remote sensing

MDPI

Article

Dual-Branch Adaptive Convolutional Transformer for Hyperspectral Image Classification

Chuanzhi Wang, Jun Huang, Mingyun Lv, Yongmei Wu and Ruiru Qin *

School of Aeronautic Science and Engineering, Beihang University, Beijing 100191, China; wangchuanzhi@buaa.edu.cn (C.W.); junh@china.com (J.H.); lv503@buaa.edu.cn (M.L.); amy5139@buaa.edu.cn (Y.W.)
* Correspondence: qinrr@buaa.edu.cn

Abstract: In hyperspectral image (HSI) classification, convolutional neural networks (CNNs) and transformer architectures have each contributed to considerable advancements. CNNs possess potent local feature representation skills, whereas transformers excel in learning global features, offering a complementary strength. Nevertheless, both architectures are limited by static receptive fields, which hinder their accuracy in delineating subtle boundary discrepancies. To mitigate the identified limitations, we introduce a novel dual-branch adaptive convolutional transformer (DBACT) network architecture featuring an adaptive multi-head self-attention mechanism. The architecture begins with a triadic parallel stem structure for shallow feature extraction and reduction of the spectral dimension. A global branch with adaptive receptive fields performs high-level global feature extraction. Simultaneously, a local branch with a cross-attention module provides detailed local insights, enriching the global perspective. This methodical integration synergizes the advantages of both branches, capturing representative spatial-spectral features from HSI. Comprehensive evaluation across three benchmark datasets reveals that the DBACT model exhibits superior classification performance compared to leading-edge models.

Keywords: hyperspectral image classification; adaptive multi-head self-attention; convolutional neural networks; transformers

Citation: Wang, C.; Huang, J.; Lv, M.; Wu, Y.; Qin, R. Dual-Branch Adaptive Convolutional Transformer for Hyperspectral Image Classification. *Remote Sens.* **2024**, *16*, 1615. https://doi.org/10.3390/rs16091615

Academic Editor: Salah Bourennane

Received: 15 March 2024
Revised: 26 April 2024
Accepted: 29 April 2024
Published: 30 April 2024

1. Introduction

Hyperspectral imaging captures both spatial and rich narrowband spectral information, playing a critical role in analyzing surface distribution, object detection, and natural resources across various fields [1–3]. Hyperspectral images (HSIs) are capable of gathering spectral data for every pixel within a given spatial region, with each pixel distinctly categorizing a land cover type [4]. Hyperspectral imagery possesses a higher spectral resolution, enabling it to offer more detailed features of terrestrial objects compared to panchromatic and multispectral images [5]. HSI classification represents a critical component within the field of hyperspectral image analysis. However, hyperspectral image classification remains challenging due to the redundancy in its spectral bands and the uneven distribution of samples.

Numerous methods have been proposed to enhance the accuracy of hyperspectral image classification. In the early phase, traditional techniques, such as support vector machines (SVM) [6], sparse representation [7], the k-nearest neighbor [8], and random forests (RF) [9], were predominantly used. Early efforts in hyperspectral classification were mainly directed towards extracting spectral information. Nevertheless, the neighboring pixels in hyperspectral imagery exhibit a high degree of correlation, suggesting that these pixels often represent the same category. Traditional approaches concentrated exclusively on spectral data, neglecting the spatial correlations present within the imagery. This oversight resulted in an incomplete exploitation of the available comprehensive hyperspectral features.

As deep learning technologies have swiftly advanced in domains like computer vision (CV) and natural language processing (NLP), their extensive application has equally proliferated into the area of hyperspectral classification [10]. Unlike traditional classification methods that rely on manually designed features, deep learning excels at identifying high-level semantic features, providing a superior capability for feature representation end to end. Due to their specialized localized receptive fields and robustness against distortion, convolutional neural networks have proven to be of considerable practical importance, especially in their capacity for local feature extraction. Characterized by a three-dimensional data structure, hyperspectral images derive significant advantages from the widespread use of 1D [11], 2D [12], and 3D [13] convolutional neural networks. 1D CNNs specialize in extracting spectral features and analyzing variations in data across one dimension. 2D CNNs are aimed at extracting spatial features and interpreting patterns and textures in two-dimensional data. 3D CNNs enable simultaneous spatial and spectral information processing at the cost of increased computational demand. Ahmad et al. [14] designed a compact 3D convolutional network that employs three-dimensional kernels across multiple adjacent spectral bands to reduce computational demand. Utilizing 3D convolution with variable kernel sizes and residual connections, Zhong et al. [15] efficiently extracted spectral and spatial features from HSI volumes. The combination of 3D CNNs and 2D CNNs in a sequential arrangement adeptly leverages the complementary strengths of both architectures [16,17]. Additionally, the integration of attention mechanisms with multi-branch convolutional network architectures efficiently extracts discriminative spatial and spectral features [18,19]. Nevertheless, the inherent limitations of localized receptive fields and invariant convolution kernel dimensions in CNN-based architectures hinder their ability to apprehend global features spanning long distances across both spatial dimensions and broad spectral bands.

Recently, the Vision Transformer (ViT) [20], along with its derivative models [21–23], has achieved remarkable advancements in the realm of image processing. By employing multi-head self-attention mechanisms, transformer-based models, such as ViT, have demonstrated proficiency in identifying long-range feature correlations. Furthermore, ViT has swiftly expanded its application scope, notably in hyperspectral classification [24–26], indicating its significant utility and adaptability in diverse imaging contexts. Hong et al. [27] successfully integrated the transformer encoder into hyperspectral classification without convolutional operations. However, pure transformer-based methods face challenges in effectively capturing local detail features within both spectral and spatial dimensions. Therefore, the naive idea is to combine the advantages of CNN structures with those of transformer structures. Sun et al. [28] observed a significant enhancement in classification performance through the pre-implementation of 3D and 2D CNN structures ahead of the transformer encoder. Qi et al. [4] combined 3D CNN and transformer structures to simultaneously extract global and local features of hyperspectral images. Through the incorporation of convolution operations in multi-head attention mechanisms, Zhang et al. [29] achieved a profound integration of CNN and transformer architectures, resulting in impressive classification performance. Yang et al. [30] developed a novel parallel multi-level feature fusion structure for integrating global and local features. Merging the benefits of CNN and transformer architectures results in improved classification performance, yet their inherent static receptive fields—from the local domain of convolution kernels to the global scope of self-attention mechanisms—pose challenges. Such fixed fields may constrain enhancements in performance, particularly with irregular and fragmented object contours, indicating opportunities for further advancements.

Through the analysis of the aforementioned work, we designed a novel dual-branch network structure that combines convolution and transformer architectures with adaptive receptive fields, namely, the dual-branch adaptive convolutional transformer (DBACT). The DBACT network is delineated into three primary components: the three-branch parallel hybrid stem module (TBPH), the local residual convolutional module (LRC), and the global adaptive transformer encoder module (GATE). Employing convolution operations with

various kernel sizes and spectral pooling, the TBPH module is designed for the extraction of shallow features and the reduction of spectral dimensionality. The LRC module with skip connection focuses on bolstering the expression of local features. The GATE module, equipped with adaptive receptive fields, serves the purpose of adaptive global context encoding. A compact module known as the cross-attention interaction module (CAI) is implemented to enhance the interaction between global features and local features within the GATE and LRC modules. The main achievements are delineated as follows:

1. Our study introduces a novel dual-branch adaptive convolutional transformer network that merges the localized feature extraction capabilities of CNNs with the global modeling advantages of transformers. This parallel dual-branch adeptly captures and fuses discriminative features across spatial and spectral dimensions, preserving the inherent structural properties of HSIs.
2. The GATE module, distinctively equipped with adaptive multi-head self-attention mechanism (AMSA), is capable of extracting global context features. It also incorporates depth wise convolution to enhance its local feature capabilities and offers implicit inference of positional information. The AMSA mechanism, armed with adaptive receptive fields, adeptly adjusts to irregular object geometries and accurately captures nuanced edge details.
3. The LRC module, tasked with capturing local information, functions in conjunction with the CAI module to augment the global information assimilated by the GATE module. The CAI module utilizes a cross-attention mechanism effectively integrating global and local features.
4. Assessments on the Salinas Valley, Pavia University, and Indian Pines datasets have established that the DBACT model exhibits superior performance relative to existing state-of-the-art models, whether CNN-based or transformer-based.

The subsequent sections of this paper are structured as follows: Section 2 describes the related works concerning the DBACT model. Section 3 presents the details of our proposed model along with its essential components. In Section 4, both qualitative and quantitative analyses were performed. Section 5 delves into the influence of hyperparameters on model performance and conducts ablation experiments to dissect the role of pivotal modules within the model. Finally, Section 6 furnishes a thorough summary of the entire paper.

2. Related Works

2.1. CNN and Transformer-Based Frameworks

Due to the advantageous properties of convolution operations, such as locality and translation equivariance, a multitude of CNN-based methodologies have been extensively explored for hyperspectral image (HSI) classification. Unlike convolutional neural networks (CNNs) with static filters, transformer-based architectures prioritize capturing long-distance relationships. To leverage the strengths of both approaches and enhance the classification performance of hyperspectral images, many researchers are dedicated to designing hybrid architectures that combine CNNs and transformers. Yang et al. [25] integrated convolution operations into the transformer framework, aiming to capture nuanced spatial-spectral discrepancies. In [31], convolution and transformer architectures were combined in a hybrid manner, utilizing both sequential and parallel processing approaches. Zhong et al. [32] designed a factorized architecture search (FAS) framework to highlight the interaction across both spatial and spectral dimensions. A dual-branch structure, utilized for separately extracting spatial and spectral features or for distinguishing between global and local characteristics, represents a significant form of hybrid structural design [33–35]. Cui et al. [36] proposed a dual-branch model that extracts spatial and spectral local context features and adaptively merges context information from various levels. In [37], two distinct branches were utilized to extract global spectral and spatial attributes from HSI, efficiently integrating these features at various scales through a dedicated feature fusion layer.

2.2. Deformable Architectures and Their Application in HSI Classification

Convolution operations employing fixed kernel sizes, while adept at extracting local features, can obscure the precise boundaries of objects, potentially diminishing classification efficacy [38]. To mitigate this issue, Zhu et al. [39] demonstrated that the integration of deformable convolution downsampling and deformable convolutions enhances adaptability to the geometric features of intricate structures. Nie et al. [40] employed a synergistic approach by combining the k-means clustering algorithm with deformable convolutions, a strategy aimed at reducing the effects of intra-cluster variance, consequently elevating the accuracy of classification tasks. Zhao et al. [38] developed a sophisticated filter mechanism for integrating superpixel regions to enhance homogeneity and applied deformable convolutions for the nuanced extraction of spatial features across diverse scales. The concept of deformable convolution was first introduced in [41]. It can be mathematically represented as follows: given an input feature map I, and a convolution kernel K, the deformable convolution at a location p_0 on the output feature map can be mathematically represented as follows:

$$Y(p_0) = \sum_{p_n \in R} K(p_n) I(p_0 + p_n + \Delta p_n) \quad (1)$$

where $Y(p_0)$ represents the output feature at the position p_0, $K(p_n)$ is the weight of the convolution kernel at the position p_n, within a regular grid R, Δp_n is the learned offset for the position p_n.

Reflecting deformable convolution principles, recent studies [42,43] have incorporated the dynamic receptive field concept into the computational frameworks of attention mechanisms in computer vision (CV). However, the deployment of architectures that synergize adaptive dynamic receptive fields with self-attention mechanisms remains relatively rare in the domain of hyperspectral classification.

3. Methodology

3.1. Overview of the Proposed Network

This study designs a novel DBACT method tailored for hyperspectral image classification tasks. As depicted in Figure 1, DBACT primarily consists of four modules: the Three-Branch Parallel Hybrid module (TBPH), the Cross-Attention Interaction module (CAI), the Local Residual Convolutional module (LRC), and the Global Adaptive Transformer Encoder module (GATE). The TBPH module, utilizing a tri-parallel branch, initially extracts shallow spectral features to diminish redundant spectral data. Following this, feature maps are segregated into two categories: a local feature maps group and a global feature maps group. The LRC module enhances local spatial feature extraction through two residual convolutional layers, while the GATE module focuses on global spectral-spatial information, incorporating a depthwise convolutional layer, an adaptive self-attention block, and an inverted residual block. The adaptive self-attention block is designed to capture multi-scale discriminative spatial features in pivotal areas through its dynamic receptive fields. Furthermore, local features and global features, derived by the LRC module and the GATE module, respectively, are synergized within the Cross-Attention Interaction module. This fusion process utilizes a multi-head cross-attention mechanism, complemented by convolutional projection and regulated via layer normalization (LayerNorm). Following the fusion, the features undergo an overlapping convolutional downsampling, which reduces the spatial resolution by half, facilitating the extraction of multi-scale feature information. Subsequent processing through the LRC and GATE modules extracts high-level semantic features. The global and local features are then merged along their dimensions and processed through a global average pooling (GAP) layer. Ultimately, the classification of hyperspectral images is accomplished using a linear classifier.

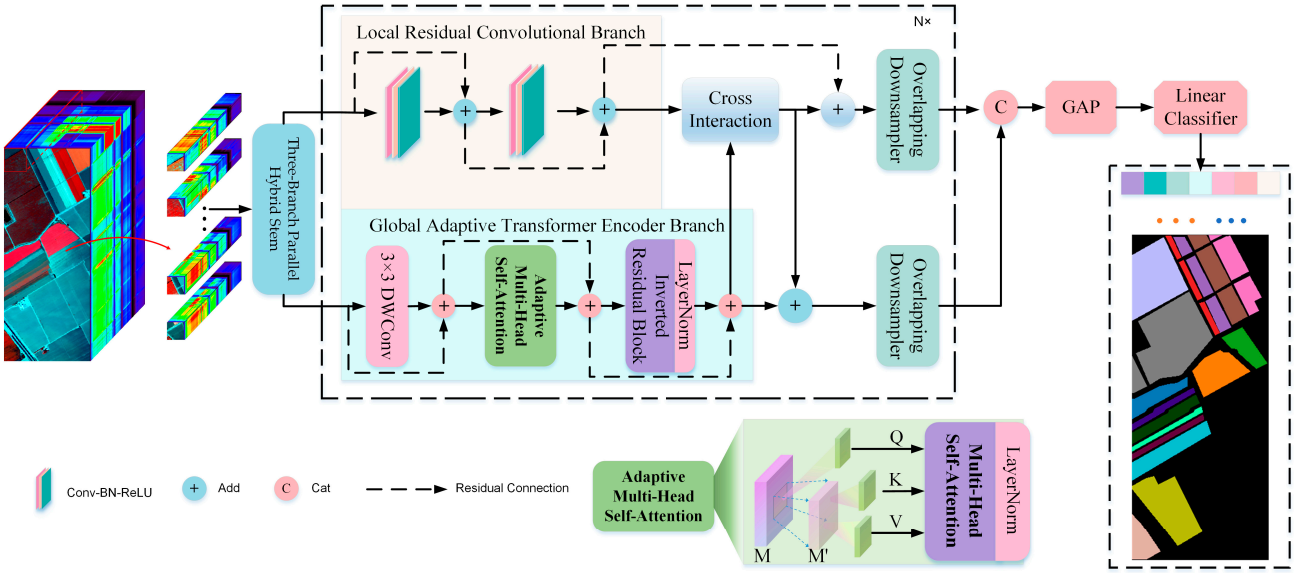


Figure 1. Overview of the Dual-Branch Adaptive Convolutional Transformer network.

Specifically, assumed $X \in \mathbb{R}^{H\times W\times B}$ represents an input original hyperspectral image, where H, W, and B refers to the height, width, and spectral dimension, respectively. X is cropped into n patches $X = [X_1, X_2, \cdots X_n]$, where $X_i \in \mathbb{R}^{p\times p\times B}$ and p is the patch size. X is randomly partitioned into subsets for training and testing. The training patches are utilized to train the DBACT network to obtain the optimal parameters while the testing patches to validate the performance of the network. TBPH module is employed to transform patch size from $X_i \in \mathbb{R}^{p\times p\times B}$ to $M_i \in \mathbb{R}^{p\times p\times c}$ to compress the spectral dimension. The LRC and GATE modules extract local features and global features from M, respectively, while preserving the original spatial and spectral dimensions. The Cross-Attention module fuses information of these two branches and an overlapping convolutional downsampling module is applied to convert the $M_i \in \mathbb{R}^{p\times p\times c}$ to $M_i \in \mathbb{R}^{p/2\times p/2\times c}$. The output feature maps of the second LRC and GATE module are concatenated in spectral dimension. A GAP layer and a linear classifier are finally applied to predict the classification map.

3.2. Three-Branch Parallel Hybrid Module

In order to extract valuable spectral bands from redundant spectral information, a triple parallel branch module is proposed. As shown in Figure 2, the TBPH module comprises two 1×1 convolutional layers, two 3×3 convolutional layers, and a spectral pooling layer. The 1×1 convolutional layers and spectral pooling layer focus on information compression in spectral dimension while 3×3 convolutional layers take spatial context information into account. The results from the three branches are subjected to a point-wise addition along the spatial dimension. Specifically, for an input HSI patch $X_i \in \mathbb{R}^{p\times p\times B}$, the output X_{TPBH} of the TBPH module can be expressed as:

$$\begin{cases} X_c = Conv_{1\times1}(g(Conv_{1\times1}(X_i))) \\ X_{pool} = SePool(X_i) \\ X_a = Conv_{3\times3}(g(Conv_{3\times3}(X_i))) \end{cases} \tag{2}$$

$$X_{TPBH} = X_c + X_{pool} + X_a \tag{3}$$

where $Conv_{1\times1}$ and $Conv_{3\times3}$ denote the convolutional layer with 1×1 kernel size and 3×3 kernel size, respectively. $g(x)$ denotes the activation function. $SePool(x)$ denotes the average pooling operation along the spectral dimension.

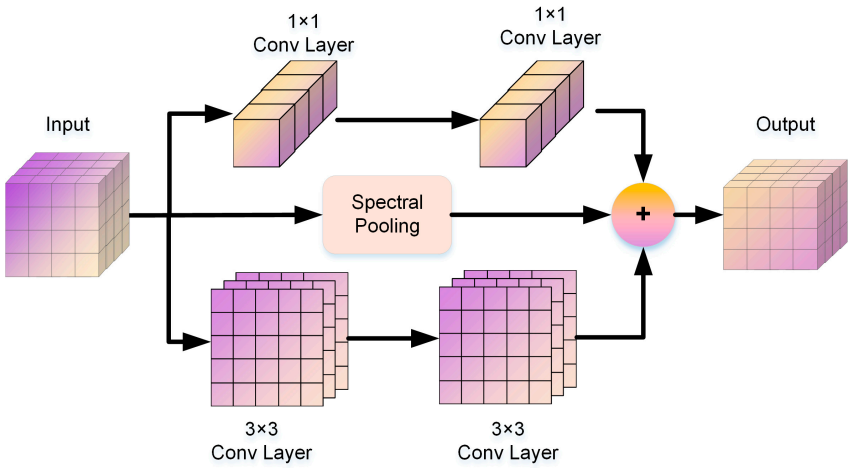


Figure 2. Three-Branch Parallel Hybrid Module.

3.3. Local Residual Convolutional Module and Cross-Attention Interaction Module

The Local Residual Convolution (LRC) module, engineered to elucidate local geometric positional relationships within the data, is depicted in Figure 3. The LRC module adopts a dual-layer convolutional framework, with each layer utilizing a 3×3 kernel size augmented by the skip connection. This architecture is specifically designed to emphasize spatial local characteristics, acting as a complement to the global information processed by the GATE. Formally, given an input HSI patch $X_i \in \mathbb{R}^{p \times p \times B}$, an output $X_{LRC} \in \mathbb{R}^{p \times p \times B}$ can be calculated as follows:

$$X_{LRC} = g(Conv_{3\times3}(g(Conv_{3\times3}(X_i)) + X_i)) \tag{4}$$

where $Conv_{3\times3}$ denotes the convolutional layer with 3×3 kernel size. $g(x)$ denotes the ReLU activation function.

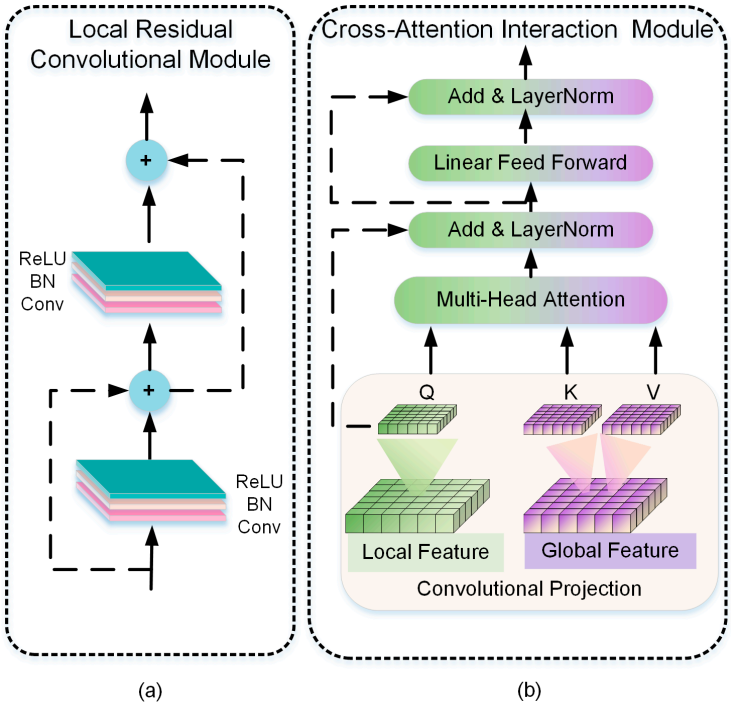


Figure 3. An overview of two parts of the DBACT model. (**a**) Local Residual Convolutional Module. (**b**) Cross-Attention Interaction Module.

As illustrated in Figure 3, the Cross-Attention Interaction module (CAI) block adeptly synthesizes local and global semantic information through an advanced multi-head cross-attention mechanism, complemented by a convolutional projection. This fusion leverages the strengths of both local feature delineation and global context integration for an all-encompassing analysis of the data. Formally, a local feature map $X_l \in \mathbb{R}^{p\times p\times B}$ and a global feature map $X_g \in \mathbb{R}^{p\times p\times B}$, possessing identical dimensions, undergo convolutional projection before being fed into the multi-head cross-attention mechanism. The process is mathematically outlined below:

$$\begin{cases} Q = Conv_{3\times 3}(X_l) \\ K = Conv_{3\times 3}(X_g) \\ V = Conv_{3\times 3}(X_g) \end{cases} \tag{5}$$

$$Atten = \text{softmax}\left(\frac{QK^T}{\sqrt{n}}\right)V \tag{6}$$

where the term softmax denotes the softmax function applied in the spectral dimension, n is the spectral dimension.

To effectively harness complex information across various subspaces, the model utilizes concatenated multi-head attention blocks.

$$CAB(Q, K, V) = concat(Atten_1, Atten_2 \cdots, Atten_h)W \tag{7}$$

where h is the head number, W is a matrix with trainable parameters.

3.4. Global Adaptive Transformer Encoder Module

We propose a novel global adaptive transformer encoder module (GATE) to accurately discern the geometric positional relationships of irregular objects along spatial dimensions. Notably, the GATE incorporates an adaptive multi-head self-attention (AMSA) mechanism, enabling a dynamic and potent receptive field tailored for hyperspectral image classification endeavors. As illustrated in Figure 4, the GATE is principally constructed of an adaptive multi-head self-attention block and an inverted residual block. The design of the GATE with global dynamic receptive fields is strategically engineered to foster the learning of robust representations.

The structure of the adaptive multi-head self-attention mechanism is illustrated in Figure 4a. It employs rectangular grid reference points, uniformly sampled and evenly spaced based on the input image's dimensions. Convolutional projection is used to generate query tokens from the input image. To obtain altered grid reference points, a streamlined, lightweight offset prediction network processes these query tokens to predict offsets. Bilinear interpolation, guided by these adjusted grid reference points, is then applied to reconstruct the input feature maps. Typically, the indices of these sampled points are fractional. For feature extraction at specific coordinates, bilinear interpolation is utilized as follows:

$$M(p_x, p_y) = \sum_{(q_x, q_y)} g(p_x, q_x)g(p_y, q_y)M(q_x, q_y) \tag{8}$$

where $g(a, b) = \max(0, 1-|a-b|)$ is the bilinear interpolation function across all spatial locations and $M(q_x, q_y)$ denotes the locations of the input feature maps $M \in \mathbb{R}^{H\times W\times C}$.

The convolutional projection is performed on the deformed feature maps M' to obtain the matrix K' and matrix V', respectively. To achieve adaptive multi-head self-attention, the attention matrix A obtained by matrix Q and matrix K' are used to obtain similarity relationships between original tokens and deformed tokens. Adaptive insights are garnered by extracting keys and values from the deformed input features, whereas the queries are obtained from the original input features. Values are then aggregated in alignment with

the specifications of the attention map. This sequence of operations can be mathematically represented as follows:

$$M' = G(M; r + \Delta r),\ \Delta r = \phi_{offset}(r) \tag{9}$$

$$Q = Conv_q(M),\ K' = Conv_k(M'),\ V' = Conv_v(M') \tag{10}$$

$$z = \text{softmax}\left(\frac{Q(K')^T}{\sqrt{n}}\right)V' \tag{11}$$

$$Z = Concat(z^1, z^2, \cdots, z^n)W_o \tag{12}$$

where M refers to the original input features, M' signifies the deformed features, G is the bilinear interpolation operation, r and Δr denote the reference points and the offsets, n represents the head number, and Z means the output feature maps $Z \in \mathbb{R}^{H \times W \times C}$.

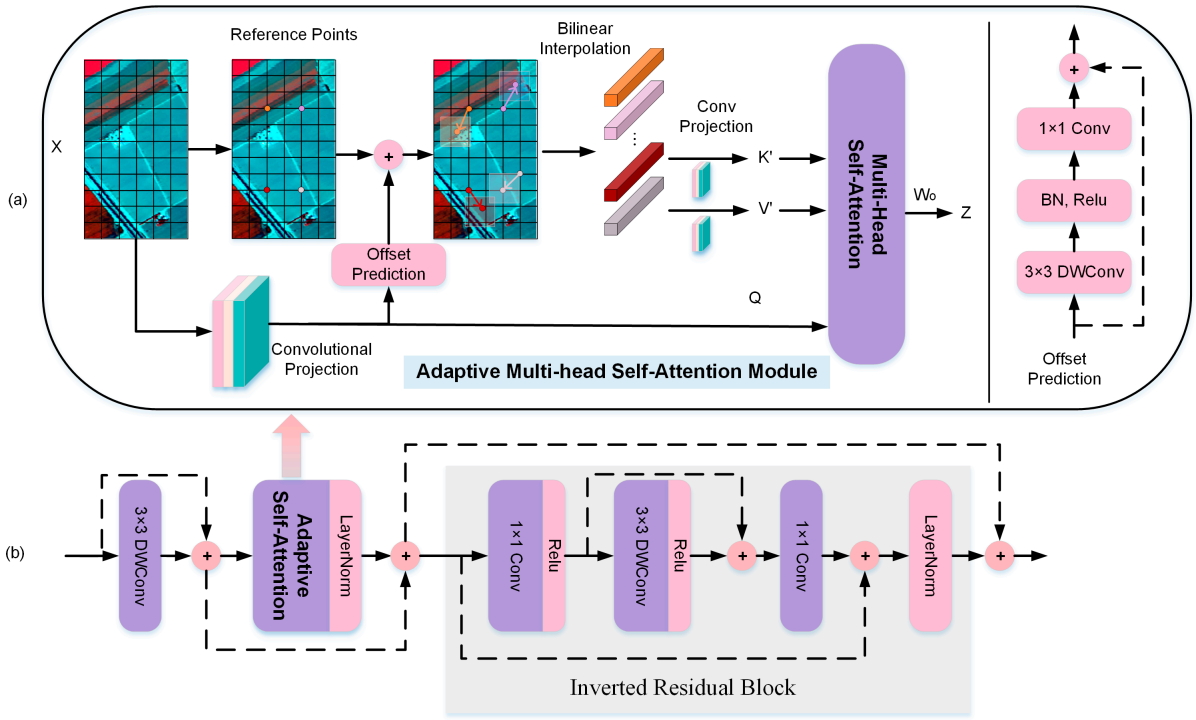


Figure 4. Detail structure of Global Adaptive Transformer Encoder. (**a**) presents the detailed implementation of adaptive multi-head self-attention. (**b**) presents the overall pipeline of the global adaptive transformer encoder module.

The overall architecture of GATE, which integrates the adaptive multi-head self-attention mechanism and an inverted residual block, is shown in Figure 4b. Depthwise convolution is applied to reinforce local feature extraction ability and impart implicit location information. The adaptive self-attention module is the pivotal component to acquire dynamic receptive fields and discriminative spatial signatures. The inverted residual block with skip connection complements the adaptive self-attention block with spatial and spectral modeling ability. The process of the l-th layer is mathematically articulated as follows:

$$Z_l' = DWConv(Z_{l-1}) + Z_{l-1} \tag{13}$$

$$Z_l'' = AMSA(Z_l') + Z_l' \tag{14}$$

$$Z_l = IR(Z_l'') + Z_l'' \tag{15}$$

where the *DWConv* denotes the depthwise convolution, the *AMSA* means the adaptive self-attention block and *IR* stands for the inverted residual block.

4. Results

In this section, to validate the performance of the proposed model DBACT, three well-known hyperspectral datasets were applied in related experiments. Comprehensive details about the experimental setup and the datasets utilized were furnished, which set the stage for both qualitative and quantitative comparative analyses with leading-edge models.

4.1. Description of Datasets

Indian Pines (IP): The scene was captured by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor over the Indian Pines test site in northwestern Indiana in 1992, consisting of 145 × 145 pixels and 224 spectral reflectance bands. The Indian Pines landscape includes two-thirds agriculture and one-third forest or other natural perennial vegetation, featuring two major dual-lane highways, a railway line, as well as some low-density housing, other buildings, and more minor roads. The existing ground truth is divided into 16 categories, which are not all mutually exclusive. After excluding spectral bands affected by noise and water absorption, the study utilized the remaining 200 spectral bands at 10 nm intervals, spanning from 400 to 2500 nm.

Pavia University (PU): The University of Pavia dataset, widely recognized and utilized within the hyperspectral image classification domain, was captured by the satellite-based ROSIS-03 sensor over Pavia, Italy's agricultural zones in 2003. This dataset is notable for its high spatial resolution and broad spectral range, providing essential data for diverse land cover and land use analyses. After abandoning 12 spectral bands compromised by noise and water absorption, the dataset comprises 103 spectral channels with spatial dimensions of 610 × 340 pixels. These channels cover a wavelength range from 430 to 860 nm and offer a spatial resolution (SR) of up to 1.3 m. The dataset is organized into 9 distinct classes, each representing a different category of land cover for classification.

Salinas Valley (SV): The Salinas dataset, derived from the Salinas Valley in California, features high-resolution hyperspectral images obtained via AVIRIS sensor in 1998. This dataset offers comprehensive spectral and spatial data pertinent to the region's agriculture. It comprises images with dimensions of 512 × 217 pixels and includes 224 spectral bands that span wavelengths from 360 to 2500 nm, with a spatial resolution (SR) of 3.7 m. The dataset categorizes the land cover into 16 distinct classes for classification purposes.

The distribution of false color maps, ground truth maps, and dataset categories for the three datasets we used, IP, PU, and SV, are illustrated in Figures 5–7, respectively. Ground truth maps are images annotated with accurate classification labels, which represent the actual categorization of the land cover. In the IP and SV datasets, the distribution of features is comparatively regular, whereas in the PU dataset, the distribution of features tends to be more scattered.

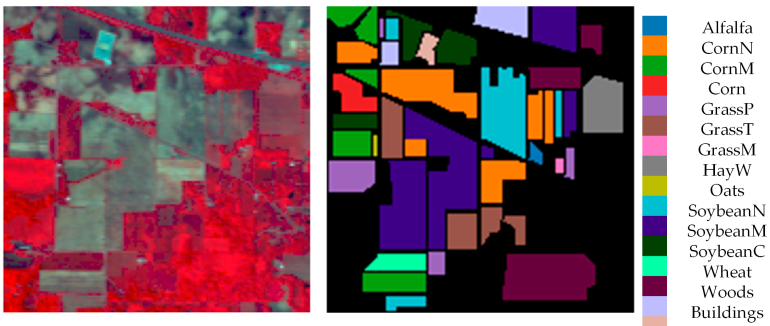


Figure 5. IP dataset. (**a**) False color map. (**b**) Ground truth map.

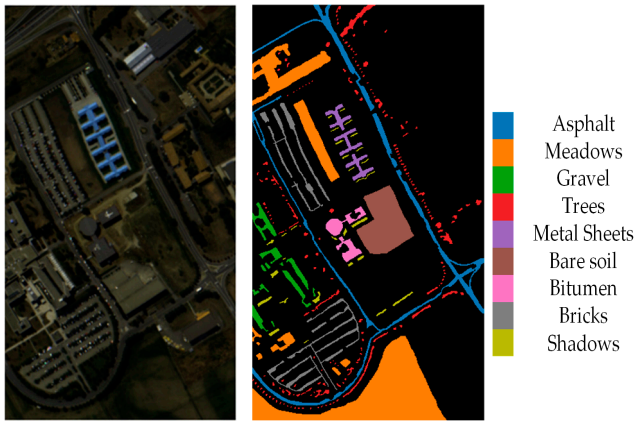


Figure 6. PU dataset. (**a**) False color map. (**b**) Ground truth map.

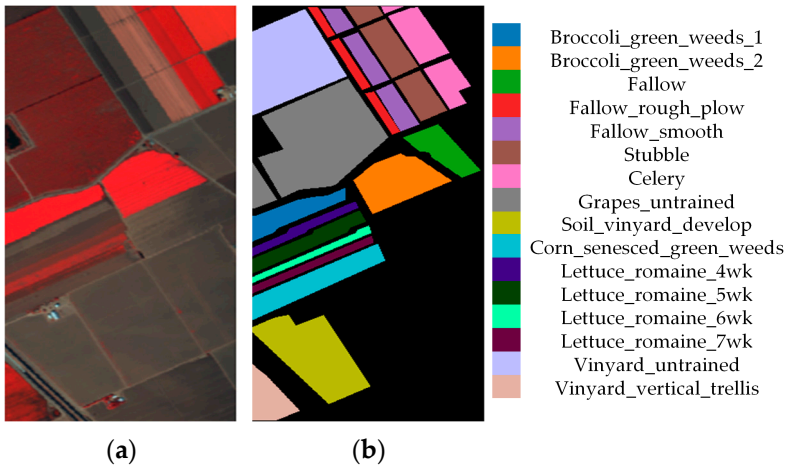


Figure 7. SV dataset. (**a**) False color map. (**b**) Ground truth map.

These three datasets are randomly divided into training and testing sets with the distribution of categories detailed in Table 1. We allocate about 8%, 5%, and 3% of the IP, PU, and SV dataset samples for training, respectively, with the remainder designated as the test set. Classification performance across all models is evaluated using three prevalent quantitative metrics: Overall Accuracy (OA), Average Accuracy (AA), and the Kappa Coefficient (K).

Table 1. Categories and the number of training/total samples of IP, PU, and SV.

NO.	Indian Pines		Pavia University		Salinas Valley	
	Class	**Training/Total**	**Class**	**Training/Total**	**Class**	**Training/Total**
1	Alfalfa	4/46	Asphalt	332/6631	Broccoli_green_weeds_1	60/2009
2	CornN	114/1428	Meadows	932/18,649	Broccoli_green_weeds_2	112/3726
3	CornM	66/830	Gravel	105/2099	Fallow	59/1976
4	Corn	19/237	Trees	153/3064	Fallow_rough_plow	42/1394
5	GrassP	39/483	Metal Sheets	67/1345	Fallow_smooth	80/2678
6	GrassT	58/730	Bare soil	251/5029	Stubble	119/3959
7	GrassM	2/28	Bitumen	67/1330	Celery	107/3579
8	HayW	38/478	Bricks	184/3682	Grapes_untrained	338/11,271
9	Oats	2/20	Shadows	47/947	Soil_vinyard_develop	186/6203

Table 1. *Cont.*

NO.	Indian Pines		Pavia University		Salinas Valley	
	Class	Training/Total	Class	Training/Total	Class	Training/Total
10	SoybeanN	78/972			Corn_senesced_green_weeds	98/3278
11	SoybeanM	196/2455			Lettuce_romaine_4wk	32/1068
12	SoybeanC	47/593			Lettuce_romaine_5wk	58/1927
13	Wheat	16/205			Lettuce_romaine_6wk	28/916
14	Woods	101/1265			Lettuce_romaine_7wk	32/1070
15	Buildings	31/386			Vinyard_untrained	218/7268
16	Stone	8/93			Vinyard_vertical_trellis	54/1807
Total Numbers		819/10,249		2138/42,776		1623/54,129

4.2. Experimental Setup

Experimental setups were executed on a hardware configuration consisting of an Intel(R) Core(TM) i7-9700 CPU at 3.00 GHz, 48 GB RAM, and a GeForce RTX 3080 GPU (10 GB VRAM) server. The software framework included Pytorch 1.8, Python 3.8, and CUDA 11.1, running on a Windows 10 operating system. Min-max scaling was applied to normalize the original HSI datasets, adjusting values to fall within the [0, 1] range. Models were trained with a batch size of 64 and a learning rate of 1×10^{-3}. The DBACT model was optimized using the AdamW optimizer, which incorporated a weight decay factor of 0.03. CosineLRScheduler is applied to dynamically adjust the learning rate. For a fair comparison, the compared models were configured according to the hyperparameters recommended in their respective publications. The epoch number is set to 100 in all experiments. All experiments were repeated 10 times under the same conditions except for random seeds, and the average value was taken as the final value.

4.3. Classification Results of Different Models

Through extensive experiments with other state-of-the-art networks, we validate the effectiveness of the proposed model. To thoroughly assess the DBACT model against leading-edge counterparts, we employ quantitative analysis and visual evaluation. Our comparative analysis involved two types of architectures to highlight the superior performance of the proposed method. Specifically, the CNN based models, including SSRN [15], DBDA [19], SPRN [18], and DCRN [44] and Transformer architecture based networks, including SSTN [32], SpectralFormer [27], SSFTT [28], and CTMixer [29], are considered.

4.3.1. Quantitative Analysis

The classification performance of these models on the three public HSI datasets is shown in Tables 2–4. These tables include a comprehensive summary of the OA, AA, and Kappa coefficients, along with their standard deviations and per-class accuracy metrics. The highest classification accuracy achieved in each category is distinguished by bold formatting in the corresponding rows. A comparison of the Overall Accuracy (OA) of different models across three distinct datasets is shown in Figure 8. CNN-based techniques exhibit a modest advantage compared to Transformer-based methods.

The results demonstrate that the DBACT model surpasses competing methods across all three datasets. For the IP dataset, characterized by its highly imbalanced and rare samples, the DBACT model achieves the highest accuracy in 7 out of 16 categories. For crops with dispersed locations, such as CornN, GrassP, GrassT, and SoybeanM, the model achieves superior classification accuracy. The DBACT model outperforms alternatives by at least 0.24% on AA, 0.21% on OA, and 0.27% on Kappa, highlighting its superior capability in extracting representative features under conditions of sample imbalance and scarcity. CNN-based models with powerful local feature-capturing abilities marginally outperform those based on Transformers. The DBDA model achieves highly competitive classification accuracy, which highlights the effectiveness of the fusion by a dual-branch structure and attention mechanisms. The pure transformer architecture, such as SpectralFormer, exhibits inadequate performance due to constrained local feature extraction capabilities. However,

merging convolutional and transformer structures in a sequential combination alone significantly enhances classification performance, such as SSFTT. Embedding convolutional operations deeply within the self-attention calculations, exemplified by the CTMixer model, significantly contributes to enhancements in ground object classification. The DBACT model, which introduces a spatial-spectral dual-branch framework, integrating convolution with self-attention, excels in capturing discriminative features on both global and local scales.

Table 2. Classification performance of different models on the IP dataset.

Class	CNNs Based					Transformer Based			
	SSRN (2018)	DBDA (2020)	SPRN (2021)	DCRN (2021)	SSTN (2021)	SF (2022)	SSFTT (2022)	CTMixer (2022)	Ours
1	96.1	97.86	**98.1**	94.63	95.95	76.67	97.62	94.15	97.38
2	97.65	98.18	98.17	98.18	98.07	81	98.03	98.24	**98.53**
3	97.91	98.74	98.36	**99.01**	98.7	80.04	97.45	98.29	98.52
4	98.07	**99.04**	97.89	98.03	98.26	70.69	97.06	98.53	98.99
5	96.91	97.82	96.78	97.55	97.39	90.63	97.09	96.89	**97.97**
6	99.64	99.46	98.88	99.49	98.97	92.9	98.42	99.29	**99.72**
7	97.2	99.62	86.92	99.6	**100**	52.31	98.46	98.8	99.62
8	99.82	**100**	99.95	**100**	**100**	97.02	99.95	99.98	**100**
9	**100**	98.33	75.56	98.82	96.67	66.11	96.11	98.82	98.33
10	96.67	97.76	98.02	97.94	97.33	82.73	97.27	97.85	**98.27**
11	98.99	99.11	99.08	99.09	99.03	90.35	99.08	99.02	**99.39**
12	98.39	**99.27**	98.08	98.96	98.63	70.71	96.74	97.75	99.07
13	99.42	99.05	97.99	99.1	99.47	98.31	99.47	**99.58**	99.37
14	99.61	99.7	99.86	99.67	99.69	95.12	99.79	99.85	**99.87**
15	95.04	97.21	**99.49**	97.3	96.59	73.35	99.24	98.7	98.48
16	**99.65**	98.35	94.82	98.35	98.82	89.29	96.59	98.59	99.41
OA (%)	98.35 (0.30)	98.82 (0.22)	98.60 (0.20)	98.79 (0.17)	98.61 (0.26)	86.29 (0.90)	98.41 (0.26)	98.70 (0.13)	**99.06 (0.15)**
AA (%)	98.19 (0.39)	98.72 (0.38)	96.12 (1.71)	98.48 (0.45)	98.35 (0.42)	81.70 (1.74)	98.02 (0.65)	98.40 (0.36)	**98.93 (0.34)**
K (%)	98.12 (0.35)	98.66 (0.25)	98.40 (0.22)	98.62 (0.19)	98.42 (0.30)	84.34 (1.05)	98.19 (0.30)	98.51 (0.15)	**98.93 (0.18)**

Note: Bold values in the table indicate the highest values.

Table 3. Classification performance of different models on the PU dataset.

Class	CNNs Based					Transformer Based			
	SSRN (2018)	DBDA (2020)	SPRN (2021)	DCRN (2021)	SSTN (2021)	SF (2022)	SSFTT (2022)	CTMixer (2022)	Ours
1	**100**	99.84	99.97	99.99	99.28	94.48	99.72	99.95	99.89
2	99.99	99.9	99.99	99.99	99.97	98.22	99.99	99.98	**100**
3	**99.8**	99.22	99.64	99.76	92.14	89.29	98.03	98.76	99.43
4	97.18	98.65	97.92	98.12	97.61	97.43	98.53	98.19	**99.13**
5	99.69	99.95	99.51	99.55	**99.98**	99.88	99.82	99.74	99.94
6	99.97	99.91	**100**	**100**	97.31	92.71	100	99.98	99.92
7	99.96	99.88	**100**	**100**	99.82	85.36	99.94	99.99	99.94
8	99.48	**99.65**	99.41	99.51	97.29	91.05	98.62	99.34	99.46
9	97.12	99.58	97.9	98.33	99.27	99.02	98.64	98.81	**99.79**
OA (%)	99.66 (0.03)	99.74 (0.07)	99.71 (0.05)	99.75 (0.04)	98.75 (0.21)	95.55 (0.45)	99.60 (0.10)	99.70 (0.09)	**99.83 (0.04)**
AA (%)	99.24 (0.08)	99.62 (0.07)	99.37 (0.10)	99.47 (0.10)	98.07 (0.27)	94.16 (0.41)	99.25 (0.23)	99.42 (0.20)	**99.72 (0.09)**
K (%)	99.55 (0.04)	99.66 (0.09)	99.62 (0.06)	99.67 (0.05)	98.34 (0.28)	94.10 (0.60)	99.46 (0.14)	99.60 (0.12)	**99.77 (0.06)**

Note: Bold values in the table indicate the highest values.

Table 4. Classification performance of different models on the SV dataset.

Class	CNNs Based				Transformer Based				
	SSRN (2018)	DBDA (2020)	SPRN (2021)	DCRN (2021)	SSTN (2021)	SF (2022)	SSFTT (2022)	CTMixer (2022)	Ours
1	**100**	**100**	**100**	**100**	99.97	98.88	**100**	**100**	99.99
2	**100**	99.98	**100**	**100**	99.98	99.1	**100**	**100**	**100**
3	99.87	99.75	**100**	99.99	99.99	99.21	**100**	99.93	**100**
4	**99.8**	99.47	99.69	99.63	99.05	98.39	99.63	99.13	99.46
5	**99.85**	99.73	99.75	99.68	99.42	99.47	99.55	99.63	99.57
6	**100**	99.99	**100**	**100**	**100**	99.97	99.99	**100**	99.99
7	99.97	99.95	99.99	99.99	99.99	99.4	99.98	**100**	99.99
8	96.49	93.55	98.26	95.52	91.37	84.61	99.14	98.75	**99.34**
9	99.98	99.97	100	99.93	99.99	99.51	**100**	**100**	**100**
10	99.37	98.24	99.73	99.43	97.45	95.73	99.81	99.84	**99.96**
11	99.51	99.03	99.86	99.82	98.88	96.77	**99.98**	99.97	99.97
12	**100**	99.99	**100**	**100**	99.98	99.99	99.99	99.95	99.99
13	**99.98**	99.95	99.64	99.92	99.77	99.74	**99.98**	99.7	99.93
14	99.39	99.42	99.74	99.5	98.82	98.82	**99.78**	99.67	99.74
15	97.35	91.67	99.35	97.6	92.84	82.95	98.66	**99.53**	99.3
16	99.06	99.26	99.7	99.49	99.2	97.37	99.33	**99.88**	99.4
OA (%)	98.80 (0.16)	97.33 (0.38)	99.49 (0.15)	98.64 (0.23)	96.95 (0.29)	93.77 (0.62)	99.57 (0.11)	99.61 (0.09)	**99.70 (0.10)**
AA (%)	99.41 (0.06)	98.75 (0.17)	99.73 (0.10)	99.41 (0.10)	98.54 (0.10)	96.87 (0.26)	99.74 (0.09)	99.75 (0.07)	**99.79 (0.06)**
K (%)	98.66 (0.18)	97.03 (0.42)	99.43 (0.17)	98.49 (0.25)	96.61 (0.32)	93.07 (0.68)	99.52 (0.13)	99.56 (0.10)	**99.67 (0.11)**

Note: Bold values in the table indicate the highest values.

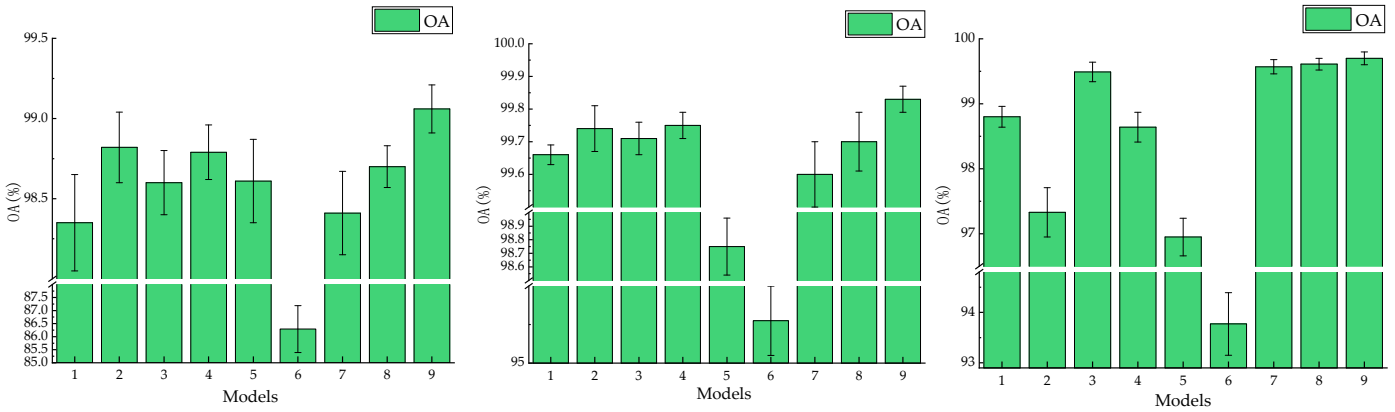


Figure 8. Comparison of OA Across Three Different Datasets. (**a**) IP dataset. (**b**) PU dataset. (**c**) SV dataset. The numbers 1–9 correspond to SSRN, DBDA, SPRN, DCRN, SSTN, SpectralFormer, SSFTT, CTMixer, and our model, respectively.

Integrating adaptive multi-head self-attention with convolution in our proposed model aids in capturing diagnostic features and subtle discrepancies. As demonstrated in Table 3, the fragmented distribution of ground objects, such as Trees and Bricks, can be effectively captured by the DBACT model. The DBACT model secures the highest classification accuracy in 3 out of 9 ground object classes and surpasses competing models by minimum advantages of 0.08% on AA, 0.1% on OA, and 0.1% on Kappa metrics. Similar to the IP dataset, models relying solely on attention architectures, such as SF and SSTN, exhibit relatively inferior performance. Integrating the attention mechanism with CNN's local feature extraction capabilities significantly enhances the model's ability to learn discriminative features. The DBACT model integrates an adaptive attention mechanism

that facilitates minor shape adjustments based on local information, thereby enhancing feature extraction from specific areas and more effectively capturing subtle features.

The DBACT model features an interactive architecture that combines branches for global and local feature information extraction, allowing for the concurrent extraction of both types of features. Such a design proves effective even for datasets with extensive and comparatively regular spatial distributions. As shown Table 4, the SV dataset is characterized by its relatively concentrated distribution of geographical features and their regular shapes. The DBACT model excels at capturing the boundary information between these regularly distributed features.

In the context of the SV dataset, where land features are notably concentrated and exhibit regular shapes, the DBACT model adeptly identifies boundary information between these systematically distributed features. As illustrated in Table 4, the SV dataset is notable for its tightly clustered geographical features and consistent patterns. The DBACT model excels in delineating the boundaries among these consistently distributed features. The DBACT model demonstrates superior classification accuracy, leading in 5 out of 16 ground object categories and exceeding other models by margins of at least 0.09% on AA, 0.04% on OA, and 0.11% on Kappa.

4.3.2. Qualitative Analysis

Alongside quantitative analysis, qualitative assessments depicted in Figures 9–11 were conducted to ensure a comprehensive evaluation. Visual analysis reveals that the DBACT model precisely delineates the boundaries between different types of ground objects, generating significantly less intra-category noise compared to competing models. For instance, in the SV dataset, distinguishing Grapes untrained from Vinyard_untrained categories is rendered difficult by their similar attributes. However, the DBACT model addresses this challenge by integrating adaptive receptive fields with the global self-attention mechanism. This combination, which adeptly captures subtle features, facilitates an effective differentiation between the two categories.

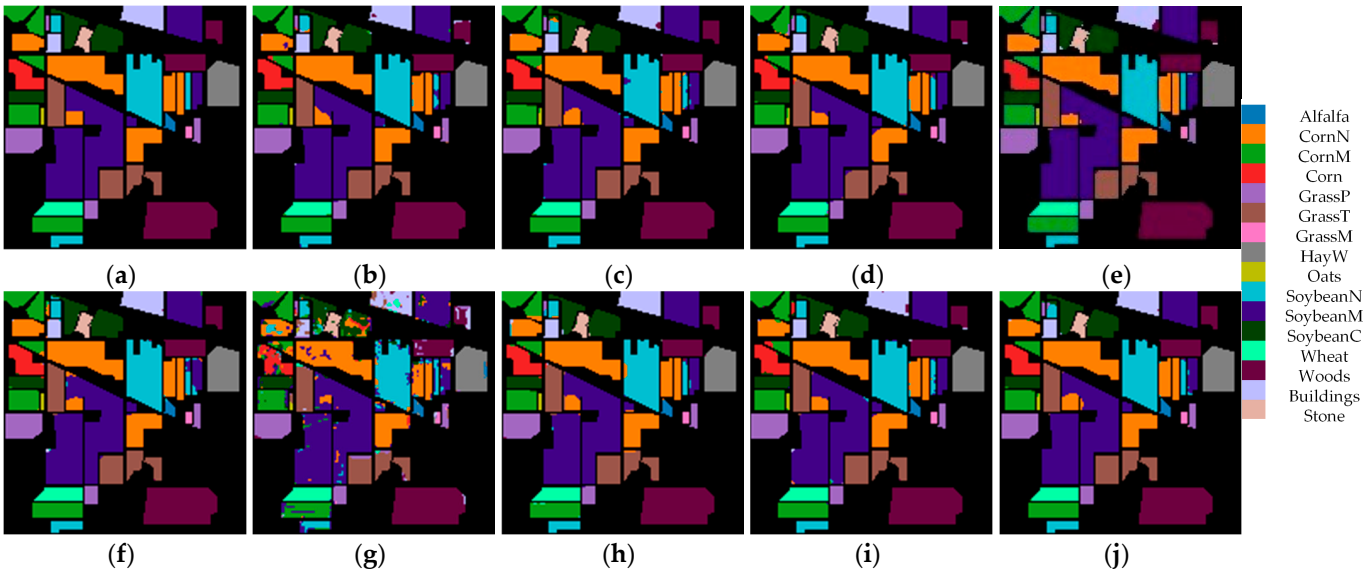


Figure 9. Classification maps on IP dataset. (**a**) Ground truth map. (**b**) SSRN. (**c**) DBDA. (**d**) SPRN. (**e**) DCRN. (**f**) SSTN. (**g**) SF. (**h**) SSFTT. (**i**) CTMixer. (**j**) Ours.

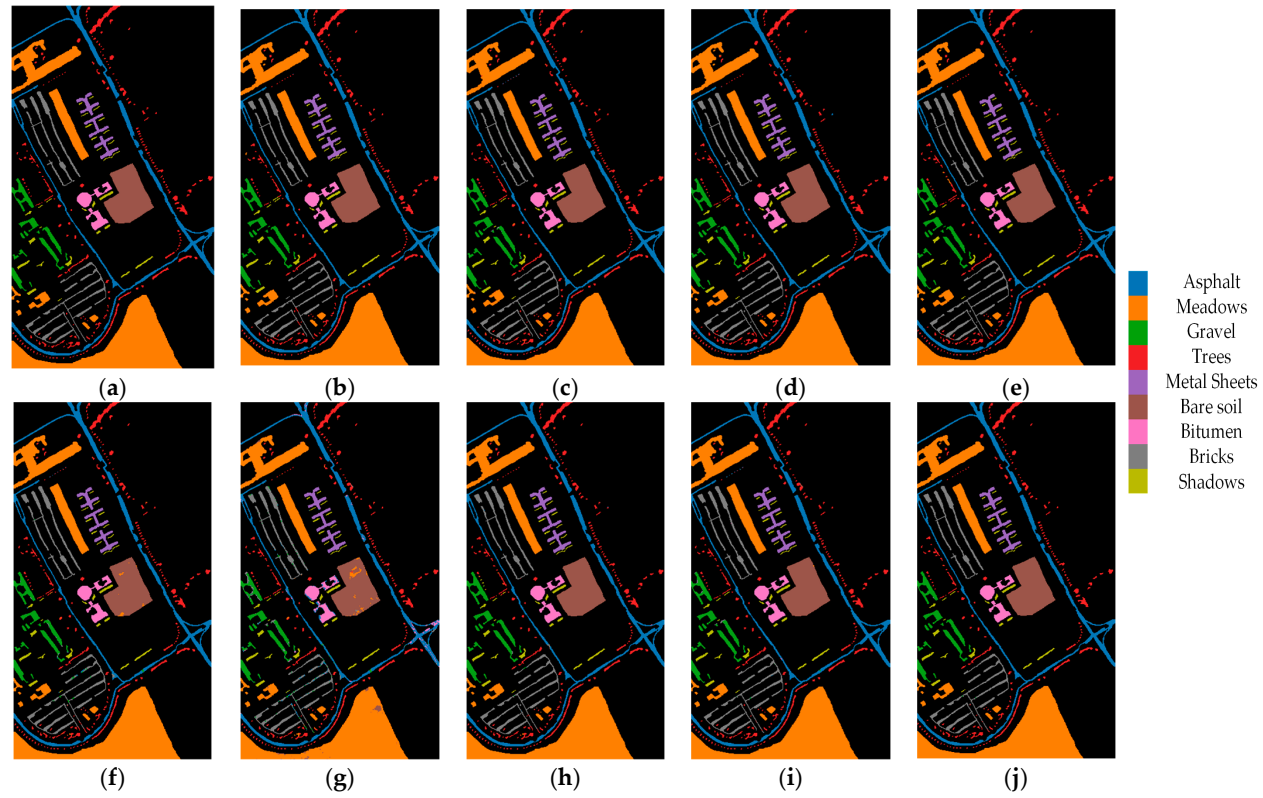


Figure 10. Classification maps on PU dataset. (**a**) Ground truth map. (**b**) SSRN. (**c**) DBDA. (**d**) SPRN. (**e**) DCRN. (**f**) SSTN. (**g**) SF. (**h**) SSFTT. (**i**) CTMixer. (**j**) Ours.

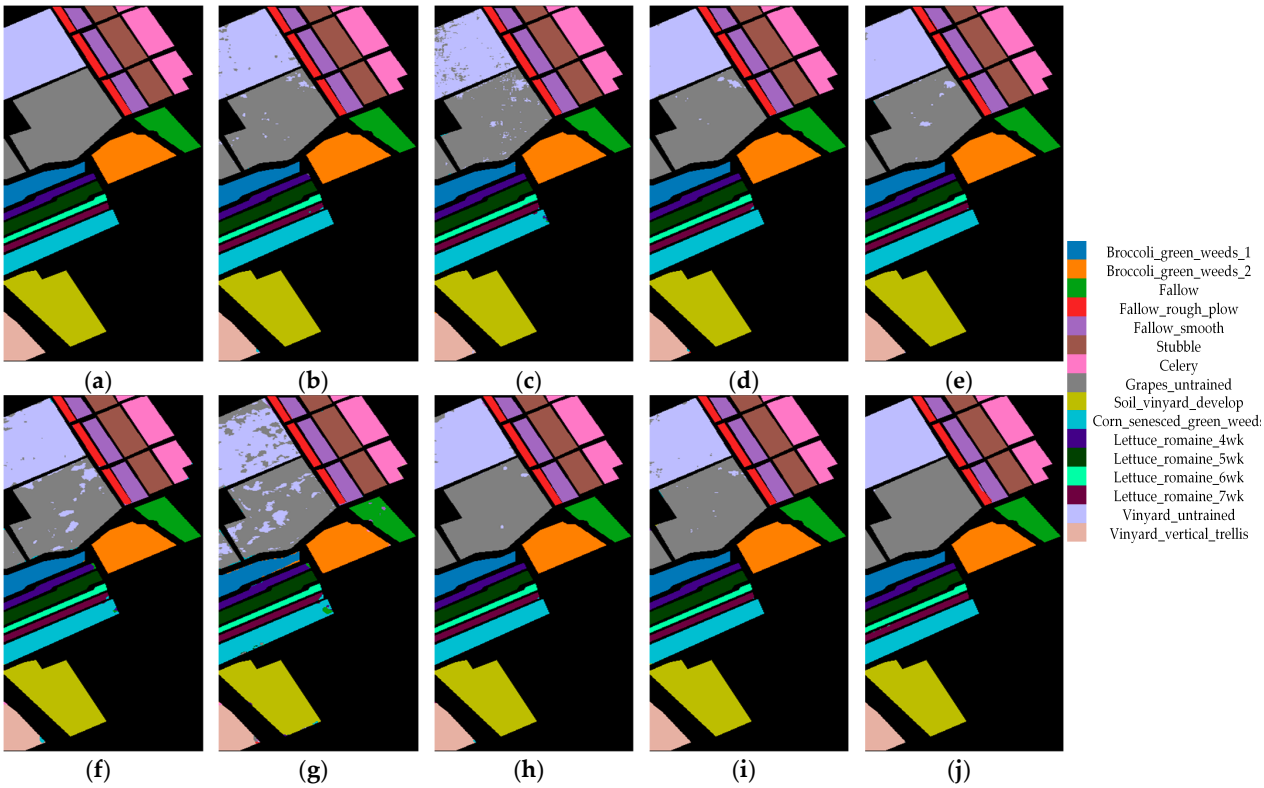


Figure 11. Classification maps on SV dataset. (**a**) Ground truth map. (**b**) SSRN. (**c**) DBDA. (**d**) SPRN. (**e**) DCRN. (**f**) SSTN. (**g**) SF. (**h**) SSFTT. (**i**) CTMixer. (**j**) Ours.

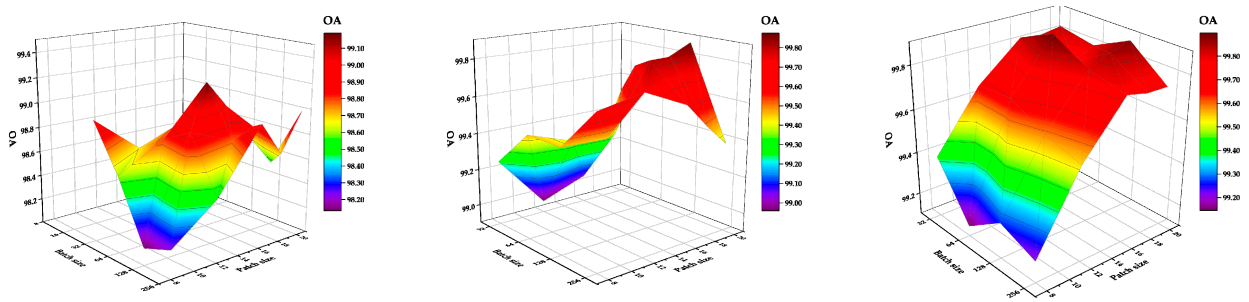
Overall, the distribution of classification maps across the three datasets is consistent with the results of quantitative analysis. The DBACT model excels in a variety of challenging scenarios: from datasets with unevenly distributed ground objects and those with small, imbalanced sample sizes (e.g., the IP dataset) to ones characterized by a fragmented distribution of geographical features (e.g., the PU dataset). It also performs well in datasets where crops are distributed orderly with distinct boundary lines (e.g., the SV dataset). The DBACT model achieves the highest classification accuracy in all cases, proving its effectiveness in identifying diverse object types.

5. Discussion

In this section, we analyze model structure parameters and ablation experiments. The parameter analysis primarily encompasses batch size, patch size, and the number of heads in self-attention. The ablation experiments aim to analyze the roles played by key components within the model.

5.1. Parameters Analysis

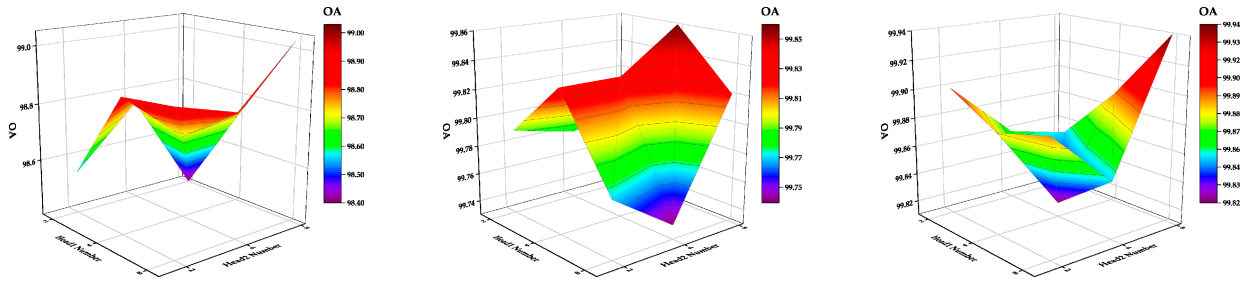
In this section, we investigate the effects of hyperparameters on Overall Accuracy (OA), including batch size, patch size, and head number in adaptive self-attention mechanism. Our objective is to determine the optimal network structure parameters. Optimal batch and patch sizes facilitate neural network convergence and performance enhancement. As illustrated in Figure 12, for the three datasets, patch size varies from 8 × 8 to 20 × 20, with an increment of 2, and batch size ranges from 32 to 256, doubling at each step. The relationship between patch size and classification accuracy demonstrates a non-linear pattern. OA improves with an increase in patch size (from 8 × 8 to 16 × 16) but begins to decrease (from 16 × 16 to 20 × 20) once the patch size crosses a certain threshold, with batch size remaining constant. Relative to patch size, batch size has a comparatively minor effect on OA. Nonetheless, setting the batch size to 64 yields the optimal classification performance. Consequently, the patch size and batch size are optimally set at 16 and 64, respectively.



(**a**) (**b**) (**c**)

Figure 12. Effects of patch size and batch size on OA (**a**) IP. (**b**) PU. (**c**) SV.

In the global adaptive transformer encoder branch, two self-attention blocks are employed, with the number of heads in the first and second blocks designated as head1 and head2, respectively. Figure 13 depicts the impact of head1 and head2 counts on OA, as the number of heads varies from 2 to 8. Compared to patch size, the effect of head number on classification effectiveness is relatively minor. Classification performance is enhanced with a rising number of heads, achieving optimal results when head1 and head2 numbers are adjusted to 8 on the three datasets. Multiple heads mean the ability to extract representative features from various subspaces of HSI. Therefore, the optimal settings for head1 and head2 numbers are both established at 8.



(**a**) (**b**) (**c**)

Figure 13. Effects of head number on OA (**a**) IP. (**b**) PU. (**c**) SV.

5.2. Ablation Study

To elucidate the contribution of each component within the model to its performance, we conducted ablation experiments on these three datasets (IP, PU, and SV). Key components of our model for ablation studies include LRC (Local Residual Convolution module), GATE (Global Adaptive Transformer Encoder module), AMSA (Adaptive Multi-head Self-Attention Mechanism), and CAI (Cross-Attention Interaction module).

In summary, each of the four critical components in the DBACT model plays a role in enhancing classification performance to some extent, with the GATE module making the most substantial contribution, as shown in Tables 5–7. Removing the GATE module results in the most significant decrease in Overall Accuracy (OA) by 0.59%, demonstrating the critical role that GATE's global feature extraction capability plays in classification tasks. The slight degradation in performance due to the removal of the AMSA module indicates that our dual-branch structure inherently possesses a high capability for feature learning, while the AMSA module serves to raise the upper limit of performance. The omission of the LRC and CAI modules led to a noticeable reduction in performance, highlighting the essential supportive function that local features serve in augmenting global features. The results of these ablation experiments confirm the effectiveness and robustness of the DBACT model.

Table 5. Ablation study on the IP dataset.

Component				IP Dataset		
LRC	**GATE**	**AMSA**	**CAI**	**OA (%)**	**AA (%)**	**K (%)**
×	√	√	×	98.64	98.54	98.45
√	×	√	×	98.47	98.44	98.26
√	√	×	√	98.78	98.26	98.61
√	√	√	×	98.63	98.66	98.44
√	√	√	√	**99.06**	**98.93**	**98.93**

Note: Bold values in the table indicate the highest values.

Table 6. Ablation study on the PU dataset.

Component				PU Dataset		
LRC	**GATE**	**AMSA**	**CAI**	**OA (%)**	**AA (%)**	**K (%)**
×	√	√	×	99.59	99.18	99.46
√	×	√	×	99.32	98.48	99.1
√	√	×	√	99.34	98.64	99.12
√	√	√	×	99.37	98.52	99.16
√	√	√	√	**99.83**	**99.72**	**99.77**

Note: Bold values in the table indicate the highest values.

Table 7. Ablation study on the SV dataset.

Component				SV Dataset		
LRC	**GATE**	**AMSA**	**CAI**	**OA (%)**	**AA (%)**	**K (%)**
×	√	√	×	99.63	99.58	99.62
√	×	√	×	99.49	99.46	99.46
√	√	×	√	99.63	99.59	99.62
√	√	√	×	99.64	99.61	99.64
√	√	√	√	**99.70**	**99.79**	**99.67**

Note: Bold values in the table indicate the highest values.

In summary, the Global Adaptive Transformer Encoder module (GATE) branch, incorporating an adaptive multi-head self-attention mechanism, efficiently extracts features from arbitrarily shaped objects. Meanwhile, leveraging residual connections, the Local Residual Convolution (LRC) module adeptly captures local representations. Combining the GATE and the LRC with cross attention, the DBACT model possesses robust feature extraction capabilities, effectively merging spatial and spectral data to extract representative high-level semantic information from hyperspectral images.

6. Conclusions

This study presents a dual-branch adaptive convolutional transformer (DBACT) network that integrates both global and local features for HSI classification. In our global feature extraction branch, we incorporated an adaptive multi-head self-attention mechanism capable of obtaining a dynamic global receptive field, thus accommodating a variety of irregular objects. The local feature extraction branch, equipped with residual connections, supplements the global feature extraction branch. The cross-attention mechanism acts as a fusion bridge between global and local features, enhancing the model's overall classification performance. By integrating the characteristics of the structures above, the DBACT model is adept at capturing global and local spatial and spectral features of hyperspectral imaging (HSI) across multiple scales. This comprehensive approach enables a nuanced understanding and representation of HSI data, facilitating superior classification performance even in complex scenarios involving diverse spatial and spectral variations. Across the three commonly used hyperspectral datasets—the IP, PU, and SV datasets—the DBACT model demonstrated superior performance in the Overall Accuracy, Average Accuracy, and Kappa Coefficient metrics compared to other state-of-the-art models, confirming its effectiveness.

In the future, we aim to investigate the implementation of dynamic receptive fields within the tri-dimensional architecture of data structures. Moreover, to tackle the challenge of limited labeled data availability, we propose developing generative models for creating synthetic hyperspectral data, thereby enriching the dataset diversity.

Author Contributions: Conceptualization, C.W. and J.H.; methodology, C.W.; validation, R.Q.; formal analysis, C.W.; writing—original draft preparation, C.W.; writing—review and editing, M.L. and Y.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The datasets are available at https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes (accessed on 12 July 2021).

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

References

1. Ahmad, M.; Shabbir, S.; Roy, S.K.; Hong, D.; Wu, X.; Yao, J.; Khan, A.M.; Mazzara, M.; Distefano, S.; Chanussot, J. Hyperspectral Image Classification—Traditional to Deep Models: A Survey for Future Prospects. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 968–999. [CrossRef]
2. Lefei, Z.; Liangpei, Z.; Dacheng, T.; Xin, H.; Bo, D. Hyperspectral Remote Sensing Image Subpixel Target Detection Based on Supervised Metric Learning. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 4955–4965. [CrossRef]

3. Liang, L.; Di, L.; Zhang, L.; Deng, M.; Qin, Z.; Zhao, S.; Lin, H. Estimation of crop LAI using hyperspectral vegetation indices and a hybrid inversion method. *Remote Sens. Environ.* **2015**, *165*, 123–134. [CrossRef]
4. Qi, W.; Huang, C.; Wang, Y.; Zhang, X.; Sun, W.; Zhang, L. Global–Local 3-D Convolutional Transformer Network for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5510820. [CrossRef]
5. Nicolas Audebert, B.L.S. Sebastien Lefevre Deep Learning for Classification of Hyperspectral Data: A Comparative Review. *IEEE Geosci. Remote Sens. Mag.* **2019**, *7*, 159–173. [CrossRef]
6. Melgani, F.; Bruzzone, L. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Trans. Geosci. Remote Sens.* **2004**, *42*, 1778–1790. [CrossRef]
7. Yu, H.; Shang, X.; Zhang, X.; Gao, L.; Song, M.; Hu, J. Hyperspectral Image Classification Based on Adjacent Constraint Representation. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 707–711. [CrossRef]
8. Cariou, C.; Chehdi, K. A New K-Nearest Neighbor Density-Based Clustering Method And Its Application To Hyperspectral Images. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium, Beijing, China, 10–15 July 2016. [CrossRef]
9. Ham, J.; Yangchi, C.; Crawford, M.M.; Ghosh, J. Investigation of the random forest framework for classification of hyperspectral data. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 492–501. [CrossRef]
10. Paoletti, M.E.; Haut, J.M.; Plaza, J.; Plaza, A. Deep learning classifiers for hyperspectral imaging: A review. *ISPRS J. Photogramm. Remote Sens.* **2019**, *158*, 279–317. [CrossRef]
11. Hu, W.; Huang, Y.; Wei, L.; Zhang, F.; Li, H. Deep Convolutional Neural Networks for Hyperspectral Image Classification. *J. Sens.* **2015**, *2015*, 258619. [CrossRef]
12. Makantasis, K.; Karantzalos, K.; Doulamis, A.; Doulamis, N. Deep supervised learning for hyperspectral data classification through convolutional neural networks. In Proceedings of the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, 26–31 July 2015.
13. 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.* **2016**, *54*, 6232–6251. [CrossRef]
14. Ahmad, M.; Khan, A.M.; Mazzara, M.; Distefano, S.; Ali, M.; Sarfraz, M.S. A Fast and Compact 3-D CNN for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 5502205. [CrossRef]
15. Zhong, Z.; Li, J.; Luo, Z.; Chapman, M. Spectral–Spatial Residual Network for Hyperspectral Image Classification: A 3-D Deep Learning Framework. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 847–858. [CrossRef]
16. Roy, S.K.; Krishna, G.; Dubey, S.R.; Chaudhuri, B.B. HybridSN: Exploring 3-D–2-D CNN Feature Hierarchy for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 277–281. [CrossRef]
17. Ghaderizadeh, S.; Abbasi-Moghadam, D.; Sharifi, A.; Zhao, N.; Tariq, A. Hyperspectral Image Classification Using a Hybrid 3D-2D Convolutional Neural Networks. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 7570–7588. [CrossRef]
18. Zhang, X.; Shang, S.; Tang, X.; Feng, J.; Jiao, L. Spectral Partitioning Residual Network with Spatial Attention Mechanism for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5507714. [CrossRef]
19. Li, R.; Zheng, S.; Duan, C.; Yang, Y.; Wang, X. Classification of Hyperspectral Image Based on Double-Branch Dual-Attention Mechanism Network. *Remote Sens.* **2020**, *12*, 582. [CrossRef]
20. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
21. Pan, X.; Ye, T.; Xia, Z.; Song, S.; Huang, G. Slide-Transformer: Hierarchical Vision Transformer with Local Self-Attention. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 2082–2091.
22. Guo, J.; Han, K.; Wu, H.; Tang, Y.; Chen, X.; Wang, Y.; Xu, C. CMT: Convolutional Neural Networks Meet Vision Transformers. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 12165–12175.
23. Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; Zhang, L. CvT: Introducing Convolutions to Vision Transformers. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021.
24. Tang, P.; Zhang, M.; Liu, Z.; Song, R. Double Attention Transformer for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 5502105. [CrossRef]
25. Yang, X.; Cao, W.; Lu, Y.; Zhou, Y. Hyperspectral Image Transformer Classification Networks. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5528715. [CrossRef]
26. Zu, B.; Li, Y.; Li, J.; He, Z.; Wang, H.; Wu, P. Cascaded Convolution-Based Transformer With Densely Connected Mechanism for Spectral–Spatial Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5513119. [CrossRef]
27. Hong, D.; Han, Z.; Yao, J.; Gao, L.; Zhang, B.; Plaza, A.; Chanussot, J. SpectralFormer: Rethinking Hyperspectral Image Classification With Transformers. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5518615. [CrossRef]
28. Sun, L.; Zhao, G.; Zheng, Y.; Wu, Z. Spectral–Spatial Feature Tokenization Transformer for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5522214. [CrossRef]
29. Zhang, J.; Meng, Z.; Zhao, F.; Liu, H.; Chang, Z. Convolution Transformer Mixer for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6014205. [CrossRef]

30. Yang, H.; Yu, H.; Zheng, K.; Hu, J.; Tao, T.; Zhang, Q. Hyperspectral Image Classification Based on Interactive Transformer and CNN With Multilevel Feature Fusion Network. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 5507905. [CrossRef]
31. Yang, L.; Yang, Y.; Yang, J.; Zhao, N.; Wu, L.; Wang, L.; Wang, T. FusionNet: A Convolution–Transformer Fusion Network for Hyperspectral Image Classification. *Remote Sens.* **2022**, *14*, 4066. [CrossRef]
32. 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.* **2022**, *60*, 5514715. [CrossRef]
33. Ghaderizadeh, S.; Abbasi-Moghadam, D.; Sharifi, A.; Tariq, A.; Qin, S. Multiscale Dual-Branch Residual Spectral–Spatial Network With Attention for Hyperspectral Image Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 5455–5467. [CrossRef]
34. Shu, Z.; Wang, Y.; Yu, Z. Dual attention transformer network for hyperspectral image classification. *Eng. Appl. Artif. Intell.* **2024**, *127*, 107351. [CrossRef]
35. Gao, H.; Zhang, Y.; Chen, Z.; Li, C. A Multiscale Dual-Branch Feature Fusion and Attention Network for Hyperspectral Images Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 8180–8192. [CrossRef]
36. Cui, Y.; Li, W.; Chen, L.; Gao, S.; Wang, L. Double-Branch Local Context Feature Extraction Network for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6011005. [CrossRef]
37. Dang, L.; Weng, L.; Hou, Y.; Zuo, X.; Liu, Y. Double-branch feature fusion transformer for hyperspectral image classification. *Sci. Rep.* **2023**, *13*, 272. [CrossRef] [PubMed]
38. Zhao, C.; Zhu, W.; Feng, S. Superpixel Guided Deformable Convolution Network for Hyperspectral Image Classification. *IEEE Trans. Image Process.* **2022**, *31*, 3838–3851. [CrossRef] [PubMed]
39. Zhu, J.; Fang, L.; Ghamisi, P. Deformable Convolutional Neural Networks for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1254–1258. [CrossRef]
40. Nie, J.; Xu, Q.; Pan, J.; Guo, M. Hyperspectral Image Classification Based on Multiscale Spectral–Spatial Deformable Network. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 5500905. [CrossRef]
41. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable Convolutional Networks. In Proceedings of the 2017 IEEE IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 764–773.
42. Xia, Z.; Pan, X.; Song, S.; Li, L.E.; Huang, G. Vision Transformer with Deformable Attention. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 4784–4793.
43. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable Transformers for End-to-End Object Detection. *arXiv* **2020**, arXiv:2010.04159.
44. Xu, Y.; Li, Z.; Li, W.; Du, Q.; Liu, C.; Fang, Z.; Zhai, L. Dual-Channel Residual Network for Hyperspectral Image Classification With Noisy Labels. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5502511. [CrossRef]