remote sensing

MDPI

Communication

Spatial-Spectral BERT for Hyperspectral Image Classification

Mahmood Ashraf [1], Xichuan Zhou [1], Gemine Vivone [2,3], Lihui Chen [1,*], Rong Chen [4] and Reza Seifi Majdard [5]

[1] School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China; mahmoodkhn24@gmail.com (M.A.); zxc@cqu.edu.cn (X.Z.)
[2] National Research Council, Institute of Methodologies for Environmental Analysis (CNR-IMAA), 85050 Tito, Italy; gemine.vivone@imaa.cnr.it
[3] NBFC—National Biodiversity Future Center, 90133 Palermo, Italy
[4] School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China; jnuc_r@163.com
[5] Department of Electrical Engineering, Ardabil Branch, Islamic Azad University, Ardabil 1477893855, Iran; re.seifim@iau.ac.ir
* Correspondence: lihui.chen@cqu.edu.cn

Abstract: Several deep learning and transformer models have been recommended in previous research to deal with the classification of hyperspectral images (HSIs). Among them, one of the most innovative is the bidirectional encoder representation from transformers (BERT), which applies a distance-independent approach to capture the global dependency among all pixels in a selected region. However, this model does not consider the local spatial-spectral and spectral sequential relations. In this paper, a dual-dimensional (i.e., spatial and spectral) BERT (the so-called D^{2}BERT) is proposed, which improves the existing BERT model by capturing more global and local dependencies between sequential spectral bands regardless of distance. In the proposed model, two BERT branches work in parallel to investigate relations among pixels and spectral bands, respectively. In addition, the layer intermediate information is used for supervision during the training phase to enhance the performance. We used two widely employed datasets for our experimental analysis. The proposed D^{2}BERT shows superior classification accuracy and computational efficiency with respect to some state-of-the-art neural networks and the previously developed BERT model for this task.

Keywords: BERT; multi-head self-attention; spatial-spectral features; convolutional neural network; hyperspectral imaging; classification; deep learning; remote sensing

Citation: Ashraf, M.; Zhou, X.; Vivone, G.; Chen, L.; Chen, R.; Majdard, R.S. Spatial-Spectral BERT for Hyperspectral Image Classification. *Remote Sens.* **2024**, *16*, 539. https://doi.org/10.3390/rs16030539

Academic Editor: Yinnian Liu

Received: 19 December 2023
Revised: 26 January 2024
Accepted: 29 January 2024
Published: 31 January 2024

1. Introduction

Since hyperspectral imagery can capture hundreds of spectral bands, it can provide richer spectral information to address the classification task. Moreover, these data have great potential for other Earth observation applications, including, but not limited to, the monitoring of the environment and the change detection in urban areas [1,2]. HSI classification (HSIC) methods exploit different spatial and spectral information to identify pixels' labels, which play a vital role in lots of applications, such as mineral exploration [3], environmental monitoring [4], and precision agriculture [5]. To achieve accurate hyperspectral image classification (HSIC), many different methods, such as classical methods [6], convolutional neural networks CNNs [7], and transformers [8,9], have been studied to identify each pixel's label in the past and current decades. State-of-the-art techniques utilize feature extraction to obtain state-of-the-art outcomes [10,11].

Classical methods usually combine typical classifiers and manual feature extractors for HSIC. This category includes approaches such as support vector machines (SVM) [12,13] regression [14], and k-nearest neighbors (KNN) [15–17], incorporating feature extraction techniques, such as kernel methods [10] and Markov random field [18]. They assume a linear relationship between the input variables and the output. In the context of hyperspectral data, linear models, such as Multiple Linear Regression (MLR) [19] and Principal

Component Analysis (PCA), have been widely used due to their computational efficiency and ease of implementation [20]. Despite their simplicity and efficiency, these classical HSIC methods have some limitations, e.g., the curse of dimensionality, the design of features in a manual way, the difficulty in managing high nonlinearity, and the sensitivity to noise/outliers.

Thanks to the overcoming of many drawbacks of classical methods, convolutional neural network (CNN)-based methods have become increasingly popular for HSIC. For example, Chen et al. [21] introduced a CNN to extract the deep hierarchical spatial-spectral features for HSIC. To extract the context-based local spatial-spectral information, Lee and Known introduced 3DCNN [22] for HSIC. CNN-based methods could improve the performance but could not extract the spatial-spectral features from the HSIs deeply; hence, a dual tunnel method was proposed by Xue et al. in [23]; 1DCNN was used for the spectral features, and 2DCNN was introduced for the spatial features to explore the deep features from spatial and spectral domains. Cao et al. [24] introduced a method using active deep learning to boost classification performance and decrease labeling costs. Although several CNNs have been proposed for HSIC, they still encounter the limitation of a local receptive field, which cannot fully use the spectral and spatial information of HSIs to classify a given pixel. Moreover, they cannot model sequential data (that is a relevant issue considering that the spectrum of targets acquired by HSIs can be viewed as a sequence of data along the wavelength). Different materials have their own spectral characteristics, such as absorption or reflectance peaks. Therefore, CNNs cannot fully use this information to identify the targets. Hence, recurrent neural networks (RNNs) have been proposed to effectively analyze hyperspectral pixels as sequential data [25] for HSIC. However, RNN is a simple sequential model that hardly has long-term memory and cannot run in parallel, resulting in a time-consuming framework for HSIs. CNN-based techniques belonging to non-linear models offer a more complex but often more accurate representation of relationships in data. Techniques like kernel-based methods and neural networks fall into this category. Their ability to model complex, nonlinear interactions makes them particularly effective for hyperspectral data, which often contains intricate spectral signatures. Comparing these two types of models, linear approaches are generally faster and require less computational resources, making them suitable for large datasets or real-time applications. However, non-linear models, despite their higher computational cost, can significantly outperform linear models in capturing the complex spectral variability inherent in hyperspectral data, thus potentially leading to more accurate classifications and predictions.

Recently, transformers have been proposed for HSIC to overcome the issue of the local receptive field of CNNs and to make full use of long-range dependence. SpectralFormer is a novel backbone network that improves hyperspectral image classification by utilizing transformers to capture spectrally local sequence information from neighboring bands, resulting in group-wise spectral embedding. BERT for HSIC, i.e., HSI-BERT [26], is one of the classical transformer-based solutions for HSIC. The method provides a robust framework for learning long-range dependencies and capturing the contextual information among pixels. However, it neglects the information of spectral dependency and spectral order. As mentioned above, different materials show absorption or reflectance peaks at different wavelengths. Thus, the spectral order is an essential cue for HSIC.

To address the above-mentioned limitations, this paper proposes a dual-dimensional BERT (the so-called D^2BERT). The goal is to improve HSI-BERT by capturing the global and local dependencies among pixels and spectral bands independently from their spatial distance and spectral orders. In D^2BERT, we use two BERT modules to learn the relations between the neighboring pixels and the spectral bands, respectively. To do it, the proposed model has two separate (and parallel) branches: one to explore the relations among neighboring pixels in the spatial domain and another to explore the relations among spectral bands in the spectral domain. The extracted features from both the spatial and spectral domains are then combined before classification. Specifically, a BERT module is applied in each dimension to explore the relations among the corresponding dimensions. For the

spatial branch, we select a square region around the target pixels to classify a pixel, while for the spectral branch, spatial features extracted from the CNN-based model for each spectral band of the selected region are fed into the spectral BERT module to explore the relations among the spectral bands. Since the input order of spectral features are important cues for classifying the target material, we also adopt position embedding in the spectral BERT module, thus enabling D^{2}BERT to distinguish the material according to the spectral features, such as absorption or reflectance peaks. In conclusion, the main contributions of this paper are as follows:

1. To make full use of spatial dependencies among neighboring pixels and spectral dependencies among spectral bands, a dual-dimension (i.e., spatial-spectral) BERT is proposed for HSIC, overcoming the limitations of merely considering the spatial dependency as in HSI-BERT.
2. To exploit long-range spectral dependence among spectral bands for HSIC, a spectral BERT branch is introduced, where a band position embedding is integrated to build a band-order-aware network.
3. To improve the learning efficiency of the proposed BERT model, a multi-supervision strategy is presented for training, which allows features from each layer to be directly supervised through the proposed loss function.

An overview of the main parts of the D^{2}BERT model is discussed in Section 2. The results of our experiments are presented in Section 3, and the conclusions are presented in Section 4.

2. Proposed D^{2}BERT Model

As shown in Figure 1, the D^{2}BERT model has two branches to extract the optimal distinctive features for HSIC. In the upper branch, the spatial BERT is used to explore the spatial dependency for the given HSI, while in the bottom branch, a spectral BERT is introduced to explore the spectral dependency. D^{2}BERT combines these spatial-spectral features from the two branches and this model is learned by multiple supervision of features from each intermediate BERT layer.

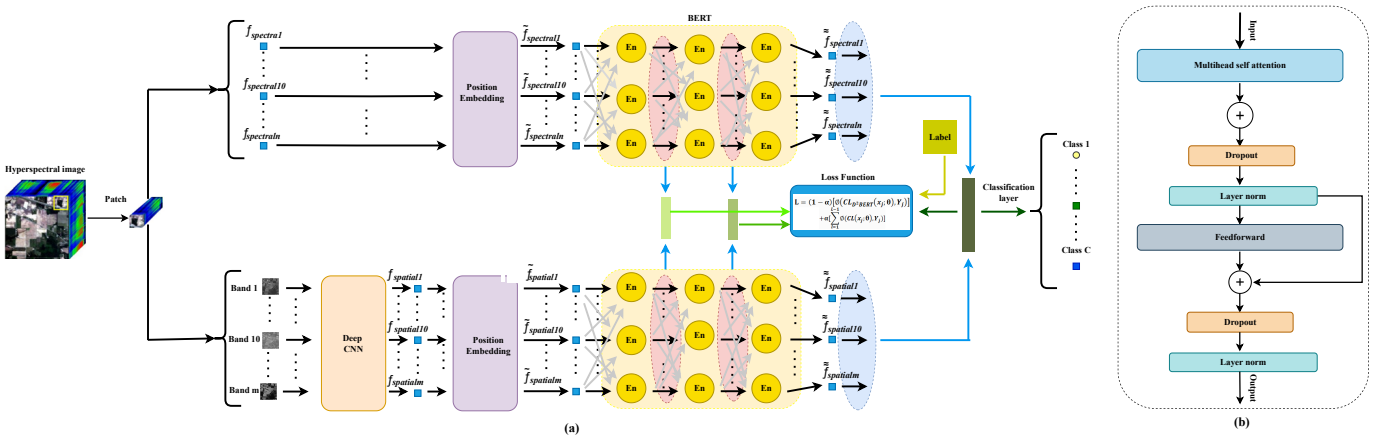

Figure 1. (**a**) The proposed D^{2}BERT model. (**b**) Encoder (En) layer.

2.1. Deep Spatial Feature Learning in Spatial BERT

The upper branch, i.e., the spatial branch, aims to determine long-range relationships among pixels in a selected region. In this branch, a square region (patch) containing a target pixel is first selected for label prediction. This area is initially flattened to form a sequential representation that passes through position embedding and stacked spatial encoders. Features extracted by these encoders are given in input to classification layers for multi-layer supervision. More in detail, each patch is flattened to create a pixel sequence ($f_{spectral1}, \ldots, f_{spectral10}, \ldots, f_{spectraln}$). The positional embedding (PE) module is fed by

the flattened patch. The PE learns positional embeddings and works independently and identically for each pixel [27]. The PE encodes the positional information. Thus, we have:

$$\tilde{f}_{\text{spectral}} = f_{\text{spectral}} + p, \tag{1}$$

where $\tilde{f}_{\text{spectral}}$ is the learned positional embedding and p is the learned positional element.

The BERT module receives these features that are enhanced by the stacked spatial BERT encoders, see Figure 1. A BERT encoder consists of a multi-head self-attention (MHSA), a feedforward network, layer norms, and dropouts. MHSA captures different aspects, by different heads, of the relationships among pixels in the patch [28]. Each attention function can be defined as a mathematical operation that takes a query vector and a set of key-value pairs as inputs and produces an output vector. In this context, vectors represent the query, the keys, the values, and the output. Different heads are related to distinct attentions. All heads operate independently and concurrently. There is a global receptive field for each head in the patch. The scaled dot-product attention is used to calculate all the attention distributions. The final result is calculated by adding all the weighted values. The importance of each value is computed using a compatibility function that compares the query to each key:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{2}$$

where Q, K, V, and d_k denote the query, the key, the value, and the dimension of the data in the input. The feedforward process affects all heads by utilizing interconnected layers to enhance and refine the learned characteristics. A ReLU activation separates two linear layers in the feedforward. All encoders share the same feedforward network. Thus, we ensure that the parameters of the feedforward layer are the same in all encoders. During the model training, layer normalization decreases the internal covariate shift. The layer normalization provides several advantages (e.g., the training becomes more efficient allowing for higher learning rates). The extracted features from the spatial BERT branch $(\tilde{\tilde{f}}_{\text{spectral}_1}, \ldots, \tilde{\tilde{f}}_{\text{spectral}_{10}}, \ldots, \tilde{\tilde{f}}_{\text{spectral}_n})$ are then injected into the classification layer.

2.2. Deep Spectral Feature Learning in Spectral BERT

The spectral branch is similar to the spatial branch. The goal of this branch is to determine long-range relationships among spectral bands in a selected region. Initially, in the patch, all HSI bands are separated. Suppose that there are m spectral bands $(Band_1, \ldots, Band_{10}, \ldots, Band_m)$. The information of each band is represented by a vector of spatial features. A deep 2-D CNN model is used to extract spatial features for each band. In this study, we used the VGG-like architecture, in which several convolutional layers are ignored to reduce overfitting [29]. The extracted features $(f_{spatial1}, \ldots, f_{spatial10}, \ldots, f_{spatialm})$ are then processed by the position embedding stage, in which the position information of the different bands is added $(\tilde{f}_{spatial_1}, \ldots, \tilde{f}_{spatial_{10}}, \ldots, \tilde{f}_{spatial_m})$. Afterwards, the output of the position embedding stage is the input of the BERT module, see Figure 1. This latter module aims to learn long-range dependencies among spectral bands in the spectral domain. As in the spatial domain, the BERT module can check different relationships among its inputs. The final features of the BERT module are injected into the classification layer.

2.3. D^{2}BERT Model Training

Most of deep-learning hyperspectral image classification models, such as HSI-BERT, are trained based on the one-hot label [26]. However, in the proposed D^{2}BERT model, the hidden information coming from the intermediate layers of the BERT modules is also exploited, thus improving the accuracy of the trained model. Accordingly, D^{2}BERT is trained based on the one-hot label and the multi-layer supervision exploiting intermediate features. Moreover, the extracted features from spatial and spectral branches are combined and then given in input to a classification layer to predict the label of the target pixel.

Suppose that the number of training samples is N and C represents the number of classes. If the i-th (target) pixel, x_i, belongs to class j then $Y_{ij} = 1$, otherwise $Y_{ij} = 0$. The BERT modules consist of L layers. The output features of each layer are transferred to an independent classification layer to predict the label. The input features for each classification layer are obtained by the concatenation of the features, obtained from the two BERT modules on the two (spatio-spectral) branches. We calculate the cross-entropy between the output of each classifier and the one-hot label. Therefore, the loss function given x_i, $\mathcal{L}^i$, is defined as follows:

$$\mathcal{L}^i = (1-\alpha)\left[\phi(CL^{(x_i;\theta)}_{D^2BERT}, Y_{ij})\right] + \alpha\left[\sum_{l=1}^{L-1}\phi(CL^{(x_i;\theta_l)}, Y_{ij})\right] \tag{3}$$

where $\phi(\cdot)$ indicates the cross-entropy, $CL(\cdot)$ is the output of the classifier of the l-th layer, θ_l indicates the network parameters of the l-th BERT block, $CL_{D^2BERT}(\cdot)$ denotes the classification output, θ shows the overall network parameters of the model, and the weight α balances the contribution of the information coming from the intermediate layers with respect to the one from the output layer.

3. Experimental Analysis

Two datasets have been considered for performance assessment. The first dataset is the Pavia University (PU) dataset containing 610×340 pixels with 103 spectral bands. Nine classes are represented in this image with 42,776 labeled pixels. The second dataset is the Indian Pines (IP). The Indian Pines dataset contains 200 bands and 16 land-cover types. This dataset has a spatial dimension of 145×145.

3.1. Experimental Setting

D^{2}BERT is implemented using PyTorch and run on a V100 GPU. For training and testing, we randomly selected 50, 100, 150, or 200 labeled pixels, dividing the data into ten sets. The learning rate was set to 3×10^{-4} with 200 training epochs and a dropout rate 0.2. The model uses three encoder layers and two attention heads to balance complexity and efficiency. Unlike the previous approach, which allowed various region shapes, the proposed method uses 32×32 patches for spatial feature extraction using CNNs. The selected metrics for comparison are the overall accuracy (OA), the average accuracy (AA), the training/testing times, and the number of parameters.

3.2. Evaluation Metrics

In evaluating the proposed model, two critical metrics are employed: average accuracy (AA) and overall accuracy (OA). The overall accuracy, OA, of the model is determined through the equation:

$$OA = \frac{\text{Sum of Correct Predictions}(\sum_i X_{eval})}{\text{Total Number of Predictions}} \tag{4}$$

$\sum_i X_{eval}$ represents the summation of correct predictions made by the model across all test instances. The denominator, 'Total Number of Predictions', corresponds to the entire set of predictions made by the model, encompassing both correct and incorrect predictions.

For assessing the accuracy of individual classes within the dataset, average accuracy (AA) is utilized. This is calculated using the following formula:

$$AA = \frac{1}{\text{Classes Count}}\sum_{i=1}^{n}\frac{\sum_{j=1} X_i^j}{C_i} \tag{5}$$

where n is the total number of classes, C_i denotes the count of instances in the ith class, and X_i^j represents the jth correct prediction for the ith class.

3.3. Ablation Study

In the proposed method, we claimed that the hidden information among spectral bands can improve the classification performance. In this experiment, this claim is examined. We performed this experiment on both datasets using 200 training samples. The classification accuracies from the D^{2}BERT and D^{2}BERT without the spectral BERT branch are reported in Tables 1 and 2. It is obvious that D^{2}BERT achieves much better results when the spectral BERT branch is considered. It can be seen that D^{2}BERT achieves the lowest confusion between classes compared to the D^{2}BERT without the spectral BERT branch model. Indeed, we have improved the classification performance, as measured by the OA, by 0.92% and 2.59% for PU and IP, respectively. Moreover, we want to analyze how the use of the intermediate layer information in the loss function impacts the classification accuracy. Hence, we evaluate our method without incorporating intermediate layer information in the loss function. The classification results are reported in Tables 1 and 2. It is clear that layer information in the loss function plays a crucial role leading to performance reduction when it is neglected. The OA is improved by 0.72% and 1.74% for PU and IP, respectively. The corresponding classification maps for these two ablation experiments are depicted in Figure 2. Overall, D^{2}BERT has shown superior performance compared to the ablated models on both datasets in terms of fewer misclassifications, especially for more challenging classes. This demonstrates the effectiveness of capturing spatial and spectral dependencies simultaneously using dual BERT branches, as well as utilizing intermediate layers during optimization. Although the performance gaps differ between datasets, they consistently suggest the significance of both proposed contributions toward achieving state-of-the-art hyperspectral classification.

Table 1. Classification results of different D^{2}BERT configurations for the Pavia University (PU) dataset.

Class	D^{2}BERT *w/o* Spectral Branch	D^{2}BERT *w/o* Multi-Supervision	D^{2}BERT
1	95.83	100	100
2	95.82	98.21	99.54
3	96.21	96.07	99.60
4	97.10	99.00	100
5	97.73	95.52	98.32
6	98.94	99.69	100
7	86.96	86.96	100
8	1.00	100	100
9	1.00	82.35	100
OA (%)	98.88 ± 0.10	99.04 ± 0.11	99.79 ± 0.06
AA (%)	98.36 ± 0.18	98.42 ± 0.14	99.68 ± 0.09

Table 2. Classification results of different D^{2}BERT configurations for the Indian Pines (IP) dataset.

Class	D^{2}BERT *w/o* Spectral Branch	D^{2}BERT *w/o* Multi-Supervision	D^{2}BERT
1	98.89	98.75	99.93
2	99.78	99.85	99.98
3	96.26	97.98	98.85
4	97.44	98.01	99.41
5	100	100	100
6	99.44	99.98	100
7	98.71	96.81	99.91
8	95.74	95.66	99.02

Table 2. *Cont.*

Class	D^{2}BERT *w/o* Spectral Branch	D^{2}BERT *w/o* Multi-Supervision	D^{2}BERT
9	98.86	98.62	100
10	94.40	97.85	99.83
11	97.30	97.87	99.86
12	94.48	98.09	100
13	99.47	98.90	100
14	99.56	100	100
15	99.70	96.60	100
16	1.00	92.68	98.25
OA (%)	99.04 ± 0.11	98.05 ± 0.27	99.76 ± 0.03
AA (%)	97.09 ± 1.26	96.22 ± 0.70	99.71 ± 0.05

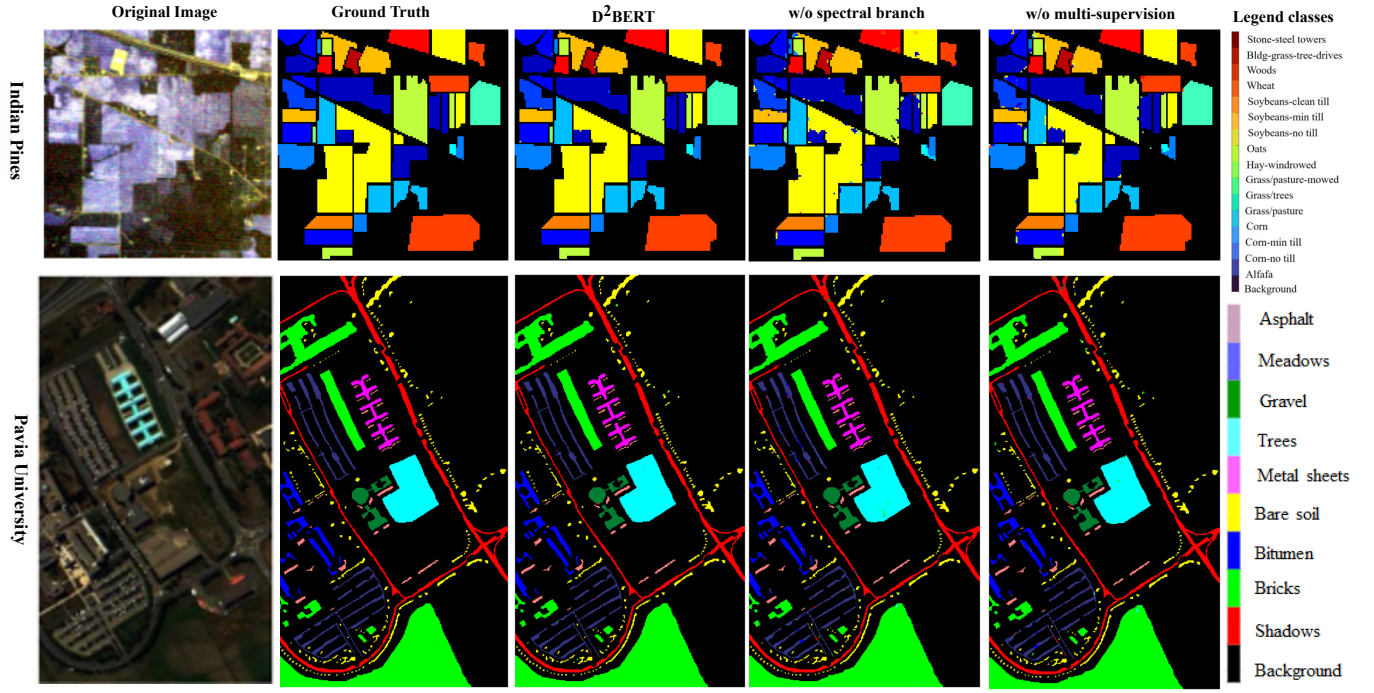

Figure 2. Classification maps achieved by D^{2}BERT in three different configurations and the ground-truth for IP and PU

3.4. Comparison with Benchmark

This section is devoted to the comparison of the proposed D^{2}BERT approach with state-of-the-art CNN-based, transformer-based, and BERT-based methods, i.e., CNN [22], CNN-PPF [30], CDCNN [31], DRCNN [31], Spa-Spe-TR [32], SSRN [33], HybridSN [34], SST [32], HFFSNet [35], GSPFormer [36], and HSI-BERT [26]. The first analysis is based on the comparison of the proposed approach varying the number of samples for training (i.e., 50, 100, 150, and 200) and using the rest of the dataset for testing. Some exemplary methods belonging to our benchmark have been selected for the sake of clarity, including the previously developed HSI BERT and some CNN-based methods. The classification performance varying the number of samples is depicted in Table 3. The better performance of the proposed method is clear, with OA always greater than one of the other techniques, whatever the number of training samples.

The proposed model consistently outperforms all five contemporary CNN-based methods, including HSI-BERT. Thus, D^{2}BERT exhibits a distinct advantage over CNN-based approaches when dealing with limited training samples.

Table 3. Classification results (OA%) of IP with different models using different numbers of training samples. Best results are in boldface

Dataset	PU				IP			
# Samples	50	100	150	200	50	100	150	200
CNN	86.39	88.5	90.89	91.41	80.43	84.32	85.30	86.81
CNN-PPF	88.14	93.35	95.5	96.48	88.34	91.72	93.14	93.90
CDCNN	92.19	93.55	95.5	96.73	84.43	88.27	92.25	94.24
DRCNN	96.91	98.67	99.21	99.56	88.74	94.94	97.49	98.54
HSI-BERT	97.43	98.78	99.38	99.75	91.31	96.86	98.03	99.56
D^2BERT	**98.58**	**99.35**	**99.73**	**99.79**	**93.09**	**98.26**	**99.14**	**99.76**

The second analysis relies upon the calculation of the OA and AA indexes for all the compared approaches, training them with 200 samples. The results are reported in Table 4, the classified maps achieved from the comparing methods are presented in Figure 3 and the OA graphs are presented in Figure 4. Among the compared methods on the Pavia University dataset, D^2BERT achieves the highest OA and AA indexes, i.e., 99.79% and 99.68%, respectively. The gap in performance is clear with respect to CNN, Spa-Spe-TR, SSRN, HybridSN, and SST, with improvements in the range from 6.07% to 8.38%. D^2BERT achieves the same remarkable level of accuracy as DRCNN, HSI-BERT, HFFSNet, and GSPFormer. These results point out the superiority of D^2BERT in accurately classifying the Pavia University data. Focusing on the Indian Pine dataset, D^2BERT achieves very high values of the OA and AA indexes, i.e., 99.76% and 99.68%, respectively, obtaining the top scores among the compared approaches. The comparison of the classification maps also shows the superiority of D^2BERT. These experiments demonstrate that D^2BERT achieves state-of-the-art classification performance on two benchmark datasets compared to recent CNN, transformer, and BERT-based methods.

Table 4. Classification accuracy (%) of the compared approaches. The best results are in boldface.

Dataset	Pavia University		Indian Pines	
Methods	OA%	AA%	OA%	AA%
CNN	91.41	81.03	86.81	63.30
CNN-PPF	96.48	97.03	93.60	96.38
CDCNN	96.73	95.77	94.24	95.75
DRCNN	99.56	98.22	98.54	99.29
Spa-Spe-TR	93.72	91.00	89.13	75.23
SSRN	91.72	87.56	83.21	68.88
HybridSN	92.18	85.16	83.77	63.18
SST	92.50	85.16	88.51	66.64
HFFSNet	98.27	97.20	86.21	83.53
GSPFormer	99.56	99.25	96.29	92.60
HSI-BERT	99.75	99.86	99.56	99.72
D^2BERT	**99.79**	**99.68**	**99.76**	**99.71**

Nonetheless, D^2BERT demonstrates exceptional classification performance, indicating its effectiveness in accurately classifying the Indian Pine dataset. D^2BERT offers advantages in hyperspectral image classification by capturing global and local dependencies, utilizing a neural language-based model, employing BERT modules, and incorporating spectral and spatial features. It explores relations between pixels and spectral bands, reduces complexity, and uses intermediate information for enhanced performance.

D^2BERT is a high-performing model for HSIC, exploiting global and local dependencies, BERT modules, and spectral/spatial features. It efficiently explores relations among pixels and spectral bands, reducing complexity and using intermediate layer information to improve performance. Despite a longer training time than HSI-BERT, D^2BERT excels in understanding data dependencies and semantic relationships, achieving a higher perfor-

mance. Results about the computational burden of the compared approaches are reported in Table 5. D^{2}BERT stands out as the most efficient model, on par with HSI-BERT, due to its relatively shallow architecture and concurrent execution of individual heads in the MHSA. Moreover, D^{2}BERT's parameter count is comparable to other methods, making it a competitive choice in hyperspectral image classification tasks. However, some limitations remain. First, D^{2}BERT incurs higher memory and computational costs than traditional CNN models due to the introduction of transformers. Second, the model may not fully capture fine-grained spatial patterns at small scales due to the use of relatively large patches. Nonetheless, D^{2}BERT also holds promising potential. Its dual-branch architecture is easily parallelizable, aiding runtime efficiency. Transformers allow modeling long-range dependencies beyond the limitations of patch-based CNN receptive fields. The main idea for future work is to investigate how lightweight transformer variants could improve efficiency while maintaining accuracy

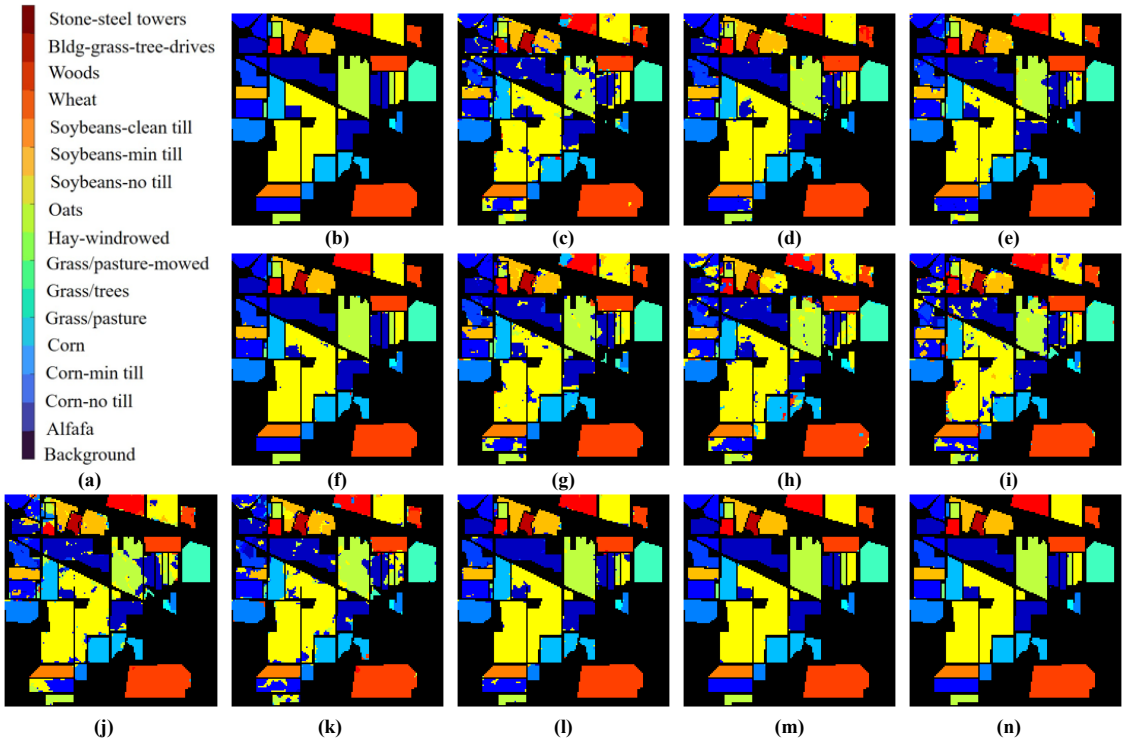

Figure 3. Classification maps of IP achieved by different methods (**a**) classes, (**b**) ground truth, (**c**) CNN, (**d**) CNN-PPF, (**e**) CDCNN, (**f**) DRCNN, (**g**) Spa-Spe-TR, (**h**) SSRN, (**i**) HybridSN, (**j**) SST, (**k**) HFFSNet, (**l**) GSPFormer, (**m**) HSI-BERT, (**n**) D^{2}BERT.

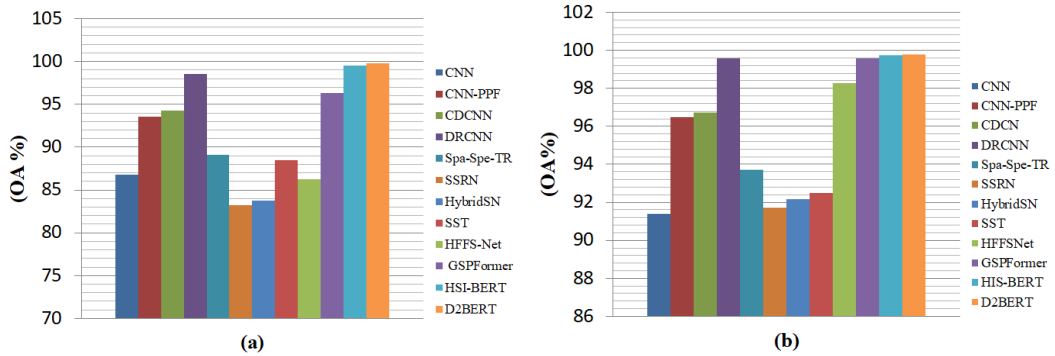

Figure 4. Classification accuracy varying the number of training samples: (**a**) PU, (**b**) IP.

Table 5. Training times, testing times, and a number of parameters for the compared approaches. H, S, and M stand for hours, seconds, and millions, respectively. The best results are in boldface.

	Training Time (Hours)				
	Pavia University		Indian Pines		# Parameters
Methods	Train (H)	Test (S)	Train (H)	Test (S)	in (M)
CNN	0.31	0.37	0.39	0.21	0.13
CNN-PPF	1.00	16.92	6.00	4.76	**0.05**
CDCNN	0.13	12.35	0.14	11.21	1.12
DRCNN	0.43	105	0.74	39	**0.05**
Spa-Spe-TR	0.16	49.2	0.14	19.80	27.65
SSRN	1.28	**0.34**	2.21	**0.06**	0.23
HybridSN	**0.02**	20.4	**0.03**	3.6	14.85
SST	16.69	0.78	21.43	0.19	29
HFFSNet	**0.02**	2.51	0.02	3.45	32.74
GSPFormer	0.47	53	0.18	12	0.68
HSI-BERT	0.07	9.28	0.12	3.52	1.21
D^2BERT	0.19	16.41	0.26	9.11	2.45

4. Conclusions

In this paper, an image classification model based on BERT, the so-called D^2BERT, has been proposed. It relies upon a dual-dimensional spatial-spectral classification in which global and local relations and dependencies among neighboring pixels are investigated, considering both the spatial and spectral domains. D^2BERT exploits two BERT modules to explore spatial and spectral dependencies among pixels belonging to a selected region. The intermediate information coming from different layers of the BERT modules has been considered in the loss function, improving the performance of the model. Experimental results demonstrated the high accuracy of the proposed model outperforming state-of-the-art CNN, transformer, and BERT methods. Experimental results on two benchmark datasets demonstrated the effectiveness of D^2BERT. On the PU dataset with 200 training samples, D^2BERT achieved an OA of 99.79%, outperforming the second-best method (HSI-BERT) by over 0.04%. On the more challenging IP dataset, D^2BERT attained an overall accuracy of 99.76%, outperforming the second-best method (HSI-BERT) by over 0.2%. When limited training data (50 samples) was used, D^2BERT improved overall accuracy over HSI-BERT by 1.18% on PU and 1.94% on IP, validating its advantages in low data regimes. Ablation studies showed removing either the spectral branch or multi-supervision lowered accuracy, demonstrating the importance of both contributions.

Author Contributions: Conceptualization, M.A. and L.C.; methodology, M.A. and R.S.M.; software, M.A.; validation, L.C. and X.Z.; formal analysis, L.C.; investigation, X.Z.; resources, X.Z.; data curation, R.S.M.; writing—original draft preparation, M.A. and L.C.; writing—review and editing, Gemine Vivone and R.C.; visualization, M.A. and L.C.; supervision, L.C.; project administration, X.Z.; funding acquisition, X.Z. and L.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported in part by the National Natural Science Foundation of China under Grant 62301093, Grant 61971072, Grant U2133211, and under Grant 62001063, in part by the Postdoctoral Fellowship Program of CPSF under Grant GZC20233336, in part by the China Postdoctoral Science Foundation under Grant 2023M730425, and in part by the Fundamental Research Funds for the Central Universities under Project No. 2023CDJXY-037.

Data Availability Statement: The datasets utilized in this research are publicly available and can be accessed via the URL https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes (accessed on 15 September 2023). The specific details regarding the datasets, including any necessary information on how to access or cite them, can be found at the provided URL(s). All data used in this study can be freely obtained from the aforementioned source.

Acknowledgments: The authors thank the Editors who handled this manuscript and the anonymous reviewers for their outstanding comments and suggestions

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

References

1. Yu, C.; Huang, J.; Song, M.; Wang, Y.; Chang, C.I. Edge-inferring graph neural network with dynamic task-guided self-diagnosis for few-shot hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [CrossRef]
2. Wang, Y.; Chen, X.; Zhao, E.; Song, M. Self-supervised Spectral-level Contrastive Learning for Hyperspectral Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–15. [CrossRef]
3. Guha, A. Mineral exploration using hyperspectral data. In *Hyperspectral Remote Sensing*; Elsevier: Amsterdam, The Netherlands 2020; pp. 293–318.
4. Aspinall, R.J.; Marcus, W.A.; Boardman, J.W. Considerations in collecting, processing, and analysing high spatial resolution hyperspectral data for environmental investigations. *J. Geogr. Syst.* **2002**, *4*, 15–29. [CrossRef]
5. Caballero, D.; Calvini, R.; Amigo, J.M. Hyperspectral imaging in crop fields: Precision agriculture. In *Data Handling in Science and Technology*; Elsevier: Amsterdam, The Netherlands, 2019; Volume 32, pp. 453–473.
6. Benediktsson, J.A.; Palmason, J.A.; Sveinsson, J.R. Classification of hyperspectral data from urban areas based on extended morphological profiles. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 480–491. [CrossRef]
7. Zhao, F.; Zhang, J.; Meng, Z.; Liu, H. Densely connected pyramidal dilated convolutional network for hyperspectral image classification. *Remote Sens.* **2021**, *13*, 3396. [CrossRef]
8. 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.* **2021**, *60*, 1–15. [CrossRef]
9. Scheibenreif, L.; Mommert, M.; Borth, D. Masked Vision Transformers for Hyperspectral Image Classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 2165–2175.
10. Fang, L.; Li, S.; Duan, W.; Ren, J.; Benediktsson, J.A. Classification of hyperspectral images by exploiting spectral-spatial information of superpixel via multiple kernels. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 6663–6674. [CrossRef]
11. Majdar, R.S.; Ghassemian, H. Improved Locality Preserving Projection for Hyperspectral Image Classification in Probabilistic Framework. *Int. J. Pattern Recognit. Artif. Intell.* **2021**, *35*, 2150042. [CrossRef]
12. Marconcini, M.; Camps-Valls, G.; Bruzzone, L. A composite semisupervised SVM for classification of hyperspectral images. *IEEE Geosci. Remote Sens. Lett.* **2009**, *6*, 234–238. [CrossRef]
13. Ye, Q.; Zhao, H.; Li, Z.; Yang, X.; Gao, S.; Yin, T.; Ye, N. L1-Norm distance minimization-based fast robust twin support vector *k*-plane clustering. *IEEE Trans. Neural Netw. Learn. Syst.* **2017**, *29*, 4494–4503. [CrossRef]
14. Pal, M. Multinomial logistic regression-based feature selection for hyperspectral data. *Int. J. Appl. Earth Obs. Geoinf.* **2012**, *14*, 214–220. [CrossRef]
15. Yang, J.M.; Yu, P.T.; Kuo, B.C. A nonparametric feature extraction and its application to nearest neighbor classification for hyperspectral image data. *IEEE Trans. Geosci. Remote Sens.* **2009**, *48*, 1279–1293. [CrossRef]
16. Samaniego, L.; Bárdossy, A.; Schulz, K. Supervised classification of remotely sensed imagery using a modified *k*-NN technique. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 2112–2125. [CrossRef]
17. Li, W.; Du, Q.; Zhang, F.; Hu, W. Collaborative-representation-based nearest neighbor classifier for hyperspectral imagery. *IEEE Geosci. Remote Sens. Lett.* **2014**, *12*, 389–393. [CrossRef]
18. Sun, L.; Wu, Z.; Liu, J.; Xiao, L.; Wei, Z. Supervised spectral–spatial hyperspectral image classification with weighted Markov random fields. *IEEE Trans. Geosci. Remote Sens.* **2014**, *53*, 1490–1503. [CrossRef]
19. Tarabalka, Y.; Fauvel, M.; Chanussot, J.; Benediktsson, J.A. SVM-and MRF-based method for accurate classification of hyperspectral images. *IEEE Geosci. Remote Sens. Lett.* **2010**, *7*, 736–740. [CrossRef]
20. Rodarmel, C.; Shan, J. Principal component analysis for hyperspectral image classification. *Surv. Land Inf. Sci.* **2002**, *62*, 115–122.
21. Cheng, G.; Li, Z.; Han, J.; Yao, X.; Guo, L. Exploring hierarchical convolutional features for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 6712–6722. [CrossRef]
22. Lee, H.; Kwon, H. Contextual deep CNN based hyperspectral classification. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 3322–3325.
23. Xu, X.; Li, W.; Ran, Q.; Du, Q.; Gao, L.; Zhang, B. Multisource remote sensing data classification based on convolutional neural network. *IEEE Trans. Geosci. Remote Sens.* **2017**, *56*, 937–949. [CrossRef]
24. Cao, X.; Yao, J.; Xu, Z.; Meng, D. Hyperspectral image classification with convolutional neural network and active learning. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 4604–4616. [CrossRef]
25. Mou, L.; Ghamisi, P.; Zhu, X.X. Deep recurrent neural networks for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3639–3655. [CrossRef]
26. 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.* **2019**, *58*, 165–178. [CrossRef]

27. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
28. Dehghani, M.; Gouws, S.; Vinyals, O.; Uszkoreit, J.; Kaiser, Ł. Universal transformers. *arXiv* **2018**, arXiv:1807.03819.
29. Windrim, L.; Melkumyan, A.; Murphy, R.J.; Chlingaryan, A.; Ramakrishnan, R. Pretraining for hyperspectral convolutional neural network classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2798–2810. [CrossRef]
30. Li, W.; Wu, G.; Zhang, F.; Du, Q. Hyperspectral image classification using deep pixel-pair features. *IEEE Trans. Geosci. Remote Sens.* **2016**, *55*, 844–853. [CrossRef]
31. Lee, H.; Kwon, H. Going deeper with contextual CNN for hyperspectral image classification. *IEEE Trans. Image Process.* **2017**, *26*, 4843–4855. [CrossRef] [PubMed]
32. He, X.; Chen, Y.; Li, Q. Two-Branch Pure Transformer for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
33. 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. [CrossRef]
34. 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.* **2019**, *17*, 277–281. [CrossRef]
35. Feng, Z.; Liu, X.; Yang, S.; Zhang, K.; Jiao, L. Hierarchical Feature Fusion and Selection for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5. [CrossRef]
36. Chen, D.; Zhang, J.; Guo, Q.; Wang, L. Hyperspectral Image Classification based on Global Spectral Projection and Space Aggregation. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5. [CrossRef]