remote sensing

MDPI

Article

A Dual Multi-Head Contextual Attention Network for Hyperspectral Image Classification

Miaomiao Liang [1], Qinghua He [1], Xiangchun Yu [1,*], Huai Wang [1], Zhe Meng [2] and Licheng Jiao [3]

[1] School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou 341000, China; liangmiaom@jxust.edu.cn (M.L.); hqh@mail.jxust.edu.cn (Q.H.); 6920190632@mail.jxust.edu.cn (H.W.)
[2] School of Telecommunication and Information Engineering, Xi'an University of Posts and Telecommunications, Xi'an 710121, China; zhemeng@xupt.edu.cn
[3] Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University, Xi'an 710071, China; lchjiao@mail.xidian.edu.cn
* Correspondence: yuxc@jxust.edu.cn

Abstract: To learn discriminative features, hyperspectral image (HSI), containing 3-D cube data, is a preferable means of capturing multi-head self-attention from both spatial and spectral domains if the burden in model optimization and computation is low. In this paper, we design a dual multi-head contextual self-attention (DMuCA) network for HSI classification with the fewest possible parameters and lower computation costs. To effectively capture rich contextual dependencies from both domains, we decouple the spatial and spectral contextual attention into two sub-blocks, SaMCA and SeMCA, where depth-wise convolution is employed to contextualize the input keys in the pure dimension. Thereafter, multi-head local attentions are implemented as group processing when the keys are alternately concatenated with the queries. In particular, in the SeMCA block, we group the spatial pixels by evenly sampling and create multi-head channel attention on each sampling set, to reduce the number of the training parameters and avoid the storage increase. In addition, the static contextual keys are fused with the dynamic attentional features in each block to strengthen the capacity of the model in data representation. Finally, the decoupled sub-blocks are weighted and summed together for 3-D attention perception of HSI. The DMuCA module is then plugged into a ResNet to perform HSI classification. Extensive experiments demonstrate that our proposed DMuCA achieves excellent results over several state-of-the-art attention mechanisms with the same backbone.

Keywords: hyperspectral image classification; dual attention; contextual keys; grouping perception; multi-head self-attention

Citation: Liang, M.; He, Q.; Yu, X.; Wang, H.; Meng, Z. and Jiao, L. A Dual Multi-Head Contextual Attention Network for Hyperspectral Image Classification. *Remote Sens.* **2022**, *14*, 3091. https://doi.org/10.3390/rs14133091

Academic Editors: Lefei Zhang, Liangpei Zhang, Qian Shi and Yanni Dong

Received: 5 May 2022
Accepted: 24 June 2022
Published: 27 June 2022

1. Introduction

Hyperspectral images (HSI) contain rich spectral information and spatial context, where the electromagnetic spectrum is approximately contiguous and covers the ultraviolet, visible, near-infrared, and even mid-to-long infrared regions. The abundant spectral-spatial information provides great opportunities for the fine identification of materials with subtle spectral discrepancies, and at the same time brings new challenges in discriminant feature learning, especially in mining the potential correlation of data with the high-dimensional nonlinear distribution.

Compared with the limitations of the shallow and handcrafted extractors in complex data representation, deep neural networks (DNNs) have proven to be more powerful in feature learning with their excellent power in layer-wise feedforward perception, and have become prevailing benchmarks in HSI classification tasks [1–3], including multilayer perceptron (MLP) [4], stacked autoencoders (SAEs) [5], deep belief networks (DBNs) [6], recurrent neural networks (RNNs) [7–9], convolutional neural networks (CNNs) [10–12], graph convolutional networks (GCNs) [13], generative adversarial networks (GANs) [14], and their variants. CNN has become populat popular due to its advantage in locally contextual perception and feature transformation with parameter sharing. To strengthen the

contribution of the spectrum as spatial information dose in CNNs, multi-branch networks, 3-D-CNNs, and other more complex models are introduced to extract spectral-spatial features. Although they improve the model ability in feature representation, new issues will also arise, such as huge computing burdens, especially difficulty in model optimization. Thus, many lightweight deep models [15–17] were proposed for HSI classification. Meanwhile, shortcut connections [18] become an almost indispensable component to avoid model degradation.

CNN shares convolution kernels among different locations in each feature map and collects diverse information encoded in all of the channels. Although the kernel sharing and feature recombination enable CNN with great performance in translation equivalence and high-level feature learning, it passively focus attention on important regions, whether in spatial or in channel dimension, while this is just exactly common in human vision. Thus, the attention mechanism has raised much concern in remote sensing (RS) fields. This can be treated as a dynamic selection of features by adaptively weighting, while CNN is a static method. To focus more on significant channels for object recognition, a squeeze-and-excitation (SE) block in the SE network (SENet) [19] calibrates the channel weights by spatial squeeze and channel excitation. Subsequently, many of its variations were presented for feature learning of HSI. For example, Zhao et al. [20] replaced the excitation part with feature capture by two 1D convolution layers and aggregation by shortcut connections, namely CBW. Wang et al. [21] performed SE on spatial and spectral dimension in parallel (namely, SSSRN), then recalibrated features by weighted summation of the two attention matrices. Convolutional block attention module (CBAM) [22] refines the feature maps by two sub-modules, which squeeze respectively the spatial and channel information by global average pooling (GAP) and max pooling, and excites channel attention via a shared MLP and spatial attention by 2-D convolution. CBAM was then embedded into diverse deep networks for attention recalibration, such as double-branch 3-D convolution network (DBMA) [23].

CNN specializes in local perception and enlarges the receptive field by a deep stack of the convolutional layers, which is relatively weak and inefficient in long-range interaction. Transformers with self-attention, which has emerged as the dominant paradigm in natural language processing (NLP) [24], are thus designed to model long-range dependencies in computer vision fields [25,26]. For example, ViT [27] and BEiT [28] treat splitting patches in an image as words in one sentence, and perform non-local operations such as self-attention in a transformer [24]. In feature learning of HSI, spatial pixels or spectral bands are usually regarded as tokens of words for long-range attention perception. For example, He et al. [29] treated pixels in one input patch as tokens, and employed BRET [30] (namely HSI-BRET) to learn the global relationships between pixel sequences by multiple Transformer layers. Sun et al. [31] insert the spatial self-attention mechanism into a deep model with sequential spectral and spatial sub-modules (SSAN). To capture subtle discrepancies of the spectrum with sequence attributes, He et al. [32] performed spectral feature embedding by a pre-trained deep CNN. It regarded one spectral band as a word, and modeled sequential spectra relationships by a modified dense transformer. Hong et al. [33] instead applied band-wise self-attention with group-wise spectral embedding and proposed a SpectralFormer model. To capture semantic dependencies in both spatial and channel dimensions, dual attention network (DANet) [26] sets, two corresponding self-attention sub-modules were used, and the outputs were combined to perform 3-D attention perception and feature calibration. This 3-D attention is appropriate to spectral-spatial feature learning of HSI. Thus, Tang et al. [34] inserted a DANet like-wise attention block into a 3-D octave convolution network. Li et al. [35] embedded the sub-modules of DANet respectively into a double-branch 3-D dense convolutional network (DBDA).

HSI contains a wealth of information in both the spatial and spectral dimension; thus, data representation will prefer 3-D attention from multiple perspectives. However, the difficulty faced by 3-D multi-head attention is the increased burden in parameter optimization, computation, and storage. As a result, the existing self-attention-based methods in HSI classification mainly insert a one-head spectral-spatial attention into a 3-D deep model.

Self-attention can be seen as non-local filter [25] that captures long-range dependencies by weighted aggregation of features at all positions, while Hu et al. [36] verified that constraining the aggregation scope to a local neighborhood will be more reasonable for feature learning in visual recognition with less computation. Thus, in this paper, we focus on building a 3-D multi-head attention with local interaction and with the fewer possible parameters and lower computation cost. Beyond that, previous designs mainly capture attention by independent pairwise query-key interaction but ignore the contextual information among neighbor keys. Li et al. [37] proposed a contextual transformer (CoT) that contextually encoded input keys and concatenated them with queries to learn dynamic attention, which is much more efficient at boosting visual representation. On this basis, we present a dual multi-head contextual attention mechanism (DMuCA) for multi-view spectral-spatial neighborhood perception.

DMuCA decouples the spatial and spectral contextual attention into two sub-modules and builds multi-head attention on groups to control model complexity. In the spatial attention module (SaMCA), we treat pixels in the input as tokens and employ depth-wise convolution to contextualize the input keys in the pure spatial domain. Then, the keys are alternately concatenated with the queries and grouped to learn multiple neighborhood relationships. The learned multi-head local attention matrices are then broadcast across group channels to aggregate the neighborhood inputs. CoT [37] can be seen as a special case of SaMCA when the number of groups is equal to 1. As for spectral attention (SeMCA), which treats each channel as a token, the feature representation of one channel involves a bidirectional dimension. Consequently, the parameters and computation for multi-head attention matrices will increase exponentially, especially when it is exposed to inputs with enlarged spatial resolution. With the neighborhood consistency assumption, we therefore group the spatial pixels by equal-interval sampling and create multi-head attention on each neighbor block, to reduce the number of the parameter and avoid the computation burden. The main contributions of this paper are summarized as follows.

- By decoupling 3-D self-attention perception into two sub-modules, SaMCA and SeMCA, we build a dual contextual self-attention mechanism, DMuCA, for dynamic spatial and spectral attention calibration.
- To avoid parameter and computation increase, we group the representation of each token by evenly sampling, and capture multi-head self-attention with an alternate concatenation of the queries and keys on each group.
- Extensive experiments on three public HSIs demonstrate that our proposed DMuCA achieves excellent results over several state-of-the-art attention mechanisms with the same backbone.

The remainder of the paper is organized as follows. Section 2 reviews the general form of self-attention mechanisms. Section 3 details the proposed DMuCA with two well-designed sub-modules SaMCA and SeMCA for HSI classification. Extensive contrast and ablation experiments are conducted and discussed in Section 4. Section 5 draws conclusions and presents a brief outlook on future work.

2. Preliminaries

Humans can focus rapidly on regions of interest to perceive an image [38]. The attention mechanism simulates human vision and guides the system to ignore irrelevant content and focus on the important regions. In this section, we will review the general architecture of attention mechanisms and their multi-head patterns.

2.1. Attention Mechanism

The attentional mechanism contains two main aspects: determine the important parts of the input and allocate limited data processing resources to the regions [38]. Given an input $\mathbf{X}$, this process can be formulated as,

$$\mathbf{Z} = f(g(\mathbf{X}), \mathbf{X}), \tag{1}$$

where $g(\mathbf{X})$ generates attention weights which are generally measured by neighborhood relationships. The resulting weights are then allocated to $\mathbf{X}$ by function $f(g(\mathbf{X}), \mathbf{X})$. $\mathbf{Z}$ is the output of the attention layer. For the local self-attention [36], a transformation layer first maps input $\mathbf{X}$ to Query ($\mathbf{Q}$), Key ($\mathbf{K}$), and Value ($\mathbf{V}$). The self-attention weights are then calculated by dot-product between the $\mathbf{Q}$ and $\mathbf{K}$,

$$g(\mathbf{X}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right), \tag{2}$$

where a scaling factor $\sqrt{d_k}$ and softmax operation are employed to normalize the weights. Finally, the attention weights are assigned to the corresponding elements of the value to yield the output.

The above method captures self-attention through pairwise query-key interaction in isolation, ignoring neighborhood keys' rich contextual information. CoT [37] contextualizes the keys by performing local convolution and concatenates them with queries to learn attention weights (see Figure 1b), which can be defined as,

$$g(\mathbf{X}) = \text{Conv}(\sigma(\text{Conv}([\mathbf{Q}, \mathbf{K}], \mathbf{W}_1)), \mathbf{W}_2), \tag{3}$$

where σ is a activation function (eg., ReLU). The keys here are obtained by a 2-D convolution and the query-key interaction is achieved through two consecutive 1×1 convolutions. The output of the attention layer is then given by

$$\mathbf{Z} = g(\mathbf{X}) \otimes \mathbf{V}, \tag{4}$$

where $\otimes$ denotes local matrix multiplication.

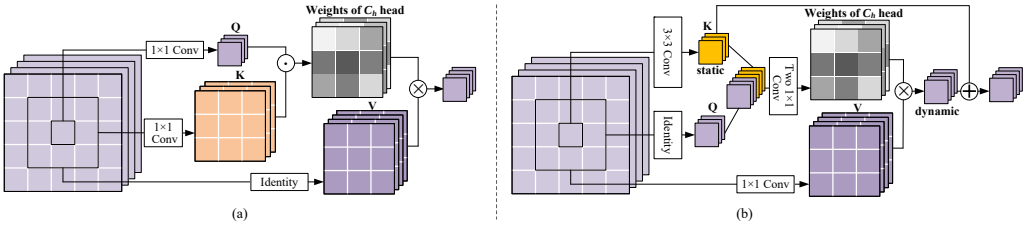

Figure 1. The multi-head structures of (**a**) local relation self-attention block and (**b**) CoT block. $\odot$ denotes dot product, $\otimes$ denotes local matrix multiplication with channel sharing, and $\oplus$ denotes element-wise sum.

2.2. Multi-Head Self-Attention

A single self-attention is coarse in mining complex relationships among neighbors in visual data. The multi-head self-attention mechanism (MHSA) [24] produces attention blocks from multiple feature subspaces to enrich the relationships. Each i-th subspace owns its $\mathbf{Q}_i$, $\mathbf{K}_i$, and $\mathbf{V}_i$ projected by the corresponding learnable parameters $\mathbf{W}_{q_i}$, $\mathbf{W}_{k_i}$, and $\mathbf{W}_{v_i}$. The obtained attention matrices are then concatenated together to aggregate the neighborhood elements in a weighted manner. For the local multi-head self-attention [36], depth-wise local dependency measurement and group weight sharing are used to perform multi-head attention with relatively fewer parameters and FLOPs (see Figure 1a). The process can be written as,

$$\mathbf{Q} = \text{Conv}_Q(\mathbf{X}), \mathbf{K} = \text{Conv}_K(\mathbf{X}), \mathbf{V} = \text{Identity}(\mathbf{X}), \tag{5}$$

where one channel in $\mathbf{Q}$ and $\mathbf{K}$ are exploited for generating one head of attention. All of the C_h aggregation weights for $\mathbf{X}_{ij}$ with $k \times k$ scope can be performed as below, if dot-product is used here for composability measurement,

$$g(\mathbf{X}_{ij}) = \text{DepthConv}\left(\mathbf{K}_{k \in \mathcal{N}(x_{ij})}, \mathbf{Q}_{ij}\right), \tag{6}$$

where $\mathcal{N}(x_{ij})$ denotes the $k \times k$ neighbors of the spatial pixel x_{ij}. The kernel of depth-wise convolution here comes from $\mathbf{Q}_{ij}$, and $g(\mathbf{X}_{ij}) \in \mathbb{R}^{k \times k \times C_h}$. Finally, the C_h multi-head attention weights that shared by C/C_h channels are allocated to the neighbors of $\mathbf{X}_{ij}$ for information aggregation. To enhance the contextual representation of keys, CoT generates multi-head attention weights by two consecutive 1×1 convolution layers, which can be written as,

$$g(\mathbf{X}_{ij}) = \mathrm{Conv}(\sigma(\mathrm{Conv}([\mathbf{Q}_{ij}, \mathbf{K}_{ij}], \mathbf{W}_1)), \mathbf{W}_2). \tag{7}$$

Figure 1b shows the reshaped multi-head attention weights $g(\mathbf{X}_{ij})$ of x_{ij} that is in size of $k \times k \times C_h$.

Self-attention performs feature filtering and aggregation with dynamic kernels generated by neighborhood interactions. Cordonnier et al. [39] empirically confirms that the attention layer with enough heads can express any convolutional layer as a special case. HSI recodes spatial samples as a spectral sequence with extremely high resolution, which is significantly superior at identifying ground objects with a subtle distinction or variation. Meanwhile, however, data structures become complex as the information increases. For feature learning of HSI as a 3-D cube data, it is preferable to capture multi-head self-attention from both spatial and spectral domains with contextual key interaction, if there is not much increase in the number of parameters and computation cost. Encoding spectral and spatial neighbors in a dynamic and decoupling fashion can be practical and effective at perceiving regions of interest and extracting discriminative features. Accordingly, we propose a dual multi-head contextual self-attention network (DMuCA) for HSI classification.

3. Proposed Method

The proposed framework of DMuCA network is illustrated in Figure 2. We build a plug-and-play 3-D attention block to guide a deep convolutional network focusing on spectral-spatial regions of interest. CoT [37] integrates both neighborhood-enriched contextual information and self-attention to enhance feature learning with dynamic local perception. We inherit the advantage of CoT and decouple the spatial and spectral neighborhood interaction into two separate blocks, SaMCA (see Figure 3) and SeMCA (see Figure 4). The former allocates more attention to important spatial regions, while the latter acts as a weighted aggregation of neighborhood spectral bands or attention-based feature recombination. To take full advantage of the two contextual information from different dimensions, we place the two blocks parallel to one another, and fuse their results by element-wise weighting summation. Thus, the output of DMuCA can be defined as,

$$\mathbf{F} = \beta \mathbf{F}^{\mathrm{spe}} + (1 - \beta)\mathbf{F}^{\mathrm{spa}}, \tag{8}$$

where $\mathbf{F}^{\mathrm{spe}}$ and $\mathbf{F}^{\mathrm{spa}}$ are the output of SeMCA and SaMCA, respectively. $\beta \in [0, 1]$ is a weighting factor that can be learned when model training, with an initial value set as 0.5.

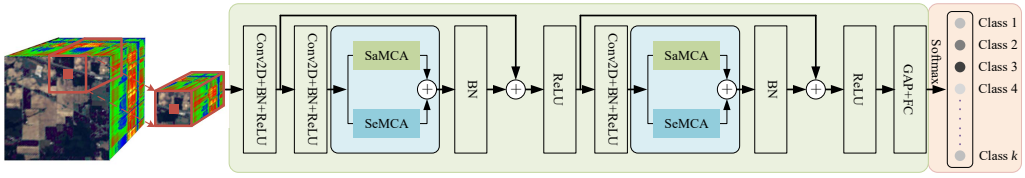

Figure 2. The overall architecture of the proposed DMuCA network, where SaMCA and SeMCA are plugged parallel into a ResNet with two residual blocks.

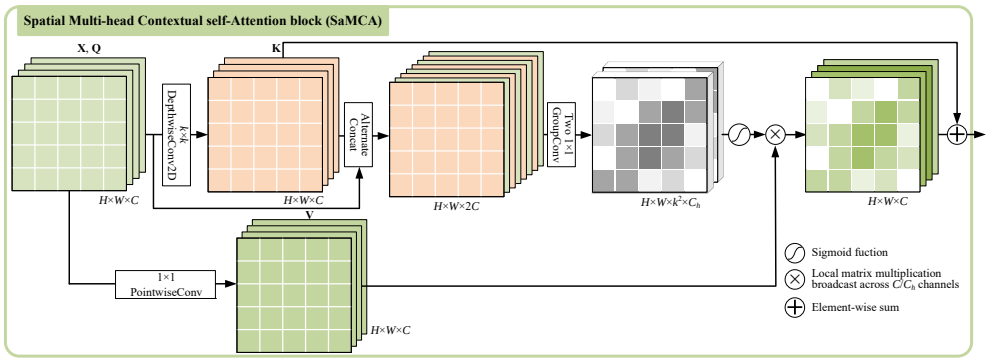

Figure 3. Architecture of the spatial multi-head contextual self-attention (SaMCA).

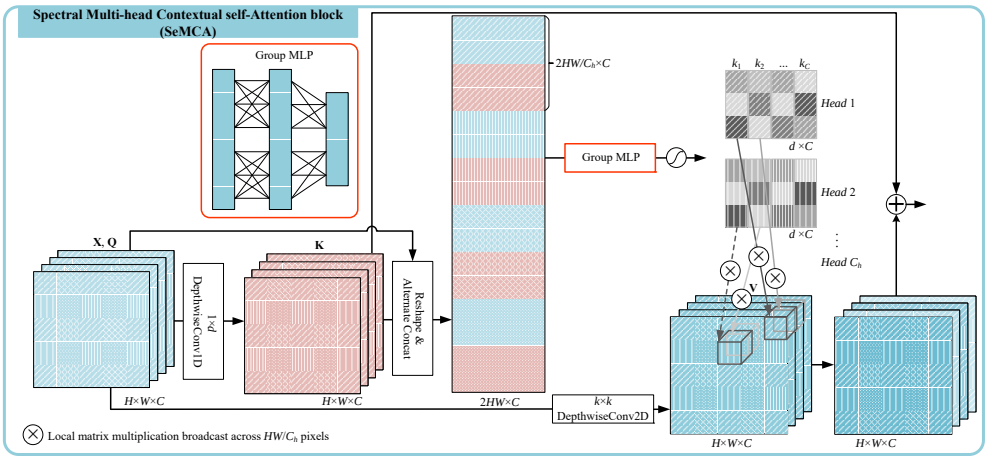

Figure 4. Architecture of the spectral multi-head contextual self-attention (SeMCA).

The integrated DMuCA can alternate any standard convolution filter. For HSI classification, we replace some convolutional layers in a ResNet backbone with our proposed DMuCA module, to strengthen the model with spectral-spatial contextual interdependencies and attention perception. As illustrated in Figure 2, the backbone contains two residual blocks, and one of the convolution layers in each block is replaced by a DMuCA module. All of the other convolution layers are followed by batch normalization and ReLU activation layer. A global average pooling (GAP) and FC layer are performed to integrate the global spatial information and to project the learned feature into label space for probability prediction of classification. Table 1 shows the detailed model structure of the DMuCA network, taking the IN dataset as an example.

Table 1. The detailed construction of our DMuCA network, a case on IN dataset.

Layer/Block	Kernel Type	Kernel Size	Input Sizes	Output Sizes
Conv1	Conv2D	3×3	$(200, 11, 11)$	$(64, 11, 11)$
Res_Block1	Conv2D	3×3	$(64, 11, 11)$	$(64, 11, 11)$
	DMuCA	$k = 5, d = 9$	$(64, 11, 11)$	$(64, 11, 11)$
Res_Block2	Conv2D	3×3	$(64, 11, 11)$	$(64, 11, 11)$
	DMuCA	$k = 5, d = 9$	$(64, 11, 11)$	$(64, 11, 11)$
GAP&FC	-	-	$(64, 11, 11)$	$(1, 16)$

3.1. Spatial Multi-Head Contextual Self-Attention

The SaMCA block aims to assign appropriate weights to the spatial neighborhood pixels and aggregates them for presentation of the central pixel. Formally, given the input $\mathbf{X} \in \mathbb{R}^{S\times S\times C}$, where $S \times S$ denotes the spatial dimension and C is the channel number. We treat each pixel as an independent token and transform $\mathbf{X}$ into queries $\mathbf{Q} = \text{Identity}(\mathbf{X})$ and values $\mathbf{V}^{\text{spa}} = \text{PointConv}(\mathbf{X}, \mathbf{W}_v)$ with kernel $\mathbf{W}_v$ of size 1×1. For absolute mining of spatial relation with relatively few parameters, we transform $\mathbf{X}$ into keys by depth-wise convolution with kernel $\mathbf{W}_k$ of size $k \times k$, defined as $\mathbf{K}^{\text{spa}} = \text{DepthConv}(\mathbf{X}, \mathbf{W}_k)$.

In HSI, neighbor bands contain similar spatial distribution but certain noise interference. So in order to avoid weight allocation suffering from noise interference, we alternately concatenate the queries and keys, and then group them to capture the local relationships for multi-head attention. This processing can be defined as

$$\tilde{g}_{\text{spa}}(\mathbf{X}) = \text{GConv}(\sigma(\text{GConv}(\text{Alter}[\mathbf{Q}, \mathbf{K}^{\text{spa}}], \mathbf{W}_1)), \mathbf{W}_2), \tag{9}$$

where $\text{Alter}[\cdot]$ and $\text{GConv}(\cdot)$ denote the alternate concatenation and group convolution operation for short, respectively. Two consecutive group convolutions with kernel of size 1×1 and group number of C_h are set here to generate multi-head attention weights. Then, the spatial weights are normalized by Softmax function as

$$g_{\text{spa}}\left(\mathbf{X}_{ij}\right)_{t,:,:} = \text{Softmax}\left(\tilde{g}_{\text{spa}}\left(\mathbf{X}_{ij}\right)_{t,:,:}\right). \tag{10}$$

$g_{\text{spa}}\left(\mathbf{X}_{ij}\right)_{t,:,:} \in \mathbb{R}^{k\times k}$ represents the attention weight of the t-th head ($t \in [1, C_h]$) for $\mathbf{X}_{ij}$ with neighbor scope of $k \times k$. The normalized attention weights $g_{\text{spa}}(\mathbf{X}) \in \mathbb{R}^{H\times W\times k^2 C_h}$ are then allocated to the corresponding elements of $\mathbf{V}^{\text{spa}} \in \mathbb{R}^{S\times S\times C}$ in a C/C_h channel-sharing manner, and get the output feature map $\mathbf{Z}^{\text{spa}} \in \mathbb{R}^{H\times W\times C}$ with dynamic weighted aggregation

$$\mathbf{Z}^{\text{spa}} = g_{\text{spa}}(\mathbf{X}) \otimes \mathbf{V}^{\text{spa}}, \tag{11}$$

where $\otimes$ denotes local matrix multiplication broadcasting across C/C_h channels. The channel sharing here can reduce the number of model parameters and facilitate memory scheduling on the GPU for efficiency.

The keys $\mathbf{K}^{\text{spa}}$ generated by depth-wise convolution, which shares weights in the spatial domain, can be seen as a static context, while pixel-wise self-attention is in a dynamic fashion. The former is adept at spatial translation invariance while the latter can adaptively capture aggregation weights. To combine both of the advantages, we fuse the static context from $\mathbf{K}^{\text{spa}}$ into the dynamic contextual representation $\mathbf{Z}^{\text{spa}}$ and get the final output $\mathbf{F}^{\text{spa}}$ of SaMCA, which focus more on rich spatial information,

$$\mathbf{F}^{\text{spa}} = \mathbf{Z}^{\text{spa}} + \mathbf{K}^{\text{spa}}. \tag{12}$$

3.2. Spectral Multi-Head Contextual Self-Attention

The far abundant spectral information of pixels in HSI, compared with the color (RGB) dataset, brings a greater challenge to finer spectral neighborhood perception. Therefore, we design a SeMCA block (see Figure 4) to mine spectral neighborhood information from multiple perspectives. Similarly, given the input feature map $\mathbf{X} \in \mathbb{R}^{S\times S\times C}$ as the same in SaMCA block, we treat each channel or band as a token. The queries ($\mathbf{Q}$), keys ($\mathbf{K}^{\text{spe}}$), and values ($\mathbf{V}^{\text{spe}}$) are defined respectively as

$$\begin{aligned} &\mathbf{Q} = \text{Identity}(\mathbf{X}), \\ &\mathbf{K}^{\text{spe}} = \text{DepthConv1D}(\mathbf{X}, \mathbf{W}_k), \\ &\mathbf{V}^{\text{spe}} = \text{DepthConv2D}(\mathbf{X}, \mathbf{W}_v). \end{aligned} \tag{13}$$

Specifically, 1-D depth-wise convolution with the kernel of size $1 \times d$ is introduced here to contextualize the keys. This means each pixel owns its convolution kernel, which helps

prevent interference between groups of samples and provides diversity to the followed multi-head self-attention. $\mathbf{X}$ is projected to values $\mathbf{V}^{\text{spe}}$ by a spatial transformation layer, a 2-D depth-wise convolution layer with a kernel of size $k \times k$.

In spectral attention, each token is featured from one two-dimensional feature map, which will sharply consume the memory if generating multi-head attention matrices by a regular convolution. Meanwhile, the parameters and computation cost will exponentially increase with the increase in spatial resolution. Fortunately, there is a fairly strong neighborhood consistency in the spatial domain. Thus, we evenly sample multiple groups of data (filled with different textures in Figure 4) to represent each channel and generate multi-head attention on the sampling groups. More specifically, pixels sampled from both queries $\mathbf{Q}$ and keys $\mathbf{V}^{\text{spe}}$ in the same location are concatenated together to learn one head of attention weights by an MLP, which can be written as

$$\tilde{g}_{\text{spe}}(\mathbf{X})_t = \text{FC}\Big(\sigma\Big(\text{FC}\Big(\Big[\mathbf{Q}_{\Delta_t,:}, \mathbf{K}^{\text{spe}}_{\Delta_t,:}\Big], \mathbf{W}_1\Big)\Big), \mathbf{W}_2\Big), \tag{14}$$

where $\Delta_t \in \mathbb{Z}^2$ ($t \in [1, 2, \cdots, C_h]$) refers to the t-th set of sampling offsets. MLP with two fully connected (FC) layers are employed for relevance learning of neighborhood channels. $\mathbf{W}_1 \in \mathbb{R}^{2HW/C_h \times HW/C_h}$ and $\mathbf{W}_2 \in \mathbb{R}^{HW/C_h \times d}$ represent the corresponding embedding matrices of FC layers, where d is the neighbor scope for channel aggregation. The final multi-head spectral attention weights $g_{\text{spe}}(\mathbf{X}) \in \mathbb{R}^{dC_h \times C}$ for all the channels can be defined as

$$g_{\text{spe}}(\mathbf{X}) = \text{Softmax}\Big(\text{GMLP}\Big(\text{Alter}\Big[\mathbf{Q}_{\Delta,:}, \mathbf{K}^{\text{spe}}_{\Delta,:}\Big], \mathbf{W}\Big)\Big), \tag{15}$$

where Δ denotes all of the sampling offsets stored in groups. GMLP denotes a grouped MLP where the group number equals C_h, the head number in self-attention. Note that the aggregation weights in each head are per channel normalized by the Softmax function.

In order to capture the spectral information from multiple perspectives without an increase in storage, we share the attention matrix in one head only to the corresponding spatial sampling set. This means that each head of attention weighted aggregate a specific set of neighborhood pixels. Thus, weighted aggregation of the neighbors of the i-th channel on the sub-sampling set Δ_t can be formulated as

$$\mathbf{Z}_{\Delta_t, i} = \text{Conv}\Big(\mathbf{V}^{\text{spe}}_{\Delta_t, \mathcal{N}(c_i)}, g_{\text{spe}}(\mathbf{X})_{(t-1)d+1:td, i}\Big), \tag{16}$$

where $\mathcal{N}(c_i) \in [i - d/2 : i + d/2]$ refers to a set which saved d offsets of neighbors around the i-th channel, $i = 1, 2, \cdots, C$. Spatial samples in Δ_t share the same filter kernel $g_{\text{spe}}(\mathbf{X})_{(t-1)d+1:td, i}$. The multi-head attention calibration in the spectral domain can be implemented on diverse sub-sampling sets, which are staggered in local space as different texture marks in Figure 4.

Similar to the SaMCA block, we fuse the static spectral context $\mathbf{K}^{\text{spe}}$ with the dynamic contextual representation $\mathbf{Z}^{\text{spe}}$, and obtain the final spectral local perception $\mathbf{F}^{\text{spe}}$,

$$\mathbf{F}^{\text{spe}} = \mathbf{Z}^{\text{spe}} + \mathbf{K}^{\text{spe}}. \tag{17}$$

4. Experiments

In this section, we conduct a comprehensive set of experiments to verify the effectiveness of DMuCA, including ablation studies for the major modules, sensitivity analysis of the hyperparameters, and comparison with some state-of-the-art classification models.

4.1. Datasets and Experimental Setup

To verify the stability of our model, we conduct experiments on three real datasets that come from different sensors and have diverse spatial resolutions.

The Indian Pines (IP) dataset, collected by the AVIRIS sensor over northwest Indiana, consists of 145 × 145 pixels with a spatial resolution of 20 m per pixel, and 200 spectral bands

after 20 noisy ones are removed due to atmospheric absorption or low SNR. The available ground truth contains 16 classes. Figure 5a,b shows its false-color image and the ground-truth map, respectively.

The University of Pavia (UP) dataset was collected by the ROSIS-03 sensor over the University of Pavia. It consists of 610 × 340 pixels with a spatial resolution of 1.3m per pixel, and 103 spectral bands after removing 12 noisy bands. The available ground reference contains 9 classes of interest. Figure 5c,d shows its false-color image and the ground-truth map, respectively.

The University of Houston (UH) dataset was acquired by the ITRES CASI-1500 sensor over the University of Houston campus and the neighboring urban area. The data consist of 349 × 1905 pixels with a spatial resolution of 2.5 m and 144 bands with a wavelength ranging from 364 nm to 1046 nm. The corresponding ground truth map consists of 15 types of land cover. Figure 6 shows its false-color image and the ground-truth map.

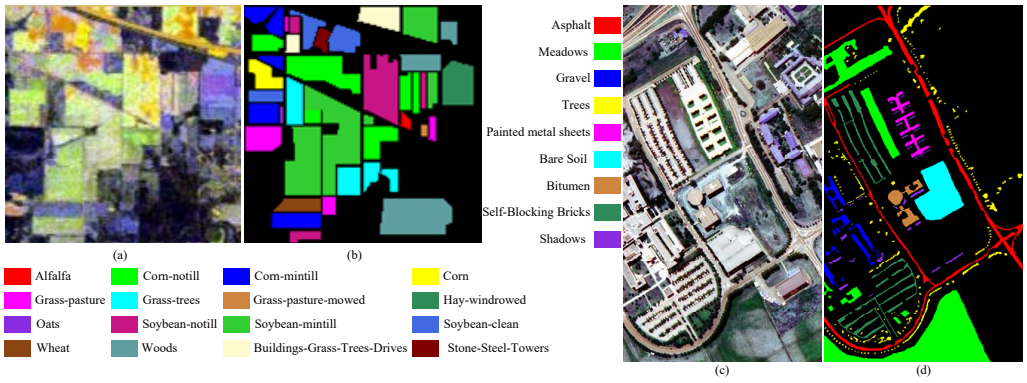

Figure 5. The false-color image (**a**) and the ground truth (**b**) for the IN dataset; The false-color image (**c**) and the ground truth (**d**) for the UP dataset.

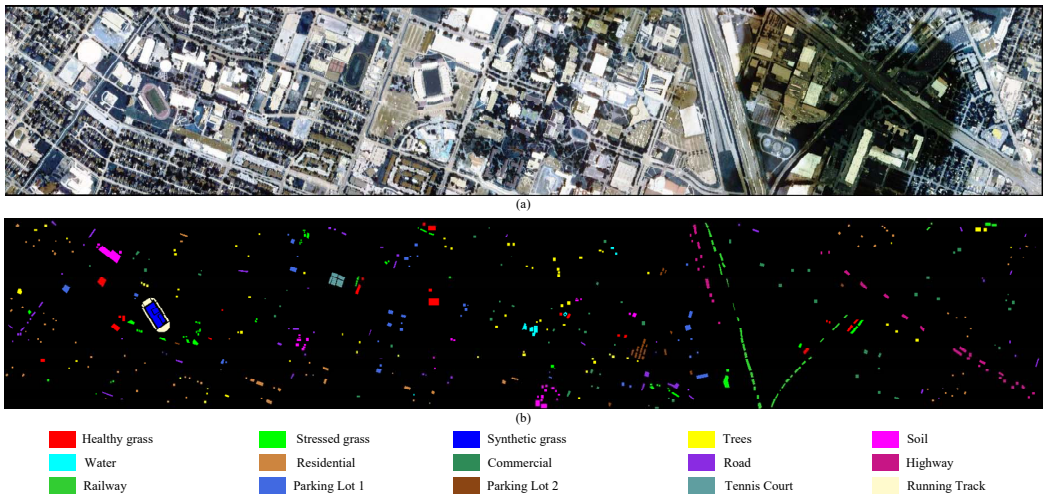

Figure 6. The false-color image (**a**) and the ground truth (**b**) for the UH dataset.

For the experiments that do not concern the effect of training sample size on model performance, we all select randomly 10%, 3%, 3% of samples per class from the ground-reference data for model training on IP, UP, and UH datasets, respectively, and the remaining samples are exploited for model testing. Before feature learning, the original HSI data is firstly normalized to [0, 1] for dimensionless transformation and acceleration of model optimization. The average accuracy (OA), overall accuracy (AA), and kappa statistic (κ) are adopted to evaluate the classification efficiency, and we provide OA results as a function of the parameters to be analyzed. All of our results are reported as the mean of ten runs. For model training, the cross-entropy loss function is used to supervise the

model prediction. Stochastic gradient descent (SGD) is adopted as the optimizer to update the model parameters with the momentum of 0.9 and the weight decay of 1×10^{-4}. Furthermore, we train models on the dataset with a batch size of 32 for 100 epochs and set the learning rate to 0.005. All of the experiments are implemented on the PyTorch platform using a workstation with i9-10900K CPU and an NVIDIA GeForce RTX 3090 graphics card. The code of our model is available at: https://github.com/mrpokere/DMuCA (accessed on 23 December 2021).

4.2. Model Analysis

In this part, we first verify the effect of the proposed components in our proposed DMuCA network by ablation study, and then analyze the model robustness at different parameter settings.

4.2.1. Ablation Study

In DMuCA, three main components are designed for discriminative feature learning, including grouping representation of tokens, contextual self-attention from spatial and spectral dimension, and fusion of features from static and dynamic perception. This part discusses the importance of different components in DMuCA through several sets of ablation studies. In addition, we try to verify the effectiveness of DMuCA in key bands perception.

Table 2 shows the model performance as the number of spatial attention heads ranging from 1 to 64, where "1" means single self-attention with no channel grouping and "64" means each channel generates an attention matrix. It can be seen that single self-attention produced over all the input is 1.3% worse than the best setting for the UH dataset, and is also slightly worse for IN and UP datasets. Meanwhile, the parameters and FLOPs decrease gradually as the group number increases. We finally set 16, 32, and 32 spatial self-attention heads for IN, UP, and UH datasets, respectively. As the same way in Table 3, we vary the number of spectral attention heads, where "1" means single self-attention with no pixel grouping or sampling and "Pixel-wise" means each pixel produces its own attention matrix. It shows a similar trend in both accuracy and computation cost. We choose the best setting, 25, 49, and 25 spatial self-attention heads, for IN, UP, and UH datasets, respectively.

Table 4 demonstrates the influence of the two self-attention sub-blocks on the classification accuracy when two convolutional layers in ResNet backbone (Base Model) are replaced respectively by just SaMCA, SeMCA, and by the final DMuCA block. The results show that the OAs are improved when we replace the convolutional layers by either SaMCA or SeMCA block, except for a slightly lower result when inserting only SeMCA for classification of the IN dataset. This is mostly due to its lower spatial resolution that exists with many mixed pixels, while SeMCA may amplify the spectral noise. However, it is worth noting that DMuCA achieves significant performance gains with a combination of the spatial and spectral self-attention blocks, illustrating the effectiveness of DMuCA in feature learning.

Table 2. Ablation study with different numbers of channel grouping and corresponding spatial attention heads ranging from 1 to 64. The best results are highlighted in bold font.

	Heads	1	2	4	8	16	32	64
IN	OA (%)	97.37	97.39	97.42	97.62	**97.79**	97.65	97.51
	Param (M)	1.378	1.362	1.353	1.349	1.347	1.346	**1.346**
	FLOPs (M)	28.14	27.15	26.66	26.41	26.29	26.23	**26.21**
UP	OA (%)	99.04	99.06	99.06	99.07	99.10	**99.18**	99.16
	Param (M)	3.655	3.638	3.630	3.626	3.624	3.623	**3.623**
	FLOPs (M)	39.88	38.03	37.11	36.65	36.43	36.32	**36.27**
UH	OA (%)	92.12	92.49	92.78	92.91	93.37	**93.49**	93.14
	Param (M)	2.239	2.223	2.214	2.210	2.208	2.207	**2.207**
	FLOPs (M)	33.51	32.13	31.44	31.09	30.92	30.84	**30.81**

Table 3. Ablation study with different numbers of channel grouping and corresponding spatial attention heads. The best results are highlighted in bold font.

	Heads	1	9	25	49	81	Pixel-Wise
IN	OA (%)	96.55	97.57	**97.79**	97.67	97.66	97.63
	Param (M)	1.443	1.345	1.347	1.337	1.348	**1.331**
	FLOPs (M)	29.33	26.22	26.29	25.95	26.32	**25.75**
UP	OA (%)	99.09	99.09	99.11	**99.18**	99.11	99.10
	Param (M)	3.985	3.625	3.597	3.623	3.598	**3.590**
	FLOPs (M)	47.87	36.36	35.46	36.32	35.49	**35.23**
UH	OA (%)	90.74	92.95	**93.49**	92.60	92.59	92.53
	Param (M)	2.416	2.236	2.207	2.197	2.208	**2.195**
	FLOPs (M)	37.52	31.74	30.84	30.51	30.87	**30.45**

Table 4. Effect of spatial and spectral attention in DMuCA (OA%). ✓ denotes the corresponding attention block present in DMuCA. The best results are highlighted in bold font.

Spatial	Spectral	IN	UP	UH
✓		97.62	99.06	92.95
	✓	97.13	98.78	92.38
✓	✓	**97.79**	**99.18**	**93.49**
Base Model		97.19	98.00	90.23

Table 5 reports the performances of utilizing the static and dynamic contextual information in the DMuCA network. Here, the solely static context means each of the two convolutional layers in ResNet backbone is replaced by a set of parallel 2-D and 1-D depth-wise convolution, while the dynamic context means no skip connections are set between the static Keys and the calibrated dynamic features. The results indicate that our model with static or dynamic context alone achieves higher performance than the baseline on the three datasets, while static and dynamic contexts complement each other and their fusion can further enhance the classification accuracy.

Table 5. Effects of different methods on contextual information exploration (OA%). The best results are highlighted in bold font.

Feature Type	IN	UP	UH
Static Context	97.37	99.10	93.02
Dynamic Context	97.68	99.04	92.86
DMuCA	**97.79**	**99.18**	**93.49**
Base Model	97.19	98.00	90.23

To reveal DMuCA in key bands perception, we further perform HSI classification by DMuCA with full or selected bands (reference the results from [40,41]) as the input, and compare it with the ResNet base (Base). The results in Table 6 show that DMuCA achieves a significant increase over the base when performing classification with the full spectral bands, and a slight increase with the selected bands. This illustrates that DMuCA can effectively focus on important bands, and band selection limits its advantages, even with a slight increase. Besides, band selection will result in some loss of information, while DMuCA can try to maximize all the information and achieve a 1% increase.

Table 6. Classification performance of DMuCA with full or selected bands as the input (OA%). The best results are highlighted in bold font.

Methods	Bands (#)	IN (15)	UP (14)	UH (22)
Base	Full bands	97.19	98.00	90.23
	Selected bands	95.68	98.11	92.25
DMuCA	Full bands	**97.79**	**99.18**	**93.49**
	Selected bands	95.87	98.64	92.36

4.2.2. Parameter Analysis

In our proposed DMuCA network, the main parameters that affect the model performance are the input patch size for neighborhood information assistance and the number of DMuCA modules for spectral-spatial contextual attention-based feature extraction.

As a pixel-wise classification task, the input patch size determines how many spatial neighborhoods are used to assist feature extraction. We verify the model performance when the patch size ranges from 7 to 19. The results in Table 7 show that a larger input patch size is beneficial to capture contextual information and thus significantly improve the recognition performance. However, the statement 'the larger the better' is not true in this case. The best scale should provide sufficient spatial texture, while also avoiding much noise interference. It can be found that the more noise interfered with the dataset, the less suitable it was for a larger patch size. Finally, we set IN, UP, and UH datasets with input sizes of 11, 15, and 13, respectively.

Table 7. OAs (%) of DMuCA network with different size of input patches. The best results are highlighted in bold font.

Patch Size	7	9	11	13	15	17	19
IN	96.60	97.27	**97.79**	97.69	96.31	95.36	94.35
UP	98.19	98.78	99.15	99.12	**99.18**	99.03	98.92
UH	90.66	92.29	93.15	**93.49**	92.67	92.17	91.31

To improve the high-level contextual perception, more residual blocks are usually stacked to deepen the model. We further explore our DMuCA network with an increasing number of residual blocks and compare it with the ResNet backbone (Base Model). As shown in Table 8, only one residual block, or too shallow a network, has limited capability in high-level feature learning. However, this does not mean 'the deeper the better', particularly for HSI datasets with much obscure high-level semantic information but detailed shallow texture. UP and UH have more distinct spatial context information than the IN dataset; thus, they benefit more from deeper backbone networks, while four residual blocks cause the OA result of IN dataset to decrease. Our DMuCA network achieves the best results by a network with just two residual blocks, presenting a more significant advantage in shallow information mining.

Table 8. Comparison of our proposed DMuCA network and the ResNet backbone with increasing number of residual blocks (OA%). The best results are highlighted in bold font.

Block Number	DMuCA			Base Model		
	IN	UP	UH	IN	UP	UH
1	96.99	98.87	92.21	96.56	97.43	86.41
2	**97.79**	**99.18**	**93.49**	97.19	98.00	90.23
3	97.68	99.07	93.26	**97.70**	98.11	91.74
4	97.39	98.97	92.54	97.53	**98.32**	**92.04**

4.3. Comparison with the State-of-the-Art Methods

The key motivation of our proposed method is to construct an efficient spectral-spatial attention mechanism for discriminative feature learning. For fair performance evaluation, we create 10 groups of training sets by randomly sampling on ground truth without replacement and the resting samples from each set for cross-validation. We firstly compare our proposed DMuCA module with some classical attention mechanisms, such as squeeze-and-excitation (SE) [19], dual attention network (DANet) [26], and convolutional block attention module (CBAM) [22]. All of the attention modules are embedded respectively into the same ResNet base model for HSI classification. Table 9 shows the mean OAs and the standard deviations of the 10 set runs with the percentage of training samples per class ranging from 3% to 20%. We can see that our DMuCA achieves the best results all the time, with either limited or more training samples. These results assure that our model has good generalization ability in feature extraction.

We further compare our proposed DMuCA network with the traditional SVM method and some of the best deep learning-based methods, such as the DRNN model with LSTM [7] that see the patch block as a sequence, the spectral-spatial residual network (SSRN) with 3-D convolution [10], the deep feature fusion network (DFFN) with spectral-spatial fusion [11], the residual spectral-spatial attention network (RSSAN) [42] that insert CBAM into a residual network, the compact band weighting-based attention network (CBW) [20], the spectral-spatial attention network (SSAN) [31] that inserts the self-attention block into a 3-D convolution network, and the ResNet backbone (Base). To be fair, we set all the state-of-the-art frameworks with the same input patch size, while the other parameters involved in the competitors are set as provided in the corresponding references.

Table 9. Comparison of the DMuCA module with some classical attention mechanisms (OA%). The best results are highlighted in bold font.

Dataset	Training Set (#%)	SE	CBAM	DANet	DMuCA
IN	3	86.23 ± 1.11	86.43 ± 0.77	86.47 ± 1.40	**86.73 ± 1.71**
	5	92.13 ± 0.79	91.85 ± 1.05	92.63 ± 1.12	**92.82 ± 0.79**
	10	96.91 ± 0.55	96.75 ± 0.60	96.63 ± 0.46	**96.95 ± 0.44**
	15	97.54 ± 0.41	97.68 ± 0.35	97.64 ± 0.43	**97.83 ± 0.37**
	20	98.28 ± 0.27	98.32 ± 0.24	98.09 ± 0.38	**98.36 ± 0.26**
UP	3	98.91 ± 0.17	98.71 ± 0.29	98.67 ± 0.20	**98.95 ± 0.27**
	5	99.30 ± 0.13	99.26 ± 0.10	99.24 ± 0.09	**99.57 ± 0.06**
	10	99.63 ± 0.07	99.58 ± 0.05	99.60 ± 0.07	**99.80 ± 0.07**
	15	99.75 ± 0.06	99.74 ± 0.05	99.75 ± 0.06	**99.92 ± 0.03**
	20	99.79 ± 0.02	99.78 ± 0.04	99.79 ± 0.06	**99.94 ± 0.02**
UH	3	90.61 ± 3.09	91.28 ± 1.02	91.65 ± 1.23	**92.60 ± 0.68**
	5	94.68 ± 1.26	94.98 ± 0.60	95.13 ± 0.55	**95.43 ± 1.17**
	10	97.89 ± 0.54	97.96 ± 0.49	98.00 ± 0.58	**98.07 ± 0.41**
	15	98.85 ± 0.19	98.89 ± 0.27	98.86 ± 0.31	**98.93 ± 0.18**
	20	99.34 ± 0.29	99.02 ± 0.91	99.37 ± 0.24	**99.41 ± 0.20**

The mean accuracy and the standard deviation of the 10 set runs are reported in Tables 10–12, including the accuracy of each class and overall quantification from our proposed model and the competitors on the IN, UP, and UH datasets, respectively. Table 13 counts the computational FLOPs, parameters, and the running time of the corresponding experiments. The results indicate that the 3-D convolution-based SSRN and the attention-based models all present comparable results. Our DMuCA network outperforms all of them in classification accuracy, especially with comparatively small variance volatility. Besides, our method shows an outstanding performance of AAs, about a 2% increase over the ResNet base, and is encouragingly competitive on the classification of the UH dataset (a 3% significant increase over the base). RSSAN has a similar backbone to ours, and the information squeeze in CBAM is indeed a lightweight means of attention perception, containing the fewest training parameters and FLOPs. However, this comes somewhat at the expense of classification accuracy. Our method is not advantageous in computation cost, with about three times the running time over the ResNet base as reported in Table 13, but the increase is acceptable relative to improving classification accuracy, especially compared with other competitors.

Table 10. Testing results over IN dataset with 10% samples per class for model training. The best results are highlighted in bold font.

Class	SVM	RNN	SSRN	DFFN	RSSAN	CBW	SSAN	DMuCA	Base
1	59.67 ± 12.66	69.91 ± 9.21	**96.72 ± 2.80**	94.34 ± 2.27	85.29 ± 14.13	94.48 ± 3.59	86.99 ± 10.15	94.59 ± 6.06	95.16 ± 4.59
2	73.35 ± 1.42	83.30 ± 5.24	94.91 ± 2.35	92.66 ± 1.30	96.29 ± 0.54	95.07 ± 0.98	95.34 ± 1.49	**97.85 ± 1.16**	96.87 ± 1.33
3	65.47 ± 1.98	77.40 ± 6.51	91.81 ± 5.21	92.90 ± 2.28	94.37 ± 2.14	92.72 ± 3.35	95.31 ± 1.62	**97.32 ± 0.84**	96.81 ± 1.63
4	53.78 ± 7.47	88.10 ± 2.87	92.40 ± 4.14	92.78 ± 1.33	91.41 ± 3.98	95.78 ± 2.22	92.93 ± 3.61	96.41 ± 1.61	**97.18 ± 1.57**
5	86.83 ± 1.41	83.27 ± 2.14	96.57 ± 1.47	97.40 ± 1.15	96.14 ± 1.08	97.39 ± 1.29	97.78 ± 0.91	**98.07 ± 1.06**	97.04 ± 2.38
6	91.31 ± 1.02	93.98 ± 2.08	98.18 ± 0.78	95.98 ± 0.68	97.99 ± 1.57	96.57 ± 1.00	**98.28 ± 0.98**	97.33 ± 1.48	97.44 ± 1.18
7	70.25 ± 13.67	74.60 ± 11.29	91.66 ± 10.40	91.68 ± 5.42	87.53 ± 11.01	91.56 ± 5.89	81.29 ± 12.42	**96.65 ± 5.49**	91.67 ± 11.60
8	96.09 ± 0.98	97.53 ± 1.03	99.51 ± 0.72	99.63 ± 0.37	99.21 ± 0.73	99.48 ± 0.46	99.21 ± 0.48	**99.80 ± 0.36**	99.71 ± 0.31
9	33.05 ± 17.64	64.53 ± 8.64	84.83 ± 13.17	78.98 ± 17.38	84.27 ± 17.05	87.25 ± 9.57	78.03 ± 14.74	**96.20 ± 4.32**	81.4 ± 13.72
10	69.98 ± 1.72	84.97 ± 4.25	93.18 ± 2.64	92.10 ± 2.33	94.21 ± 1.10	93.52 ± 1.63	94.36 ± 2.04	**96.59 ± 1.58**	95.17 ± 1.22
11	75.72 ± 0.98	91.30 ± 3.03	95.23 ± 1.96	94.68 ± 0.40	97.33 ± 1.14	96.92 ± 0.74	97.27 ± 0.83	**98.28 ± 0.73**	97.46 ± 0.59
12	71.75 ± 2.73	79.28 ± 3.05	89.40 ± 4.57	92.50 ± 1.94	91.75 ± 2.11	88.01 ± 4.25	92.66 ± 2.10	**95.12 ± 2.05**	92.36 ± 2.45
13	93.30 ± 1.53	94.52 ± 2.20	99.01 ± 0.77	99.19 ± 0.49	97.32 ± 3.03	**99.52 ± 0.50**	98.43 ± 2.87	99.38 ± 0.66	99.19 ± 1.10
14	91.89 ± 1.15	93.35 ± 1.19	97.41 ± 1.19	96.46 ± 0.90	95.94 ± 1.40	**98.70 ± 0.90**	96.22 ± 1.00	96.46 ± 0.68	96.41 ± 0.62
15	57.48 ± 4.63	78.05 ± 6.38	**92.02 ± 3.82**	81.67 ± 4.13	81.70 ± 7.03	77.48 ± 4.49	82.05 ± 5.04	82.13 ± 3.27	84.00 ± 4.20
16	**94.01 ± 2.94**	78.78 ± 4.98	91.93 ± 6.14	91.11 ± 3.34	89.89 ± 3.13	90.57 ± 4.14	91.12 ± 5.82	93.80 ± 2.62	91.32 ± 2.76
OA (%)	77.64 ± 0.56	87.45 ± 2.39	95.02 ± 1.45	94.10 ± 0.55	95.41 ± 0.71	96.28 ± 0.72	95.91 ± 0.53	**96.95 ± 0.44**	96.26 ± 0.55
AA (%)	74.00 ± 1.67	83.30 ± 1.44	94.06 ± 2.48	92.75 ± 1.11	92.54 ± 1.49	94.65 ± 1.00	92.66 ± 2.00	**96.00 ± 0.79**	94.33 ± 1.66
Kappa (%)	74.52 ± 0.65	85.67 ± 2.72	94.30 ± 1.66	93.27 ± 0.63	94.76 ± 0.81	95.00 ± 0.84	95.37 ± 0.95	**96.52 ± 0.59**	95.73 ± 0.63

Table 11. Testing results over UP dataset with 3% samples per class for model training. The best results are highlighted in bold font.

Class	SVM	RNN	SSRN	DFFN	RSSAN	CBW	SSAN	DMuCA	Base
1	90.83 ± 0.65	95.72 ± 1.56	97.42 ± 0.46	99.03 ± 0.62	98.06 ± 0.83	99.05 ± 0.34	99.21 ± 0.22	**99.26 ± 0.36**	97.13 ± 0.71
2	94.22 ± 0.38	97.97 ± 0.53	99.12 ± 0.50	99.17 ± 0.23	99.46 ± 0.29	**99.72 ± 0.15**	99.43 ± 0.03	99.61 ± 0.21	99.67 ± 0.11
3	72.14 ± 1.62	82.48 ± 4.46	93.66 ± 1.49	96.64 ± 1.56	95.26 ± 1.57	92.17 ± 7.06	93.98 ± 1.62	**96.86 ± 1.67**	92.71 ± 2.63
4	91.41 ± 0.81	95.51 ± 0.66	96.39 ± 0.74	96.86 ± 0.60	93.98 ± 1.56	97.76 ± 0.95	**98.93 ± 0.27**	97.49 ± 1.20	95.64 ± 0.66
5	98.20 ± 0.74	98.43 ± 1.00	99.35 ± 0.24	98.75 ± 0.18	97.77 ± 1.09	99.62 ± 0.18	99.24 ± 0.05	**99.65 ± 0.11**	98.69 ± 0.68
6	82.28 ± 1.35	93.48 ± 1.90	97.14 ± 1.61	97.32 ± 0.85	98.96 ± 0.62	99.23 ± 0.8	99.83 ± 0.07	**99.87 ± 0.35**	99.47 ± 0.21
7	75.85 ± 3.58	89.57 ± 1.28	99.17 ± 0.26	98.48 ± 1.10	97.61 ± 2.47	99.04 ± 0.73	99.29 ± 0.37	**99.35 ± 0.84**	92.72 ± 2.20
8	80.39 ± 0.97	91.57 ± 1.65	92.89 ± 1.02	**97.68 ± 1.07**	96.23 ± 1.21	95.22 ± 2.29	96.10 ± 0.90	97.43 ± 0.80	96.76 ± 0.84
9	**99.83 ± 0.05**	92.63 ± 3.44	97.99 ± 0.74	97.27 ± 0.68	91.47 ± 2.55	97.60 ± 1.02	98.96 ± 0.85	98.10 ± 0.62	93.11 ± 1.33
OA (%)	89.53 ± 0.49	95.20 ± 1.17	97.64 ± 0.36	98.01 ± 0.45	98.32 ± 0.50	98.61 ± 0.47	98.82 ± 0.25	**98.95 ± 0.27**	97.98 ± 0.36
AA (%)	87.24 ± 0.64	93.04 ± 1.54	97.11 ± 0.30	98.13 ± 0.59	96.91 ± 0.79	97.73 ± 1.06	97.51 ± 0.37	**98.50 ± 0.47**	96.21 ± 0.65
Kappa (%)	86.11 ± 0.64	93.66 ± 1.53	96.86 ± 0.47	98.04 ± 0.60	97.85 ± 0.66	98.16 ± 0.62	97.83 ± 0.26	**98.61 ± 0.36**	97.33 ± 0.48

Table 12. Testing results over UH dataset with 3% samples per class for model training. The best results are highlighted in bold font.

Class	SVM	RNN	SSRN	DFFN	RSSAN	CBW	SSAN	DMuCA	Base
1	90.10 ± 3.64	90.38 ± 1.59	94.25 ± 2.49	82.68 ± 10.26	89.79 ± 2.37	93.13 ± 2.47	93.79 ± 2.54	**94.48 ± 3.96**	91.79 ± 2.23
2	95.82 ± 1.97	89.26 ± 3.75	96.04 ± 2.27	90.74 ± 7.68	94.50 ± 2.33	97.08 ± 1.51	97.57 ± 1.17	**98.28 ± 1.25**	90.38 ± 3.11
3	98.91 ± 0.36	95.00 ± 2.72	99.06 ± 1.13	98.17 ± 1.18	97.24 ± 3.04	98.81 ± 1.33	**99.48 ± 0.57**	97.22 ± 1.95	96.48 ± 2.55
4	92.43 ± 2.83	89.21 ± 2.15	96.29 ± 1.52	86.00 ± 8.38	89.05 ± 2.76	95.35 ± 2.92	**96.49 ± 0.77**	93.51 ± 2.84	91.55 ± 2.37
5	93.53 ± 1.20	94.12 ± 2.08	97.74 ± 0.90	98.34 ± 0.91	96.22 ± 1.29	97.42 ± 1.03	**98.47 ± 1.17**	98.18 ± 1.37	95.34 ± 1.12
6	89.46 ± 2.01	86.05 ± 4.55	**93.60 ± 3.60**	86.91 ± 4.68	83.00 ± 7.05	84.94 ± 5.39	89.52 ± 3.76	83.22 ± 5.75	89.59 ± 2.84
7	82.34 ± 2.01	85.94 ± 3.63	90.38 ± 2.13	89.37 ± 3.66	90.17 ± 1.85	**93.32 ± 1.89**	90.14 ± 2.38	92.79 ± 1.82	89.67 ± 2.43
8	67.09 ± 4.83	78.31 ± 6.06	86.30 ± 3.36	83.46 ± 4.89	86.26 ± 3.75	90.78 ± 2.98	89.88 ± 2.22	**91.45 ± 1.98**	79.34 ± 3.01
9	70.65 ± 2.85	81.84 ± 2.51	88.92 ± 5.30	83.44 ± 5.51	85.12 ± 3.15	88.69 ± 1.99	84.68 ± 3.42	**89.34 ± 2.33**	86.15 ± 1.65
10	77.13 ± 2.26	78.88 ± 3.80	85.65 ± 3.08	85.55 ± 2.79	87.82 ± 2.56	87.51 ± 3.03	89.20 ± 2.51	**89.41 ± 2.61**	86.39 ± 2.90
11	76.85 ± 3.02	83.53 ± 5.61	87.73 ± 3.37	88.76 ± 3.74	91.79 ± 0.98	95.03 ± 1.75	93.03 ± 2.84	**95.19 ± 2.13**	92.87 ± 1.78
12	66.43 ± 2.16	78.68 ± 4.31	76.70 ± 9.83	78.77 ± 4.72	85.84 ± 3.16	**89.63 ± 2.28**	87.25 ± 3.16	88.08 ± 3.47	87.47 ± 1.81
13	33.19 ± 7.75	80.58 ± 6.49	81.40 ± 8.63	84.70 ± 6.45	86.84 ± 3.88	94.49 ± 1.66	73.95 ± 4.13	**95.86 ± 3.50**	87.97 ± 3.18
14	93.93 ± 3.15	86.55 ± 5.90	99.18 ± 1.23	98.48 ± 0.94	97.72 ± 1.35	97.99 ± 0.04	97.84 ± 1.60	**99.31 ± 0.48**	96.58 ± 2.08
15	97.87 ± 0.35	89.92 ± 4.13	**97.91 ± 1.44**	96.79 ± 1.49	95.01 ± 3.09	96.96 ± 0.57	97.34 ± 0.60	94.10 ± 3.01	92.59 ± 2.26
OA (%)	82.21 ± 0.62	85.65 ± 2.81	91.09 ± 1.83	88.05 ± 2.45	91.20 ± 1.26	92.32 ± 1.85	92.12 ± 0.99	**92.60 ± 0.68**	89.82 ± 1.13
AA (%)	81.71 ± 0.69	85.88 ± 2.71	91.56 ± 1.92	88.88 ± 2.03	90.42 ± 0.98	92.18 ± 0.87	92.01 ± 1.04	**92.29 ± 1.50**	90.28 ± 0.99
Kappa (%)	80.76 ± 0.67	84.49 ± 3.04	90.36 ± 1.99	87.09 ± 2.65	89.41 ± 1.37	91.97 ± 0.86	91.39 ± 1.07	**91.70 ± 1.65**	89.00 ± 1.22

Table 13. The number of training parameters, computational FLOPs, and the running time of the corresponding experiments from the deep frameworks. The best results are highlighted in bold font.

Dataset	Evaluations	RNN	SSRN	DFFN	RSSAN	CBW	SSAN	DMuCA	Base
IN	Parameters (M)	0.23	0.34	0.38	**0.10**	2.76	2.10	1.32	1.29
	FLOPs (M)	27.54	235.54	45.43	**9.49**	11.41	59.96	25.52	23.45
	Training time (S)	54.94	104.81	66.78	47.79	34.12	46.34	64.17	**23.77**
	Testing time (S)	2.96	3.47	3.54	1.95	1.12	1.49	2.82	**0.89**
UP	Parameters (M)	0.19	0.19	0.38	**0.06**	2.75	3.78	3.58	3.53
	FLOPs (M)	42.83	225.83	84.48	**11.28**	16.12	62.76	34.81	31.03
	Training time (S)	88.94	187.86	89.24	52.53	75.64	57.62	71.91	**24.04**
	Testing time (S)	49.33	31.12	42.93	19.08	10.52	15.99	28.04	**8.37**
UH	Parameters (M)	0.21	0.25	0.38	**0.08**	2.76	2.88	2.18	2.15
	FLOPs (M)	34.83	236.01	63.48	**10.49**	12.40	62.39	30.07	27.30
	Training time (S)	22.01	73.65	24.19	20.31	29.01	18.89	21.22	**7.35**
	Testing time (S)	125.76	106.75	112.27	61.95	35.73	49.83	87.60	**28.43**

Figures 7–9 visualize the best classification maps from all the competitors on IN, UP, and UH datasets, respectively. Although SVM produces a noisy map, it keeps clear contour information. RNN, SSRN, DFFN, and RSSAN all overly smooth the boundaries, particularly apparent in the region of "Self-Blocking Bricks" on the UP dataset. CBW and SSAN do better in boundaries, which will benefit greatly from the special spectral attention in CBW and the self-attention-based feature calibration in SSAN. Our DMuCA module decouples the 3-D self-attention into two parallel blocks, and perceives key region of spectral and spatial feature from multiple perspectives. This structure performs well in capturing local relationships, and generate more accurate boundary location, achieving the expected effect.

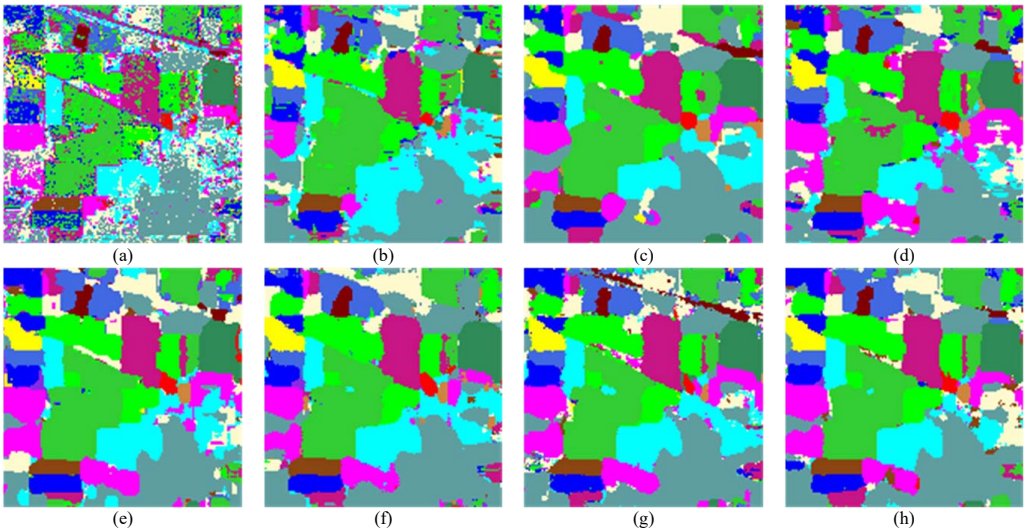

Figure 7. Classification maps of the IN data set obtained by: (**a**) SVM, (**b**) RNN, (**c**) SSRN, (**d**) DFFN, (**e**) RSSAN, (**f**) CBW, (**g**) SSAN, (**h**) DMuCA.

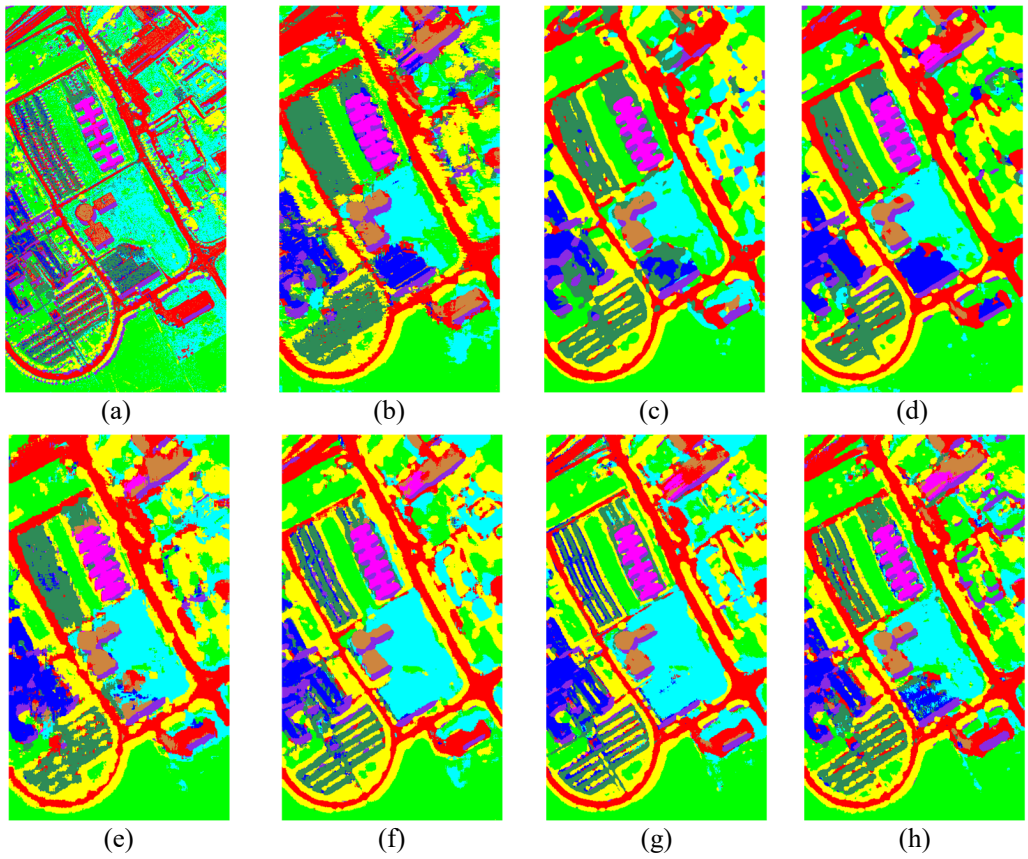

Figure 8. Classification maps of the UP data set obtained by: (**a**) SVM, (**b**) RNN, (**c**) SSRN, (**d**) DFFN, (**e**) RSSAN, (**f**) CBW, (**g**) SSAN, (**h**) DMuCA.

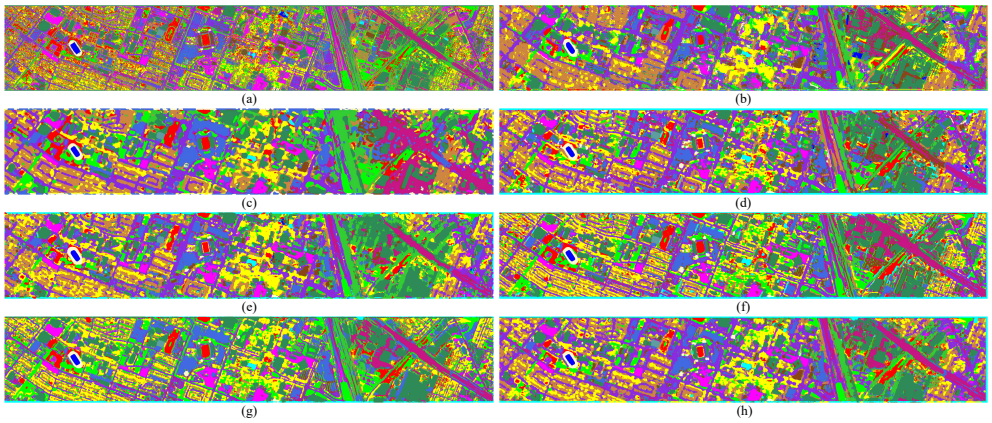

Figure 9. Classification maps of the UH data set obtained by: (**a**) SVM, (**b**) RNN, (**c**) SSRN, (**d**) DFFN, (**e**) RSSAN, (**f**) CBW, (**g**) SSAN, (**h**) DMuCA.

5. Conclusions

In this paper, we presented a dual multi-head contextual self-attention (DMuCA) network, which decouples the spectral-spatial contextual attention into SaMCA and SeMCA sub-modules, for contextual dependencies learning of HSI with fewer possible parameters and lower computation costs. The former serves to filter and aggregate the local information, while the latter leads to weighted aggregates of the neighborhood channels. Grouping tokens by channels or pixel sampling, and performing multi-head attention on each group effectively prevents the increase of parameters and computation burden, with no accuracy drop. A careful study of those proposed components demonstrates the effectiveness of discriminant feature learning and accurate classification.

Some questions still exist, such as why DMuCA achieves a noticeable improvement on ground object recognition of the UH dataset, but with a large deviation. Adaptive boundary perceptual location and neighborhood smoothing may possibly ameliorate the situation. Thus, we would like to build multi-scale attention on different semantic layers in the future, with the purpose of refining region perceptions and reducing noise interference.

Author Contributions: Conceptualization, M.L. and X.Y.; methodology, M.L., Q.H. and H.W.; software, Q.H.; validation, Q.H. and H.W.; data curation, Z.M.; writing—original draft preparation, Q.H.; writing—review and editing, M.L. and X.Y.; visualization, Q.H.; supervision, L.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Nos. 61901198, 61862031, 62066018); the Natural Science Basic Research Plan in Shaanxi Province of China (No. 2022JQ-704); and the Program of Qingjiang Excellent Young Talents, Jiangxi University of Science and Technology (No. JXUSTQJYX2020019).

Data Availability Statement: The IN and UP datasets are available at: http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes. The UH dataset is available at: https://hyperspectral.ee.uh.edu/. All these links last accessed on 23 June 2022.

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

References

1. Li, S.; Song, W.; Fang, L.; Chen, Y.; Ghamisi, P.; Benediktsson, J.A. Deep learning for hyperspectral image classification: An overview. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 6690–6709.
2. Rasti, B.; Hong, D.; Hang, R.; Ghamisi, P.; Kang, X.; Chanussot, J.; Benediktsson, J.A. Feature extraction for hyperspectral imagery: The evolution from shallow to deep: Overview and toolbox. *IEEE Geosci. Remote Sens. Mag.* **2020**, *8*, 60–88.
3. Ghamisi, P.; Yokoya, N.; Li, J.; Liao, W.; Liu, S.; Plaza, J.; Rasti, B.; Plaza, A. Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 37–78.
4. Lokman, G.; Çelik, H.H.; Topuz, V. Hyperspectral Image Classification Based on Multilayer Perceptron Trained with Eigenvalue Decay. *Can. J. Remote Sens.* **2020**, *46*, 253–271.
5. Zhou, P.; Han, J.; Cheng, G.; Zhang, B. Learning compact and discriminative stacked autoencoder for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 4823–4833.
6. Chen, Y.; Zhao, X.; Jia, X. Spectral–spatial classification of hyperspectral data based on deep belief network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 2381–2392.
7. Mou, L.; Ghamisi, P.; Zhu, X.X. Deep recurrent neural networks for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3639–3655.
8. Shi, C.; Pun, C.M. Multiscale superpixel-based hyperspectral image classification using recurrent neural networks with stacked autoencoders. *IEEE Trans. Multimed.* **2019**, *22*, 487–501.
9. Zhou, W.; Kamata, S.i.; Luo, Z.; Wang, H. Multiscanning Strategy-Based Recurrent Neural Network for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5521018.
10. 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.* **2017**, *56*, 847–858.
11. Song, W.; Li, S.; Fang, L.; Lu, T. Hyperspectral image classification with deep feature fusion network. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 3173–3184.
12. Wang, W.; Dou, S.; Jiang, Z.; Sun, L. A Fast Dense Spectral–Spatial Convolution Network Framework for Hyperspectral Images Classification. *Remote Sens.* **2018**, *10*, 1068.
13. Zhang, X.; Chen, S.; Zhu, P.; Tang, X.; Feng, J.; Jiao, L. Spatial Pooling Graph Convolutional Network for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5521315.
14. Zhu, L.; Chen, Y.; Ghamisi, P.; Benediktsson, J.A. Generative adversarial networks for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 5046–5063.
15. Cui, B.; Dong, X.M.; Zhan, Q.; Peng, J.; Sun, W. LiteDepthwiseNet: A Lightweight Network for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5502915.
16. Wang, J.; Huang, R.; Guo, S.; Li, L.; Zhu, M.; Yang, S.; Jiao, L. NAS-guided lightweight multiscale attention fusion network for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 8754–8767.
17. Liang, M.; Wang, H.; Yu, X.; Meng, Z.; Yi, J.; Jiao, L. Lightweight Multilevel Feature Fusion Network for Hyperspectral Image Classification. *Remote Sens.* **2021**, *14*, 79.
18. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June –1 July 2016; pp. 770–778.

19. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
20. Zhao, L.; Yi, J.; Li, X.; Hu, W.; Wu, J.; Zhang, G. Compact Band Weighting Module Based on Attention-Driven for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 9540–9552.
21. Wang, L.; Peng, J.; Sun, W. Spatial–spectral squeeze-and-excitation residual network for hyperspectral image classification. *Remote Sens.* **2019**, *11*, 884.
22. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018, pp. 3–19.
23. Ma, W.; Yang, Q.; Wu, Y.; Zhao, W.; Zhang, X. Double-branch multi-attention mechanism network for hyperspectral image classification. *Remote Sens.* **2019**, *11*, 1307.
24. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2017; pp. 5998–6008.
25. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
26. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
27. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
28. Bao, H.; Dong, L.; Wei, F. BEiT: Bert pre-training of image transformers. *arXiv* **2021**, arXiv:2106.08254.
29. He, J.; Zhao, L.; Yang, H.; Zhang, M.; Li, W. HSI-BERT: Hyperspectral Image Classification Using the Bidirectional Encoder Representation From Transformers. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 165–178.
30. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
31. Sun, H.; Zheng, X.; Lu, X.; Wu, S. Spectral–spatial attention network for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 3232–3245.
32. He, X.; Chen, Y.; Lin, Z. Spatial-Spectral Transformer for Hyperspectral Image Classification. *Remote Sens.* **2021**, *13*, 498.
33. 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.
34. Tang, X.; Meng, F.; Zhang, X.; Cheung, Y.M.; Jiao, L. Hyperspectral Image Classification Based on 3-D Octave Convolution With Spatial-Spectral Attention Network. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 2430–2447.
35. 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.
36. Hu, H.; Zhang, Z.; Xie, Z.; Lin, S. Local relation networks for image recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Long Beach, CA, USA, 15–20 June 2019; pp. 3464–3473.
37. Li, Y.; Yao, T.; Pan, Y.; Mei, T. Contextual transformer networks for visual recognition. *arXiv* **2021**, arXiv:2107.12292.
38. Guo, M.H.; Xu, T.X.; Liu, J.J.; Liu, Z.N.; Jiang, P.T.; Mu, T.J.; Zhang, S.H.; Martin, R.R.; Cheng, M.M.; Hu, S.M. Attention Mechanisms in Computer Vision: A Survey. *arXiv* **2021**, arXiv:2111.07624.
39. Cordonnier, J.B.; Loukas, A.; Jaggi, M. On the relationship between self-attention and convolutional layers. *arXiv* **2019**, arXiv:1911.03584.
40. Li, T.; Cai, Y.; Cai, Z.; Liu, X.; Hu, Q. Nonlocal band attention network for hyperspectral image band selection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 3462–3474.
41. Yu, C.; Zhou, S.; Song, M.; Chang, C.I. Semisupervised hyperspectral band selection based on dual-constrained low-rank representation. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 5503005.
42. Zhu, M.; Jiao, L.; Liu, F.; Yang, S.; Wang, J. Residual spectral–spatial attention network for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 449–462.