Visual Attention-Driven Hyperspectral Image Classification

Juan M. Haut, *Student Member, IEEE,* Mercedes E. Paoletti, *Student Member, IEEE,* Javier Plaza, *Senior Member, IEEE,* Antonio Plaza, *Fellow, IEEE,* and Jun Li, *Senior Member, IEEE*

***Abstract*—Deep neural networks (DNNs), including convolutional (CNNs) and residual (ResNets) models, are able to learn abstract representations from the input data by considering a deep hierarchy of layers that performs advanced feature extraction. The combination of these models with visual attention techniques can assist with the identification of the most representative parts of the data from a visual standpoint, obtained through a more detailed filtering of the features extracted by the operational layers of the network. This is of significant interest for analyzing remotely sensed hyperspectral images (HSIs), characterized by their very high spectral dimensionality. However, few efforts have been conducted in the literature in order to adapt visual attention methods to remotely sensed HSI data analysis. In this paper, we introduce a new visual attention-driven technique for HSI classification. Specifically, we incorporate attention mechanisms to a ResNet in order to better characterize the spectral-spatial information contained in the data. Our newly proposed method calculates a mask that is applied on the features obtained by the network in order to identify the most desirable ones for classification purposes. Our experiments, conducted using four widely used HSI datasets, reveal that the proposed deep attention model provides competitive advantages in terms of classification accuracy when compared to other state-of-the-art methods.**

***Index Terms*—Hyperspectral image classification, visual attention, feature extraction, deep learning, residual neural networks.**

I. Introduction

This paper was supported by Ministerio de Educación (Resolución de 26 de diciembre de 2014 y de 19 de noviembre de 2015, de la Secretaría de Estado de Educación, Formación Profesional y Universidades, por la que se convocan ayudas para la formación de profesorado universitario, de los subprogramas de Formación y de Movilidad incluidos en el Programa Estatal de Promoción del Talento y su Empleabilidad, en el marco del Plan Estatal de Investigación Científica y Técnica y de Innovación 2013-2016. This work has also been supported by Junta de Extremadura (Decreto 14/2018, de 6 de febrero, por el que se establecen las bases reguladoras de las ayudas para la realización de actividades de investigación y desarrollo tecnológico, de divulgación y de transferencia de conocimiento por los Grupos de Investigación de Extremadura, Ref. GR18060) and the European Union's Horizon 2020 research and innovation programme under grant agreement No. 734541 (EOXPOSURE). This work was supported in part by the National Natural Science Foundation of China under Grant 61771496, in part by the Guangdong Provincial Natural Science Foundation under Grant 2016A030313254, and in part by the National Key Research and Development Program of China under Grant 2017YFB0502900. *(Corresponding author: Jun Li.)*

J. M. Haut, M. E. Paoletti, J. Plaza and A. Plaza are with the Hyperspectral Computing Laboratory, Department of Technology of Computers and Communications, Escuela Politécnica, University of Extremadura, 10003 Cáceres, Spain.(e-mail: juanmariohaut@unex.es; mpaoletti@unex.es; jplaza@unex.es; aplaza@unex.es).

J. Li is with the Guangdong Provincial Key Laboratory of Urbanization and Geosimulation, Center of Integrated Geographic Information Analysis, School of Geography and Planning, Sun Yat-sen University, Guangzhou 510275, China (e-mail: lijun48@mail.sysu.edu.cn).

Hyperspectral image (HSI) classification is a very active research field in remote sensing and Earth observation [1], [2]. This is due to the excellent characterization that HSI instruments can provide for large areas on the surface of the Earth. HSI data are often collected by imaging spectrometers mounted on aerial or satellite platforms, and comprise hundreds of images at different (continuous and narrow) wavelengths, usually from the visible to the near infrared regions of the electromagnetic spectrum. As a result, high-dimensional data cubes are obtained in which each pixel captures the emitted, reflected and transmitted light over the observed land cover materials. Each pixel (vector) in the data cube can be interpreted as a spectral signature or *fingerprint* that uniquely characterizes the observed materials of the target area [3]. Such data cubes provide a wealth of spectral and spatial information, a property that is very useful for monitoring the surface of the Earth [4], [5] in a wide range of applications, such as precision agriculture [6]–[8], environmental and natural resources management [9], surveillance [10]–[12], and others [13].

HSI classification has been usually tackled as an optimization problem, trying to assign each pixel of the scene to a certain land cover class by adapting traditional image analysis methods to HSI data [14]. For instance, standard machine learning methods assume that the HSI data cube is a collection of spectral vectors with no spatial arrangement, exploiting only the spectral information to discriminate and classify the pixels. Several unsupervised and supervised spectral-based approaches have been applied to interpret HSI data, including k-means clustering [15], k-nearest neighbors (KNN) [16], support vector machines (SVMs) [17], [18] and other kernel-based methods [19], [20], logistic regression (LR) [21] or random forest (RF) [22], among many others. However, the classification of HSI data involves certain difficulties not to be found in other kinds of image data (in addition to the huge amount of information contained in HSI data cubes [2]). Specifically, traditional supervised classification approaches are largely affected by the *curse of dimensionality* [23], which may hamper the accuracy of the classifier when the number of available labeled training samples is limited in relation to the (high) dimensionality of the data. This is also due to the high cost and effort involved in expert annotation of labeled data, a fact that can result in an under-complete training process which is prone to over-fitting (this is also known as the *Hughes phenomenon* [24]). Moreover, HSI data sets present high intra-class variability and inter-class similarity, resulting from atmospheric interferers, spectral variability and the configuration of the sensor. These aspects bring additional difficulties when characterizing the data and call for new techniques that can better exploit the rich spatial and spectral information contained in HSI scenes.

To address some of the aforementioned issues, several deep neural network (DNN) models have been developed in the literature [25]. These flexible architectures, composed by a stack of layers, allow multiple techniques to include and process not only the spectral signatures, but also the spatial-contextual information contained in the captured scenes. Based on the idea that spatially adjacent pixels often belong to the same class, these classifiers take advantage of spatial information to reduce sample variability. In fact, it is well-known that the extraction of spectral-spatial features is very useful to improve the classification process, helping to reduce label uncertainty and intra-class variance. As a result, joint spectral-spatial methods can often perform better than purely spectral or spatial-based ones. However, in deep learning (DL) methods there is a problem of how to fuse the spectral and spatial information. Focusing on stacked autoencoders (SAEs) [26] and deep belief networks (DBNs) [27] we can find several techniques that concatenate the spectral signatures and the spatial information extracted from neighbouring pixels by taking advantage of simple dimensionality reduction methods such as the principal component analysis (PCA) [28]–[31], or more sophisticated methods such as superpixels [32], guided filtering [33] or morphological profiles [34], [35], among others. Traditional fully connected architectures admit vector inputs, so the spatial structure is usually lost. In this sense convolutional neural networks (CNNs) [36] are a powerful tool for the analysis of HSI images due to their capacity to accurately characterize both the spectral and spatial-contextual information contained in HSI data cubes [37], being able to effectively extract features with a high-level of abstraction from the raw data and achieving excellent classification results [38].

However, DL-based models are not totally immune to the curse of dimensionality and the Hughes phenomenon. In fact, CNNs tend to quickly overfit when few labeled samples are available. To overcome this limitation, several techniques have been developed, including: i) semi-supervised and active learning (AL) techniques [39], able to deal with overfitting when very few training samples are available, ii) residual learning (e.g., using residual networks or *ResNets*) [40], [41] and dense connections (e.g. using dense networks or *DenseNets*) [42], [43], which can alleviate the loss of information and vanishing gradient problems of very deep and complex architectures, and iii) the development of new information routing techniques such as capsule modules (e.g., using capsule networks or *CapsNets*) [44], [45]. Despite these advances, CNN-based models still present a main limitation when dealing with HSI data. In fact, they can be hindered by the mode-operation of their own convolution filters, which treat the input content completely equally, while probably not all spectral-spatial information provided by the input hyperspectral pixels is equally interesting, informative, relevant and/or predictive for classification purposes [46].

In the area of computer vision, several efforts are now being made to improve DL techniques, overcoming the equal treatment of the convolution kernel by incorporating visual attention mechanisms. The goal of these techniques is to explore in detail the objects or regions that stand out in a given scene [47], as opposed to convolutional methods, whose kernels treat equally the whole content in the image. The main idea is to simulate the human behavior, as we try to understand an image by selecting a subset of features that contain the most relevant characteristics instead of treating the full scene equally. In fact, the human brain focuses on the most valuable and informative stimulus perceived by the eyes, ignoring other irrelevant information. Such visual attention mechanisms are based on two kinds of components [48]: i) *bottom-up* (stimulus-driven) components, which are traditionally related with automatic/involuntary processing of salient visual features in raw sensory information and are performed in a feed-forward way, and ii) *top-down* (goal-oriented) components, that modulate bottom-up component behavior through voluntary attention to certain characteristics, objects, or regions in the space. The study of these components, together with their characteristics, has resulted in a great variety of attention-driven techniques [49], turning visual attention into a hot research topic.

In the remote sensing literature, several attention-driven techniques have been developed for detecting salient regions [50]–[56] and target objects [57]–[60]. However, their application to HSI data has been quite sparse [61], [62]. Although the adaptation of visual attention techniques to deep models is demonstrating excellent performance in several classification tasks [63]–[65], there is still room for contributions in the area of HSI classification.

In this paper, we develop a new spectral-spatial visual attention-driven technique for HSI classification. Our newly developed technique combines the use of advanced visual attention mechanisms with powerful feature extraction approaches based on DNNs for spectral-spatial HSI classification. As a case study, we introduce visual attention mechanisms in the ResNet architecture (A-ResNet). The translation of a visual attention working mode to DNNs for HSI data processing allows to increase the sensitivity of the network to those features that contain the most important and useful information for classification purposes. In this regard, the main innovative contributions of our work can be summarized as follows:

- The development, for the first time in the literature, of a visual attention-driven mechanism (incorporated into a ResNet architecture) for spatial-spectral HSI classification. This is done by introducing a dual data-path attentional module as the basic building module, considering both bottom-up and top-down visual factors to improve the feature extraction capability of the network.
- A detailed comparison between our attention-driven model and traditional pixel-based machine learning and spectra-spatial DL-based techniques for HSI classification, demonstrating that the proposed model is able to outperform current state-of-the-art classifiers.
- A study of how the performance of the considered classifiers is affected by perturbations in the data, introducing

controlled noise in the samples. To this end, four well-known and publicly available HSIs are considered in our experiments: Indian Pines (IP), University of Pavia (UP), Salinas Valley (SV), and University of Houston (UH)

The remainder of this paper is organized as follows. In section II we introduce the basic principles of CNNs and the ResNet model. Section III describes in detail our newly proposed A-ResNet methodology. Section IV discusses our experimental results. Finally, Section V concludes the paper with some remarks and hints at plausible future research lines.

II. Related Work

A. Convolutional Neural Networks (CNNs)

DNNs are characterized by a hierarchical structure composed by a deep stack of processing layers, placed one after the other. Such deep structure allows these models to learn representations of the original input data with multiple levels of abstraction, from the most concise ones (at the first layers) to the most abstract ones (at the end of the architecture). Such multi-level representations of the data allow for a powerful mechanism of feature extraction, in which each layer is able to discover (or reinforce) different relations, distributions and structures in the data, supported by features extracted by previous layers. In this sense, the architecture of CNNs is based on receptive fields, and follows the behaviour of neurons in the primary visual cortex of a biological brain [66], [67]. These models have become a state-of-the-art in remote sensing data analysis, outperforming many algorithms [68]. CNNs are typically composed of two main parts: i) the *feature extractor net*, and ii) the *classifier*.

The *feature extractor* is composed by several kinds of n-dimensional blocks or layers, depending on how the information is used and how it is processed by these blocks. A HSI dataset $\mathbf{X}$ can be seen as a collection of spectral vectors $\mathbf{X} \in \mathbb{R}^{(n_1 \cdot n_2) \times n_{bands}}$, where $n_1 \cdot n_2$ denotes the number of spectral pixels in the scene and n_{bands} is the number of spectral bands. Each pixel in the scene is given by $\mathbf{x}_i \in \mathbb{R}^{n_{bands}} = [x_{i,1}, x_{i,2}, \cdots x_{i,n_{bands}}]$. CNN models composed by one-dimensional (1-D) blocks process only the spectral information in the data, and are also known as spectral-based CNNs. These models exhibit similar disadvantages as traditional pixel-based processing methods. On the contrary, if we apply a spectral dimension reduction technique over $\mathbf{X}$, for example PCA [69], [70] and retain only the first PC, the hyperspectral image can be treated as a 2-D matrix of spatial information $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$, where $n_1 \times n_2$ denotes the number of rows and columns in the scene. Traditional CNNs employ 2-D blocks to process the spatial information contained in the input data, which in RGB data corresponds with the whole image. However, to process the HSI $\mathbf{X}$ using both spatial and spectral information, we need to extract, for each pixel $\mathbf{x}_{i,j} \in \mathbb{R}^{n_{bands}}$, a neighborhood window or spatial patch $\mathbf{p}_{i,j} \in \mathbb{R}^{d \times d}$, which comprises the set of $d \times d$ pixels that surround the central sample $\mathbf{x}_{i,j}$. The usual way to perform the classification is to assign the label $y_{i,j}$ of the central pixel $\mathbf{x}_{i,j}$ to the entire patch $\mathbf{p}_{i,j}$. Although such spatial-based classification strategy can achieve good results, the loss of significant spectral information is often critical in many applications [37], [38]. A third way to classify the HSI scene $\mathbf{X}$ is to exploit the spatial-contextual information together with the full or filtered spectra, retaining the full spectral information from the original bands (or a significant percentage of it, by means of an appropriate number of principal components) and creating spectral-spatial patches or data sub-cubes $\mathbf{p}_{i,j} \in \mathbb{R}^{d \times d \times n_{channels}}$. In this sense, the spectral-spatial CNN model allows to treat the data in 3-D fashion by combining both sources of information (spatial and spectral) in a most natural and simple way, by considering 3-D sub-blocks extracted from the input data cube.

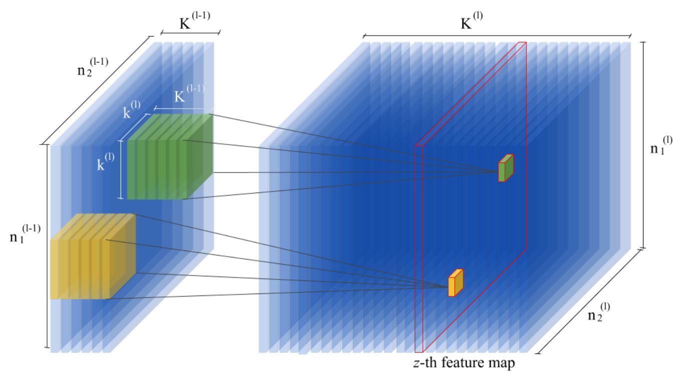


Fig. 1. Visualization of a convolutional layer operation with 2D kernel. Unlike fully-connected layers, the l-th convolutional block presents *local connectivity* to small regions of the whole input volume, that is, the z-th filter's weights $\mathbf{W}^{(l)}$ are applied over windows of the input volume $\mathbf{X}^{(l-1)} \in \mathbb{R}^{n_1^{(l-1)} \times n_2^{(l-1)} \times K^{(l-1)}}$ defined by the receptive field of size $k^{(l)} \times k^{(l)}$, taking into account the full depth $K^{(l-1)}$ of the input data (highlighted as green and yellow patches), slipped by a stride determined by $s^{(l)}$. It can be observed that the z-th kernel produces, for each region, a scalar value (represented as a smaller rectangle) which is allocated into the z-th feature map, giving as a result an output volume $\mathbf{X}^{(l)} \in \mathbb{R}^{n_1^{(l)} \times n_2^{(l)} \times K^{(l)}}$ that comprises $K^{(l)}$ feature maps of $n_1^{(l)} \times n_2^{(l)}$ features each.

Using spectral-spatial patches as inputs, the feature extractor net of the spectral-spatial CNN model hierarchically applies three kinds of operations: i) *convolution*, ii) non-linear *activation*, and iii) donwsampling by *pooling*. The convolutional layer is the main processing block, composed by K filters defined by their receptive field. In this sense, regarding the dimension of the filters, the CNN can be understood as 1D, 2D or 3D depending on whether its receptive field is of dimensions $K \times q$, $K \times k \times k$ or $K \times k \times k \times q$, respectively, being q and k the spectral and spatial components of the kernel (in this context, the proposed model implements a spectral-spatial convolutional-based model with 2D kernels). In fact, the convolutional layer can be interpreted as a sliding-window method, where the windows/kernel of the block slide over the spatial and spectral dimensions of the input volume using a stride $s^{(l)}$:

$$\mathbf{X}^{(l)} = \mathbf{W}^{(l)} * \mathbf{X}^{(l-1)} + \mathbf{b}^{(l)}, \quad (1)$$

where $\mathbf{X}^{(l)}$ is the output volume of the l-th layer, composed by K feature maps and obtained as the convolution $(*)$ of the input volume $\mathbf{X}^{(l-1)}$ and the layer weights $\mathbf{W}^{(l)}$ and biases $\mathbf{b}^{(l)}$. More specifically, each feature of the $\mathbf{X}^{(l)}$ in Eq. (1) is obtained as follows:

$$x_{i,j}^{(l)_z} = (\mathbf{W}^{(l)} * \mathbf{X}^{(l-1)} + \mathbf{b}^{(l)})_{i,j} = \sum_{\hat{i}=0}^{k^{(l)}-1} \sum_{\hat{j}=0}^{k^{(l)}-1} \left(\mathbf{x}_{(i \cdot s^{(l)}+\hat{i}),(j \cdot s^{(l)}+\hat{j})}^{(l-1)} \cdot \mathbf{w}_{\hat{i},\hat{j}}^{(l)} \right) + \mathbf{b}^{(l)}, \quad (2)$$

where $x_{i,j}^{(l)_z} \in \mathbb{R}$ is the (i,j)-th element of the z-th feature map of $\mathbf{X}^{(l)}$ (with $z = 0, 1, \cdots, K^{(l)} - 1$ and $K^{(l)}$ being the number of filters of the layer), $\mathbf{x}_{i,j}^{(l-1)} \in \mathbb{R}^{K^{(l-1)}}$ is the (i,j)-th element of the input volume $\mathbf{X}^{(l-1)}$, $\mathbf{w}_{\hat{i},\hat{j}}^{(l)}$ is the $(\hat{i},\hat{j})$-th weight of the layer weights $\mathbf{W}^{(l)}$, $\mathbf{b}^{(l)}$ denotes the biases, and $s^{(l)}$ is the stride, being $k^{(l)} \times k^{(l)}$ the receptive field of the l-th layer. Fig. 1 presents a graphical visualization of the operations conducted by Eqs. (1) and (2).

Convolutional blocks extract the features contained in the input volume by applying a linear dot product. In order to learn non-linear relationships present in the data, a non-linear activation function is adopted before sending the resulting output volume to the following layer $\mathbf{X}^{(l)} = \mathcal{H}\left(\mathbf{X}^{(l)}\right)$, being $\mathcal{H}(\cdot)$ usually implemented by the Rectified Linear Unit (ReLU) [71]. Also, with the aim of reducing the spatial dimensions of the output volume, and also to summarize the obtained features and obtain a certain invariability to geometric transformations, a non-linear sub-sampling strategy is implemented by the pooling layer. In fact, the pooling layer applies a sample-based discretization process, selecting from small windows of the input volume those values that satisfy the selection criteria, being the max-pooling one of the most widely used methods for this purpose. It simply slides a spatial kernel $k \times k$ over the input volume, selecting the maximum value for each region, as Eq. (3) indicates:

$$pool_{i,j}^{(l)_z} = \max_{(a,b) \in \mathcal{R}_{i,j}} x_{a,b}^{(l)_z}, \quad (3)$$

where $pool_{i,j}^{(l)_z}$ represents the (i,j)-th output value of the pooling associated with the z-th feature map, and $x_{a,b}^{(l)_z}$ denotes the (a,b)-th element contained by the pooling region $\mathcal{R}_{i,j}$ that encapsulates a spatial receptive field around the position (i,j) [72].

At the end of the feature extractor net, a final output $\mathbf{X}^{(l)}$ is obtained that contains an abstract representation of the original input data. Usually, this output is flattened in order to allow the classifier to perform the final categorization of the input data. Normally, the classifier is implemented by one or more fully-connected layers of a multilayer perceptron (MLP), creating an end-to-end structure.

B. Residual Neural Networks (ResNets)

CNNs present several problems when processing HSI data. In particular, they tend to overfit when very few labeled samples are available to perform the training procedure, and they also can suffer from loss of information when deep structures are implemented. To overcome the first problem, several strategies have been developed in the literature, such as the use of data regularization and dropout techniques, data augmenting, or semi-supervised and active learning approaches.

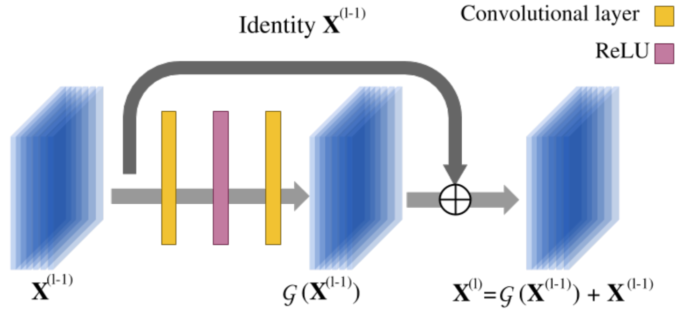


Fig. 2. Graphic visualization of a standard residual unit. The final output volume is obtained as the aggregation of the original input volume $\mathbf{X}^{(l-1)}$ and the resulting output volume of the hidden stack of layers, $\mathcal{G}\left(\mathbf{X}^{(l-1)}\right)$, where $\mathcal{G}(\cdot)$ refers to the convolutions, normalizations, pooling steps and activation functions applied along the stack over the input data. As a result, the architecture reinforces the learning process of the entire model by reusing previous information in the following layers: $\mathbf{X}^{(l)} = \mathcal{G}\left(\mathbf{X}^{(l-1)}\right) + \mathbf{X}^{(l-1)}$.

However, the loss of information is produced by the vanishing gradient problem [73]. In this case, for very deep architectures, the errors become quite hard to propagate back correctly, and the gradient signal tends to zero [74]. Several strategies have also been developed to deal with this problem, such as data normalization techniques [75] or new optimizer/activation functions [76], [77]. However, the accuracy of deep CNNs still can saturate due to the complexity of the mapping function of the convolutional blocks and the hard learning of these functions [78]. In this sense, the architectural modifications introduced by ResNets can improve the learning process of convolutional layers by learning small residuals and adding them to the input volume of each layer, instead of transforming the whole input volume directly. In order to differentiate the CNN and ResNet models, we note that the main building block of a CNN is composed by the convolutional layer and the non-linear activation function, so Eq. (1) with $\mathcal{H}(\cdot)$ can be re-written as:

$$\mathbf{X}^{(l)} = \mathcal{H}\left(\mathbf{W}^{(l)} * \mathbf{X}^{(l-1)} + \mathbf{b}^{(l)}\right)$$
$$\text{simplifying } \mathbf{X}^{(l)} = \mathcal{H}\left(\mathbf{X}^{(l-1)}\right) \quad (4)$$

Eq. (4) indicates that the CNN hierarchically extracts the features, processing them by the successive layers that compose the architecture. Instead of that, the ResNet uses the residual unit as a building block [79] and is composed by a stack of several layers, normally convolutional layers stacked with ReLUs and batch-normalization layers, and with two types of connections allowing different kinds of data streams (see Fig. 2):

- The traditional forward connection, that connects the current layer with the previous and the following ones, extracting from the original input volume $\mathbf{X}^{(l-1)}$ a representation $\mathcal{G}\left(\mathbf{X}^{(l-1)}, \mathcal{W}^{(l)}, \mathcal{B}^{(l)}\right)$, where $\mathcal{G}(\cdot)$ approximates the residual function referring to those operations that are applied over the input data by all the stacked layers of the residual unit, which depends on the weight matrices $\mathcal{W}^{(l)} = \{\mathbf{W}^{(i)}\}_{i=0}^{N-1}$ of the N convolutional layers associated to the l-th residual unit, and the corresponding biases $\mathcal{B}^{(l)} = \{\mathbf{b}^{(i)}\}_{i=0}^{N-1}$.

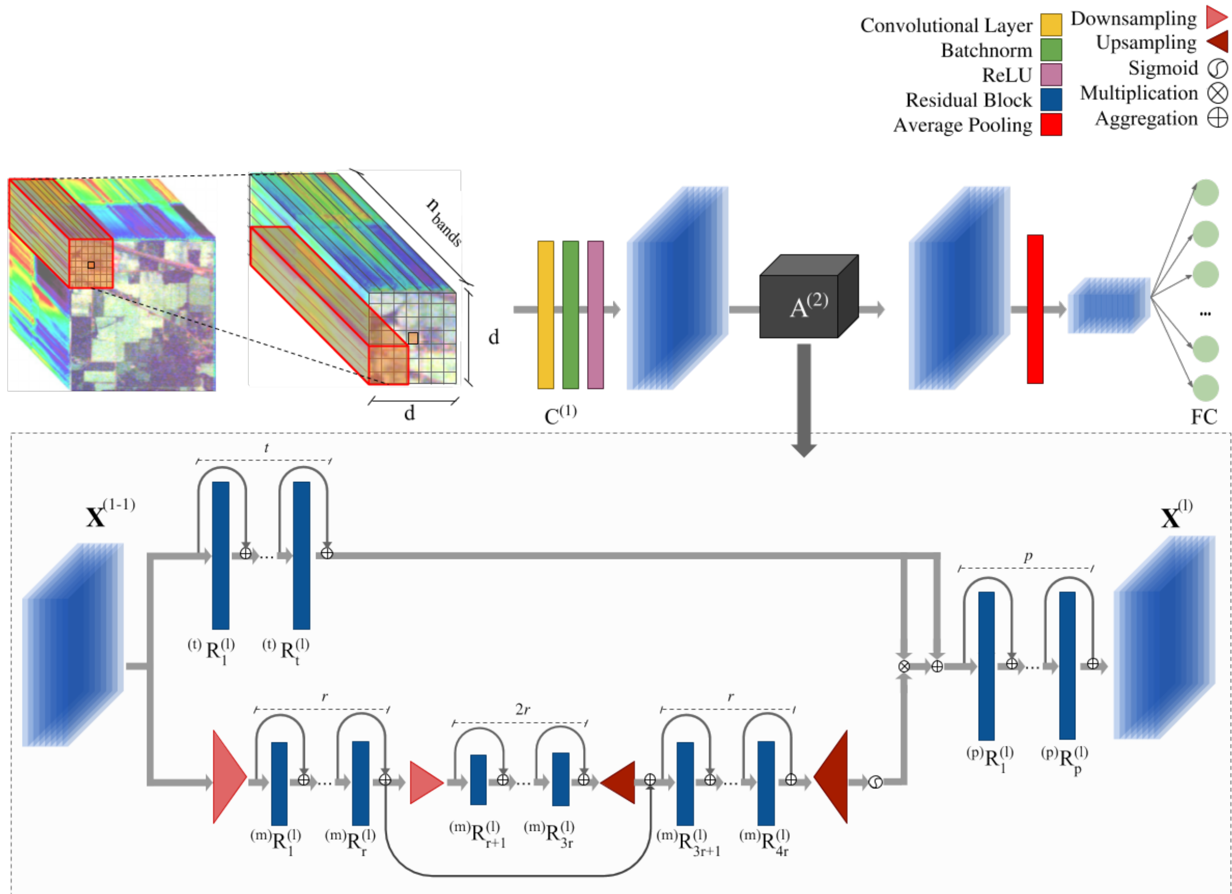


Fig. 3. Standard architecture of the proposed network, with i) the network's head, composed by a convolutional layer $C^{(1)}$ that presents the input volume data $\mathbf{X}$ to the ii) network's body, composed by the residual attention module, $A^{(2)}$, whose output is finally vectored through an average pooling and sent to the iii) network's tail, composed by one fully connected layer that performs the final classification. Two branches: trunk and mask, compose the attentional module: i) the trunk branch (upper path), composed by t residual blocks that perform feature extraction from the data; ii) the mask branch (bottom path), composed by a symmetrical downsampler-upsampler structure in which r residual blocks are allocated (in between each down/up-sampling step) to extract information from the current scale, adding a shortcut connection to link the downsampling step $(/2)$ with its corresponding upsampling $(\times 2)$ counterpart to combine both kinds of data (instead of the bottleneck part, where only $2 \cdot r$ residual blocks are stacked one after the other), and followed by a sigmoid function to prepare the mask, which is applied over the trunk feature data. The resulting output is sent to iii) a final group of p residual blocks located at the end of the module.

- The shortcut connection, that communicates the original input volume with the end of the residual unit, performing an identity mapping that allows to reuse the previous information to reinforce the learning of the residual block.

At the end, residual learning is introduced into Eq. (1) as:

$$\begin{aligned} \mathbf{X}^{(l)} &= \mathcal{G}\left(\mathbf{X}^{(l-1)}, \mathcal{W}^{(l)}, \mathcal{B}^{(l)}\right) + \mathbf{X}^{(l-1)} \\ \text{simplifying: } \mathbf{X}^{(l)} &= \mathcal{G}\left(\mathbf{X}^{(l-1)}\right) + \mathbf{X}^{(l-1)} \end{aligned} \tag{5}$$

where the previous features are exploited once again by the next unit, which reinforces the learning and allows the gradient to be transmitted.

III. Attentional Residual Network (A-ResNet) for Hyperspectral Image Classification

The combination of convolutional kernels and residual connections make the ResNet a very powerful and efficient model for image analysis, in general, and for HSI processing in particular. Based on this architecture, this section develops a new architecture for HSI classification that incorporates visual attention mechanisms in order to extract more discriminatory features, improving the model performance and enhancing its accuracy. In this sense, analogous to the original ResNet, the proposed spectral-spatial A-ResNet for HSI classification adopts a basic building block, called *attentional module* [65], that contains two data-paths or branches: i) the *trunk branch*, and ii) the *mask branch*. Fig. 3 presents the overall architecture of the proposed attentional neural network for HSI data classification. Focussing on the attentional module, the specifications of each part are discussed in detail below.

A. Attentional Module → Trunk Branch

The attentional module can be denoted as $A^{(l)}$, with l being the number of layers, and receives the volume $\mathbf{X}^{(l-1)}$ as input data, which is forward-propagated through two different paths, being the trunk branch the simplest and easiest one to implement. It is composed by t residual blocks, which are stacked one by one, performing a feature extraction and processing task. These residual blocks can be implemented following previous works such as the basic residual block and its bottleneck implementation [78], the wide residual block [80], and the pyramidal residual block and its bottleneck variation [41], [81], among other complex structures [79], [82], [83]. The obtained features can be denoted as $\mathbf{X}^{(l_{trunk})} = \text{Trunk}\left(\mathbf{X}^{(l-1)}\right)$, and contain the high-level data representation of the module. At this point, and following visual attention principles, the next step is to single out the most relevant features from all of the available information contained into $\mathbf{X}^{(l_{trunk})}$, masking the least interesting parts for the learning procedure. In this sense, an attention mask $\mathbf{X}^{(l_{mask})}$ must be calculated and applied over the processed features of the trunk branch.

B. Attentional Module → Mask Branch

As mentioned before, the input module $\mathbf{X}^{(l-1)}$ is propagated through two paths, with the mask branch being in charge of calculating and applying the attention mask $\mathbf{X}^{(l_{mask})}$ over the output features obtained by the trunk branch, $\mathbf{X}^{(l_{trunk})}$. In fact, its goal is to obtain a weight matrix with the same dimensions of $\mathbf{X}^{(l_{trunk})}$, that softly weights the trunk's output features to highlight the most important ones, simulating the element-wise soft attention mechanism.

In order to obtain the final $\mathbf{X}^{(l_{mask})}$, the mask branch applies a network architecture over $\mathbf{X}^{(l-1)}$. It is based on a spatial downsampler-upsampler structure with r residual blocks, allocated between each pair of down/up-sampling steps and with skip connections between each downsampling step and its upsampling counterpart (similar to the hourglass network [84]), following the anatomical connections of cortical processing [85] where feedforward connections transform the input into fast behavioural responses, whereas skip/feedback connections modulate these responses using perceptual context or attention. Moreover, each sampling step (coupled with its corresponding r residual blocks) provides semantic information about the input data, from low-level cues (edges, color, intensity) to high-level cues which, coupled with the forward connections (aimed at collecting global information from the data) and skip connections (which allow to combine multi-scale data taking into account global information and original features) simulate the bottom-up and the top-down attention selection of the visual cortex [86]. In this sense, the downsampler-upsampler structure stacks as many down/up-sampling steps as possible, until the smallest feasible spatial resolution of the data is reached.

In the attention module $A^{(l)}$, the naive application of the attentional mask over the trunk features in the spatial-spectral domain gives the following output:

$$\mathbf{X}^{(l)} = \mathbf{X}^{(l_{mask})} \cdot \mathbf{X}^{(l_{trunk})} \tag{6}$$

However, Eq. (6) presents several limitations. Considering the mask $\mathbf{X}^{(l_{mask})}$ as a collection of values in the range $[0, 1]$, its application over trunk features may degrade them in deeper layers. Also, if the mask contains in most of its elements a value that is equal or close to 0, it may disregard relevant features of the trunk branch. In order to overcome these problems, Eq. (6) is reformulated as follows:

$$\mathbf{X}^{(l)} = \left(1 + \mathbf{X}^{(l_{mask})}\right) \cdot \mathbf{X}^{(l_{trunk})} \tag{7}$$

In this case, Eq. (7) allows to propagate the characteristics extracted from the trunk branch, where the mask branch suppresses the least significant features to facilitate the detection of important features. The combination of both allows to single out the salient features.

Finally, the masked output volume is passed through a tail composed by p residual blocks that performs a final feature extraction step, taking into account the features that have been highlighted in the previous phase.

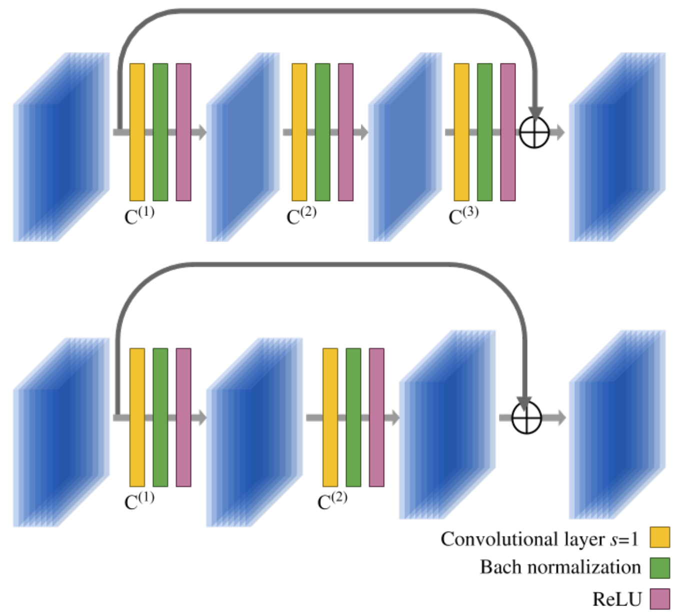


Fig. 4. Graphic visualization of the architecture of the internal residual blocks that conform the trunk branch of the attentional module (top row) and those that conform the mask branch (bottom row). Convolutional details are given in Table I.

C. Proposed Network Topology

The proposed network for spectral-spatial HSI data classification has been developed to work with 3-D sub-cubes $\mathbf{p}_{i,j}\mathbb{R}^{d\times d\times n_{channels}}$ extracted around each spectral pixel $\mathbf{x}_{i,j}$ of the original scene, taking $d = 11$ as the spatial height and width dimensions [40]. These input patches are passed through the network, which is composed by the network's head, attentional body, and classification tail (see Fig. 3) in order to extract relevant features and perform their corresponding classification. The head of the network is given by a convolutional layer $C^{(1)}$ with batch-normalization and ReLU, which prepares the data to be processed by the rest of the network, followed by one or several attentional modules, depending on the complexity of the problem. As mentioned above, the l-th attentional module $A^{(l)}$ is, in turn, composed by several residual blocks ${}^{*}R_i^{(l)}$ (see Fig. 3):

- t residual blocks, denoted as ${}^{(t)}R_i^{(l)}$, with $i = 1, \cdots, t$, for extracting features in the trunk branch.

TABLE I
BASIC ARCHITECTURE OF THE RESIDUAL BLOCKS OF THE TRUNK AND MASK BRANCHES, WHERE $K_{middle} = K_{input}/2$.

Layer ID	Kernel size	Stride	Padding
Bottleneck residual block from trunk branch			
$C^{(1)}$	$K_{middle} \times 3 \times 3 \times K_{input}$	$s = 1$	$p = 1$
$C^{(2)}$	$K_{middle} \times 3 \times 3 \times K_{middle}$	$s = 1$	$p = 1$
$C^{(3)}$	$K_{input} \times 3 \times 3 \times K_{middle}$	$s = 1$	$p = 1$
Residual blocks from trunk branch			
$C^{(1)}$	$K_{input} \times 3 \times 3 \times K_{input}$	$s = 1$	$p = 1$
$C^{(2)}$	$K_{input} \times 3 \times 3 \times K_{input}$	$s = 1$	$p = 1$

- $r(2DU)$ residual blocks, denoted as ${}^{(m)}R_i^{(l)}$, being DU the number of down/up-sampling steps for processing multi-scale data and obtain the attention module mask. For instance, in Fig. 3, with $DU = 2$ down/up-sampling steps, there are $4r$ residual blocks
- p residual blocks denoted as ${}^{p}R_i^{(l)}$ with $i = 1, \cdots, p$, located at the end of the module for post-processing the filtered data.

In total, the attention module is composed by $t+r(2DU)+p$ residual blocks, being $t = 2$, $r = 1$ and $p = 1$, while DU depends on the spatial size of the input volume. The residual block architecture of the trunk branch is composed by three sub-blocks of convolutional layers, batch-normalization and ReLU (see Fig. 4), whose kernels are defined in Table I, creating a spectral-bottleneck architecture in order to better analyze the spectral-spatial domains [87], while the residual blocks of the mask and the ending of the module follow the simple residual unit designed in [78]. Kernels are defined in Table I. As we can observe, each kernel performs a convolution operation using windows of size 3×3, with padding $p = 1$. In this context, the output of the attention module, $\mathbf{X}^{(l)}$, maintains the same spatial-spectral dimensions as the input, $\mathbf{X}^{(l-1)}$, in the sense that all its residual blocks keep the volume dimensions constant. This allows us to add a lot of flexibility to the model, which is able to stack modules one after another (as *plug-&-play* structures). In order to avoid the overfitting problem caused by the large number of parameters that must be trained, we propose a simple architecture with one attentional module. Details can be found on Table II.

Furthermore, the network has been optimized using the Adam optimizer [76] with 300 epochs, where the learning rate decays half of its value on epochs 50, 100 and 200, using a batch size of 100. Also $n_{channels} = 40$ principal components have been considered as the input spectral bands, being $d = 11$.

IV. EXPERIMENTAL RESULTS

A. Experimental Configuration

With the aim of testing the performance of the proposed attentional network for spectral-spatial HSI classification, a battery of experiments have been performed on a desktop computer equipped with a 6th Generation Intel® Core™i7-6700K processor, with 8M of Cache, clock speed of 4.20GHz and 4 cores/8 way multi-task processing. From the point of view of memory, it is equipped with 40GB of DDR4 RAM, with serial speed of 2400MHz, and a Toshiba DT01ACA HDD with 7200RPM and 2TB of storage capacity. Also, it is equipped with a graphic processing unit (GPU) NVIDIA GeForce GTX 1080 with 8GB GDDR5X of video memory and 10 Gbps of memory frequency, and an ASUS Z170 pro-gaming motherboard. The operating system is Ubuntu 18.04. In order to efficiently implement the proposed approach, our models have been parallelized on the available GPU using Pytorch.

TABLE II
TOPOLOGY OF THE PROPOSED ATTENTION NETWORK, WHERE $n_{channels}$ INDICATES THE NUMBER OF CONSIDERED SPECTRAL BANDS.

Input convolutional layer		
ID	**Kernel size**	**Stride**
$C^{(1)}$	$64 \times 1 \times 1 \times n_{channels}$	$s = 1$
Attention module		
ID	**Processed data**	**Parameters**
$A^{(2)}$	$11 \times 11 \times 64$	$t = 2$ $r = 1$ $p = 1$ $DU = 2$
Average pool		
ID	**Kernel**	
$AVPOOL$	2×2	
Fully connected layer		
ID	**Input $\times$ output neurons**	**Activation**
FC	$576 \times n_{classes}$	Softmax

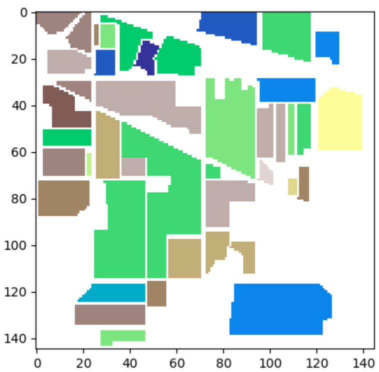
B. Hyperspectral Datasets

Four public and widely used HSI data sets have been considered in our experiments: Indian Pines (IP), University of Pavia (UP), Salinas Valley (SV), and Kennedy Space Center (KSC). Table III shows, for each dataset, its corresponding ground-truth with the number of samples per class. In the following, we summarize the characteristics of each dataset:

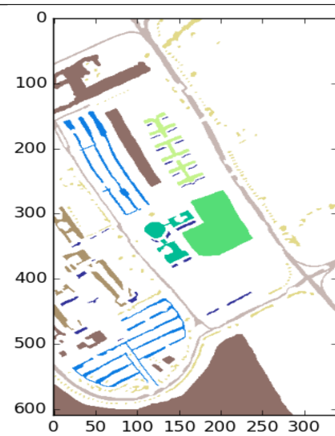
- **Indian Pines (IP)** dataset was collected by the Airborne Visible InfraRed Imaging Spectrometer (AVIRIS) [88] in 1992, over an agricultural area in Northwestern Indiana using 145×145 pixels with spatial resolution of 20 meters per pixel (mpp), and 224 spectral bands in the wavelength range from 0.4 to 2.5μm. After deleting 24 bands due to water absorption and null values, a total of 200 spectral bands are considered for experimental purposes. The ground-truth is divided into 16 different classes (Table III).
- **University of Pavia (UP)** dataset was collected by the reflective optics system imaging spectrometer (ROSIS) [89] in 2002, over the Engineering School at the University of Pavia, Northern Italy, using 610×340 pixels with spatial resolution of 1.3mpp, and 103 spectral bands in the wavelength range from 0.43 to 0.86μm. The ground-truth is divided into 9 different classes (Table III).
- **Salinas Valley (SV)** dataset was collected by the AVIRIS sensor in 1998 over an agricultural field in Salinas

TABLE III
NUMBER OF SAMPLES OF THE INDIAN PINES (IP), UNIVERSITY OF PAVIA (UP), SALINAS VALLEY (SV) AND UNIVERSITY OF HOUSTON (UH) DATASETS.

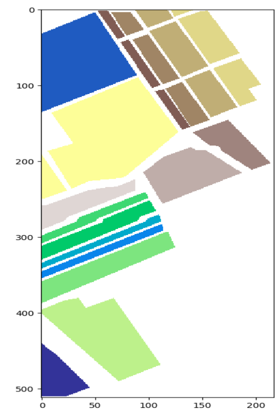
INDIAN PINES (IP)



UNIVERSITY OF PAVIA (UP)

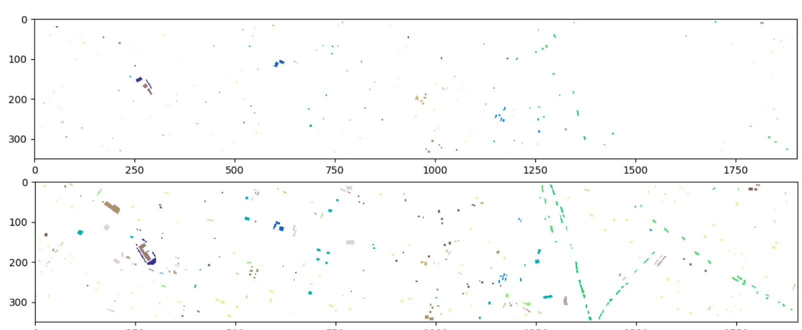


SALINAS (SV)



Color	Land-cover type	Samples	Color	Land-cover type	Samples	Color	Land-cover type	Samples
	Background	10776		Background	164624		Background	56975
	Alfalfa	46		Asphalt	6631		Brocoli-green-weeds-1	2009
	Corn-notill	1428		Meadows	18649		Brocoli-green-weeds-2	3726
	Corn-min	830		Gravel	2099		Fallow	1976
	Corn	237		Trees	3064		Fallow-rough-plow	1394
	Grass/Pasture	483		Painted metal sheets	1345		Fallow-smooth	2678
	Grass/Trees	730		Bare Soil	5029		Stubble	3959
	Grass/pasture-mowed	28		Bitumen	1330		Celery	3579
	Hay-windrowed	478		Self-Blocking Bricks	3682		Grapes-untrained	11271
	Oats	20		Shadows	947		Soil-vinyard-develop	6203
	Soybeans-notill	972					Corn-senesced-green-weeds	3278
	Soybeans-min	2455					Lettuce-romaine-4wk	1068
	Soybean-clean	593					Lettuce-romaine-5wk	1927
	Wheat	205					Lettuce-romaine-6wk	916
	Woods	1265					Lettuce-romaine-7wk	1070
	Bldg-Grass-Tree-Drives	386					Vinyard-untrained	7268
	Stone-steel towers	93					Vinyard-vertical-trellis	1807
	Total samples	21025		Total samples	207400		Total samples	111104

UNIVERSITY OF HOUSTON (UH)



Color	Land cover type	Samples train	Samples test
	Background	649816	
	Grass-healthy	198	1053
	Grass-stressed	190	1064
	Grass-synthetic	192	505
	Tree	188	1056
	Soil	186	1056
	Water	182	143
	Residential	196	1072
	Commercial	191	1053
	Road	193	1059
	Highway	191	1036
	Railway	181	1054
	Parking-lot1	192	1041
	Parking-lot2	184	285
	Tennis-court	181	247
	Running-track	187	473
	Total samples	2832	12197

Valley, California, using 512×217 spectral samples with 224 spectral bands (20 of which were discarded due to water absorption and noise). The ground-truth contains 16 classes (Table III).

- **University of Houston** (UH) dataset [90] provides an interesting benchmark, first presented by the IEEE Geoscience and Remote Sensing Society Image Analysis and Data Fusion Technical Committee during the 2013 data fusion contest [91]. It was gathered by the Compact Airborne Spectrographic Imager (CASI) in June 2012 over the campus of the University of Houston and the neighboring urban area, forming a ata cube of dimensions

$349 \times 1905 \times 144$, with spatial resolution of 2.5m and spectral information captured in the range from 0.38 to 1.05μm, containing 15 ground-truth classes divided in two categories: training (top UH map in Table III) and testing (bottom UH map in Table III).

C. Results and Discussion

In order to test the performance of proposed attention-guided network for spectral-spatial HSI data classification, four main experiments have been carried out:

1) Our first experiment one performs a comparison between the proposed attention-driven network and seven different and widely-used HSI classifiers available in the literature: i) random forest (RF), ii) multinomial logistic regression (MLR), iii) support vector machine (SVM), iv) multilayer perceptron (MLP), v) spectral CNN (CNN1D), vi) spatial CNN (CNN2D), and vii) spectral-spatial residual network (ResNet). In this context, the four HSI datasets described in the previous subsection have been used. We extracted patches of size $11 \times 11 \times 40$. For the IP scene, we used 15% of the available labeled data per class for training (and the rest of the available labeled data for testing). For the UP and SV scenes, we used 10% of the available labeled data for training. Finally, for the UH scene we used the available (fixed) training set (see Table III).
2) Our second experiment expands the initial comparison carried out in the first experiment using different classifiers and particularly focusing on different spectral-spatial methods carried out on the UP dataset with the fixed training set adopted in [92]. In this case, the following classifiers have been considered: i) Markov random field combined with Gaussian class-conditional model (MRF-Gauss), ii) contextual SVM (CSVM) [93], iii) CNN with extinction profiles (EP-CNN) [94], iv) CNN with a previously applied PCA (PCA), v) CNN with extended morphological profiles (EMP-CNN), and vi) CNN with Gabor filter (Gabor-CNN). Focusing on convolutional models, the EP-CNN is fed by patches of size $27 \times 27 \times n_{bands}$, while the proposed attentional model and PCA-CNN, EMP-CNN and Gabor-CNN employ input patches of size $27 \times 27 \times 3$.
3) Our third experiment performs a comparison between the original spectral-spatial ResNet and the proposed A-ResNet, evaluating the evolution of the overall accuracy of both classifiers when different training ratios are considered for the IP, UP and SV scenes. In particular, 5%, 10% and 15% ratios have been considered for the IP scene, and 1%, 5% and 10% ratios have been considered for the UP and SV scenes. Again, the input patches have been extracted with size of $11 \times 11 \times 40$.
4) Finally, our fourth experiment analyzes in detail the performance of the proposed network as compared with the original ResNet model in the presence of noisy data. In this case, several levels of noise have been tested, with noise being modeled as a normal distribution with $\mu = 0$ and $\sigma = \{0.10, 0.20, 0.40, 0.80, 1.60, 3.20, 6.40\}$.

In order to carry out the aforementioned comparisons, some widely-used measures have been considered, including the overall (OA) and average (AA) accuracy, the kappa coeficient (K) and the execution times (in seconds).

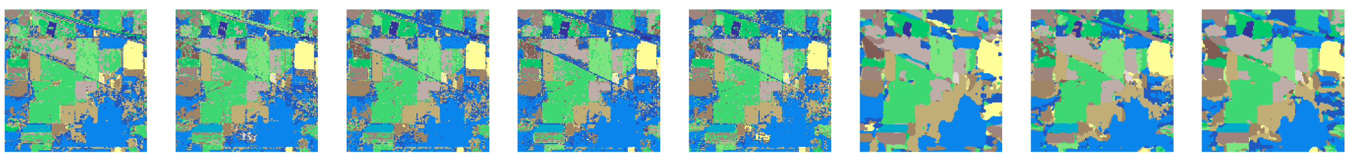
1) Experiment 1: Comparison between standard HSI classifiers and the proposed methods: First experiment performs a comparison between the proposed network and some of the most well-known HSI classifiers available in the literature. These methods can be divided into spectral-based ones (RF, MLR, SVM, MLP and CNN1D), spatial classifiers (CNN2D), and spectral-spatial classifiers (ResNet and A-ResNet). For all the spectral-spatial methods, the input patch size has been set to $11 \times 11 \times 40$. In order to perform a fair comparison, the ResNet has been implemented with the basic architecture of the proposed network in Table II, where the ResNet is composed by the same network's head and tail, and the same architecture of the trunk brach inside the network's body.

The obtained results are reported in Tables IV-VII, where the corresponding average and standard deviation values (obtained after five Monte Carlo runs) are also displayed. Focusing on the obtained OA values, we can observe that spatial and spectral-spatial methods are, in general, able to outperform pixel-based methods (RF, MLR, SVM, MLP and CNN1D), being residual based models (i.e. ResNet and A-Resnet) able to outperform the results obtained by the CNN2D. Focusing on the ResNet and the proposed A-Resnet, the performance of the latter is better than the performance of the former, being able to reach higher OA values than the original ResNet, in particular, in the classification of the IP and SV scenes. Another interesting aspect is the AA, which is higher in the proposed A-ResNet than in the original ResNet, indicating that, on average, the high OA achieved is not due to peaks in, say, very well ranked classes, but to a generally better rank for all classes. This is also supported by the smaller standard deviation values exhibited by our A-ResNet. In particular, we can highlight the good performance of the proposed model in small classes, (for instance Alfalfa and Oats in the IP scene or Lettuce romaine 6wk in the SV scene), where the A-ResNet is able to reach better accuracy values than the basic ResNet. Focusing on SV and UH scenes (Tables VI and VII, respectively), the obtained OA values may lead us to think that both ResNet and A-ResNet exhibit a similar behaviour. However, the standard deviation of A-ResNet is significantly smaller, indicating more robust and stable results (as the AA scores also suggest).

In addition, some of the obtained classification maps are shown in Figs. 5-7. It can be observed that the classification maps obtained by pixel-based classifiers show salt-and-pepper noise in almost the full IP dataset and in some classes of SV, particularly, Vinyard-untrained and Grapes-untrained. In the UP scene, the RF missclassifies a large amount of pixels in the Bare Soil class, for instance. In contrast, spectral-spatial methods greatly reduce these effects, with ResNet and A-ResNet being able to obtain classification maps that are close to the original ground-truth. In addition, if we compare the original ResNet to our A-ResNet, we can see that the classification maps produced by the latter exhibit borders between classes that are more sharply defined and clean than

TABLE IV
CLASSIFICATION RESULTS FOR INDIAN PINES (IP) DATASET USING 15% OF THE AVAILABLE LABELED DATA

Class	RF	MLR	SVM	MLP	CNN1D	CNN2D	ResNet	A-ResNet
Alfalfa	20.00±8.01	32.82±13.51	62.05±12.07	50.77±9.65	44.61±5.28	75.38±10.20	84.62±3.62	**89.23**±1.92
Corn-notill	61.53±1.95	75.07±0.99	81.45±1.21	78.90±3.01	81.04±2.19	91.54±0.76	94.64±1.48	**97.69**±0.85
Corn-min	53.62±2.61	57.96±2.32	70.55±2.40	66.27±2.53	70.69±0.36	86.95±3.59	95.26±3.32	**99.29**±0.48
Corn	35.12±2.90	45.67±6.03	72.93±4.93	61.19±6.40	60.10±2.79	88.56±4.02	84.48±6.83	**92.24**±2.61
Grass/Pasture	84.39±4.29	86.98±1.94	93.17±2.26	89.61±2.65	92.34±0.88	86.05±2.40	96.49±0.55	**99.02**±0.74
Grass/Trees	96.10±0.93	96.36±0.94	97.32±0.26	96.55±0.39	97.29±1.24	96.13±1.92	98.06±0.97	**99.77**±0.38
Grass/pasture-mowed	29.57±10.79	47.83±15.80	84.35±3.48	75.65±5.90	69.57±11.99	82.61±11.00	85.22±5.22	**93.04**±5.90
Hay-windrowed	96.11±2.98	99.16±0.71	98.32±0.66	97.54±1.36	98.18±0.79	97.88±0.40	100.00±0.00	100.00±0.00
Oats	1.18±2.35	18.82±6.86	51.76±12.56	61.18±15.16	44.70±16.89	65.88±21.18	68.24±10.91	**90.59**±7.98
Soybeans-notill	65.96±2.63	66.54±1.77	77.87±2.13	78.18±5.23	78.67±1.92	89.85±2.91	94.65±1.65	**98.57**±0.51
Soybeans-min	89.13±3.07	79.53±1.71	85.10±0.72	86.10±2.71	83.42±3.44	95.28±1.45	97.57±0.77	**99.37**±0.18
Soybean-clean	46.59±4.62	58.25±3.33	79.09±0.99	78.85±3.36	83.97±1.05	88.65±2.04	90.28±3.77	**97.14**±0.87
Wheat	92.18±4.20	98.51±0.59	98.39±1.23	98.74±0.67	98.62±0.28	97.82±2.42	99.89±0.23	**100.00**±0.00
Woods	94.53±0.59	95.31±0.75	95.59±0.54	94.55±1.30	94.51±0.97	98.40±0.57	99.14±0.32	**99.57**±0.31
Bldg-Grass-Tree-Drives	40.55±5.01	63.90±2.81	61.28±3.42	65.55±3.48	67.44±4.86	89.21±6.02	93.54±3.53	**99.58**±0.41
Stone-steel towers	83.54±1.96	85.06±2.58	87.60±5.74	89.37±4.43	87.59±3.53	82.53±6.27	89.87±6.46	**97.72**±1.68
OA	75.31±0.48	77.76±0.48	84.48±0.23	83.50±0.47	84.02±0.83	92.69±0.53	95.94±1.32	**98.75**±0.31
AA	61.88±0.98	69.24±1.51	81.05±1.44	79.31±1.23	78.30±1.01	88.29±2.01	92.00±2.27	**97.05**±1.01
K(x100)	71.41±0.54	74.46±0.56	82.26±0.28	81.13±0.54	81.75±0.90	91.65±0.60	95.37±1.51	**98.58**±0.36
Time (s.)	1.29±0.54	6.05±0.56	**0.25**±0.28	26.46±0.54	53.91±0.90	59.28±0.60	61.57±1.51	92.56±0.36

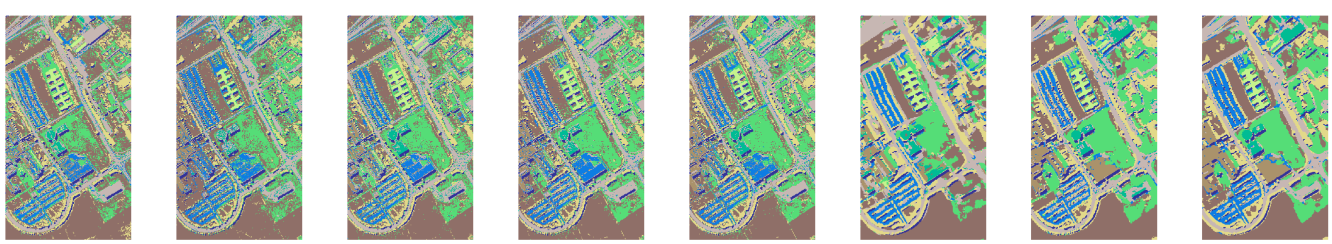


(a) RF (75.31%) (b) MLR (77.76%) (c) SVM (84.48%) (d) MLP (83.50%) (e) CNN1D (84.02%) (f) CNN2D (92.69%) (g) ResNet (95.94%) (h) A-ResNet (98.75%)

Fig. 5. Classification maps provided for the Indian Pines (IP) dataset by different methods (see Table IV).

TABLE V
CLASSIFICATION RESULTS FOR UNIVERSITY OF PAVIA (UP) DATASET USING 10% OF THE AVAILABLE LABELED DATA

Class	RF	MLR	SVM	MLP	CNN1D	CNN2D	ResNet	A-ResNet
Asphalt	91.63±0.58	92.39±0.59	94.29±0.47	93.81±1.33	95.85±0.52	98.01±0.65	99.01±0.27	**99.80**±0.09
Meadows	97.71±0.36	96.09±0.48	97.49±0.12	97.58±0.45	98.13±0.41	99.41±0.15	99.91±0.03	**99.97**±0.03
Gravel	66.88±2.70	73.27±0.98	80.84±1.30	78.11±3.87	81.48±1.98	93.90±1.73	97.82±0.42	**99.56**±0.24
Trees	89.10±1.25	86.90±1.34	94.21±1.18	93.59±1.25	94.15±1.34	98.14±0.36	99.28±0.15	**99.74**±0.07
Painted metal sheets	98.60±0.39	99.59±0.31	99.22±0.31	99.52±0.16	99.82±0.08	99.57±0.35	99.92±0.13	**99.97**±0.04
Bare Soil	64.35±1.30	77.83±0.77	90.91±0.71	91.64±1.27	91.71±1.66	98.08±0.49	99.99±0.02	**100.00**±0.00
Bitumen	77.66±1.29	56.34±4.95	87.35±1.12	85.53±2.34	87.52±0.88	89.72±2.86	96.86±0.53	**99.16**±0.32
Self-Blocking Bricks	88.52±0.77	86.68±1.18	87.47±0.48	88.92±1.25	85.68±2.32	98.28±0.69	98.13±0.26	**99.73**±0.20
Shadows	99.74±0.23	99.67±0.12	99.86±0.09	99.53±0.25	99.88±0.07	98.87±0.51	**99.95**±0.06	99.88±0.10
OA	89.37±0.15	89.73±0.31	94.10±0.10	94.04±0.22	94.61±0.21	98.27±0.14	99.39±0.06	**99.86**±0.04
AA	86.02±0.29	85.41±0.63	92.40±0.16	92.02±0.45	92.69±0.10	97.11±0.25	98.99±0.12	**99.76**±0.05
K(x100)	85.67±0.20	86.27±0.41	92.17±0.14	92.09±0.28	92.84±0.28	97.71±0.18	99.19±0.09	**99.82**±0.05
Time (s.)	4.29±0.20	8.63±0.41	**0.44**±0.14	68.22±0.28	139.58±0.28	139.82±0.18	93.63±0.09	205.89±0.05



(a) RF (89.37%) (b) MLR (89.73%) (c) SVM (94.10%) (d) MLP (94.04%) (e) CNN1D (94.61%) (f) CNN2D (98.27%) (g) ResNet (99.39%) (h) A-ResNet (99.86%)

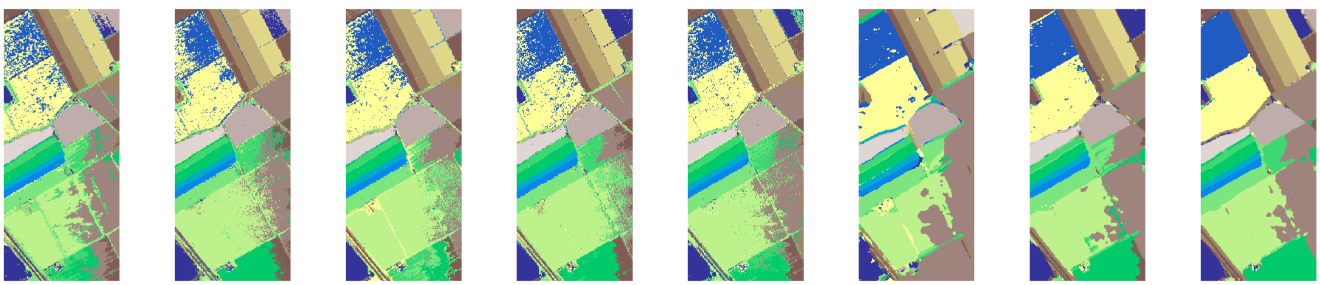
Fig. 6. Classification maps provided for the University of Pavia (UP) dataset by different methods (see Table V).

those obtained by the original ResNet (for instance, in the SV scene, the A-ResNet provides a better separation between the Fallow-rough-plow field and the Vinyard-vertical-trellis and Grapes-untrained classes).

2) Experiment 2: Comparison between advanced spectral-spatial HSI classifiers and the proposed method: In order to focus in more details on spectral-spatial classifiers, this experiment compares the proposed attentional model with

TABLE VI
CLASSIFICATION RESULTS FOR SALINAS VALLEY (SV) DATASET USING 10% OF THE AVAILABLE LABELED DATA

Class	RF	MLR	SVM	MLP	CNN1D	CNN2D	ResNet	A-ResNet
Brocoli green weeds 1	99.46±0.14	99.47±0.16	99.63±0.20	99.57±0.12	99.88±0.10	99.45±0.32	99.61±0.47	**99.95**±0.04
Brocoli green weeds 2	99.83±0.05	99.94±0.06	99.91±0.08	99.87±0.09	99.96±0.02	99.51±0.38	99.99±0.01	99.99±0.01
Fallow	99.15±0.42	98.60±0.77	99.68±0.09	99.44±0.28	**99.85**±0.16	99.62±0.21	98.75±0.48	99.01±0.35
Fallow rough plow	99.42±0.25	99.28±0.29	99.31±0.33	99.25±0.58	99.57±0.15	99.89±0.16	99.76±0.20	**99.92**±0.07
Fallow smooth	97.87±0.38	99.12±0.32	99.35±0.17	99.09±0.42	99.05±0.47	**99.88**±0.10	99.25±0.61	99.86±0.19
Stubble	99.68±0.10	99.92±0.06	99.80±0.17	99.85±0.09	99.85±0.07	99.78±0.26	100.00±0.00	100.00±0.00
Celery	99.39±0.09	**99.89**±0.06	99.54±0.17	99.57±0.22	99.84±0.06	99.64±0.10	99.82±0.09	99.88±0.13
Grapes untrained	84.42±0.93	87.98±0.50	90.51±0.40	86.88±1.75	90.98±1.33	95.60±0.42	97.45±2.51	**99.77**±0.09
Soil vinyard develop	99.07±0.17	99.73±0.17	99.92±0.03	99.73±0.23	99.83±0.18	99.54±0.20	99.98±0.02	**99.99**±0.01
Corn senesced green weeds	91.56±1.09	95.79±0.54	97.71±0.48	96.56±1.05	98.03±0.22	98.45±0.84	99.38±0.39	**99.92**±0.07
Lettuce romaine 4wk	94.13±0.69	95.90±1.02	98.88±0.39	97.81±0.34	98.33±0.94	98.73±0.90	98.96±0.43	**99.60**±0.16
Lettuce romaine 5wk	98.79±0.23	99.63±0.15	99.79±0.07	99.65±0.12	99.96±0.03	99.58±0.51	100.00±0.00	100.00±0.00
Lettuce romaine 6wk	97.86±0.92	99.03±0.45	98.88±0.98	99.03±0.20	99.17±0.58	99.13±0.95	98.91±0.39	**99.76**±0.28
Lettuce romaine 7wk	91.34±1.77	96.03±0.70	97.65±1.34	96.80±0.82	97.34±0.80	97.53±0.84	99.48±0.52	**99.94**±0.05
Vinyard untrained	60.46±2.51	66.63±0.91	70.54±1.26	77.81±2.20	79.52±1.99	95.01±1.21	97.47±2.06	**99.84**±0.08
Vinyard vertical trellis	97.06±0.84	98.89±0.52	99.18±0.28	99.08±0.26	99.00±0.30	97.00±1.35	**99.91**±0.14	99.84±0.13
OA	90.12±0.43	92.35±0.13	93.67±0.15	93.73±0.11	95.01±0.22	97.94±0.20	98.92±0.87	**99.85**±0.04
AA	94.34±0.31	95.99±0.13	96.89±0.20	96.87±0.06	97.51±0.17	98.65±0.25	99.29±0.41	**99.83**±0.05
K(x100)	88.98±0.48	91.47±0.14	92.94±0.17	93.02±0.11	94.44±0.24	97.71±0.22	98.80±0.97	**99.83**±0.04
Time (s.)	2.85±0.48	65.21±0.14	**0.94**±0.17	86.63±0.11	177.78±0.24	177.29±0.22	203.93±0.97	287.58±0.04



(a) RF (90.12%) (b) MLR (92.35%) (c) SVM (93.67%) (d) MLP (93.73%) (e) CNN1D (95.01%) (f) CNN2D (97.94%) (g) ResNet (98.92%) (h) A-ResNet (99.85%)

Fig. 7. Classification maps provided for the Salinas Valley (SV) dataset by different methods (see Table VI).

TABLE VII
CLASSIFICATION RESULTS FOR UNIVERSITY OF HOUSTON (UH) DATASET

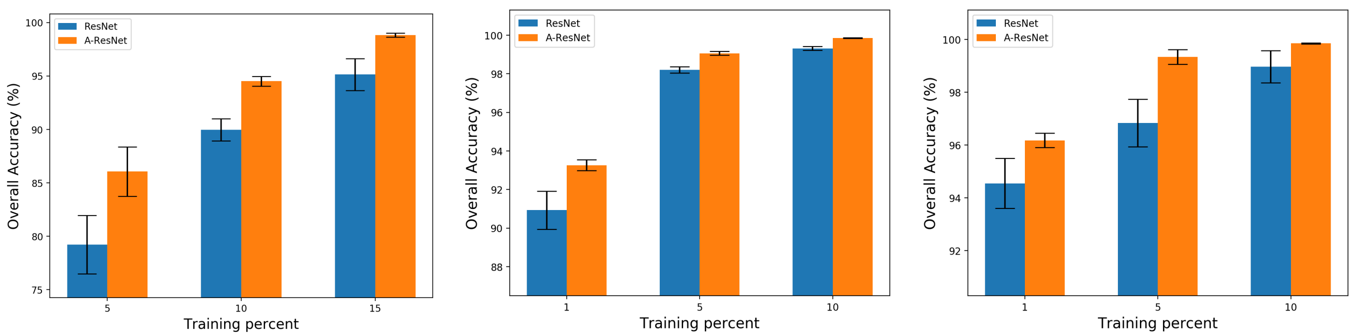
Class	RF	MLR	SVM	MLP	CNN1D	CNN2D	ResNet	A-ResNet
Grass healthy	82.49±0.05	**82.62**±0.00	82.34±0.00	81.58±0.38	81.75±0.69	80.48±2.48	82.15±0.47	81.39±1.05
Grass stressed	83.36±0.15	83.93±0.00	83.36±0.00	81.67±0.67	**95.04**±5.33	85.49±2.45	85.09±0.08	84.91±0.49
Grass synthetic	97.82±0.25	99.80±0.00	99.80±0.00	99.64±0.08	**99.88**±0.10	88.99±7.40	98.26±0.52	98.38±0.36
Tree	91.74±0.31	98.01±0.00	**98.96**±0.00	88.69±1.11	89.45±0.59	83.66±3.02	89.55±1.69	86.14±2.20
Soil	96.80±0.20	97.16±0.00	98.77±0.00	97.08±0.43	98.63±0.56	100.00±0.00	100.00±0.00	100.00±0.00
Water	**99.16**±0.28	94.41±0.00	97.90±0.00	94.41±0.00	95.94±1.68	92.59±2.33	95.80±0.00	97.34±1.90
Residential	75.28±0.47	74.25±0.00	77.43±0.00	76.79±2.03	**80.88**±3.59	74.65±3.56	77.28±0.93	76.96±5.42
Commercial	33.01±0.32	65.15±0.00	60.30±0.00	55.82±4.08	80.32±6.54	**80.85**±5.07	79.09±1.14	77.45±3.91
Road	69.40±0.35	69.12±0.00	76.77±0.00	69.91±5.40	77.09±5.76	81.34±3.26	**88.63**±2.35	88.35±1.06
Highway	43.86±0.31	54.44±0.00	61.29±0.00	49.71±3.46	72.57±13.83	63.69±1.40	71.47±10.58	**86.89**±12.40
Railway	70.36±0.25	76.09±0.00	80.55±0.00	75.67±1.37	86.36±6.43	93.74±3.18	**98.14**±1.16	96.28±1.45
Parking lot1	54.77±0.81	73.39±0.00	79.92±0.00	77.16±5.41	91.91±1.68	96.96±2.01	**98.79**±0.31	98.04±1.40
Parking lot2	60.14±0.36	68.42±0.00	70.88±0.00	72.21±2.98	74.74±3.34	**82.88**±3.09	80.42±3.24	79.37±5.21
Tennis court	98.87±0.40	98.79±0.00	100.00±0.00	99.03±0.20	99.36±0.32	98.79±1.33	100.00±0.00	100.00±0.00
Running track	97.50±0.21	95.98±0.00	96.41±0.00	98.31±0.33	98.14±0.49	97.34±3.23	**99.96**±0.08	99.87±0.25
OA	73.09±0.11	79.53±0.00	81.86±0.00	77.98±0.79	86.66±0.44	85.18±0.42	88.20±0.86	**88.71**±0.67
AA	72.16±0.08	76.97±0.00	79.04±0.00	81.18±0.68	88.14±0.35	86.76±0.21	89.64±0.75	**90.09**±0.37
K(x100)	71.09±0.11	77.89±0.00	80.43±0.00	76.29±0.85	85.53±0.47	83.90±0.45	87.18±0.93	**87.73**±0.73
Time (s.)	2.68±0.11	21.25±0.00	**0.37**±0.00	46.09±0.85	94.41±0.47	165.33±0.45	10.44±0.93	34.23±0.73

several spectral-spatial methods discussed in [92]. In this context, the proposed A-Resnet has been adapted to receive the same input data as PCA-CNN, EMP-CNN and Gabor-CNN, extracting from a fixed training set available for the UP scene [92] the same patches with size $27 \times 27 \times 3$.

The obtained results can be observed on Table VIII. Focusing on the methods described in [92], it is interesting to note that the convolution-based ones are able to reach the highest OA scores, being Gabor-CNN the best one in [92] (thanks to the ability of the Gabor filter to extract and encode highly

TABLE VIII
CLASSIFICATION RESULTS FOR UNIVERSITY OF PAVIA (UP) DATASET WITH THE FIXED TRAINING SET USED IN [92].

Class	MRF-Gauss	CSVM	EP-CNN	PCA-CNN	EMP-CNN	Gabor-CNN	ResNet	A-ResNet
Asphalt	84.84	92.56	88.43	92.23	**95.87**	87.75	86.53	90.74
Meadows	72.56	73.60	91.64	97.72	**99.50**	97.25	96.96	99.10
Gravel	65.12	71.68	75.95	52.85	61.12	70.92	89.31	**92.51**
Trees	96.63	**98.97**	96.53	89.46	94.81	97.09	93.03	92.89
Painted	99.91	**100.00**	98.56	99.46	95.15	98.83	98.38	97.21
Bare	92.34	**96.35**	57.87	57.66	64.84	64.62	55.36	66.16
Bitumen	91.95	**92.46**	80.43	91.42	80.63	76.66	85.12	82.06
Self-Blocking	94.59	97.41	98.10	98.06	97.26	**99.05**	97.32	96.88
Shadows	**98.99**	95.09	96.84	98.48	96.08	98.36	82.52	81.51
OA	81.78	84.58	87.01	88.93	91.37	91.62	89.45	**92.06**
AA	88.55	**90.90**	87.15	86.37	87.25	87.83	87.03	88.68
K(x100)	76.76	80.31	83.08	85.44	88.67	**89.14**	85.52	89.11

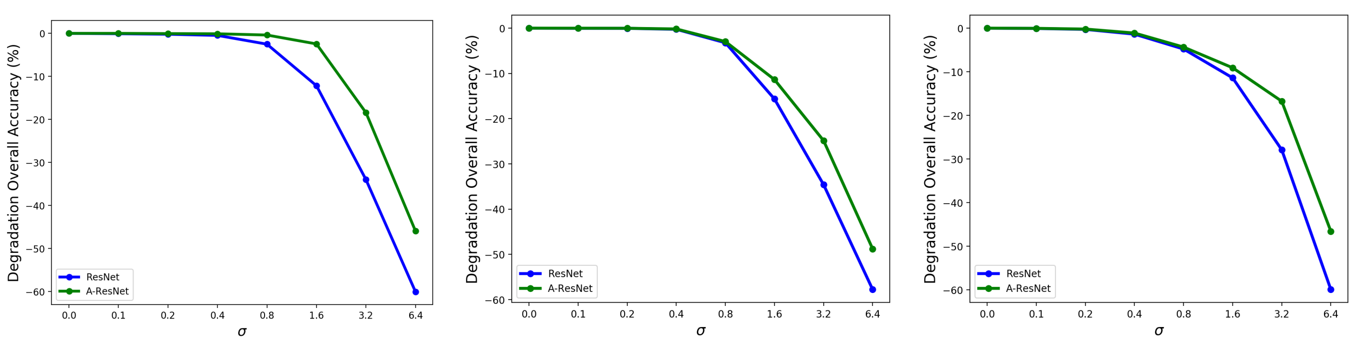


(a) IP scene (b) UP scene (c) SV scene

Fig. 8. Evolution of the overall accuracy (Y-axis) for the ResNet and the proposed model (A-ResNet) when classifying the IP (a), UP (b) and SV (c) hyperspectral scenes, using different training ratios.

TABLE IX
OVERALL ACCURACY OF RESNET AND THE PROPOSED MODEL (A-RESNET) OVER THE IP, UP AND SV DATASETS WHEN DIFFERENT NORMAL RANDOM PERTURBATIONS ARE INSERTED INTO THE DATA

Normal Perturbation	$\mu = 0$, $\sigma = 0.10$		$\mu = 0$, $\sigma = 0.20$		$\mu = 0$, $\sigma = 0.40$		$\mu = 0$, $\sigma = 0.80$		$\mu = 0$, $\sigma = 1.60$		$\mu = 0$, $\sigma = 3.20$		$\mu = 0$, $\sigma = 6.40$	
Dataset	**ResNet**	**A-ResNet**	**ResNet**	**A-ResNet**	**ResNet**	**A-ResNet**	**ResNet**	**A-ResNet**	**ResNet**	**A-ResNet**	**ResNet**	**A-ResNet**	**ResNet**	**A-ResNet**
IP	95.05	**98.84**	94.93	**98.78**	94.67	**98.74**	92.63	**98.45**	82.97	**96.38**	61.13	**80.40**	35.06	**52.87**
UP	99.29	**99.85**	99.28	**99.84**	99.07	**99.69**	96.09	**96.9**	83.67	**88.49**	64.74	**75.01**	41.57	**51.08**
SV	98.90	**99.81**	98.69	**99.64**	97.60	**98.73**	94.19	**95.51**	87.57	**90.77**	71.08	**83.1**	39.05	**53.28**



(a) IP scene (b) UP scene (c) SV scene

Fig. 9. Degradation of the overall accuracy (Y-axis) of ResNet and the proposed model A-ResNet for IP (a), UP (b), SV (c) and KSC (d), comparing the accuracy reached with the original data ($\sigma = 0$) and the accuracy reached with perturbed data, being $\sigma = \{0.01, 0.02, 0.03, 0.04, 0.05\}$ (X-axis)

discriminant spatial features). However, the A-ResNet is able to outperform the OA values of the methods reported in [92], exhibiting 92.06% OA, which is around 0.44 percentage points higher than the Gabor-CNN.

3) Experiment 3: Evolution of overall accuracy of ResNet and A-Resnet when different training ratios are considered:

Focusing on residual models, the original ResNet and the proposed A-ResNet, this experiment studies the behaviour of both models when different amounts of labeled data are available to perform the training step. The IP, UP and SV scenes have been considered, training the models with 5%, 10% and 15% of the available labeled samples for the IP scene, and 1%, 5% and 10% of the available labeled samples for the UP and SV scenes, respectively.

The obtained results are graphically displayed in Fig. 8. We can observe that, when few training samples are used (5% for IP and 1% for UP and SV, respectively), the proposed A-ResNet model is able to reach the best OA values with the lowest standard deviation, suggesting that the proposed method is able to better address the problem of overfitting when few training samples are provided to the network, obtaining robust results. As we feed more samples to the network, the accuracy gap between the original ResNet and proposed A-ResNet becomes smaller, although the deviation of the attentional network is always much smaller than that of the standard ResNet. This indicates that the proposed method is able to improve the standard ResNet when few training samples are employed, achieving at least the same result when a reasonable amount of training samples are used (see Fig. 8(c), obtained using 10% of the available labeled samples for the SV scene).

4) Experiment 4: Comparison between the basic ResNet and the proposed method: Motivated by the previous experiment, the fourth experiment studies in more detail the behavior of the basic ResNet and the proposed model A-ResNet. The goal of this experiment is to validate the performance and robustness of the proposed method with respect to ResNet when the test data is corrupted. In remote sensing, it is desirable to generate models that process data in a robust manner, for instance training and testing the classifier model with data obtained at different temporal acquisitions, or after different captures of the same area. These situations introduce certain disturbances or changes in the training and testing data to which the models must be able to respond in a reliable manner. As a result, this experiment evaluates how Resnet and A-ResNet behave when they have to deal with perturbed data.

In order to simulate perturbed data, the original IP, UP and SV datasets have been modified through a random normal distribution with mean $\mu = 0$ and seven different standard deviation values $\sigma = \{0.10, 0.20, 0.40, 0.80, 1.60, 3.20, 6.40\}$. Neural models have been trained over the original datasets using 15% of the available labeled samples from IP, and 10% of the available labeled samples from UP and SV. Again, patches of $11 \times 11 \times 40$ have been employed as the input data. The obtained results are given in Table IX.

With slight disturbances ($\sigma = 0.10$), we can observe that the ResNet exhibits a small decay of OA values in comparison with the case that no perturbations are present in the IP (-0.89) and UP (-0.1) datasets, while in the SV dataset the difference is very small (-0.02), as we can observe in Fig. 9. In turn, the A-ResNet is not significantly affected by the introduced perturbations. For instance, in the IP scene, it is even able to outperform the ResNet in terms of OA, being 0.09 percentual points better when noise is not included.

However, as the noise level increases, we can see how the OA of the standard ResNet decreases significantly, in particular from $\sigma = 1.6$. Therefore, the features extracted by the standard ResNet from these datasets are not relevant or generic enough to be applied in scenarios with perturbations. Instead of that, the performance of the proposed models remains more stable. For instance, for the IP dataset, the A-ResNet exhibits a degradation of 2.37 percentage points, while the ResNet exhibits a degradation of 12.97 points. Also, in the experiments with the UP and SV scenes, the ResNet is more affected than the A-ResNet, although the gap between the two seems smaller. However, with greater σ values, the gap becomes larger. This behaviour can be also observed for the rest of σ values (see Fig. 9): ResNet reaches the lowest OA and exhibits the worse degradation of performance with perturbed data, while the proposed model maintains a high OA and significantly lower degradation.

The OA values in Table IX and the degradation performance in Fig. 9 indicate that the proposed model is more robust to perturbations in the data, achieving high OA values. Also, it is able to extract more discriminative features from the original training data in comparison with ResNet, being the A-ResNet the most robust architecture for all datasets (even in presence of significant distortions).

V. Conclusions and Future Lines

In this work, a new model for spatial-spectral HSI classification has been proposed by combining a deep learning architecture (ResNet) and visual attention techniques. The filtering system introduced by the visual attention model, following bottom-up and top-down visual selection, allows for a post-processing of the extracted data, enhancing the quality of the feature extraction process as well as obtaining more representative and significant features, leading to a more precise and robust classification of HSI data.

Our experimental comparisons have been conducted using four publicly available HSI datasets, evaluating the proposed visual attention-driven model (A-ResNet) versus seven standard machine learning and deep learning classifiers and six advanced spectral-spatial methods, revealing that the proposed networks exhibit competitive results when compared to state-of-the-art techniques such as CNNs (combined with different techniques) and ResNets. Also, a deeper comparison between the ResNet and the proposed model with different amounts of training data and perturbed data revealed that our newly proposed model is able to extract more relevant, discriminative and complete features from HSI scenes, exhibiting robustness to network degradation when very limited training samples and/or highly disturbed data are considered.

As future work, we intend to improve the parameter optimization mechanism of the proposed network (particularly when very few labeled samples are available) in order to reduce the effect of overfitting. Also, we are planning to combine additional visual attention techniques with other deep models, with the aim of enhancing the quality of the extracted features and the final classification results.

Acknowledgement

The authors would like to gratefully thank the Associate Editor and the three Anonymous Reviewers for their outstanding comments and suggestions, which greatly helped us to improve the technical quality and presentation of our work.

References

[1] P. Ghamisi, N. Yokoya, J. Li, W. Liao, S. Liu, J. Plaza, B. Rasti, and A. Plaza, "Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 37–78, Dec 2017.

[2] D. Chutia, D. K. Bhattacharyya, K. K. Sarma, R. Kalita, and S. Sudhakar, "Hyperspectral remote sensing classifications: A perspective survey," *Transactions in GIS*, vol. 20, no. 4, pp. 463–490, 2015.

[3] M. Fauvel, Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton, "Advances in spectral-spatial classification of hyperspectral images," *Proceedings of the IEEE*, vol. 101, no. 3, pp. 652–675, 2013.

[4] A. F. Goetz, G. Vane, J. E. Solomon, and B. N. Rock, "Imaging spectrometry for earth remote sensing," *Science*, vol. 228, no. 4704, pp. 1147–1153, 1985.

[5] D. Landgrebe, "Hyperspectral image data analysis," *IEEE Signal processing magazine*, vol. 19, no. 1, pp. 17–28, 2002.

[6] D. Haboudane, J. R. Miller, E. Pattey, P. J. Zarco-Tejada, and I. B. Strachan, "Hyperspectral vegetation indices and novel algorithms for predicting green lai of crop canopies: Modeling and validation in the context of precision agriculture," *Remote sensing of environment*, vol. 90, no. 3, pp. 337–352, 2004.

[7] X. Zhang, Y. Sun, K. Shang, L. Zhang, and S. Wang, "Crop classification based on feature band set construction and object-oriented approach using hyperspectral images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 9, pp. 4117–4128, 2016.

[8] S. L. Martin and T. George, "Applications of hyperspectral image analysis for precision agriculture," in *Micro-and Nanotechnology Sensors, Systems, and Applications X*, vol. 10639. International Society for Optics and Photonics, 2018, p. 1063916.

[9] J. M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. Nasrabadi, and J. Chanussot, "Hyperspectral remote sensing data analysis and future challenges," *IEEE Geoscience and remote sensing magazine*, vol. 1, no. 2, pp. 6–36, 2013.

[10] P. W. Yuen and M. Richardson, "An introduction to hyperspectral imaging and its application for security, surveillance and target acquisition," *The Imaging Science Journal*, vol. 58, no. 5, pp. 241–253, 2010.

[11] E. Puckrin, C. S. Turcotte, M.-A. Gagnon, J. Bastedo, V. Farley, and M. Chamberland, "Airborne infrared hyperspectral imager for intelligence, surveillance, and reconnaissance applications," in *Airborne Intelligence, Surveillance, Reconnaissance (ISR) Systems and Applications IX*, vol. 8360. International Society for Optics and Photonics, 2012, p. 836004.

[12] B. Uzkent, A. Rangnekar, and M. J. Hoffman, "Aerial vehicle tracking by adaptive fusion of hyperspectral likelihood maps," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*. IEEE, 2017, pp. 233–242.

[13] G. A. Carter, K. L. Lucas, G. A. Blossom, C. L. Lassitter, D. M. Holiday, D. S. Mooneyhan, D. R. Fastring, T. R. Holcombe, and J. A. Griffith, "Remote sensing and mapping of tamarisk along the colorado river, usa: a comparative use of summer-acquired hyperion, thematic mapper and quickbird data," *Remote Sensing*, vol. 1, no. 3, pp. 318–329, 2009.

[14] P. Ghamisi, J. Plaza, Y. Chen, J. Li, and A. J. Plaza, "Advanced spectral classifiers for hyperspectral images: A review," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 1, pp. 8–32, 2017.

[15] J. Haut, M. Paoletti, J. Plaza, and A. Plaza, "Cloud implementation of the K-means algorithm for hyperspectral image analysis," *Journal of Supercomputing*, vol. 73, no. 1, 2017.

[16] J.-M. Yang, P.-T. Yu, and B.-C. Kuo, "A nonparametric feature extraction and its application to nearest neighbor classification for hyperspectral image data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 3, pp. 1279–1293, 2010.

[17] J. A. Gualtieri and R. F. Cromp, "Support vector machines for hyperspectral remote sensing classification," in *27th AIPR Workshop: Advances in Computer-Assisted Recognition*, vol. 3584. International Society for Optics and Photonics, 1999, pp. 221–233.

[18] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Transactions on geoscience and remote sensing*, vol. 42, no. 8, pp. 1778–1790, 2004.

[19] G. Mercier and M. Lennon, "Support vector machines for hyperspectral image classification with spectral-based kernels," in *Geoscience and Remote Sensing Symposium, 2003. IGARSS'03. Proceedings. 2003 IEEE International*, vol. 1. IEEE, 2003, pp. 288–290.

[20] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 6, pp. 1351–1362, 2005.

[21] J. Haut, M. Paoletti, A. Paz-Gallardo, J. Plaza, and A. Plaza, "Cloud implementation of logistic regression for hyperspectral image classification," in *Proceedings of the 17th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2017*, J. Vigo-Aguiar, Ed., Costa Ballena (Rota), Cádiz, Spain, 2017, pp. 1063–2321.

[22] J. Ham, Y. Chen, M. M. Crawford, and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 3, pp. 492–501, 2005.

[23] R. E. Bellman, *Adaptive control processes: a guided tour*. Princeton university press, 2015, vol. 2045.

[24] D. L. Donoho *et al.*, "High-dimensional data analysis: The curses and blessings of dimensionality," *AMS Math Challenges Lecture*, vol. 1, p. 32, 2000.

[25] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[26] W. Li, H. Fu, L. Yu, P. Gong, D. Feng, C. Li, and N. Clinton, "Stacked autoencoder-based deep learning for remote-sensing image classification: a case study of african land-cover mapping," *International Journal of Remote Sensing*, vol. 37, no. 23, pp. 5632–5646, 2016.

[27] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[28] Z. Lin, Y. Chen, X. Zhao, and G. Wang, "Spectral-spatial classification of hyperspectral image using autoencoders," in *2013 9th International Conference on Information, Communications & Signal Processing*. IEEE, 2013, pp. 1–5.

[29] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE Journal of Selected topics in applied earth observations and remote sensing*, vol. 7, no. 6, pp. 2094–2107, 2014.

[30] Y. Chen, X. Zhao, and X. Jia, "Spectral-spatial classification of hyperspectral data based on deep belief network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 6, pp. 2381–2392, 2015.

[31] C. Tao, H. Pan, Y. Li, and Z. Zou, "Unsupervised spectral–spatial feature learning with stacked sparse autoencoder for hyperspectral imagery classification," *IEEE Geoscience and remote sensing letters*, vol. 12, no. 12, pp. 2438–2442, 2015.

[32] Y. Liu, G. Cao, Q. Sun, and M. Siegel, "Hyperspectral classification via deep networks and superpixel segmentation," *International Journal of Remote Sensing*, vol. 36, no. 13, pp. 3459–3482, 2015.

[33] L. Wang, J. Zhang, P. Liu, K.-K. R. Choo, and F. Huang, "Spectral–spatial multi-feature-based deep learning for hyperspectral remote sensing image classification," *Soft Computing*, vol. 21, no. 1, pp. 213–221, 2017.

[34] H. Luo, Y. Y. Tang, X. Yang, L. Yang, and H. Li, "Autoencoder with extended morphological profile for hyperspectral image classification," in *2017 3rd IEEE International Conference on Cybernetics (CYBCONF)*. IEEE, 2017, pp. 1–4.

[35] S. Paul and D. N. Kumar, "Spectral-spatial classification of hyperspectral data with mutual information based segmented stacked autoencoder approach," *ISPRS journal of photogrammetry and remote sensing*, vol. 138, pp. 265–280, 2018.

[36] Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio, "Object recognition with gradient-based learning," in *Shape, contour and grouping in computer vision*. Springer, 1999, pp. 319–345.

[37] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, "A new deep convolutional neural network for fast hyperspectral image classification," *ISPRS J. Photogrammetry Remote Sens.*, to be appeared, doi:10.1016/j.isprsjprs.2017.11.021, 2017.

[38] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, Oct. 2016.

[39] J. M. Haut, M. E. Paoletti, J. Plaza, J. Li, and A. Plaza, "Active learning with convolutional neural networks for hyperspectral image classification using a new bayesian approach," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–22, 2018.

[40] Z. Zhong, J. Li, Z. Luo, and M. Chapman, "Spectral-spatial residual network for hyperspectral image classification: A 3-d deep learning framework," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 847–858, Feb 2018.

[41] M. E. Paoletti, J. M. Haut, R. Fernandez-Beltran, J. Plaza, A. J. Plaza, and F. Pla, "Deep pyramidal residual networks for spectral-spatial hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–15, 2018.

[42] W. Wang, S. Dou, Z. Jiang, and L. Sun, "A fast dense spectral–spatial convolution network framework for hyperspectral images classification," *Remote Sensing*, vol. 10, no. 7, p. 1068, 2018.

[43] M. Paoletti, J. Haut, J. Plaza, and A. Plaza, "Deep&dense convolutional neural network for hyperspectral image classification," *Remote Sensing*, vol. 10, no. 9, p. 1454, 2018.

[44] F. Deng, S. Pu, X. Chen, Y. Shi, T. Yuan, and S. Pu, "Hyperspectral image classification with capsule network using limited training samples," *Sensors*, no. 18, p. 3153, 2018.

[45] M. E. Paoletti, J. M. Haut, R. Fernandez-Beltran, J. Plaza, A. J. Plaza, and F. Pla, "Capsule networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–16, 2018.

[46] T. S. Nazaré, G. B. P. da Costa, W. A. Contato, and M. Ponti, "Deep convolutional neural networks and noisy images," in *Iberoamerican Congress on Pattern Recognition*. Springer, 2017, pp. 416–424.

[47] N. Imamoglu, Y. Oishi, X. Zhang, G. Ding, Y. Fang, T. Kouyama, and R. Nakamura, "Hyperspectral image dataset for benchmarking on salient object detection," in *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2018, pp. 1–3.

[48] A. Borji, H. R. Tavakoli, and Z. Bylinskii, "Bottom-up attention, models of," *arXiv preprint arXiv:1810.05680*, 2018.

[49] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, Nov 1998.

[50] X. Chen, H. Huoa, F. Taoa, D. Lib, and Z. Lia, "A computational method to emulate bottom-up attention to remote sensing images," in *Proc. XXI Congress International Society for Photogrammetry and Remote Sensing (ISPRS)*, vol. 37, 2008, pp. 244–249.

[51] L. Zhang, H. Li, P. Wang, and X. Yu, "Detection of regions of interest in a high-spatial-resolution remote sensing image based on an adaptive spatial subsampling visual attention model," *GIScience & remote sensing*, vol. 50, no. 1, pp. 112–132, 2013.

[52] L. Zhang and K. Yang, "Region-of-interest extraction based on frequency domain analysis and salient region detection for remote sensing image," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 5, pp. 916–920, 2014.

[53] L. Zhang, K. Yang, and H. Li, "Regions of interest detection in panchromatic remote sensing images based on multiscale feature fusion," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 12, pp. 4704–4716, Dec 2014.

[54] D. Zhu, B. Wang, and L. Zhang, "Airport target detection in remote sensing images: A new method based on two-way saliency," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 5, pp. 1096–1100, 2015.

[55] L. Zhang and A. Li, "Region-of-interest extraction based on saliency analysis of co-occurrence histogram in high spatial resolution remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 5, pp. 2111–2124, May 2015.

[56] T. Li, J. Zhang, X. Lu, and Y. Zhang, "Sdbd: A hierarchical region-of-interest detection approach in large-scale remote sensing image," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 5, pp. 699–703, May 2017.

[57] Z. Li and L. Itti, "Saliency and gist features for target detection in satellite images," *IEEE Transactions on Image Processing*, vol. 20, no. 7, pp. 2017–2029, July 2011.

[58] F. Bi, B. Zhu, L. Gao, and M. Bian, "A visual search inspired computational model for ship detection in optical satellite images," *IEEE Geoscience and Remote Sensing Letters*, vol. 9, no. 4, pp. 749–753, July 2012.

[59] X. Ke and G. He, "Visual attention based model for target detection in high resolution remote sensing images," in *2012 International Conference on Computer Vision in Remote Sensing*, Dec 2012, pp. 84–89.

[60] C. Wang, X. Bai, S. Wang, J. Zhou, and P. Ren, "Multiscale visual attention networks for object detection in vhr remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2018.

[61] S. Kumar, J. Ghosh, and M. M. Crawford, "Best-bases feature extraction algorithms for classification of hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 7, pp. 1368–1379, July 2001.

[62] Q. Wang, J. Lin, and Y. Yuan, "Salient band selection for hyperspectral image classification via manifold ranking," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 6, pp. 1279–1289, June 2016.

[63] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," *arXiv preprint arXiv:1412.7755*, 2014.

[64] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[65] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2017, pp. 6450–6458.

[66] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[67] D. H. Hubel and T. N. Wiesel, "Receptive fields of single neurones in the cat's striate cortex," *The Journal of physiology*, vol. 148, no. 3, pp. 574–591, 1959.

[68] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 2, pp. 22–40, June 2016.

[69] Q. Du and J. E. Fowler, "Hyperspectral image compression using jpeg2000 and principal component analysis," *IEEE Geoscience and Remote sensing letters*, vol. 4, no. 2, pp. 201–205, 2007.

[70] J. M. Haut, M. E. Paoletti, J. Plaza, and A. Plaza, "Fast dimensionality reduction and classification of hyperspectral images with extreme learning machines," *Journal of Real-Time Image Processing*, pp. 1–24, 2018.

[71] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[72] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural computation*, vol. 29, no. 9, pp. 2352–2449, 2017.

[73] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.

[74] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 2377–2385.

[75] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[76] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[77] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. icml*, vol. 30, no. 1, 2013, p. 3.

[78] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[79] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2017, pp. 5987–5995.

[80] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[81] D. Han, J. Kim, and J. Kim, "Deep pyramidal residual networks," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2017, pp. 6307–6315.

[82] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*. Springer, 2016, pp. 630–645.

[83] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," *arXiv preprint arXiv:1602.07261*, 2016.

[84] J. M. Haut, R. Fernandez-Beltran, M. E. Paoletti, J. Plaza, A. Plaza, and F. Pla, "A new deep generative network for unsupervised remote sensing single-image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–19, 2018.

[85] V. A. Lamme, H. Super, H. Spekreijse *et al.*, "Feedforward, horizontal, and feedback processing in the visual cortex," *Current opinion in neurobiology*, vol. 8, no. 4, pp. 529–535, 1998.

[86] A. Mahdi and J. Qin, "Deepfeat: A bottom up and top down saliency model based on deep features of convolutional neural nets," *arXiv preprint arXiv:1709.02495*, 2017.

[87] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[88] R. O. Green, M. L. Eastwood, C. M. Sarture, T. G. Chrien, M. Aronsson, B. J. Chippendale, J. A. Faust, B. E. Pavri, C. J. Chovit, M. Solis, M. R. Olah, and O. Williams, "Imaging spectroscopy and the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS)," *Remote Sensing of Environment*, vol. 65, no. 3, pp. 227–248, 1998. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0034425798000649

[89] B. Kunkel, F. Blechinger, R. Lutz, R. Doerffer, H. van der Piepen, and M. Schroder, "ROSIS (Reflective Optics System Imaging Spectrometer) - A candidate instrument for polar platform missions," in *Proc. SPIE 0868 Optoelectronic technologies for remote sensing from space*, J. Seeley and S. Bowyer, Eds., 1988, p. 8.

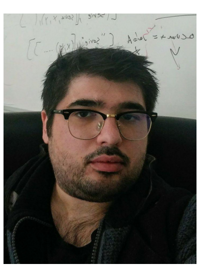
[90] X. Xu, f. Lil, and A. Plaza, "Fusion of hyperspectral and LiDAR data using morphological component analysis," in *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2016, pp. 3575–3578.

[91] C. Debes, A. Merentitis, R. Heremans, J. Hahn, N. Frangiadakis, T. van Kasteren, W. Liao, R. Bellens, A. Piurica, S. Gautama, W. Philips, S. Prasad, Q. Du, and F. Pacifici, "Hyperspectral and lidar data fusion: Outcome of the 2013 grss data fusion contest," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 2405–2418, June 2014.

[92] P. Ghamisi, E. Maggiori, S. Li, R. Souza, Y. Tarablaka, G. Moser, A. De Giorgi, L. Fang, Y. Chen, M. Chi *et al.*, "New frontiers in spectral-spatial hyperspectral image classification: the latest advances based on mathematical morphology, markov random fields, segmentation, sparse representation, and deep learning," *IEEE Geoscience and Remote Sensing Magazine*, vol. 6, no. 3, pp. 10–43, 2018.

[93] P. Gurram and H. Kwon, "Contextual svm using hilbert space embedding for hyperspectral classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 5, pp. 1031–1035, 2013.

[94] P. Ghamisi, B. Höfle, and X. X. Zhu, "Hyperspectral and lidar data fusion using extinction profiles and deep convolutional neural network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 6, pp. 3011–3024, 2017.



Juan Mario Haut (S17) received the B.Sc and M.Sc. degrees in computer engineering from the University of Extremadura, Caceres, Spain, in 2011 and 2014, respectively. He is currently a Member of the Hyperspectral Computing Laboratory, Department of Computers and Communications, University of Extremadura, as Ph.D. student, with an University Teacher Training Programme from the Spanish Ministry of Education. His research interests include remote sensing and analysis of very high spectral resolution with the current focus on deep learning and cloud computing.

Mercedes Paoletti (S17) received the B.Sc and M.Sc. degrees in computer engineering from the University of Extremadura, Caceres, Spain, in 2014 and 2016, respectively. She is currently a Member of the Hyperspectral Computing Laboratory, Department of Technology of Computers and Communications, University of Extremadura, as Ph.D. student, with an University Teacher Training Programme from the Spanish Ministry of Education. Her research interests include remote sensing and analysis of very high spectral resolution with the current focus on deep learning and high performance computing.

Javier Plaza (M09-SM15) is a member of the Hyperspectral Computing Laboratory at the Department of Technology of Computers and Communications, University of Extremadura, where he received the M.Sc. degree in 2004 and the PhD degree in 2008, both in Computer Engineering. He was the recipient of the Outstanding Ph.D. Dissertation Award at the University of Extremadura in 2008. His main research interests comprise hyperspectral data processing and parallel computing of remote sensing data. He has authored more than 150 publications, including over 50 JCR journal papers, 10 book chapters, and 90 peer-reviewed conference proceeding papers. He has guest edited 4 special issues on hyperspectral remote sensing for different journals. He is an Associate Editor for IEEE Geoscience and Remote Sensing Letters and an Associate Editor of the IEEE Remote Sensing Code Library. He is a recipient of the Best Column Award of the IEEE Signal Processing Magazine in 2015 and the most highly cited paper (2005-2010) in the Journal of Parallel and Distributed Computing. He received best paper awards at the IEEE International Conference on Space Technology and the IEEE Symposium on Signal Processing and Information Technology. http://www.umbc.edu/rssipl/people/jplaza

Antonio Plaza (M05-SM07-F15) is the Head of the Hyperspectral Computing Laboratory at the Department of Technology of Computers and Communications, University of Extremadura, where he received the M.Sc. degree in 1999 and the PhD degree in 2002, both in Computer Engineering. His main research interests comprise hyperspectral data processing and parallel computing of remote sensing data. He has authored more than 600 publications, including over 200 JCR journal papers (over 160 in IEEE journals), 23 book chapters, and around 300 peer-reviewed conference proceeding papers. He has guest edited 10 special issues on hyperspectral remote sensing for different journals. Prof. Plaza is a Fellow of IEEE for contributions to hyperspectral data processing and parallel computing of Earth observation data. He is a recipient of the recognition of Best Reviewers of the IEEE Geoscience and Remote Sensing Letters (in 2009) and a recipient of the recognition of Best Reviewers of the IEEE Transactions on Geoscience and Remote Sensing (in 2010), for which he served as Associate Editor in 2007-2012. He is also an Associate Editor for IEEE Access (receiving a recognition as outstanding Associate Editor of the journal in 2017), and was a member of the Editorial Board of the IEEE Geoscience and Remote Sensing Newsletter (2011-2012) and the IEEE Geoscience and Remote Sensing Magazine (2013). He was also a member of the steering committee of the IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS). He is a recipient of the Best Column Award of the IEEE Signal Processing Magazine in 2015, the 2013 Best Paper Award of the JSTARS journal, and the most highly cited paper (2005-2010) in the Journal of Parallel and Distributed Computing. He received best paper awards at the IEEE International Conference on Space Technology and the IEEE Symposium on Signal Processing and Information Technology. He served as the Director of Education Activities for the IEEE Geoscience and Remote Sensing Society (GRSS) in 2011-2012, and as President of the Spanish Chapter of IEEE GRSS in 2012-2016. He has reviewed more than 500 manuscripts for over 50 different journals. He served as Editor-in-Chief of the IEEE Transactions on Geoscience and Remote Sensing from 2013 to 2017. Additional information: http://www.umbc.edu/rssipl/people/aplaza

Jun Li was born in Hunan, China, in Feburary 1982. She received the Geographical Information Systems degree from Hunan Normal University in 2004, the M.Sc. degree in Remote sensing and Photogrammetry from Peking University in 2007, and the Ph.D. degree in Electrical and Computer Engineering from Instituto Superior Tecnico, Technical University of Lisbon in 2011. From 2011 to 2012, she was a Post-Doc Researcher with the Department of Technology of Computers and Communications, University of Extremadura, Spain. Dr Li is currently a Professor with the School of Geography and Planning, Sun Yat-Sen University, China, where she founded her own research group on hyperspectral image analysis in 2013. Since then, she has obtained several prestigious funding grants at the national and international level. She has published a total of 69 journal citation report (JCR) papers, 48 conference international conference papers, and 1 book chapter. She has received a significant number of citations to her published works, with several papers distinguished as Highly Cited Papers in Thomson Reuters Web of Science - Essential Science Indicators (WoS-ESI). She has served as a Guest Editor of a Special Issue in the prestigious Proceedings of the IEEE journal. She has also served as a Guest Editor of a Special Issue in the prestigious ISPRS Journal of Photogrammetry and Remote Sensing journal. She was elevated to IEEE Senior Member status in 2016, and is currently serving as Associate Editor of the IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS) since 2014. Her students have also obtained important distinctions and awards at international conferences and symposia. Her main research interests comprise remotely sensed hyperspectral image analysis, signal processing, supervised/semisupervised learning and active learning.