhibit very similar spectral characteristics, leading to spectral confounding. In certain bands, different crops may have similar spectral responses, making it difficult to accurately distinguish their boundaries based solely on spectral information. This similarity can be attributed to factors such as similar chlorophyll content, vegetation structures, or light reflection properties among the crops.
   b) Spatial mixing: Boundaries between different crops often lack clear demarcation and are spatially mixed within a certain range. This mixing can occur due to factors such as field layout, farming techniques, or topography. As a result, adjacent crop boundaries in HSIs may have blurry transition regions, making boundary classification more challenging.

Therefore, in agricultural scenarios, in order to improve crop classification, it is necessary to comprehensively utilize multiple sources of information, such as spectral data and contextual information. We design SANet which combines multiple self-attention mechanisms and Transformer encoders for agricultural HSI classification.

The main contributions of this article include the following points.

1) In the spatial–spectral information extraction stage, we propose a spatial–spectral self-attention (SSA) structure called Criss-Cross-3-D (CC3D), which not only focuses on the correlation of spatial context but also enhances the modeling of spectral dimension correlation in 3-D spectral space, so as to better improve the effect of crop boundary classification in agricultural HSIs. Meanwhile, we iterate CC3D three times to achieve correlation modeling between long-distance positions in 3-D space.
2) To enhance the discriminability of spectra between crops of similar categories, we propose a multiscale spectral feature fusion (MSFF) that combines SSA and feature compression (FC) structure to effectively capture the correlations between different spectra, ranging from global to local scales.
3) We propose a spatial–spectral semantic feature extraction method by a two-branch structure and the adaptive fusion (AF) operation. This module is better for agricultural HSIs to learn and distinguish between small differences in spatial and spectral semantic features.
4) We conduct experiments on multiple challenging agricultural HSI classification datasets to evaluate the performance of our proposed network. The results demonstrate that the proposed SANet performs well in extracting the inter-correlation of spectral features and understanding rich semantic information in the HSIs.

The structure of this article is arranged as follows. Section II discusses some related work on attention mechanisms in hyperspectral classification. Section III outlines our proposed approach to address the agricultural HSI classification problem. Section IV presents the experiments conducted to validate our proposed method. Finally, Section V concludes the article and suggests future research directions.

II. Related Works

Here, we mainly distinguish attention into local attention, nonlocal attention, and composite attention for related work. We believe that the nonlocal attention mechanism is able to construct dependencies between long distances, and the use of composite attention is more effective than using single attention, which is one of the important reasons why we propose SANet.

A. Local Attention

1) Spatial Attention: HSIs typically contain rich spatial information. By employing spatial attention (SA), the model can dynamically adjust its focus on different regions of the image based on the current attention window. Woo et al. [35] proposed the CBAM structure which combines SA mechanism to focus on more important spatial areas. And Zhang et al. [21] proposed an SA used in HSIs, and it helps the model concentrate on important spatial locations such as object edges, textures.

2) Channel Attention: HSIs typically consist of multiple spectral channels, and each channel plays a different role in the classification task. Channel attention allows the model to dynamically adjust the weights of each channel based on the current attention window. Hu et al. [36] introduced the SELayer, which adaptively adjusts the feature graph's weights by learning the relationships among channels. Drawing inspiration from [36], Roy et al. [37] proposed the scSE method to enhance the effectiveness of channel attention. Furthermore, Wang et al. [38] developed a more efficient channel attention mechanism by computing inter-channel correlations locally instead of globally. These approaches enable the model to focus on the most relevant channel information for the specific classification task at hand, thereby enhancing its capability to express important features. Channel attention assists the model in distinguishing key features across different spectral bands, leading to improved classification performance.

3) Spectral Attention: In HSI classification, spectral attention mechanisms can adaptively learn the importance weights of each band according to the task demands. Wang et al. [39] proposed a twinned network structure and spectral attention mechanism to learn the importance of bands. In addition, Shi et al. [40] multiscale-spectral-attention (MSA) module as to reduce the spatial and spectral redundancies simultaneously and provide strong discrimination. Also, Zhang et al. [41] proposed a spectral attention dense block to learn long-range spectral information. By calculating the spectral attention maps, the model is able to focus on those bands that are more useful for the classification task. In this way, the model will focus more on the bands with discrimination, thus improving the classification accuracy.

B. Nonlocal Attention

Nonlocal attention is an attention mechanism used in computer vision tasks to enhance the model's representation ability by considering the correlation among different positions in

Authorized licensed use limited to: Jaypee Insituite of Information Technology-Noida Sec 128 (L3). Downloaded on January 15,2025 at 10:36:02 UTC from IEEE Xplore. Restrictions apply.

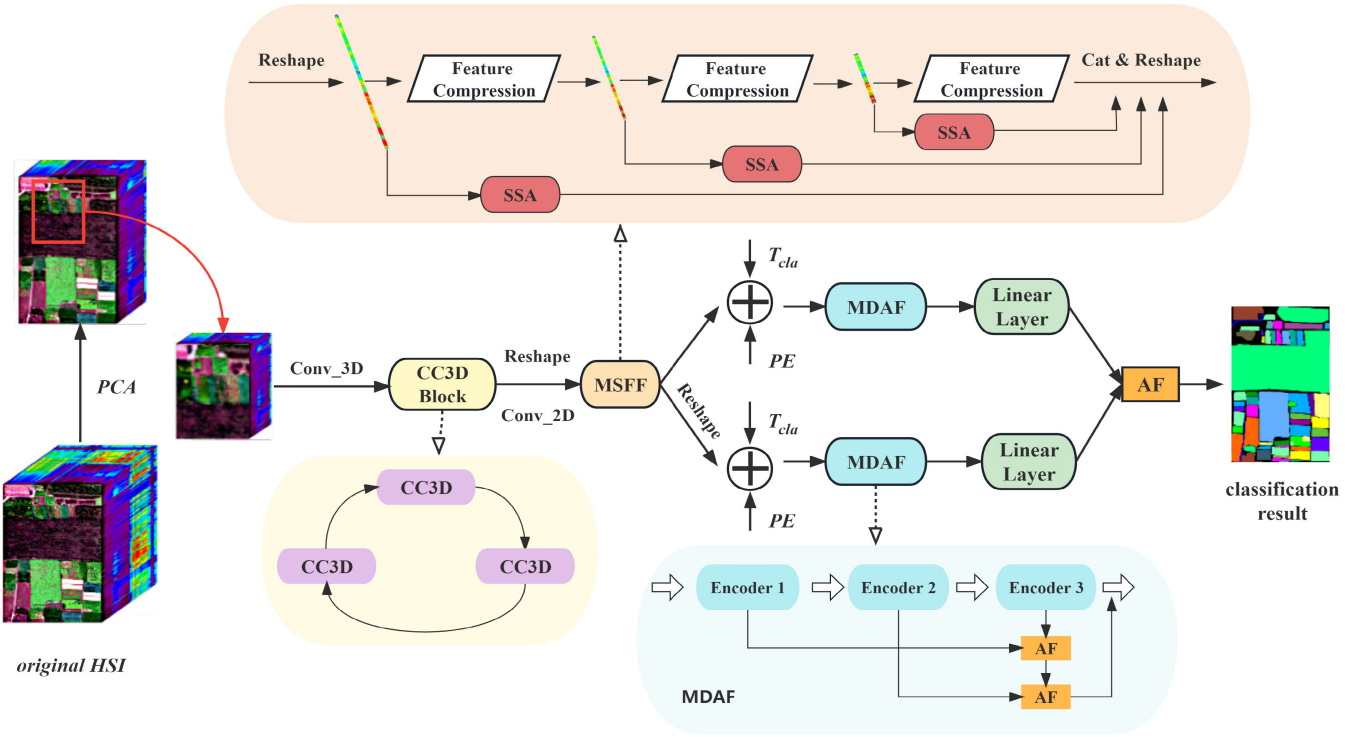


Fig. 2. Overview of the proposed SANet framework for the agricultural HSI classification. The proposed SANet mainly includes, in turn, CC3D self-attention block, multiscale spectral feature fusion module, and a spatial–spectral two-branched semantic feature extraction structure composed of transformer.

an image. Wang et al. [42] first proposed the concept of nonlocal operation and applied it in CNNs to model pixel dependence over long distances. By introducing nonlocal operations, the method is able to capture the global context information in the HSIs, thus improving the classification accuracy. Wang et al. [43] utilized nonlocal operations to capture global information in HSIs and achieves excellent classification results. Mou et al. [44] propose an HSI classification method based on nonlocal graph convolutional networks. It builds graph convolution networks using nonlocal neighborhood information and extracts spectral features through graph convolution operations to enhance global context information. Shen et al. [45] integrated nonlocal block or attention mechanisms into the CNN architecture to capture nonlocal dependencies in the hyperspectral data. These studies further validate the effectiveness of nonlocal neural networks in HSI classification. In a word, nonlocal attention mechanisms can better establish correlations between long-distance features for HSIs.

C. Composite Attention

Composite attention leverages the power of multiple attention mechanisms to enhance feature representation. Dong et al. [46] proposed a network structure called Dense Connection that incorporates both spectral attention and SA mechanisms. Shu et al. [47] introduced Spatial Spectral Split Attention, which dynamically learns the relevance of each dimension by separately weighting the spatial and spectral dimensions. Liang et al. [48] designed a multiscale spectral SA module that adaptively learns correlations between different scales and spectral bands, effectively weighting the features. Zhao et al. [49] presented a method that utilizes both self-attention and cross-channel attention mechanisms to learn the relationships and importance between features. By adjusting and combining attention weights from different types of attention, more accurate and informative features can be selected and strengthened in hyperspectral data. This integration of diverse attention mechanisms enables the model to effectively capture relevant information and improve the performance of tasks such as classification or segmentation in hyperspectral analysis.

III. Proposed Method

We will introduce our proposed SANet according to the order of the network structure shown in Fig. 2, mainly including CC3D Self-Attention Block, Multiscale SSA Module and a Spatial–Spectral two-branched Semantic Feature Extraction structure. The specific methods are as follows.

A. CC3D Self-Attention Block

HSIs are often rich in spatial information. Different from the 2-D spatial nonlocal attention mechanism proposed in [50], in order to effectively model the relationship between different positions in the 3-D spectral space, we apply nonlocal attention in depth, height and width dimensions simultaneously. The model can capture the long-range dependencies on different spatial dimensions, so as to better understand the relationship between the characteristics of different positions in the image. In addition, the attention on detail information, context relationship, and global structure in the image can effectively improve the problem of boundary feature confounding in agricultural HSI classification, and improve the accuracy and robustness of classification. In addition, Our proposed method saves more computational resources and memory than [42].

Authorized licensed use limited to: Jaypee Insituite of Information Technology-Noida Sec 128 (L3). Downloaded on January 15,2025 at 10:36:02 UTC from IEEE Xplore. Restrictions apply.

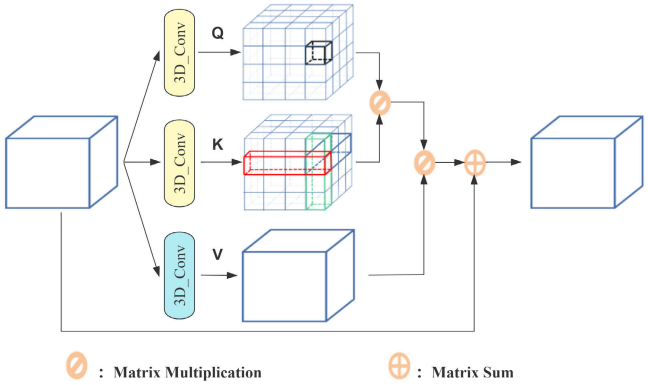


Fig. 3. CC3D self-attention module which mainly reflects the idea of nonlocal blocks being applied to different directions of 3-D space. A single CC3D operation can learn the correlation between a pixel and its pixels in the same height, the same width, and the same depth.

1) CC3D Self-Attention: The proposed CC3D is an attentional mechanism used in computer vision for models to capture long-range dependencies between different dimensions. This module can calculate the nonlocal attention in three directions including height, width, and depth to obtain a more global feature representation.

Assume that query ($\boldsymbol{Q}$), key ($\boldsymbol{K}$), and value ($\boldsymbol{V}$) are the three inputs used to calculate attention weights. In detail, $\boldsymbol{Q}$ representing the current position is used to calculate the correlation with other positions. $\boldsymbol{K}$ represents the feature vector of other locations to calculate the degree of correlation with the query. And $\boldsymbol{V}$ represents the feature vector of other locations used to provide weighted context information to the query based on the correlation degree. In contrast to Transformer, $\boldsymbol{K}$ and $\boldsymbol{V}$ in CC3D are not merely acting on the current position, but on all positions. This design can capture a wider range of contextual information to improve the representation power of the model. The specific execution process of this module is shown in Fig. 3. And the algorithm steps of CC3D are as follows.

1) First assuming the shape of the input tensor is $X \in R^{B\times C\times D\times H\times W}$, and performing a series of convolution and transformation operations on $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$, transforming them into projected representations in the depth, height, and width directions, respectively.
2) Then, through the matrix multiplication and addition operation of the projection representation, the energy matrix in three directions is calculated: energy_D, energy_H, and energy_W. These energy matrices reflect the similarity between the $\boldsymbol{Q}$ and the $\boldsymbol{K}$.
3) Moreover, the energy matrix is spliced in a certain manner and converted to attention weights using the softmax operation.
4) Finally, the attention weights are matrix-multiply with the $\boldsymbol{V}$ to obtain the output tensor (out_D, out_H and out_W), which is added to the input tensor and multiplied by the learned parameter γ, as shown in (1). Unlike the SSA module in (11), CC3D requires adding the attention weights from three directions: depth, height, and width

$$\text{Output} = \gamma \times (\text{out\_}D + \text{out\_}H + \text{out\_}W) + X \quad (1)$$

where γ represents a learnable scaling factor.

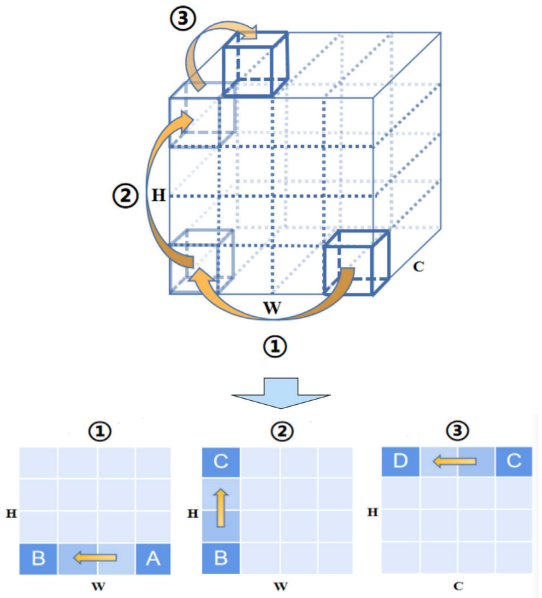


Fig. 4. Process of the CT. The transfer of correlation is achieved through three iterations of the CC3D operations.

By the way, the calculation process of each step is from (6) to (11), which will be elaborated when introducing the SSA module.

2) Correlation Transfer: The CC3D simultaneously applies nonlocal attention to the depth, height, and width dimensions, enabling the model to model the spatial relationships between different locations in the input feature map. However, due to the complex computational overhead, we focus on the association between the current pixel and the same row, the same column, and the pixel in the same spectrum, rather than the association between any two pixels in the 3-D spectral space. Therefore, we achieved the transfer of correlation through multiple iterations using the CC3D module to achieve the desired effect. The conceptual diagram of the correlation transfer (CT) is shown in Fig. 4.

Assuming point A coordinates (x_i, y_j, z_k), B $(x_{i'}, y_j, z_k)$, C $(x_{i'}, y_{j'}, z_k)$, and D $(x_{i'}, y_{j'}, z_{k'})$. Then the process of ① in Fig. 4 is shown in the following equation:

$$(x_i, y_j, z_k) \longrightarrow (x_{i'}, y_j, z_k). \quad (2)$$

The process of ② in Fig. 4 is shown in the following equation:

$$(x_{i'}, y_j, z_k) \longrightarrow (x_{i'}, y_{j'}, z_k). \quad (3)$$

The process of ③ in Fig. 4 is shown in the following equation:

$$(x_{i'}, y_{j'}, z_k) \longrightarrow (x_{i'}, y_{j'}, z_{k'}). \quad (4)$$

Finally, the process of CT can be expressed as follows:

$$\textbf{Relation}((x_i, y_j, z_k) \longrightarrow (x_{i'}, y_{j'}, z_{k'})) = \sum_{R=1}^{3} \text{CT}(x_i, y_j, z_k) \quad (5)$$

where **Relation** represents the correlation modeling between different points, R represents the number of recurrences.

Authorized licensed use limited to: Jaypee Insituite of Information Technology-Noida Sec 128 (L3). Downloaded on January 15,2025 at 10:36:02 UTC from IEEE Xplore. Restrictions apply.

B. MSFF Module

In agricultural HSIs, pixels of the same crop category exhibit a strong correlation, indicating a relationship between the spectral vectors of a pixel and those of other pixels. By identifying these correlations, more representative features specific to a certain category can be extracted. Furthermore, land cover classes in HSIs often display diversity and complexity, with visually similar crop categories exhibiting distinct features across different spectral scales. By considering spectral information at various scales, we can capture the variations and discrepancies of similar crop categories within different spectral ranges, thus further enhancing the robustness of classification.

1) Spectral Self-Attention: SSA implements an SSA module based on nonlocal attention mechanism. Different from the self-attention mechanism in Transformers, SSA calculates the correlations between each vector in the input spectral feature vector sequence and other vectors, and computes the attention weights for each vector. While the self-attention mechanism in Transformer [51] captures the long-range dependence between the different positions in the sequence, rather than the correlation between the sequences. The output is obtained by performing a weighted sum. In summary, The main function of this module is to calculate the attention values for each element in the input sequence, weight and merge their information, and generate a comprehensive feature representation output. By applying the SSA module at different levels of features, the features from different levels can interact and reinforce each other, thereby improving the expressiveness and discriminability of the features.

The specific calculation process of SSA shown in Fig. 5 is described as follows: the input $X \in R^{(B\times C\times L)}$, and we can define query ($\boldsymbol{Q}$), key ($\boldsymbol{K}$) and value ($\boldsymbol{V}$). By the way, the meanings of $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ representatives are consistent with those in CC3D, with the difference in the calculation procedure and the resulting tensor sizes. They can be calculated as follows:

$$\boldsymbol{Q} = \text{query\_conv}(X) \in R^{B\times C/\lambda\times L_Q} \tag{6}$$

$$\boldsymbol{K} = \text{key\_conv}(X) \in R^{B\times C/\lambda\times L_K} \tag{7}$$

$$\boldsymbol{V} = \text{value\_conv}(X) \in R^{B\times C\times L_V} \tag{8}$$

where L_Q, L_K, and L_V are the sequence lengths of the $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$, respectively. And λ is the channel compression rate.

And the attention is calculated as follows:

$$\begin{aligned}\text{Attention\_}H &= \text{softmax}(\text{energy\_}H)\\ &= \text{softmax}[(Q\_H * K\_H^T + \text{INF})]\end{aligned} \tag{9}$$

where $*$ represents the matrix multiplication, INF is a matrix where the elements on the diagonal are of negative infinity. And Q_H K_H are the transpose form of the corresponding representation, used to compute the attention matrix.

Finally the weight is calculated as (10) and the Output is calculated as (11)

$$\text{out\_}H = V\_H^T * \text{Attention\_}H \tag{10}$$

$$\text{Output} = \gamma \times \text{out\_}H + X \tag{11}$$

where $*$ represents the matrix multiplication, γ is a learnable scaling factor, and V_H is the transpose form of the corresponding representation.

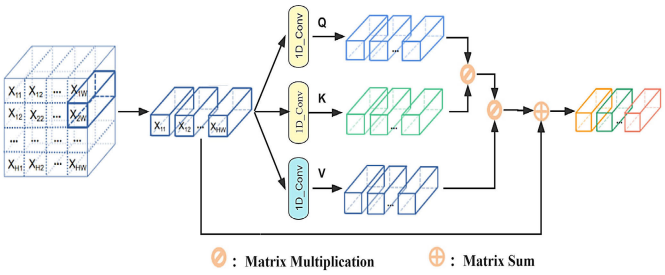


Fig. 5. SSA module can learn the association between different spectral vectors within the hyperspectral data cube.

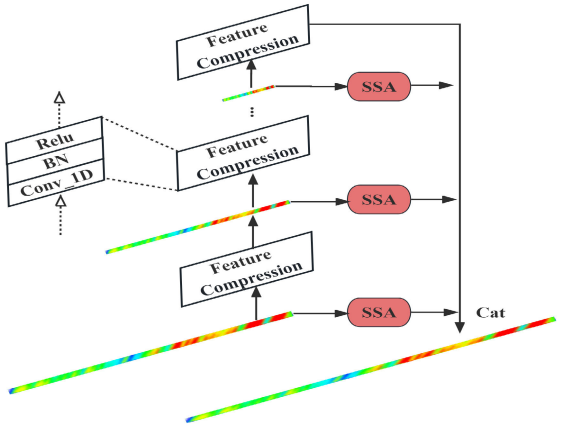


Fig. 6. Structure of MSFF. The MSFF mainly includes SSA, FC, and fusion (Cat) operations.

2) Feature Compression: Usage of FC: Convolution operation smooths and extracts features in the spectral dimension. By applying convolution operations to the input features, features of different scales and abstraction levels can be extracted. This helps to remove noise and redundant information while retaining important features, achieving feature extraction and dimensionality reduction. The FC is formulated as follows:

$$\text{FC}(S) = \text{Relu}\{\text{BN}[\text{Conv\_1D}(S)]\} \tag{12}$$

where the kernel size of Conv_1D is 4, and the S represents the input spectral feature.

In summary, MSFF shown in Fig. 6 is formulated as follows:

$$\begin{aligned}S' = {} & \text{SSA}(S) + \text{SSA}[\text{FC}(S)] + \text{SSA}\{\text{FC}[\text{FC}(S)]\}\\ & + \text{FC}\{\text{FC}[\text{FC}(S)]\}\end{aligned} \tag{13}$$

where S' represents the output enhanced spectral feature, and $+$ represents the Cat operation in the graph, indicating connecting the vectors along the spectral dimension.

C. spatial–spectral Semantic Feature Extraction

We use two identical branches to extract multilevel spatial and spectral semantic features separately, and the structure framework is shown in Fig. 7.

1) AF Operation: In order to make the network better adaptively learn the importance of different semantic features, we use the AF module to integrate the results obtained by two different semantic features. The specific fusion method is shown in Fig. 8, and finally the results obtained by fusion are used as the basis for label discrimination.

Authorized licensed use limited to: Jaypee Insituite of Information Technology-Noida Sec 128 (L3). Downloaded on January 15,2025 at 10:36:02 UTC from IEEE Xplore. Restrictions apply.

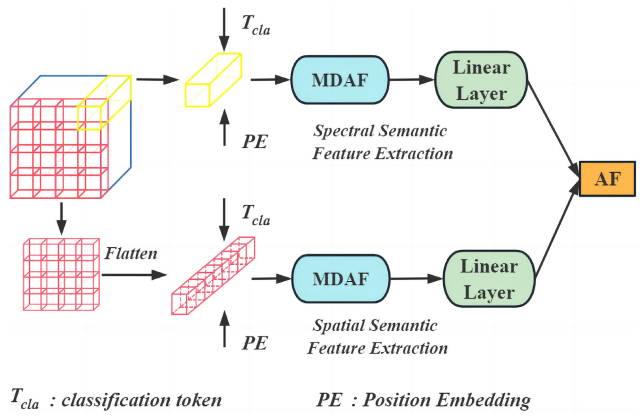


Fig. 7. Two-branch structure of spatial–spectral semantic extraction. The input feature vector size of the two branches is inconsistent, but the output feature vector size is the same after performing the MDAF module.

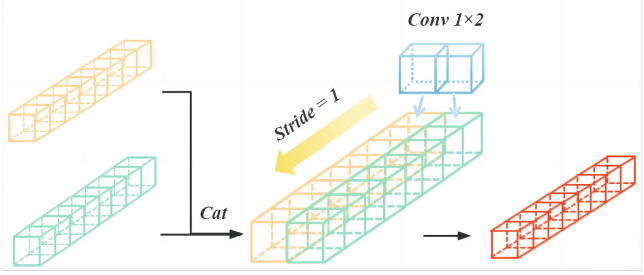


Fig. 8. Structure of AF. The 1-D feature vector size of the output remains consistent with the input features.

The mathematical expression of AF is the following equation:

$$\mathrm{AF} = \mathrm{Conv}_{1\times 2}[\mathrm{Cat}(E_i, E_j)] \tag{14}$$

where E is the input feature which is a 1-D-vector, E_i represents the output feature of different encoder.

2) Spatial Semantic Feature Extraction: Similar to common panchromatic RGB images, HSIs also contain abundant spatial context information. The distinctive texture features in HSIs play a crucial role in differentiating between various crops. These texture features can be categorized into different levels of semantic features, including low-level spatial semantic features such as shape and color, as well as high-level semantic features encompassing various texture details. To explore the spatial semantic features of HSIs, we employ the multiscale dense AF (MDAF) structure. This structure facilitates the fusion of semantic features at different levels, thereby enhancing the richness of the final semantic information obtained. By effectively integrating these semantic features, we can gain a better understanding of the HSIs' spatial characteristics and improve crop classification performance.

3) Spectral Semantic Feature Extraction: Unlike panchromatic RGB images, HSIs possess not only rich spatial semantic features but also rich spectral semantic features. The low-level spectral semantic features consist of spectral reflectance, while the higher-level semantic features include spectral curves and spectral indices, among others. Spatial semantic features alone may not always be sufficient to differentiate between similar crop categories, thus highlighting the need to explore the richer spectral semantic information.

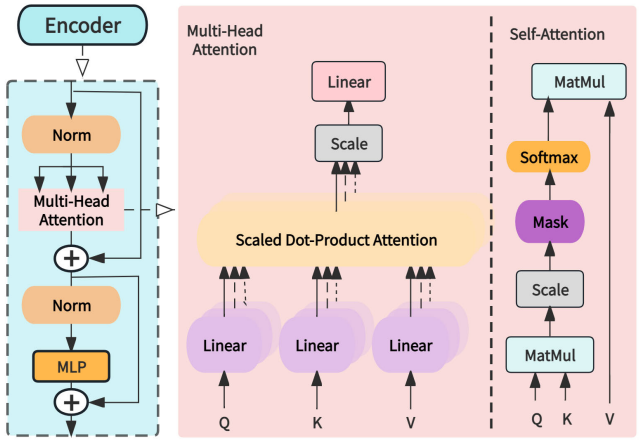


Fig. 9. Encoder with multihead self-attention mechanism which captures the long-range dependence between the different positions in the sequence.

To address this, we utilize the MDAF structure to extract multilevel spectral semantic information. By leveraging this structure, we can effectively capture and analyze the various spectral characteristics present in HSIs, enabling more accurate classification of similar crop categories based on their unique spectral signatures.

The mathematical expression of MDAF is the following equation:

$$\mathrm{MDAF}(E) = \mathrm{AF}[\mathrm{AF}(E_3, E_1), E_2] \tag{15}$$

where E is the input feature which is a 1-D-vector, E_i represents the output feature of different encoder.

4) Transformer Encoder: In the Encoder structure, the multihead self-attention mechanism refers to the application of the self-attention mechanism to multiple independent attention heads and integrates them through certain linear transformations. The self-attention mechanism is a mechanism used to process sequence data that can assign different weights to each element in a sequence in order to establish dependencies between elements in the model [51]. The detailed structure of the encoder is shown in Fig. 9.

5) Multihead Self-Attention: Assuming that X is an input feature, the corresponding calculation formulas of query ($\boldsymbol{Q}$), key ($\boldsymbol{K}$), and value ($\boldsymbol{V}$) are shown in the following equation:

$$\boldsymbol{Q} = X * W_Q \tag{16}$$
$$\boldsymbol{K} = X * W_K \tag{17}$$
$$\boldsymbol{V} = X * W_V \tag{18}$$

where W_Q, W_K, W_V are the weight matrix of a linear transformation of the input, and the $*$ represents the matrix multiplication.

The Attention Output is expressed by a mathematical formula as follows:

$$\mathrm{SA}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_K}}\right)\boldsymbol{V} \tag{19}$$

where d_K represents the dimension of $\boldsymbol{K}$. And $\mathrm{Softmax}(\boldsymbol{Q}\boldsymbol{K}^T/\sqrt{d_K})$ represents the Attention Scores.

The above calculation process is conducted for each attention head, which can have multiple parallel attention heads.

Authorized licensed use limited to: Jaypee Insituite of Information Technology-Noida Sec 128 (L3). Downloaded on January 15,2025 at 10:36:02 UTC from IEEE Xplore. Restrictions apply.

D. Implementation

We introduce the step flow of the network in Algorithm 1 to facilitate a better understanding of the network structure.

Algorithm 1 Self-Attention Network for Agricultural HSI Classification

Input: The agricultural hyperspectral data.
Output: Predicted labels of different classes.
Initialization:
batch size is 64; PCA bands number is 30; patch size is 13; training sample rate is 1%; learning rate to $1e-3$;
Repeat:
1: Perform a $3D$ convolution layer;
2: Perform the CC3D Self-Attention block which recurrences with three CC3D operations.;
3: Reshape the $3D$ feature maps;
4: Perform a $2D$ convolution layer;
5: Reshape the $2D$ feature maps;
6: Generate enhanced semantic tokens via the MSFF module which combines SSA modules and FC blocks;
7: Extract the spatial and spectral semantic information on the enhanced tokens through two branches;
8: Concatenate the classification tokens $\boldsymbol{T}_{cla}$ to get new spatial and spectral semantic tokens;
9: Add the position information on the new spatial and spectral semantic tokens;
10: Perform the MDAF module separately;
11: Take the classification token as the input to the linear layer;
12: Perform the AF module on the outputs from two linear layers to get labels;
end for: All test samples are predicted.

Take the WHU-Hi-HongHu dataset as the input to our proposed SANet. The detailed steps and sizes are shown in Table I.

TABLE I
SHOW THE DETAILS OF SANET ON THE WHU-HI-HONGHU DATASET

Step	Step Name	Size
1	Original Agriculture HSI	$940 \times 475 \times 270$
2	Reduce Dimension (PCA)	$940 \times 475 \times 30$
3	Extract Patch	$13 \times 13 \times 30$
4	Perform $3D$ convolution	$8\ 11 \times 11 \times 28$
5	Perform CC3D Block	$8\ 11 \times 11 \times 28$
6	Reshape	$224\ 11 \times 11$
7	Perform $2D$ convolution	$64\ 9 \times 9$
8	Flatten	$64\ 1 \times 81$
9	Reshape	64×81
10	Perform MSFF	85×81
11.1	Add Classification Token $\boldsymbol{T}_{cla}^{spe}$	$\boldsymbol{T}_{spe} \in \boldsymbol{R}^{86\times 81}$
11.2	Add Classification Token $\boldsymbol{T}_{cla}^{spa}$	$\boldsymbol{T}_{spa} \in \boldsymbol{R}^{82\times 85}$
12.1	Position Embedding	$\boldsymbol{T}_{spe} \in \boldsymbol{R}^{86\times 81}$
12.2	Position Embedding	$\boldsymbol{T}_{spa} \in \boldsymbol{R}^{82\times 85}$
13.1	Perform MDAF	$\boldsymbol{T}_{spe} \in \boldsymbol{R}^{86\times 81}$
13.2	Perform MDAF	$\boldsymbol{T}_{spa} \in \boldsymbol{R}^{82\times 85}$
14.1	Take Out Classification token	$\boldsymbol{T}_{cla}^{spe} \in \boldsymbol{R}^{1\times 81}$
14.2	Take Out Classification token	$\boldsymbol{T}_{cla}^{spa} \in \boldsymbol{R}^{1\times 85}$
15.1	Perform Linear Layer	$\boldsymbol{P}_{spe} \in \boldsymbol{R}^{1\times 16}$
15.2	Perform Linear Layer	$\boldsymbol{P}_{spa} \in \boldsymbol{R}^{1\times 16}$
16	Perform AF	$\boldsymbol{P}_{i} \in \boldsymbol{R}^{1\times 16}$

IV. EXPERIMENTS

In this section, we list three datasets for HSI classification used in the experiments. Then we verify the effective performance of the network, and the ablation experiments and comparison experiments are performed on our proposed modules.

A. HSI Datasets

To validate our proposed SANet for good classification in agricultural hyperspectral scenarios with dense spatial–spectral features, we perform experiments on three typical agricultural HSI datasets. They are WHU-Hi-HongHu, WHU-Hi-HanChuan and WHU-Hi-LongKou. The details of the datasets are shown in Table II. Due to the dense feature distribution in agricultural hyperspectral scenarios, we select 1% of the samples in each category for training, and the remaining samples are used for testing. And the detailed category names and sample numbers of three datasets are shown in Tables III–V.

B. Experimental Setup

1) Evaluation Metrics: We use three convincing classification performance indicators in the experiment, including overall accuracy (OA), average accuracy (AA), Kappa Coefficient (κ). At the same time, we show the classification accuracy of each category in the results.

2) Configuration: The SANet proposed is implemented under PyTorch and Mindspore. Also, we test all networks with the workstation which combines Intel Xeon[1] Silver 4210R CPU, 64GB RAM, and an NVIDIA Quadro RTX 5000 16GB GPU. The optimizer used in SANet is Adam [52]. By the way, the learning rate is set to $1e-3$.

3) Comparison With Some State-of-the-Art DL-Based Networks: To prove the performance of the proposed SANet, we selected some popular methods for comparison: Resnet [53], PyResnet [5], ContextualNet [54], SSRN [6], A2S2KResNet [7], SSFTT [55], SFormer [56], LSFAT [20], MATNet [21], and our proposed SANet.

1) For Resnet [53], PyResnet [5], ContextualNet [54], SSRN [6], A2S2KResNet [7], and SFormer [56], we keep the specific settings of these network structures consistent with original papers. All networks are trained 200 epochs on the three datasets.
2) For the SSFTT [55], MATNet [21], and LSFAT [20], we keep the specific settings of the network structure consistent with the original papers. And we train 100 epochs on the three datasets.
3) For the proposed SANet, we set the size of the convolution kernel of the 3-D-convolution to $3 \times 3 \times 3$ and the number of output channels to 8. In addition, the size of the 2-D-convolution kernel is 3×3 and the number of output channels is 64. We train 100 epochs on the three datasets.
4) For the patch size of the input, we set it to 13 except LSFAT. And we set it to 15 which is consistent with [20].
5) For loss function, to better compare the powerful learning ability of the proposed SANet, we adapt the Lpoly

[1]Registered trademark.

Authorized licensed use limited to: Jaypee Insituite of Information Technology-Noida Sec 128 (L3). Downloaded on January 15,2025 at 10:36:02 UTC from IEEE Xplore. Restrictions apply.

TABLE II
DETAILS OF THREE DATASETS

Dataset Name	Spectral Resolution	Spatial Resolution	Bands Number	Classes Number	Pixel Size	Training Proportion
WHU-Hi-LongKou	6nm	0.463m	270	9	550×400	1%
WHU-Hi-HongHu	6nm	0.043m	270	22	940×475	1%
WHU-Hi-HanChuan	6nm	0.109m	274	16	1217×303	1%

TABLE III
DETAILS AND THE NUMBER OF SAMPLES FOR WHU-HI-HANCHUAN

No.	Class Name	Training	Testing
1	Strawberry	456	44279
2	Cowpea	231	22522
3	Soybean	105	10182
4	Sorghum	54	5299
5	Water spinach	12	1188
6	Watermelon	46	4487
7	Greens	59	5844
8	Trees	183	17795
9	Grass	95	9374
10	Red roof	106	10410
11	Gray roof	174	16737
12	Plastic	39	3640
13	Bare soil	91	9025
14	Road	188	18372
15	Bright object	11	1125
16	Water	768	74633
	Total	2618	254912

TABLE IV
DETAILS AND THE NUMBER OF SAMPLES FOR WHU-HI-LONGKOU

No.	Class Name	Training	Testing
1	Corn	345	34166
2	Cotton	84	8290
3	Sesame	30	3001
4	Broad-leaf soybean	633	62579
5	Narrow-leaf soybean	42	4109
6	Rice	119	11735
7	Water	670	66386
8	Roads and houses	71	7053
9	Mixed weed	52	5177
	Total	2046	202496

loss function proposed by MATNet [21], and we all keep the parameters in the loss function consistent with MATNet on different datasets.

We maintain the 1% training proportion of samples trained during all network training stage, and the remaining 99% of samples are used for testing. We set all of the batch size to 32, set the reduction dimension of PCA to 30 and set learning rate (lr) to $1e-3$. In addition, all network models are performed ten experiments under the Adam optimizer.

4) Parameter Experiment:

1) To verify whether the PCA dimensionality reduction method can affect the correlation of the original spectral space, we conduct comparative tests on the Indian Pines dataset, including the absence of dimensionality reduction preprocessing (NoPCA) and setting the number of bands after dimension reduction to 30, and the results are shown in Table VI. We can see from the results that both the OA and κ are improved by 6% and 6.74%, respectively, after using PCA. Although AA has

TABLE V
DETAILS AND THE NUMBER OF SAMPLES FOR WHU-HI-HONGHU

No.	Class Name	Training	Testing
1	Red roof	140	13901
2	Road	35	3477
3	Bare soil	218	21603
4	Cotton	1639	161646
5	Cotton firewood	62	6156
6	Rape	447	44110
7	Chinese cabbage	243	23860
8	Pakchoi	41	4013
9	Cabbage	108	10711
10	Tuber mustard	124	12270
11	Brassica parachinensis	111	10904
12	Brassica chinensis	90	8864
13	Small Brassica chinensis	225	22282
14	Lactuca sativa	74	7282
15	Celtuce	10	992
16	Film covered lettuce	73	7189
17	Romaine lettuce	30	2980
18	Carrot	32	3185
19	White radish	87	8625
20	Garlic sprout	35	3451
21	Broad bean	13	1315
22	Tree	40	4000
	Total	3877	382816

TABLE VI
CLASSIFICATION RESULTS OF THE SPECIFIC CATEGORIES AT 5% TRAINING SAMPLESNON THE INDIAN PINES DATASET. THE BOLD ONE IS THE OPTIMAL RESULT

Classes	Sample Size		Result	
	Training	Testing	NoPCA	PCA_{30}
Alfalfa	3	43	34.88	6.98
Corn Notill	71	1357	87.40	93.00
Corn Mintill	41	789	98.08	97.85
Corn	13	224	83.04	89.29
Grass Pasture	24	459	95.64	98.69
Grass Trees	38	692	99.86	99.28
Grass Pasture Mowed	1	27	29.63	0.00
Hay Windrowed	25	453	99.78	99.56
Oats	1	19	63.16	21.05
Soybean Notill	50	922	96.31	94.47
Soybean Mintill	125	2330	80.56	98.71
Soybean Clean	30	563	71.76	88.28
Wheat	10	195	99.49	99.49
Woods	64	1201	98.92	98.44
Buildings Grass Trees	19	367	87.47	91.04
Stone Steel Towers	6	87	91.95	92.12
OA (%)	-	-	89.50	**95.50**
AA (%)	-	-	**82.31**	78.82
κ*100	-	-	88.11	**94.85**

Authorized licensed use limited to: Jaypee Insituite of Information Technology-Noida Sec 128 (L3). Downloaded on January 15,2025 at 10:36:02 UTC from IEEE Xplore. Restrictions apply.

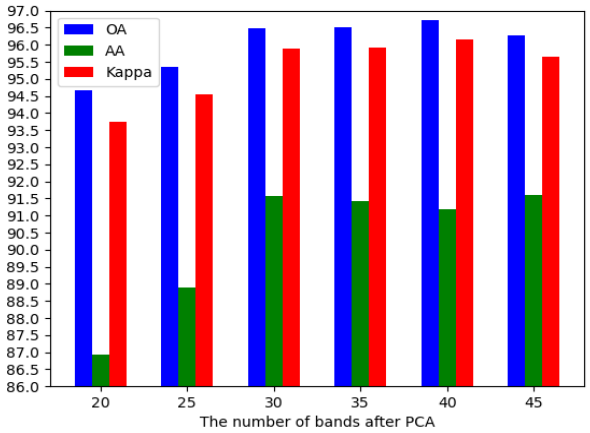


Fig. 10. Classification results when different amounts of PCA dimension reduction are set on WHU-Hi-HanChuan dataset.

TABLE VII

ABLATION RESULTS OF CC3D-BLOCK ON WHU-HI-HANCHUAN DATASET. THE OPTIMAL RESULTS ARE INDICATED IN BOLD

Metric	CC3D*1	CC3D*2	CC3D*3
OA (%)	94.39	96.03	**96.49**
AA (%)	87.37	**91.56**	**91.56**
κ*100	93.42	95.35	**95.88**

TABLE VIII

ABLATION RESULTS OF CONVOLUTIONAL KERNEL SIZE IN FC ON WHU-HI-HANCHUAN DATASET. THE OPTIMAL RESULTS ARE INDICATED IN BOLD

Metric	The size of kernel				
	3	4	5	6	7
OA (%)	95.01	96.49	**96.51**	96.02	95.71
AA (%)	86.96	**91.56**	91.45	89.98	90.17
κ*100	94.15	95.88	**95.91**	95.33	94.97

decreased, we analyze that this is the very few categories of training samples, and the PCA method may lose a small amount of useful spectral information. In a word, the significant increase in OA has proved that PCA does not destroy the correlation between spectra. It is worth mentioning that after using PCA, we save great computational memory consumption, so we think it is necessary and meaningful to reduce the dimension of HSIs.

2) Due to the huge amount of data in the original HSIs, there is a large amount of redundant spectral information between adjacent spectra that is easy to interfere with classification, so we, like most methods, we first adopted PCA method to reduce the dimension. In order to find the appropriate dimension reduction, we tried on the WHU-Hi Hanchuan dataset to [20, 25, 30, 35, 40, 45], and the experimental results are shown in Fig. 10. We can intuitively see that the classification results remain stable at the dimensionality reduction degree greater than 30. And OA and κ reach the optimal at 40, which are 96.71%, 0.9615%, respectively. By the way, AA reaches the optimal at 45, which is 91.61%. Given that the number of pca dimension reduction coincides with the other contrast experiments, we finally set it to 30. Even so, our method is still the best.
3) For the CC3D module, to verify the effectiveness of CT, we have tried to iterate the CC3D module for different times during the experiment, and the specific experimental results are shown in Table VII. From the results, we can find that it is necessary to construct correlations between long-distance pixels in 3-D-space, as stated by our method, the classification with three iterations performs best. Moreover, we found that in two CC3D iterations, the accuracy improvement is most obvious. We analyze that this is because the 2-D spatial context information is very rich in agricultural hyperspectral scenarios, especially the learning of texture features in the crop boundary area will significantly improve the classification results.
4) For the setting of the Conv_1D kernel size used in the FC of the MSFF module, we conduct a large number of experiments on the WHU-Hi-HanChuan dataset. We set the size of the convolution kernel to [3, 4, 5, 6, 7], and compare the accuracy of the three indicators. The experimental results are shown in Table VIII. Combining the three indexes, we finally set the size of the convolution kernel to 4. By the way, the size setting of the convolution kernel will affect the spectral dimension of the final fusion, and the appropriate small convolution nuclear energy to ensure that the output spectral dimension is larger. Also, the smaller convolution kernel has a stronger nonlinear expression ability, so we do not try a larger convolution kernel.
5) To fully understand the rich spatial spectral semantic information in the HSIs, we have tried different numbers of encoders structure in MDAF, and the specific experimental results are shown in Table IX. Considering the three experimental metrics and the time overhead

TABLE IX

ABLATION RESULTS OF MDAF ON WHU-HI-HANCHUAN DATASET. THE OPTIMAL RESULTS ARE INDICATED IN BOLD

Metric	The number of encoders				
	1	2	3	4	5
OA (%)	94.71	95.29	**96.49**	95.01	96.39
AA (%)	87.95	89.86	91.56	88.23	**92.13**
κ*100	93.79	94.48	**95.88**	94.16	95.77
training(s)	293.37	326.29	344.52	392.56	536.70
test(s)	84.23	75.87	93.68	124.88	149.57

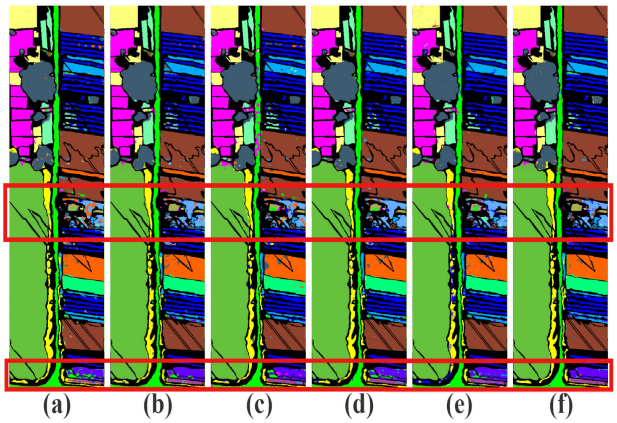


Fig. 11. Results visualization of adding CC3D to SSFTT, LSFAT, and MATNet on WHU-Hi-HanChuan dataset. (a) SSFTT. (b) SSFTT + CC3D. (c) LSFAT. (d) LSFAT + CC3D. (e) MATNet. (f) MATNet + CC3D.

Authorized licensed use limited to: Jaypee Insituite of Information Technology-Noida Sec 128 (L3). Downloaded on January 15,2025 at 10:36:02 UTC from IEEE Xplore. Restrictions apply.

TABLE X

ABLATION RESULTS OF CC3D BLOCK ON THREE DATASETS. THE OPTIMAL RESULTS ARE INDICATED IN BOLD

Metric	WHU-Hi-HongHu		WHU-Hi-HanChuan		WHU-Hi-LongKou	
	No_CC3D	Ours	No_CC3D	Ours	No_CC3D	Ours
OA (%)	96.35	**97.12**	92.96	**96.49**	99.73	**99.78**
AA (%)	90.86	**92.80**	81.79	**91.56**	98.98	**99.23**
κ*100	95.39	**96.35**	91.74	**95.88**	99.65	**99.71**

TABLE XI

IMPACT OF ADDING CC3D TO OTHER NETWORKS ON THE WHU-HI-HANCHUAN DATASET. THE OPTIMAL RESULTS ARE INDICATED IN BOLD

Metric	SSFTT	SSFTT+	LSFAT	LSFAT+	MATNet	MATNet+
OA (%)	95.16	**95.92**	95.00	**96.66**	95.50	**96.23**
AA (%)	87.94	**92.05**	89.00	**91.85**	89.28	**92.70**
κ*100	94.34	**95.23**	94.13	**96.09**	94.72	**95.58**

TABLE XII

RESULTS OF DIFFERENT ATTENTION ON THE WHU-HI-HANCHUAN DATASET. THE OPTIMAL RESULTS ARE INDICATED IN BOLD

Metric	NL_3D	SE	scSE	ECA	CC3D
OA (%)	94.54	94.29	94.75	92.76	**96.49**
AA (%)	85.76	86.17	88.00	82.23	**91.56**
κ*100	93.60	93.31	93.85	91.53	**95.88**

TABLE XIII

ABLATION RESULTS OF SSA ON THREE DATASETS. THE OPTIMAL RESULTS ARE INDICATED IN BOLD

Metric	WHU-Hi-HongHu		WHU-Hi-HanChuan		WHU-Hi-LongKou	
	No_SSA	Ours	No_SSA	Ours	No_SSA	Ours
OA (%)	96.26	**97.12**	94.56	**96.49**	99.71	**99.78**
AA (%)	90.35	**92.80**	87.66	**91.56**	98.89	**99.23**
κ*100	95.26	**96.35**	93.63	**95.88**	99.62	**99.71**

TABLE XIV

ABLATION RESULTS OF SPATIAL–SPECTRAL SEMANTIC EXTRACTION STRUCTURE ON THREE DATASETS. THE OPTIMAL RESULTS ARE INDICATED IN BOLD

Metric	WHU-Hi-HongHu			WHU-Hi-HanChuan			WHU-Hi-LongKou		
	Spa	Spe	Ours	Spa	Spe	Ours	Spa	Spe	Ours
OA (%)	96.09	95.65	**97.12**	94.39	94.65	**96.49**	99.77	99.72	**99.78**
AA (%)	87.39	86.60	**92.80**	87.37	86.60	**91.56**	99.20	99.03	**99.23**
κ*100	95.05	94.48	**96.35**	93.42	94.48	**95.88**	99.70	99.63	**99.71**

TABLE XV

COMPARISON OF SANET USING CE LOSS ON THE WHU-HI-HANCHUAN DATASET. THE OPTIMAL RESULTS ARE INDICATED IN BOLD

Metric	SSFTT(CE)	LSFAT(CE)	MATNet(Lpoly)	SANet(CE)
OA (%)	94.02	94.62	95.36	**95.66**
AA (%)	89.95	91.32	90.70	**93.06**
κ*100	93.00	93.69	94.57	**95.09**

of sample training, we set the number of layers of the encoder to 3.

C. Ablation Study

We perform sufficient ablation experiments on the proposed method to support our view.

1) CC3D Block Module: To demonstrate the effectiveness of our proposed CC3D module, we removed the CC3D block in experiments and performed experimental comparisons on three datasets with the results shown in Table X.

In addition, we further validate the effectiveness of the spatial–spectral nonlocal self-attention mechanism in the CC3D module in SSFTT [55], LSFAT [20], and MATNet [21]. Specifically, after adding the CC3D module to the first 3-D convolutional layer in the network, there is no impact on the original network structure due to the same output size of the CC3D module input. We retrained the network for 100 epochs on the WHU-Hi-HanChuan dataset and obtained the experimental results as shown in Table XI. The + represents adding the CC3D module to the original network.

Finally, to better compare our proposed CC3D module with other popular attention mechanisms, the attention we used during the spatial–spectral feature extraction phase is replaced with Nonlocal [42], SELayer [36], scSE [37] and ECALayer [38] separately. And the other network structures remain consistent. The experimental results of these attention mechanisms on the WHU-Hi-HanChuan dataset are shown in Table XII.

2) SSA Module: In the MSFF module, we propose a spectral vector-based SSA module to model the connection between the spectral vectors. To verify the effectiveness of the SSA module, we only retained the multiscale fusion structure of the spectrum while removing the SSA module, and the experimental results are shown in Table XIII.

3) spatial–spectral Semantic Feature Extraction: To verify the effectiveness of AF between the spatial semantic features and the spectral semantic features, we use only the spatial semantic feature extraction branch (Spa) and the spectral semantic feature extraction branch (Spe), respectively. The classification accuracy is shown in Table XIV, and experimentally proved that the spatial–spectral semantic features have a large impact on the classification results. The network structure we use can achieve the ideal spatial–spectral semantic feature extraction results, and the results perform optimally. By the way, we find that in the process of extracting spatial features and spectral features on HSI, spatial features often have a greater impact on classification accuracy than spectral features. And extracting only spatial semantic features will be more better effective than extracting only spectral semantic features.

4) Loss Function: To more fairly compare the performance of the SANet itself, we compare the three most competitive network models with our proposed SANet under the CE loss function, and the experimental results are shown in Table XV. The results show that our proposed method still performs best, and we finally chose to use Lpoly loss as the loss function of SANet because the spatial–spectral information between similar categories in agricultural hyperspectral scenarios, so a more robust loss function is needed to guide network learning. Lpoly loss considers not only the correct category but also the error category, while further adjusting the learning of different classification results, such a method is more suitable for the agricultural HSI classification task.

D. Classification Results and Visualization

1) Classification Results: As presented in Tables XVI–XVIII, our proposed SANet demonstrates the best classification performance for both conventional and

Authorized licensed use limited to: Jaypee Insituite of Information Technology-Noida Sec 128 (L3). Downloaded on January 15,2025 at 10:36:02 UTC from IEEE Xplore. Restrictions apply.

TABLE XVI

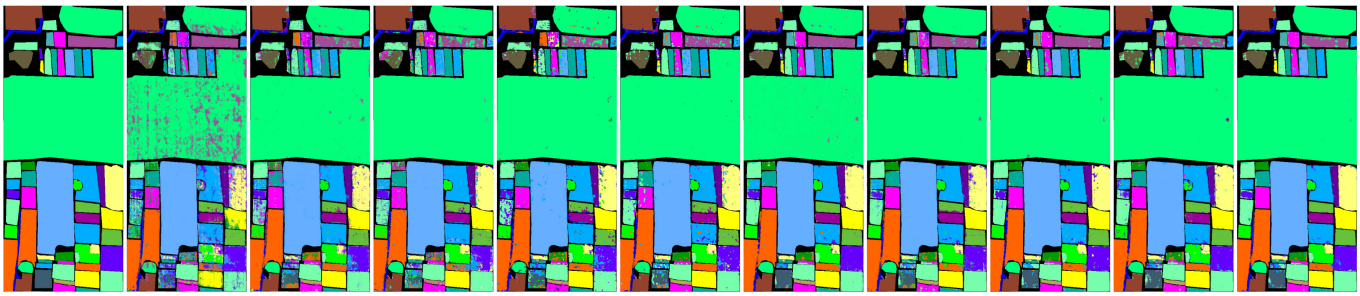
CLASSIFICATION RESULTS OF THE WHU-HI-HONGHU DATASET ON DIFFERENT NETWORKS. THE OPTIMAL RESULTS ARE INDICATED IN BOLD

Class No.	ResNet	PyResNet	ContextualNet	SSRN	A2S2KResNet	SSFTT	SpectralFormer	LSFAT	MATNet	SANet
1	96.86±2.74	96.39±3.15	96.42±3.33	98.31±0.97	**98.48±1.03**	96.93±0.92	97.48±0.45	97.99±1.10	97.37±0.49	98.14±0.38
2	73.79±9.82	75.58±9.19	76.49±8.03	84.56±4.57	**86.93±4.92**	82.22±5.42	76.71±7.61	86.11±5.18	83.78±5.38	85.50±3.31
3	85.68±6.50	82.35±4.57	87.80±2.97	91.33±2.88	93.36±1.87	96.40±0.59	96.60±1.17	96.61±0.88	95.95±1.39	**97.35±0.89**
4	96.41±1.75	96.61±3.65	97.96±0.78	98.51±0.69	98.74±0.43	99.67±0.31	99.65±0.30	99.70±0.14	99.69±0.24	**99.89±0.11**
5	77.14±21.46	78.36±10.39	88.20±4.01	87.95±7.03	89.83±5.82	91.30±2.05	91.49±2.07	**92.25±2.40**	92.06±4.91	84.39±3.87
6	95.55±3.89	96.83±2.85	96.67±1.12	97.20±1.09	97.51±1.56	99.23±0.34	98.43±0.61	**99.31±0.38**	99.20±0.35	99.29±0.52
7	78.76±11.59	84.65±9.03	84.51±4.11	89.81±3.76	92.93±2.20	94.84±0.72	95.20±0.89	95.11±1.45	94.83±1.93	**97.59±0.77**
8	57.74±18.36	68.01±17.31	56.62±12.30	67.82±10.66	82.54±8.29	83.54±4.79	69.65±9.13	**86.23±1.86**	86.08±4.53	85.57±5.37
9	97.40±2.75	97.16±5.19	96.71±2.42	98.84±0.63	98.88±0.45	97.76±0.93	98.07±0.85	**98.93±0.90**	98.42±0.62	98.85±0.60
10	85.93±16.86	86.51±10.26	87.01±5.39	92.37±4.98	90.61±5.63	93.92±1.34	89.13±2.68	91.38±4.29	93.35±1.93	**95.51±1.42**
11	77.08±11.90	79.68±11.87	82.77±7.65	82.67±5.79	90.11±4.02	91.61±1.81	87.55±5.04	89.06±4.14	91.57±4.79	**93.66±1.35**
12	61.12±13.32	66.96±10.74	78.68±5.58	78.42±9.42	80.68±7.77	**90.04±2.62**	82.20±4.21	88.77±3.21	87.36±4.35	89.11±4.08
13	75.46±12.00	76.30±9.87	80.46±5.16	88.34±4.36	88.01±4.09	91.18±2.15	89.17±2.29	91.29±1.91	92.79±2.03	**93.47±1.68**
14	87.79±10.36	88.06±5.71	88.89±5.84	90.05±5.82	92.80±3.30	93.09±1.75	91.60±3.09	93.61±4.03	91.90±4.31	**94.19±1.53**
15	92.48±17.76	94.23±6.32	84.28±14.90	93.23±12.87	**96.04±1.82**	83.80±6.61	85.57±3.49	87.65±4.88	85.70±9.36	80.65±5.22
16	96.38±2.57	95.84±2.77	93.46±5.02	95.47±8.30	98.16±0.88	97.75±1.64	95.73±2.21	97.67±1.01	97.99±1.54	**98.83±1.05**
17	79.94±7.98	70.77±6.36	80.11±8.35	80.82±14.82	89.23±5.78	97.48±1.36	90.97±4.94	93.50±5.95	96.28±1.81	**97.89±1.58**
18	83.76±14.20	85.45±10.62	84.06±10.33	84.41±6.69	92.11±6.32	87.20±5.79	89.55±9.80	92.70±4.47	89.26±5.37	**95.20±2.42**
19	84.97±9.57	86.75±6.79	88.33±6.92	92.19±4.91	91.41±4.81	92.66±2.56	91.75±3.11	91.79±4.17	93.02±1.31	**93.11±1.24**
20	57.27±17.64	71.71±14.18	75.31±10.49	89.51±5.87	**91.69±4.16**	85.57±3.88	83.57±8.88	88.58±4.09	86.44±3.95	89.63±3.72
21	27.04±13.92	64.97±20.52	31.06±8.18	24.82±19.70	79.50±11.04	**93.09±3.38**	63.54±15.19	82.93±8.65	85.19±9.57	86.69±6.94
22	86.90±12.19	89.80±4.06	85.23±3.90	86.62±6.97	91.42±4.06	91.48±3.17	92.18±3.77	94.45±4.82	92.74±4.07	**95.23±3.18**
OA (%)	87.51±2.74	89.62±2.51	92.05±0.44	93.98±0.70	95.25±0.74	96.67±0.23	95.65±0.82	96.73±0.50	96.75±0.62	**97.40±0.32**
AA (%)	79.79±3.22	83.32±2.39	82.77±1.49	86.06±2.73	91.41±1.33	92.30±0.63	88.90±2.54	92.53±1.39	92.32±1.44	**93.17±1.23**
κ*100	84.17±3.33	86.79±3.31	89.91±0.58	92.37±0.89	93.99±0.93	95.79±0.29	94.50±1.04	95.87±0.63	95.89±0.78	**96.71±0.40**
Avg Training Time (s)	11644.06	12784.93	9316.36	6486.96	2346.57	**877.35**	1466.95	1963.22	2070.06	3699.02
Avg Testing Time (s)	5209.34	5157.84	2004.81	1041.69	808.66	**246.58**	563.46	531.07	768.09	1266.90
Trainable Params	21380630	21865026	554128	364168	373184	288136	616929	**278276**	1708400	1086370

TABLE XVII

CLASSIFICATION RESULTS OF THE WHU-HI-HANCHUAN DATASET ON DIFFERENT NETWORKS. THE OPTIMAL RESULTS ARE INDICATED IN BOLD

Class No.	ResNet	PyResNet	ContextualNet	SSRN	A2S2KResNet	SSFTT	SpectralFormer	LSFAT	MATNet	SANet
1	93.09±2.62	89.44±11.08	93.51±1.66	93.70±2.49	94.82±1.32	93.66±1.65	96.47±1.26	95.97±1.43	**96.85±0.54**	94.95±1.96
2	94.66±2.44	93.16±2.62	91.05±3.33	91.16±5.01	91.48±2.96	93.90±2.54	95.25±1.92	93.88±3.01	94.74±2.90	**95.46±1.37**
3	84.31±8.58	83.72±10.25	85.12±4.44	88.37±6.33	89.21±4.80	93.33±2.53	93.00±2.59	90.31±7.26	**95.86±2.42**	95.67±2.00
4	92.01±6.58	93.67±5.57	93.57±5.48	94.36±6.98	93.03±2.79	98.25±1.65	97.56±1.10	96.24±2.63	97.92±1.10	**98.53±0.62**
5	80.94±12.03	83.28±9.73	62.56±18.94	41.08±36.02	49.67±12.06	81.58±9.26	86.80±5.78	90.47±6.85	76.82±16.78	**88.48±8.35**
6	77.29±7.88	67.41±12.86	61.92±9.39	64.29±11.57	67.17±6.39	78.47±6.72	84.28±3.54	84.27±6.74	80.11±9.08	**87.67±2.75**
7	78.61±10.49	74.65±10.95	81.82±9.75	82.02±9.24	67.40±4.82	84.52±7.85	83.19±4.97	**90.83±5.37**	88.62±9.05	89.45±5.11
8	85.79±6.13	88.21±7.96	84.16±3.96	85.99±4.31	85.80±3.44	90.48±2.99	93.30±2.91	92.19±5.46	93.27±3.75	**94.54±1.09**
9	77.49±5.48	78.87±6.74	77.68±8.08	82.80±5.57	78.11±4.68	84.96±6.83	87.41±4.14	88.79±2.94	88.75±7.40	**89.92±2.88**
10	97.91±2.09	92.76±12.83	96.69±1.90	98.23±1.21	94.14±3.19	98.04±0.98	97.59±1.19	97.23±3.00	**98.62±0.87**	97.79±2.00
11	85.50±7.26	81.53±11.11	92.23±4.24	93.25±2.42	90.83±2.68	94.19±2.14	94.37±1.89	**96.54±1.60**	93.99±3.04	95.51±1.54
12	69.63±16.63	75.41±6.90	59.89±8.07	43.83±23.68	71.01±7.50	81.87±5.11	85.68±4.90	84.48±8.94	85.14±5.75	**86.89±8.09**
13	73.83±10.05	77.38±5.30	72.15±12.51	74.63±11.45	73.85±5.44	84.06±4.45	81.84±5.76	78.97±8.79	86.61±4.83	**89.07±4.15**
14	84.63±5.82	84.30±6.62	87.17±1.93	87.44±5.96	89.22±3.03	93.55±2.73	92.71±1.21	93.02±2.53	**95.55±3.50**	94.65±1.75
15	83.02±10.75	78.76±17.11	79.37±11.56	67.95±39.47	65.47±11.00	88.63±14.55	88.45±8.66	88.35±8.00	78.84±17.47	**93.41±6.78**
16	96.75±4.08	98.51±1.13	99.34±0.50	**99.86±0.10**	99.24±3.3	99.77±0.07	99.58±0.22	99.61±0.20	99.51±0.57	99.76±0.13
OA (%)	89.88± 1.35	88.70± 3.74	90.69±1.02	91.58±1.38	91.04±0.47	94.02±0.57	94.91±0.59	94.62±0.83	95.36±1.34	**95.81±0.66**
AA (%)	84.72± 1.49	83.82± 2.88	82.39±2.10	80.56±5.70	81.28±1.51	89.95±1.20	91.09±1.45	91.32±1.24	90.70±3.34	**93.23±1.23**
κ*100	88.11± 1.64	86.72±4.47	89.09±1.19	90.13±1.63	89.50±0.56	93.00±0.66	94.04±0.70	93.69±0.98	94.57±1.57	**95.66±0.78**
Avg Training Time (s)	6265.33	2635.91	2196.63	2325.89	1828.56	**500.51**	821.28	1128.59	1607.19	2436.61
Avg Testing Time (s)	3247.94	1125.42	1328.85	672.70	535.90	**236.21**	587.53	570.98	909.33	1492.72



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k)

Fig. 12. Visualization results of different DL-based Networks on WHU-Hi-HongHu dataset. (a) Ground truth. (b) Resnet [53]. (c) PyResnet [5]. (d) ContextualNet [54]. (e) SSRN [6]. (f) A2S2KResNet [7]. (g) SSFTT [55]. (h) SFormer [56]. (i) LSFAT [20]. (j) MATNet [21]. (k) Proposed SANet.

Authorized licensed use limited to: Jaypee Insituite of Information Technology-Noida Sec 128 (L3). Downloaded on January 15,2025 at 10:36:02 UTC from IEEE Xplore. Restrictions apply.

TABLE XVIII

CLASSIFICATION RESULTS OF THE WHU-HI-LONGKOU DATASET ON DIFFERENT NETWORKS. THE OPTIMAL RESULTS ARE INDICATED IN BOLD

Class No.	ResNet	PyResNet	ContextualNet	SSRN	A2S2KResNet	SSFTT	SpectralFormer	LSFAT	MATNet	SANet
1	99.46±0.33	99.82±0.27	99.68±0.46	99.86±0.10	99.87±0.09	99.68±0.30	99.80±0.16	99.78±0.23	99.85±0.09	**99.91±0.04**
2	96.09±7.06	98.02±1.64	99.21±0.36	98.07±2.36	**99.55±0.41**	98.38±0.64	97.68±2.29	98.29±1.75	99.01±0.79	99.49±0.63
3	96.97±6.97	97.99±3.63	99.20±0.89	95.87±9.25	**99.39±1.02**	94.86±2.76	98.98±1.13	99.02±0.98	99.21±0.98	98.93±1.40
4	98.96±0.60	98.98±0.48	99.20±0.22	99.57±0.34	99.71±0.23	98.71±0.43	99.47±0.22	99.24±1.02	99.48±0.36	**99.73±0.11**
5	96.21±4.77	95.43±4.11	98.98±0.70	97.82±2.09	98.82±0.63	95.21±1.49	93.94±4.41	97.47±2.34	97.87±1.97	**99.14±0.73**
6	99.28±0.65	99.23±0.46	99.10±0.67	99.70±0.26	99.74±0.28	98.56±1.11	98.80±1.64	98.02±4.19	99.41±0.51	**99.77±0.12**
7	99.74±0.33	99.90±0.06	99.86±0.12	99.96±0.04	**99.97±0.03**	99.75±0.22	99.92±0.04	99.90±0.08	99.88±0.04	99.95±0.02
8	97.05±1.44	96.27±3.00	96.64±2.65	96.26±2.05	**97.51±1.28**	91.29±2.73	96.72±1.29	96.16±1.27	93.57±3.24	97.20±1.41
9	90.00±0.19	95.02±4.94	95.42±1.36	96.44±1.24	94.19±6.97	88.26±4.26	**97.64±1.37**	95.45±3.46	94.74±3.64	97.42±1.70
OA (%)	98.46± 1.66	99.09± 0.27	99.30± 0.15	99.40± 0.39	99.55± 0.26	98.52± 0.18	99.29± 0.38	99.18± 0.81	99.27± 0.17	**99.65± 0.05**
AA (%)	97.08± 2.37	97.85± 1.18	98.59± 0.34	98.17± 1.33	98.75± 0.88	96.08± 0.58	98.11± 1.09	98.15± 1.37	98.11± 0.56	**99.06± 0.18**
κ*100	97.99± 2.16	98.81± 0.35	99.07± 0.20	99.18± 0.52	99.41± 0.34	98.06± 0.24	99.06± 0.49	98.92± 1.07	99.04± 0.22	**99.54± 0.07**
Avg Training Time	6632.91	6309.97	7080.64	2713.32	1984.77	**1162.10**	1538.85	1565.77	1520.33	4996.23
Avg Testing Time (s)	2507.70	3597.92	1890.48	515.60	423.54	**154.09**	431.57	433.38	736.03	1210.45

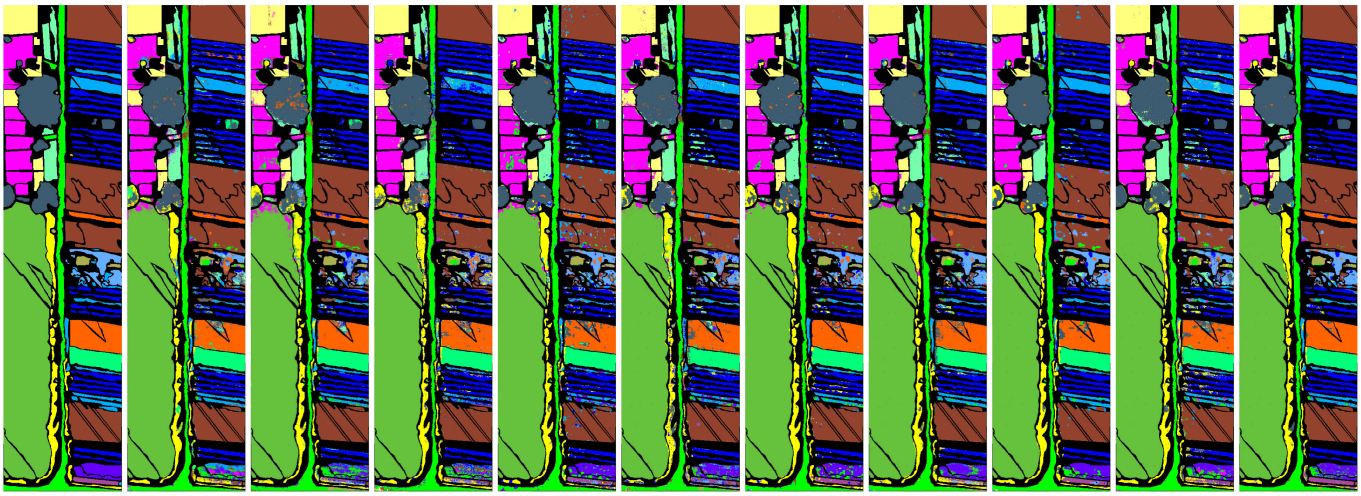


Fig. 13. Visualization results of different DL-based Networks on WHU-Hi-HanChuan dataset. (a) Ground truth. (b) Resnet [53]. (c) PyResnet [5]. (d) ContextualNet [54]. (e) SSRN [6]. (f) A2S2KResNet [7]. (g) SSFTT [55]. (h) SFormer [56]. (i) LSFAT [20]. (j) MATNet [21]. (k) Proposed SANet.

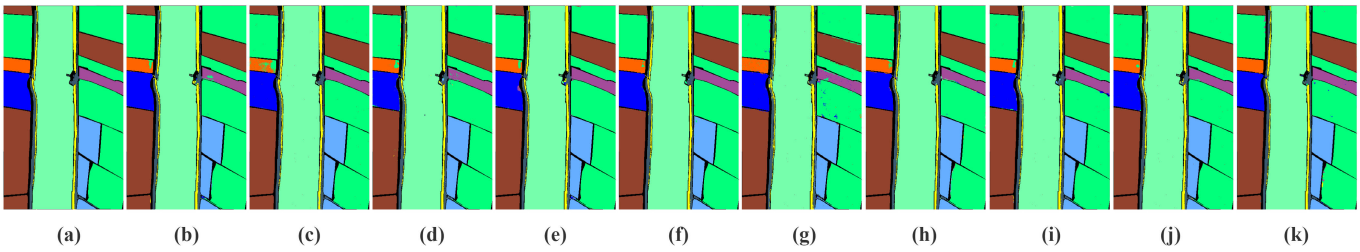


Fig. 14. Visualization results of different DL-based Networks on WHU-Hi-LongKou dataset. (a) Ground truth. (b) Resnet [53]. (c) PyResnet [5]. (d) ContextualNet [54]. (e) SSRN [6]. (f) A2S2KResNet [7]. (g) SSFTT [55]. (h) SFormer [56]. (i) LSFAT [20]. (j) MATNet [21]. (k) Proposed SANet.

popular DL methods. The results in Tables XVI–XVIII show the mean and standard deviation. The optimal results of OA, AA, and κ also prove that our method also demonstrates good robustness in the face of significant differences in the number of samples. And we find from the results that our proposed method is the best classification result on many categories. The standard deviation of our experimental results is smaller compared to the other networks, which also proves that our method performs very steadily in different categories. Incidentally, we, although not optimal in some categories, are very close to optimal. Of course, our method requires more training time than other state-of-the-art methods, but we consider it worthwhile to obtain a stable classification performance with an acceptable time cost. Finally, We list the number of trainable parameters for all the networks in Table XVI.

2) Visual Results: To more intuitively observe the effects of attention mechanisms in the CC3D module, we visually present the results in Table XI, as shown in Fig. 11 Through careful observation, we can find that adding CC3D module to the network can significantly improve the classification effect of areas with dense ground distribution.

We present the visualization results of all the network models in Figs. 12–14. At the same time, we can carefully observe that the classification results of our proposed method perform well.

Authorized licensed use limited to: Jaypee Insituite of Information Technology-Noida Sec 128 (L3). Downloaded on January 15,2025 at 10:36:02 UTC from IEEE Xplore. Restrictions apply.

V. Conclusion

In this article, we discuss the challenges faced in agricultural HSI classification, such as the dense spatial distribution of land cover categories and spatial–spectral feature mixing at crop boundaries. To tackle these challenges, we propose a network called SANet that integrates spectral data and contextual information while emphasizing self-correlation within the HSI. The SANet combines spatial–spectral nonlocal block structure and multiscale SSA structure to allocate attention resources to spatial and spectral dimensions and model correlations within the spectral–spatial space. Finally, we use the same structure to simultaneously identify and fuse spatial–spectral semantic features to improve the robustness of SANet to the understanding of semantic information at different levels. Rich experimental results demonstrate the stable performance of our proposed method in handling densely distributed hyperspectral scene classification tasks.

Currently, there is extensive research on HSI classification in the field of crop classification, covering almost all types of crops. Concurrently, technologies such as machine learning, DL, and artificial intelligence have been widely applied in HSI processing. Moreover, we are actively exploring the integration of multimodel data to enhance the automation and accuracy of crop classification in agriculture, catering to the requirements of large-scale and efficient agricultural production.

References

[1] C. Sammut and G. I. Webb, *Principal Component Analysis*. Boston, MA, USA: Springer, 2010, p. 795, doi: 10.1007/978-0-387-30164-8_665.

[2] J. Xia, N. Falco, J. A. Benediktsson, P. Du, and J. Chanussot, "Hyperspectral image classification with rotation random forest via KPCA," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 4, pp. 1601–1609, Apr. 2017.

[3] Y. Qian, Y. Shen, M. Ye, and Q. Wang, "3-D nonlocal means filter with noise estimation for hyperspectral imagery denoising," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2012, pp. 1345–1348.

[4] A. Mahmood, A. Robin, and M. Sears, "Modified residual method for the estimation of noise in hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 3, pp. 1451–1460, Mar. 2017.

[5] M. E. Paoletti, J. M. Haut, R. Fernandez-Beltran, J. Plaza, A. J. Plaza, and F. Pla, "Deep pyramidal residual networks for spectral–spatial hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 740–754, Feb. 2019.

[6] Z. Zhong, J. Li, Z. Luo, and M. Chapman, "Spectral–spatial residual network for hyperspectral image classification: A 3-D deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 847–858, Feb. 2018.

[7] S. K. Roy, S. Manna, T. Song, and L. Bruzzone, "Attention-based adaptive spectral–spatial kernel ResNet for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7831–7843, Sep. 2021.

[8] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.

[9] W. Luo et al., "A new algorithm for bilinear spectral unmixing of hyperspectral images using particle swarm optimization," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 12, pp. 5776–5790, Dec. 2016.

[10] S. Khazai, A. Safari, B. Mojaradi, and S. Homayouni, "Improving the SVDD approach to hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 4, pp. 594–598, Jul. 2012.

[11] H. Yu, L. Gao, J. Li, S. Li, B. Zhang, and J. Benediktsson, "Spectral–spatial hyperspectral image classification using subspace-based support vector machines and adaptive Markov random fields," *Remote Sens.*, vol. 8, no. 4, p. 355, Apr. 2016. [Online]. Available: https://www.mdpi.com/2072-4292/8/4/355

[12] H. Yu, L. Gao, W. Li, Q. Du, and B. Zhang, "Locality sensitive discriminant analysis for group sparse representation-based hyperspectral imagery classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 8, pp. 1358–1362, Aug. 2017.

[13] Y. Chen, X. Lu, and X. Li, "Supervised deep hashing with a joint deep network," *Pattern Recognit.*, vol. 105, Sep. 2020, Art. no. 107368, doi: 10.1016/j.patcog.2020.107368.

[14] Y. Chen, X. Lu, and S. Wang, "Deep cross-modal image–voice retrieval in remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7049–7061, Oct. 2020, doi: 10.1109/TGRS.2020.2979273.

[15] Y. Chen and X. Lu, "Deep category-level and regularized hashing with global semantic similarity learning," *IEEE Trans. Cybern.*, vol. 51, no. 12, pp. 6240–6252, Dec. 2021, doi: 10.1109/TCYB.2020.2964993.

[16] Y. Chen, S. Xiong, L. Mou, and X. X. Zhu, "DEEP QUADRUPLE-BASED HASHING FOR REMOTE SENSING IMAGE-SOUND RETRIEVAL," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022, doi: 10.1109/TGRS.2022.3155283.

[17] H. Guo, J. Liu, J. Yang, Z. Xiao, and Z. Wu, "Deep collaborative attention network for hyperspectral image classification by combining 2-D CNN and 3-D CNN," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4789–4802, 2020.

[18] Z. Xu, H. Yu, K. Zheng, L. Gao, and M. Song, "A novel classification framework for hyperspectral image classification based on multiscale spectral–spatial convolutional network," in *Proc. 11th Workshop Hyperspectral Imag. Signal Process., Evol. Remote Sens. (WHISPERS)*, Mar. 2021, pp. 1–5.

[19] C. Yu, R. Han, M. Song, C. Liu, and C.-I. Chang, "Feedback attention-based dense CNN for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022, doi: 10.1109/TGRS.2021.3058549.

[20] B. Tu, X. Liao, Q. Li, Y. Peng, and A. Plaza, "Local semantic feature aggregation-based transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022, doi: 10.1109/TGRS.2022.3201145.

[21] B. Zhang, Y. Chen, Y. Rong, S. Xiong, and X. Lu, "MATNet: A combining multi-attention and transformer network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023, doi: 10.1109/TGRS.2023.3254523.

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

[23] T. Guo, R. Wang, F. Luo, X. Gong, L. Zhang, and X. Gao, "Dual-view spectral and global spatial feature fusion network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–13, 2023, doi: 10.1109/TGRS.2023.3277467.

[24] F. Luo, T. Zhou, J. Liu, T. Guo, X. Gong, and J. Ren, "Multiscale diff-changed feature fusion network for hyperspectral image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–13, 2023, doi: 10.1109/TGRS.2023.3241097.

[25] M. Ma, S. Mei, F. Li, Y. Ge, and Q. Du, "Spectral correlation-based diverse band selection for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–13, 2023, doi: 10.1109/TGRS.2023.3263580.

[26] H. Yang, H. Yu, D. Hong, Z. Xu, Y. Wang, and M. Song, "Hyperspectral image classification based on multi-level spectral–spatial transformer network," in *Proc. 12th Workshop Hyperspectral Imag. Signal Process., Evol. Remote Sens. (WHISPERS)*, Sep. 2022, pp. 1–4.

[27] Q. Zhou, S. Zhou, F. Shen, J. Yin, and D. Xu, "Hyperspectral image classification based on 3-D multihead self-attention spectral–spatial feature fusion network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 1072–1084, 2023, doi: 10.1109/JSTARS.2022.3226758.

[28] H. Liu, W. Li, X.-G. Xia, M. Zhang, C.-Z. Gao, and R. Tao, "Central attention network for hyperspectral imagery classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 11, pp. 8989–9003, 2023, doi: 10.1109/TNNLS.2022.3155114.

[29] D. Wang, B. Du, L. Zhang, and D. Tao, "HKNAS: Classification of hyperspectral imagery based on hyper kernel neural architecture search," *CoRR*, vol. abs/2304.11701, 2023, doi: 10.48550/ARXIV.2304.11701.

[30] M. Zhang, W. Li, Y. Zhang, R. Tao, and Q. Du, "Hyperspectral and LiDAR data classification based on structural optimization transmission," *IEEE Trans. Cybern.*, vol. 53, no. 5, pp. 3153–3164, May 2023.

[31] Y. Zhang, W. Li, M. Zhang, S. Wang, R. Tao, and Q. Du, "Graph information aggregation cross-domain few-shot learning for hyperspectral image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–14, 2022, doi: 10.1109/TNNLS.2022.3185795.

[32] Y. Zhang, M. Zhang, W. Li, S. Wang, and R. Tao, "Language-aware domain generalization network for cross-scene hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–12, 2023, doi: 10.1109/TGRS.2022.3233885.

[33] Y. Zhang, W. Li, W. Sun, R. Tao, and Q. Du, "Single-source domain expansion network for cross-scene hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 32, pp. 1498–1512, 2023.

Authorized licensed use limited to: Jaypee Insituite of Information Technology-Noida Sec 128 (L3). Downloaded on January 15,2025 at 10:36:02 UTC from IEEE Xplore. Restrictions apply.

[34] Y. Zhang, W. Li, M. Zhang, Y. Qu, R. Tao, and H. Qi, "Topological structure and semantic information transfer network for cross-scene hyperspectral image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 6, pp. 2817–2830, Jun. 2023.

[35] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, 2018, pp. 3–19.

[36] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[37] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks," in *Proc. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds. Cham, Switzerland: Springer, 2018, pp. 421–429.

[38] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.

[39] L. Wang, Y. Lin, J. Liu, Z. Li, and C. Wu, "Siamese spectral attention with channel consistency for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10226–10241, 2021.

[40] Y. Shi, J. Li, Y. Zheng, B. Xi, and Y. Li, "Hyperspectral target detection with RoI feature transformation and multiscale spectral attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5071–5084, Jun. 2021.

[41] X. Zhang et al., "Spectral–spatial self-attention networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022, doi: 10.1109/TGRS.2021.3102143.

[42] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[43] C. Wang, X. Bai, L. Zhou, and J. Zhou, "Hyperspectral image classification based on non-local neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2019, pp. 584–587.

[44] L. Mou, X. Lu, X. Li, and X. X. Zhu, "Nonlocal graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8246–8257, Dec. 2020.

[45] Y. Shen et al., "Efficient deep learning of nonlocal features for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6029–6043, Jul. 2021.

[46] Z. Dong, Y. Cai, Z. Cai, X. Liu, Z. Yang, and M. Zhuge, "Cooperative spectral–spatial attention dense network for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 866–870, May 2021.

[47] Z. Shu, Z. Liu, J. Zhou, S. Tang, Z. Yu, and X.-J. Wu, "Spatial–spectral split attention residual network for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 419–430, 2023.

[48] L. Liang, S. Zhang, J. Li, A. Plaza, and Z. Cui, "Multi-scale spectral–spatial attention network for hyperspectral image classification combining 2D octave and 3D convolutional neural networks," *Remote Sens.*, vol. 15, no. 7, p. 1758, Mar. 2023. [Online]. Available: https://www.mdpi.com/2072-4292/15/7/1758

[49] C. Zhao et al., "Hyperspectral image classification with multi-attention transformer and adaptive superpixel segmentation-based active learning," *IEEE Trans. Image Process.*, vol. 32, pp. 3606–3621, 2023.

[50] Z. Huang et al., "CCNet: Criss-cross attention for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 6896–6908, Jun. 2023.

[51] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, I. Guyon et al., Eds. Red Hook, NY, USA: Curran Associates, 2017, pp. 1–11. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[52] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, May 2015, pp. 1–15.

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[54] Z. Wang, Q. She, P. Zhang, and J. Zhang, "ContextNet: A click-through rate prediction framework using contextual information to refine feature embedding," 2021, *arXiv:2107.12025*.

[55] L. Sun, G. Zhao, Y. Zheng, and Z. Wu, "Spectral–spatial feature tokenization transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022, doi: 10.1109/TGRS.2022.3144158.

[56] D. Hong et al., "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022, doi: 10.1109/TGRS.2021.3130716.

Bo Zhang received the B.S. degree in engineering from Jiangxi Normal University, Nanchang, China, in 2021. He is currently pursuing the master's degree with the School of Computer and Artificial Intelligence, Wuhan University of Technology, Wuhan, China.

His research interests include pattern recognition and image classification in remote sensing scenarios.

Yaxiong Chen is currently an Associate Professor with the School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan, China.

His research interests include pattern recognition, machine learning, hyperspectral image analysis, and medical imaging.

Zhiheng Li received the B.Sc. degree in computer science and technology from the Jiangsu University of Science and Technology, Zhenjiang, China, in 2021. He is currently pursuing the master's degree with the School of Computer and Artificial Intelligence, Wuhan University of Technology, Wuhan, China.

His research interests include pattern recognition and image segmentation in remote sensing scenarios.

Shengwu Xiong received the B.Sc. degree in computational mathematics and the M.Sc. and Ph.D. degrees in computer software and theory from Wuhan University, Wuhan, China, in 1987, 1997, and 2003, respectively.

He is currently a Professor with the School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan. His research interests include intelligent computing, machine learning, and pattern recognition.

Xiaoqiang Lu (Senior Member, IEEE) is currently a Full Professor with the College of Physics and Information Engineering, Fuzhou University, Fuzhou, China.

His research interests include intelligent optical sensing, pattern recognition, machine learing, and hyperspectral image analysis.

Authorized licensed use limited to: Jaypee Insituite of Information Technology-Noida Sec 128 (L3). Downloaded on January 15,2025 at 10:36:02 UTC from IEEE Xplore. Restrictions apply.