






Hybrid Dense Network With Attention Mechanism for Hyperspectral Image Classification

Muhammad Ahmad , Adil Mehmood Khan , *Member, IEEE*, Manuel Mazzara , Salvatore Distefano, Swalpa Kumar Roy , *Student Member, IEEE*, and Xin Wu , *Member, IEEE*

Abstract—The nonlinear relation between the spectral information and the corresponding objects (complex physiognomies) makes pixelwise classification challenging for conventional methods. To deal with nonlinearity issues in hyperspectral image classification (HSIC), convolutional neural networks (CNN) are more suitable, indeed. However, fixed kernel sizes make traditional CNN too specific, neither flexible nor conducive to feature learning, thus impacting on the classification accuracy. The convolution of different kernel size networks may overcome this problem by capturing more discriminating and relevant information. In light of this, the proposed solution aims at combining the core idea of 3-D and 2-D inception net with the attention mechanism to boost the HSIC CNN performance in a hybrid scenario. The resulting attention-fused hybrid network (AfNet) is based on three attention-fused parallel hybrid subnets with different kernels in each block repeatedly using high-level features to enhance the final ground-truth maps. In short, AfNet is able to selectively filter out the discriminative features critical for classification. Several tests on HSI datasets provided competitive results for AfNet compared to state-of-the-art models.

Index Terms—Attention mechanism, convolutional neural network (CNN), hyperspectral images classification (HSIC), inception network.

I. INTRODUCTION

HYPERSPECTRAL imaging (HSI) systems based on collections of electromagnetic spectrum, ranging from visible to near-infrared region, reflected by the objects of interest [1]. The images thus obtained are usually generated by a preconfigured HSI camera installed on either mobile (e.g., satellites, drones, air-crafts) or static (e.g., indoor, rooms, labs) setups depending upon the problem at hand [2]–[11]. Thereby, HSI sensors gather a huge amount of data from hundreds of contiguous spectral bands [12], [13].

Such big and rich HSI dataset, including the different spectral bands data related by the (spatial) geo-located position, may contain hidden information and patterns. HSI classification (HSIC) [14], [15] aims to discover, detect, identify, and recognize such patterns. However, the spectral dataset size usually increases combinatorially with the problem size (e.g., the area, the resolution), leading to the curse of dimensionality, and, thus, making traditional HSIC methods inefficient [16]. To mitigate such issues, several dimensionality reduction and feature selection techniques have been proposed [17], [18]. Conventional feature extraction/selection methods rely on hand-crafted features, however, due to the spatial variability of spectral information [19], the extraction of discriminative and most informative features is still a big challenge [20].

Hand-crafted features may be insubstantial in the case of HSI data, therefore, it is challenging to achieve a tradeoff between discriminability and robustness on features, which also considerably differ on the different HSI datasets. Furthermore, the process of feature design and selection is strongly affected by the designer and architect knowledge and skills [20]–[23]. To such a purpose, an automatic approach to hierarchically identify the features was developed by Hinton back in 2006 [21], [24], [25], based on a deep learning model developed in growing semantic layers until a desirable representation is achieved. Similarly, other models have been proposed on feature learning and classification as well [21], [26]–[28], automatically learning and improving the underlying system representation from the available data without any prior knowledge. They can extract both linear and nonlinear features, thus capable of handling HSI data in both spatial and spectral domains [29].

In nature, HSI datasets are usually nonlinear due to the undesired light-scattering phenomena given by land cover objects and particles in the atmosphere [30]. Thus rendering the use of linear transformation or feature learning methods [31] for

Manuscript received January 16, 2022; revised March 13, 2022 and April 19, 2022; accepted April 22, 2022. Date of publication May 3, 2022; date of current version May 25, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62101045, in part by the China Postdoctoral Science Foundation Funded Project 2021M690385, in part by the Natural Science Foundation of Guangxi, China under Grant 2021GXNSFBA220056, and in part by The Analytical Center for the Government of the Russian Federation Agreement 70-2021-00143 dd. 01.11.2021, under Grant I GK 000000D730321P5Q0002. (*Corresponding author: Xin Wu*)

Muhammad Ahmad is with the Department of Computer Science, National University of Computer and Emerging Sciences, Islamabad, Chiniot 35400, Pakistan (e-mail: mahmad00@gmail.com).

Adil Mehmood Khan is with the Institute of Data Science and Artificial Intelligence, Innopolis University, 420500 Innopolis, Russia (e-mail: a.khan@innopolis.ru).

Manuel Mazzara is with the Institute of Software Development and Engineering, Innopolis University, 420500 Innopolis, Russia (e-mail: m.mazzara@innopolis.ru).

Salvatore Distefano is with the Dipartimento di Matematica e Informatica—MIFT, University of Messina, 98121 Messina, Italy (e-mail: sdistefano@unime.it).

Swalpa Kumar Roy is with the Department of Computer Science and Engineering, Jalpaiguri Government Engineering College, West Bengal 735102, India (e-mail: swalpa@cse.jgec.ac.in).

Xin Wu is with the School of Computer Science (National Pilot Software Engineering School), Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: 040251522wuxin@163.com).

Digital Object Identifier 10.1109/JSTARS.2022.3171586

HSIC. To overcome the nonlinearity issues, convolutional neural network (CNN) was proposed to extract both high as well as low-level features which ultimately lead to the extraction of abstract and invariant features [32], [33]. As a result, 2-D CNN achieved remarkable performance but unfortunately not so good for HSIC due to the missing channel-related information, i.e., 2-D CNN are not able to learn spectrally discriminative features. Unlike 2-D CNN, 3-D ones can jointly extract the spatial–spectral information for HSIC providing higher accuracy than 2-D CNN [34]. However, 3-D CNN models are both computationally and time-intensive due to the high number of parameters involved by 3-D convolutional filters on each layer.

For instance, in [35], a spatial–spectral residual network (SSRN) implemented a 3-D CNN residual network based on ResNet [36]. Despite SSRN achieves remarkable classification results, the summation method used to aggregate features at each layer requires output feature maps to have consistent scale as the residual feature maps. Hence each layer has self weights which overall lead to an explosion of network parameters [30]. Thereby, high accuracy comes at expense of the computational power, i.e., SSRN is more complex than traditional 3-D CNN. Similarly, Ahmad *et al.* [21] proposed a fast and compact 3-D CNN (FC-3D-CNN) model to overcome the limitations of computational cost and reduce the number of training parameters. FC-3D-CNN achieves better results in a computationally efficient manner than SSRN due to the reduced spectral information used in the experimental process. Recently, Roy *et al.* introduced trainable kernel for dilation and erosion operation to extract more meaningful morphological features from HSI and classify them in various land used and land covers [37].

In general, CNN models tend to poorly perform especially in the case of pixels of different classes but similar texture over contiguous spectral bands [38]. To deal with that, the authors in [18], [39], and [40] proposed hybrid models that combine the power of 2-D and 3-D convolutional layers to extract high and low-level features, i.e., extraction of abstract and invariant features. The hybrid models achieve better accuracy than state-of-the-art 2-D and 3-D CNN solutions. Despite the high accuracy, hybrid models still require a large number of parameters as compared to the 3-D CNN network [21] while, on the other hand, 3-D CNN/SRNN has a longer processing time than hybrid models. Therefore, inception models have proved that the network topology significantly affects the complexity and the accuracy [41]. Recently, graph convolutional network (GCN) shows the superiority in HSIC, Hong *et al.* [42] proposed a novel GCN in a mini-batch fashion, called miniGCN, which solves the problem of large-scale graph computation and learning. Apart from the complexity and accuracy tradeoff, all the inception models have one common property, i.e., a split–transform–merge strategy which proved to be a good strategy for HSIC. Traditional 2-D, 3-D, hybrid, and inception models exploit the fixed convolution kernel size, however, HSI class distribution is complicated thus conventional CNN with fixed kernel size is not flexible enough. Convolutions with different spatial sizes may capture more discriminative and important information for pixel-based HSIC.

Nowadays, an attention mechanism has been extensively used to suppress redundant information, while extracting features for classification. SENet [43] was the first network proposed to suppress the redundant features by weighting channel direction features. The work [44] proposed CBAM (convolutional block attention module) combines spatial attention with channel attention through pooling, whereas the work [45] proposed a NLNet which combines the convolution operations with matrix multiplication operations to capture the long-range relationship in the global space. A combination of SENet and NLNet was proposed in [46], which consists of a simplified lightweight module GCNet for effectively extracting global context. Very recently, a novel transformer framework was rapidly developed and its spectral version, called SpectralFormer, was for the first time proposed with the application to HSIC, yielding state-of-the-art classification performance [47]. In addition, multimodal fusion transformer (MFT) opens another direction in research for joint hyperspectral and LiDAR classification [48].

Though, attention networks have achieved remarkable results for HSIC based on the internal architecture of the attention module. These works, to some extent, put attention weights in either one or two dimensions and ignore the rest of the HSI dimensions. For instance, single and double attention networks were proposed in [49] and [50], in which the work [49] only consider the spectral information and ignore the spatial information, whereas, the work [50] was proposed to reduce the interference between the spatial and channel information. The work [51] was proposed to jointly explore the spectral and spatial information, where spectral–spatial dimensions were weighted by the spectral–spatial attention module. The combination of attention in more or less in one or two dimensions may improve the performance, however, it is highly recommended to integrate all channel information for better classification.

Wang *et al.* [52] significantly improved the squeeze and excitation structure attention mechanism proposed in [43], reducing the model complexity by a local cross-channel interaction strategy without any preprocessing, i.e., dimensional reduction. Zheng *et al.* [53] worked to overcome the limitations of inconsistent class ratio and overparameterization using a stratified sample-based training strategy. While the spectral attention module was proposed to render the soft band selection process to further refine the redundant spectrum information. However, all these spatial or spectral attention models are to some extent isolated in which the spatial attention ignores to make a connection between spectral dimension, whereas the spectral attention ignores the spatial connection and uses only the correlation between different spectral bands.

Moreover, it has been proven fact that accuracy improves, while increasing the depth of the model, thus requiring more parameters and higher computational burden, which makes optimization an NP-hard problem. Therefore, to overcome the afore-said issues, this work explicitly investigates the possibilities of combining the core idea of 2-D as well as 3-D inception models into a hybrid attention architecture to boost the pixel-based HSIC performance. We tested the model on several publicly available HSI datasets, which shows competitive results compared to the

state-of-the-art methods. Our proposed pipeline achieved an overall accuracy of 97% for the Indian Pines dataset, 100% for Botswana, 99% for both Pavia University, and Salinas datasets, respectively. In a nutshell, the contributions made in this article are summarized as follows.

- 1) A novel densely connected hybrid inception net is proposed to enrich the spatial–spectral feature learning process. Different from the conventional inception model which consists of a single branch in each block, the proposed densely connected hybrid blocks are composed of parallel multiple sized filters which significantly improves the propagation and reuse of features with less number of tuneable parameters. Moreover, the proposed hybrid inception net comprehensively utilizes features of different scales from HSI dataset.
- 2) A dual-branch attention fusion block is introduced which boosts the robustness of the discriminative network. As compared with recently published attention blocks for HSIC, the proposed pipeline simultaneously model interactions across different spectral bands and spatial regions by reweighting the significance of features. The triple-branch attention block adaptively emphasizes the important information and significantly suppresses the redundant and ineffective information.
- 3) A hybrid spectral convolutional block is used to reduce the required number of parameters for the HSIC model. Moreover, the activation inducted convolutional layer can further improve the nonlinear representation capacity of the whole network.

The rest of the article is organized as follows. Section II presents the pipeline proposed in this article. Section III describes the experimental settings along the metrics used to compute the accuracies. Sections III-A–III-C exhibits the experimental datasets used to conduct the experiments and to validate the proposed methodology. Moreover, Sections III-A–III-C demonstrates the experimental results as compared with the state-of-the-art methods proposed in the literature. Finally, Section IV concludes this article.

II. PROBLEM FORMULATION

Lets assume $R = [r_1, r_2, r_3, \dots, r_L]^T \in R^{L \times n}$ be the HSI cube, where $n = N \times M$ samples associated with C classes and L band images. Each $r_i = (r_i, c_j)$ where c_j be the class associated with r_i sample. In nature, r_i exhibit high intraclass variability and interclass similarity, overlapping and nested regions. Therefore, HSI cube has been first divided into small spatial patches to overcome the aforesaid issues. For each patch, the ground truths are formed based on the central pixel of the patch. Principle component analysis (PCA) has been used before creating the patches which eliminate the redundancy among the band images, i.e., $L \rightarrow B$, where $B \ll L$.

The patching process creates neighboring patches $P \in R^{S \times S \times B}$ centered at the spatial location (a, b) covering $S \times S$ spatial windows [21], [39]. The total of Z patches given by $(U - S + 1) \times (V - S + 1)$. Thus, in total, these patches cover the width from $\frac{a+(S-1)}{2}$ to $\frac{a-(S-1)}{2}$ and height from $\frac{b+(S-1)}{2}$

to $\frac{b-(S-1)}{2}$ [39]. The Z patches are first convolved with a kernel function which computes the sum of the dot product between the input patch and kernel function to introduce the nonlinearity [21], [25], [39], [54]. The activation maps for spatial–spectral position (x, y, z) at i th feature map and j th layer can be represented as $v_{i,j}^{(x,y,z)}$

$$v_{i,j}^{x,y,z} = ReLu \left(\sum_{\tau=1}^{d_{i-1}} \sum_{\rho=-\gamma}^{\gamma} \sum_{\phi=-\delta}^{\delta} \sum_{\lambda=-\nu}^{\nu} w_{i,j,\tau}^{\rho,\phi,\lambda} \times v_{(i-1),\tau}^{(x+\rho),(y+\phi),(z+\lambda)} + b_{i,j} \right) \quad (1)$$

where d_{i-1} be the total number of feature maps at $(i-1)$ -th layer, $w_{i,j}$ and $b_{i,j}$ be the depth of the kernel and bias, respectively. Moreover, $2\gamma + 1$, $2\delta + 1$, and $2\nu + 1$ be the height, width, and depth of the kernel [21]. Similarly, 2-D modules do the same process with 2-D input as well as the 2-D kernel function. In both 3-D and 2-D layers, the kernel is striding over the input to cover the whole spatial dimension. More specifically, as the proposed model combines the power of 3-D and 2-D kernel functions, thus, 2-D convolution $V_{i,j}^{x,y}$ represents the activation value of i th feature map at (x, y) spatial position on j th layer and can be formulated as $v_{i,j}^{x,y}$ and finally can be formed as follows:

$$v_{i,j}^{x,y} = ReLu \left(\sum_{\tau=1}^{d_{i-1}} \sum_{\rho=-\gamma}^{\gamma} \sum_{\phi=-\delta}^{\delta} w_{i,j,\tau}^{\rho,\phi} \times v_{i-1,\tau}^{(x+\rho),(y+\phi)} + b_{i,j} \right) \quad (2)$$

where $2\gamma + 1$ and $2\delta + 1$ be the height and width of the kernel, respectively. In short, the proposed hybrid attention-fused hybrid network (AfNet) convolutional filters are as follows with the input of $9 \times 9 \times 15$. The size of 3-D filters are $3D_1 = (7 \times 7 \times 9)$, $K_1^1 = 7$, $K_1^2 = 7$, $K_1^3 = 9$, $3D_2 = (5 \times 5 \times 7)$, $K_2^1 = 5$, $K_2^2 = 5$, $K_2^3 = 7$, and $3D_3 = (3 \times 3 \times 5)$, $K_3^1 = 3$, $K_3^2 = 3$ and $K_3^3 = 5$ for each layer on each block with different number of filters, i.e., (30, 20, 10) for first block, (40, 20, 10) for second block and (60, 30, 10) for third block. Similarly the size of 2-D filters are $2D_1 = (3 \times 3)$, $K_1^1 = 3$, $K_1^2 = 3$, $2D_2 = (3 \times 3)$, $K_2^1 = 3$, $K_2^2 = 3$, and $2D_3 = (1 \times 1)$, $K_3^1 = 1$, $K_3^2 = 1$ for each layer on each block with different number of filters, i.e., (16, 32, 64) for first block, (16, 32, 64) for second block, and (16, 32, 64) for third block. A 2-D fusion module has been used to incorporate the information learned hierarchically at different blocks. Finally, a 2-D convolutional layer is used with (1×1) kernel size with total of 128 filters to better represent the low to high level information.

To decrease the number of spectral–spatial feature maps, nineteen densely connected 3-D and 2-D convolutional layers are used prior to the flatten layer to make sure the convolutional process discriminate the spatial information, while considering different spectral bands with no loss and less number of parameters to boost the performance [21]. The weights are randomized initially and optimized using Adam optimizer based on back-propagation with a soft-max loss function. Later the

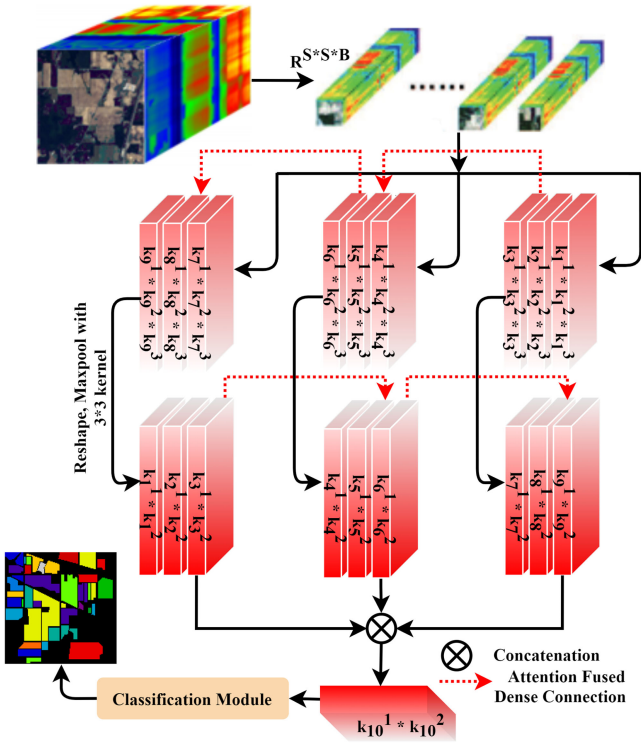


Fig. 1. Irrespective of the traditional dense connections, this work proposes dense connections along with the attention mechanism among different inception blocks highlighted in red dotted lines. AfNet for HSIC, where $S \times S$ be the height and width of the patch, B refer to the number of Bands, $(k_{1 \rightarrow 9}^1 \times k_{1 \rightarrow 9}^2 \times k_{1 \rightarrow 9}^3)$ are 3-D Conv layers and $(k_{1 \rightarrow 10}^1 \times k_{1 \rightarrow 10}^2)$ are 2-D Conv layers. The size of 3-D filters are $3D_1 = (7 \times 7 \times 9)$, $3D_2 = (5 \times 5 \times 7)$, $3D_3 = (3 \times 3 \times 5)$ with different number of filters, i.e., (30, 20, 10) for first block, (40, 20, 10) for second block and (60, 30, 10) for third block. Similarly the size of 2-D filters are $2D_1 = (3 \times 3)$, $2D_2 = (3 \times 3)$, and $2D_3 = (1 \times 1)$. A 2-D fusion module has been used to incorporate the information learned hierarchically at different blocks. Finally, a 2-D convolutional layer is used with (1×1) kernel size with total of 128 filters to better represent the low to high level information.

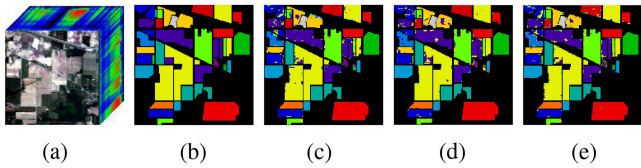


Fig. 2. Indian Pines Dataset illustration in-terms of HSI cube and ground truth (GT) label maps. The classification maps obtained using AfNet with overall accuracy of 92.82 (%), 2-D inception net with overall accuracy of 94.18 (%), and 3-D inception net with overall accuracy of 81.51 (%). (a) IP Cube. (b) GT. (c) AfNet. (d) 2D. (e) 3D.

randomized weights are updated using a mini-batch size of 256 for 50 epochs. The overall structure of the proposed hybrid AfNet using the Indian Pines dataset as an example is presented in Fig. 1.

A. Dense Connections and Attention Blocks

CNN extracts different features with different characteristics on each layer in which lower and middle layer features have relatively high resolution and encompass more location and detailed information. However, it may have lower semantics and more noise due to fewer convolutional layers passing through.

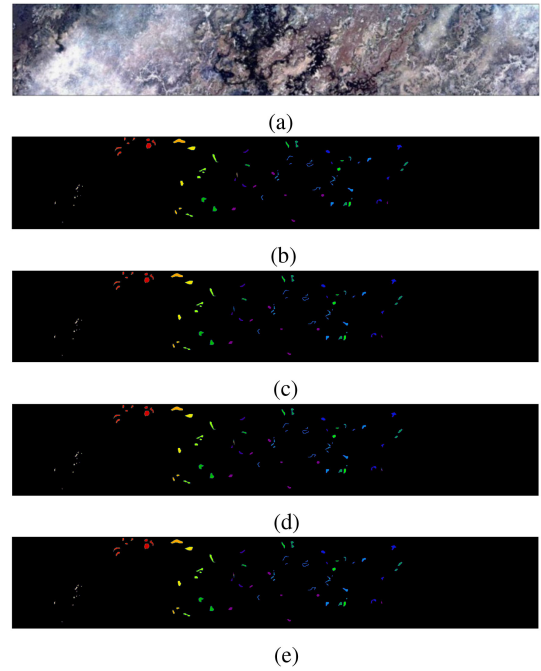


Fig. 3. Botswana Dataset illustration in-terms of HSI Cube and Ground-Truth (GT) label maps. The classification maps obtained using AfNet with overall accuracy of **98.94 (%)**, 2-D Inception Net with overall accuracy of 98.81 (%) and 3-D Inception Net with overall accuracy of 98.46 (%). (a) BS Cube. (b) GT. (c) AfNet. (d) 2D. (e) 3D.

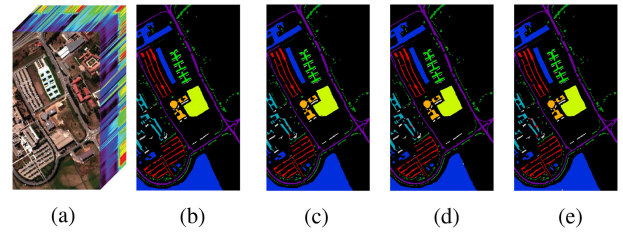


Fig. 4. Pavia University Dataset illustration in-terms of HSI Cube and Ground Truth (GT) label maps. The classification maps obtained using AfNet with overall accuracy of **99.27 (%)**, 2-D inception net with overall accuracy of 99.09 (%) and 3-D inception net with overall accuracy of 98.55 (%). (a) PU Cube. (b) GT. (c) AfNet. (d) 2D. (e) 3D.

Since we know that the high-level features hold strong semantic information with low resolution and poor perception capability. Therefore, cross-layer feature fusion can be considered as an effective strategy to preserve the quality features and ultimately improve classification performance [55].

Dense connectivity (e.g., different kinds of connectivity patterns irrespective of the traditional network) has been first proposed as DenseNet and widely used framework in many real-life applications. Traditionally, all layers are connected one after another, in order to maximize the feature information flow between layers. In this hierarchy, each layer accepts the features of all previous layers in front of it as input and passes its output to the subsequent layer. However, irrespective of the traditional dense connections, this article proposes dense connections along with the attention mechanism among different inception network blocks, as shown in Fig. 1. From Fig. 1, one can see that the output of the second convolutional layer of block-1 is densely

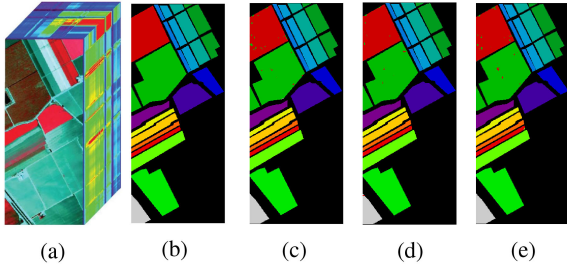


Fig. 5. Salinas Dataset illustration in-terms of HSI Cube and Ground Truth (GT) label maps. The classification maps obtained using AfNet with overall accuracy of 99.59 (%), 2-D inception net with overall accuracy of **99.76 (%)** and 3-D inception net with overall accuracy of 99.29 (%). (a) SA Cube. (b) GT. (c) AfNet. (d) 2D. (e) 3D.

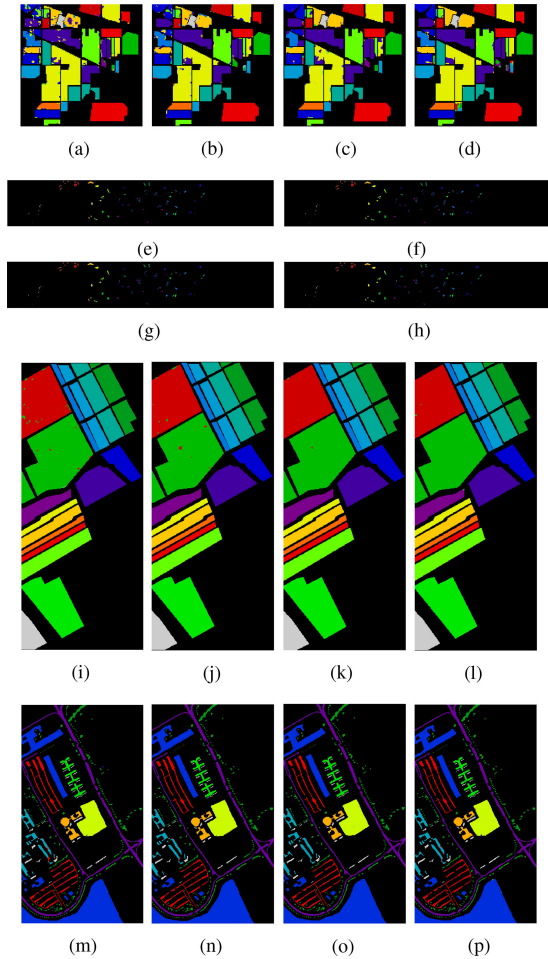


Fig. 6. Classification performance in terms of ground truth maps on different spatial dimensions. (a) IP:9×9. (b) IP:11×11. (c) IP:13×13. (d) IP:15×15. (e) BS:9×9. (f) BS:11×11. (g) BS:13×13. (h) BS:15×15. (i) SA:9×9. (j) SA:11×11. (k) SA:13×13. (l) SA:15×15. (m) PU:9×9. (n) PU:11×11. (o) PU:13×13. (p) PU:15×15.

connected with the second layer of block-2, where the other layers of each block are densely connected traditionally as well followed by the nonlinear transformation in both cases.

Let us assume X_i be the output of the i th block and X_0 be the output of the previous convolutional block. Thus the output of the i th convolutional block is not only related to the output of ($i-1$)th block but also includes the middle output of all previous

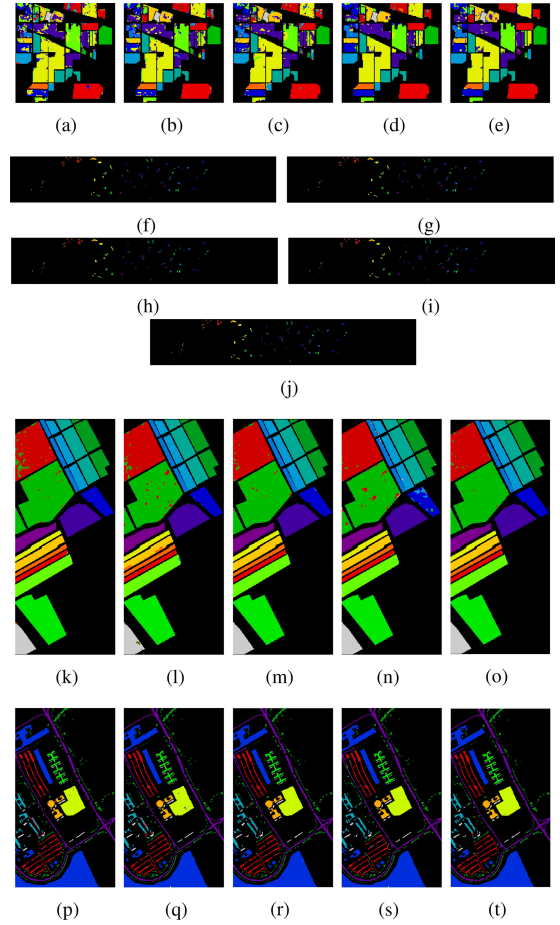


Fig. 7. Classification performance of different percentages of training samples in terms of ground truth maps. (a) IP:5%. (b) IP:7%. (c) IP:10%. (d) IP:12%. (e) IP:15%. (f) BS:5%. (g) BS:7%. (h) BS:10%. (i) BS:12%. (j) BS:15%. (k) SA:5%. (l) SA:7%. (m) SA:10%. (n) SA:12%. (o) SA:15%. (p) PU:5%. (q) PU:7%. (r) PU:10%. (s) PU:12%. (t) PU:15%.

blocks. Similarly, each 3-D CNN block is densely connected along with the attention mechanism with the 2-D CNN blocks, respectively, as just explained above. However, while connecting the 3-D feature maps with 2-D feature maps, a reshape and max-pooling with 3×3 kernel is used. Finally, a concatenation (fusion) layer is deployed to fuse the output maps obtained from all three blocks, and subsequently, a 2-D convolutional layer is used to further refine the feature maps obtained from the densely connected network. The attention blocks are flexible in the proposed model and can be positioned anywhere in the network as explained in Fig. 1.

B. Overview

For high-level intuition of the proposed model, the overall structure has been illustrated in Fig. 1 in which each block of the network is densely connected with the help of an attention mechanism. The proposed AfNet is an end-to-end framework for HSI in which the input $R^{L \times n}$ is the original HSI dataset and output is regarded as the probability of each HSI pixel for c_j classes.

Since HSI composed of hundreds of contiguous spectral bands, and some of these are highly correlated with each other, which provides no new information for classification. Moreover, some noisy bands exist in HSI, therefore, to eliminate the noisy and redundant bands, PCA transformation has been applied before feature learning and classification, which significantly reduces the processing time and memory capacity as well.

The HSI cube has been divided into overlapping 3-D cubes to take full advantage of spatial–spectral information present in the HSI dataset. These 3-D cubes are composed of the target pixel and its neighboring pixels to perform pixel-level feature learning and classification. Let us assume each 3-D patch formed from HSI is $R^{S \times S \times B}$, where $S \times S$ denotes the spatial size of each patch and B be the number of PCs preserved in the spectral dimension. In AfNet, a densely connected convolutional (3-D and 2-D) layers with rectified linear unit (ReLU) and without normalization is a basic building unit of the network.

The 3-D patches are first to go through nine attention-based 3-D interconnected convolutional layers. The obtained feature maps are further fed to nine attention-based 2-D interconnected convolutional layers which are designed to improve the propagation and reuse of features with a fewer number of tuneable parameters. Moreover, a hybrid structure can discriminate the spatial information, while considering different spectral bands with no loss to increase the generalization performance. On the other hand, the proposed network is built by stacking multiscale parallel filters of various sizes. On top of that, the attention module has been added to handle the skip connections from the first block to the second and all other subsequent blocks, which improves the flow of information. As a result, the fused features provide surpass feature space as compared to the stacked single-sized convolutional layers.

Another way around, attention blocks are incorporated to selectively filter out the information (features) which are critical for classification, i.e., weakening information that is useless for classification has been eliminated which leads to obtaining a feature representation that is more discriminative to get a higher class probability for each pixel. Following the fusion, a 2-D convolutional layer is employed to aggregate the obtained features once again. Afterward, feature maps are converted into feature vectors by flattening, and finally, class labels are generated using Softmax.

III. EXPERIMENTAL SETTINGS

The experimental results explained in this article have been obtained through Google Colab, an online platform to execute any python environment with graphical process unit (GPU), up to 358+ GB of cloud storage, and up to 25 GB of random access memory (RAM). In all the stated experiments (not only for the proposed method but also for all comparative methods), the training, validation, and test sets are randomly divided into three parts; 15%/15%/70% (i.e., Training/Validation/Test sets).

For fair comparative analysis and to make the claims more reliable, the learning rate for all experimental methods are set to 0.001, Relu as the activation function for all hidden layers except the final (output) later on which Softmax activation function is used. Patch size is set to 9×9 for all experimental results, and 15 most informative spectral dimensions were selected using

PCA to reduce the computational burden in terms of time and space. All the models are evaluated on 100 epochs without batch normalization.

To compute the experimental results, several metrics, such as average accuracy (AA), overall accuracy (OA), and Kappa (κ) have been used, where κ metric is known as a statistical metric that considered the mutual information regarding a strong agreement among the classification and ground-truth maps. AA represents the average classwise classification performance, whereas the OA is computed as the number of correctly classified examples out of the total test examples.

A. Experimental Datasets and Initial Experiments

Several publicly available hyperspectral datasets have been used to evaluate the performance of the proposed AfNet. These datasets are acquired at different times and locations using different sensors, such as Hyperion NASA EO-1 satellite, reflective optics system imaging spectrometer (ROSIS), and airborne visible/infrared imaging spectrometer (AVIRIS) sensor. Further information regarding the experimental datasets can be found from [21], [56]–[58]. As earlier explained, the performance of AfNet is evaluated using four publicly available HSI datasets, namely, Indian Pines, Pavia University, Botswana, and Salinas. For each of the above datasets, the samples are randomly split into three subsets, i.e., training, validation, and test sets. Table I provides a summary or description of each dataset used in the following experiments.

1) *Indian Pines*: Indian Pines dataset was acquired back in 1992, June 12 over the Purdue University Agronomy farms, northwest of West Lafayette and the surrounding area using AVIRIS sensor. This dataset was mainly acquired to facilitate soil research being initiated by Prof. Marion Baumgardner and his graduate students. Indian Creek and Pine Creek watersheds contain most of the part of the dataset thus known by Indian Pines and include two flight lines: 1) Flown East-West and 2) Flown North-South. There are three 2×2 miles intensive test sites with the area as; 1) northern portion of north-south flight line, 2) near the center, and 3) southern portion.

Indian Pines dataset consists of 145×145 spatial dimensions per spectral band and in total 224 spectral bands in the wavelength range $0.4 - 2.510^{-6}$ m. The scene used in this research is a subset of a larger scene. It consists of 1/3 forest, 2/3 agriculture, and natural perennial vegetation, a rail line, two major dual-lane highways, low-density housing, other structures, and small roads as earlier explained. Some crops, e.g., corn, soybeans are in the early stages of growth, i.e., less than 5% coverage due to the reason that the dataset was acquired in June. The ground truths are available and distinguished into 16 nonmutual exclusive classes. The image cube and true ground truths label maps are shown in Fig. 2(a) and (b), whereas Fig. 2(c)–(e) show the classification performance in terms of classification maps (ground-truth label maps) for three different models, i.e., AfNet, 3-D attention inception net, and 2-D attention inception net. These maps clearly show that the proposed method performs better than 3-D as well as 2-D attention inception networks. The higher accuracies are emphasized.

2) *Botswana*: The Hyperion NASA EO-1 satellite acquired the Botswana dataset over OkavangoDelta, Botswana back in

TABLE I
SUMMARY OF THE HIS DATASETS USED IN THE FOLLOWING EXPERIMENTS

Dataset	Year	Source	Spatial dimensions	Spectral	Wavelength	Samples	Classes	Sensor	Resolution
Botswana	2001-2004	NASA EO-1	1496 × 256	242 bands	400-2500	3248	14	Satellite	30
Indian Pines	1992	NASA AVIRIS	145 × 145	220 bands	400 - 2500	10249	16	Aerial	20
Salinas	1998	NASA AVIRIS	512 × 217	224 bands	360 - 2500	54129	16	Aerial	3.7
Pavia University	2001	ROSIS-03 sensor	610 × 610	115 bands	430 - 860	42776	9	Aerial	1.3

2001–2004. The WO-1 sensor acquired the subject data at 30-m pixel resolution over a 7.7-km strip in 242 spectral bands covering the 400–2500-nm portion of the spectrum in 10-nm window.

To mitigate the effects of bad detectors, inter detector mis-calibration, and intermittent anomalies, extensive preprocessing has been carried out by UT Center for space research. While processing, uncalibrated and noisy bands (i.e., water absorption features) were removed and the remaining 145 spectral bands are used for experimental purposes. The data used in this article was acquired back on May 31, 2001, and it consist of observations from 14 mutually exclusive classes, which represent the land cover types in seasonal swamps, occasional swamps, and drier woodlands located in the distal portion of the Delta. The image cube and true ground truths label maps are shown in Fig. 3(a) and (b), whereas Fig. 3(c) and (d) show the classification performance in terms of classification maps (ground-truth label maps) for three different models, i.e., AfNet, 3-D attention inception net and 2-D attention inception net. These maps clearly show that the proposed method performs better than 3-D as well as 2-D attention inception networks. The higher accuracies are emphasized.

3) *Pavia University*: Pavia university dataset was acquired using ROSIS sensor during a flight campaign over Pavia, northern Italy. The total number of spectral bands is 103 in which each spectral band covers 610×610 spatial dimensions per spectral band. Some of the samples contain no information in the above spatial dimensions, thus have to be discarded before the analysis. The geometric resolution is 1.3 m. Pavia Center image ground truths differentiate nine mutually exclusive classes. The image cube and true ground truths label maps are shown in Fig. 4(a) and (b), whereas Fig. 4(c) and (d) show the classification performance in terms of classification maps (ground truth label maps) for three different models, i.e., AfNet, 3D attention inception net and 2-D attention inception net. These maps clearly show that the proposed method performs better than 3-D as well as 2-D attention inception networks. The higher accuracies are emphasized.

4) *Salinas*: The Salinas dataset was acquired using the AVIRIS sensor over Salinas Valley, California, and is characterized by high spatial resolution with 3.7-m per pixel with 224 spectral bands. The area covered by each spectral band is 512×217 samples. As with the Indian Pines dataset, 20 water absorption bands which are [108–112], [154–167], and 224 were discarded. Salinas dataset is only available as sensor radiance data. It includes vegetables, bare soils, and vineyard fields. Salinas ground truths contain 16 mutually exclusive classes. The image cube and true ground truths label maps are shown in Fig. 5(a) and (b), whereas Fig. 5(c) and (d) show

TABLE II
CLASSIFICATION PERFORMANCE AND TRAINING (Tr) AND TESTING (Te) TIME (IN SECONDS) ON DIFFERENT SPATIAL DIMENSIONS, I.E., $9 \times 9 \times B$, $11 \times 11 \times B$, $13 \times 13 \times B$, AND $15 \times 15 \times B$

Dataset	Measure	Spatial Dimensions			
		9×9	11×11	13×13	15×15
IP	κ (%)	91.79	94.34	95.04	95.97
	OA (%)	92.82	95.04	95.65	96.47
	AA (%)	83.26	79.53	86.62	86.13
	Tr Time	83.92	83.88	143.23	263.93
	Te Time	1.60	1.83	2.62	3.39
BS	κ (%)	98.57	98.76	99.18	98.90
	OA (%)	98.68	98.85	99.25	98.98
	AA (%)	97.83	98.01	98.85	98.774
	Tr Time	22.42	42.14	83.95	144.14
	Te Time	0.94	0.91	1.30	1.62
SA	κ (%)	99.54	99.72	99.94	99.85
	OA (%)	99.59	99.74	99.94	99.86
	AA (%)	99.76	99.85	99.96	99.84
	Tr Time	248.53	443.90	716.78	1043.87
	Te Time	10.77	10.90	21.23	21.40
PU	κ (%)	99.27	99.57	99.79	99.47
	OA (%)	99.45	99.67	99.84	99.60
	AA (%)	98.92	99.38	99.68	99.28
	Tr Time	203.92	323.88	552.19	821.60
	Te Time	4.74	10.71	10.81	11.74

The training, validation, and test sets are randomly divided into three parts, i.e., 15%/15%/70% (Train/Validation/Test).

the classification performance in terms of classification maps (ground truth label maps) for three different models, i.e., AfNet, 3-D attention inception net, and 2-D attention inception net. These maps clearly show that the proposed method performs better than 3-D as well as 2-D attention inception networks. The higher accuracies are emphasized.

B. Artefacts of Spatial Dimensions

To process the HSI cube in any CNN, the Spatial dimensions is being considered an important component and have an important impact on classification results [21]. This section experimentally illustrates the impact of spatial dimensions on classification results, i.e., OA, AA, and Kappa (κ) accuracy irrespective of the processing time and the computational cost which gradually increases as the spatial dimensions increase. The OA, AA, and κ accuracy is presented on all four experimental datasets to explore the impact of spatial dimensions on our proposed hybrid AfNet. All the experimental settings and tuning parameters for this particular experiment remain the same except spatial dimensions in which we tested the models on several different sizes, i.e., $9 \times 9 \times B$, $11 \times 11 \times B$, $13 \times 13 \times B$, and $15 \times 15 \times B$.

All these experimental results are presented in Fig. 6 and Table II in which one can observe that the classification accuracy improves as the spatial size improves. The reason behind this trend is that “the larger spatial dimensions contain more samples.” However, this trend does not remain the same for all the spatial dimensions, as it may contain redundant samples,

TABLE III
CLASSIFICATION PERFORMANCE (κ , OA, AND AA IN PERCENTAGE) ON
DIFFERENT PERCENTAGES OF TRAINING SAMPLES ALONG WITH THE
PROCESSING TIME (IN SECONDS) FOR BOTH TRAINING (TR) AND TESTING (TE)
PROCESS

Dataset	Measure	Percentage of Training Samples				
		5%	7%	10%	12%	15%
IP	κ (%)	77.41	84.32	84.06	90.15	91.79
	OA(%)	80.34	86.26	86.04	91.37	92.82
	AA(%)	65.78	74.08	80.33	88.48	83.26
	Tr Time	22.41	42.88	33.44	42.88	83.93
	Te Time	1.95	1.87	2.89	2.91	1.60
BS	κ (%)	95.25	96.97	98.16	98.06	98.57
	OA(%)	95.62	97.20	98.30	98.21	98.68
	AA(%)	96.10	96.62	98.32	97.15	97.83
	Tr Time	16.69	22.42	20.26	22.47	22.42
	Te Time	1.07	1.08	0.90	0.91	0.94
SA	κ (%)	98.39	97.72	99.19	97.79	99.54
	OA(%)	98.56	97.95	99.27	98.01	99.59
	AA(%)	99.16	98.07	99.56	98.25	99.76
	Tr Time	83.81	2243.88	143.90	208.87	248.53
	Te Time	7.33	82.28	10.77	10.82	10.77
PU	κ (%)	98.14	98.36	99.26	98.26	99.27
	OA(%)	98.59	98.76	99.44	98.68	99.45
	AA(%)	97.39	97.90	99.05	98.19	98.92
	Tr Time	1164.25	1583.49	2243.48	166.92	203.92
	Te Time	57.44	82.54	49.62	5.58	4.74

or by increasing the spatial dimension may contain interfering samples in a spatial patch or may contain the overlapping regions which bring nothing new to the classifier but just confuse the classifier and deteriorate the classification performance with redundant samples. Thus, in a nutshell, an appropriate size of spatial dimension with respect to the characteristics of the data is quite important to attain reliable accuracy.

C. Artefacts of Training Samples

CNN's have been extensively utilized for HSIC, however, deep CNN requires a large number of annotated training samples for appropriate learning. However, the collection of annotated samples for HSI is expensive and critical, which demands human experts or exploration of real-time scenarios. Limited availability of annotated training samples hinders HSIC performance. Thus, an appropriate size of annotated training samples is an important factor for HSIC performance. This section provides the performance evaluation in terms of OA, AA, and κ accuracy for different percentages of training samples. These percentages include 5/5/90% (Train/Validation/Test), 7/7/86%, 10/10/80%, 12/12/76%, and 15/15/70% respectively. We intentionally did not use below 5% training samples as several classes of different datasets do not have enough training samples to include, for instance, Oats class of IP dataset which only have 20 samples in total, thus selecting 1–4% of training samples for this class would only include one sample from this class, which is not enough to train the model. There is another option to avoid such limitation is to select the number of training samples rather using the percentage. However, this process will lead to another issue which is known as the ‘‘Class Imbalance’’ issue, which is not the problem under study. This may be considered as a potential future research direction.

Table III and Fig. 7 show the classification performance of AfNet with different percentages of annotated training samples. One can observe from these results, as the number of annotated training samples increases the classification performance significantly improves. However, the trend is for some certain

stage not for the entire group of percentages. This is due to the redundancy among the training samples, as the higher number of annotated samples may contain samples spectrally similar to each other, brings no new information, or may lead to confusion for learning. The performance evaluation indicates the quality of spatial–spectral features learned by our proposed model, i.e., the features learned by AfNet. From these results, one can also conclude that the 7–10% training samples are enough to get satisfactory results. For all these experimental results, $9 \times 9 \times B$ spatial dimension are used, rest of the experimental protocols remains the same except the number of training samples, which are further explained in Table III. From the computational time, similar observations can be made, as the number of training samples increases, the training, and testing time significantly increase as well the accuracy increases.

IV. CONCLUSION

CNNs are known to overcome the nonlinearity issues with fixed kernel sizes which are not flexible enough because these kernels are specific and not conducive to feature learning thus impairs classification accuracy. However, CNN with different kernel sizes may capture more discriminative and important features. Thus, taking into account the aforesaid advantages, this work proposed a hybrid (3-D–2-D) inception net with an attention mechanism to boost the classification performance. The proposed AfNet used attention-based six parallel hybrid subnets with different kernels in each subblock to enhance the final ground-truth maps. The proposed AfNet selectively filters out the discriminative feature, i.e., the critical features for classification. AfNet has been tested on several hyperspectral datasets and shows competitive results as compared to the state-of-the-art models except for a few expensive choices. The possible future research directions include incorporating the attention mechanism and hybrid process into the exception net which can produce better accuracy and generalization performance of CNNs. Other possible research directions could include the utilization of different convolutional processes in hybrid scenarios.

REFERENCES

- [1] D. Hong *et al.*, ‘‘Interpretable hyperspectral artificial intelligence: When nonconvex modeling meets hyperspectral remote sensing,’’ *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 2, pp. 52–87, Jun. 2021.
- [2] M. H. Khan, Z. Saleem, M. Ahmad, A. Sohaib, H. Ayaz, and M. Mazzara, ‘‘Hyperspectral imaging for color adulteration detection in red chili,’’ *Appl. Sci.*, vol. 10, no. 17, 2020.
- [3] Z. Saleem, M. H. Khan, M. Ahmad, A. Sohaib, H. Ayaz, and M. Mazzara, ‘‘Prediction of microbial spoilage and shelf-life of bakery products through hyperspectral imaging,’’ *IEEE Access*, vol. 8, pp. 176 986–176 996, 2020.
- [4] D. Hong, J. Yao, D. Meng, Z. Xu, and J. Chanussot, ‘‘Multimodal GANs: Toward crossmodal hyperspectral–multispectral image segmentation,’’ *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5103–5113, Jun. 2020.
- [5] S. Liu, Q. Du, X. Tong, A. Samat, and L. Bruzzone, ‘‘Unsupervised change detection in multispectral remote sensing images via spectral–spatial band expansion,’’ *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 9, pp. 3578–3587, Sep. 2019.
- [6] H. Ayaz *et al.*, ‘‘Myoglobin-based classification of minced meat using hyperspectral imaging,’’ *Appl. Sci.*, vol. 10, no. 19, 2020, Art. no. 6862.
- [7] X. Liu, D. Hong, J. Chanussot, B. Zhao, and P. Ghamisi, ‘‘Modality translation in remote sensing time series,’’ *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5401614.

- [8] H. Ayaz, M. Ahmad, M. Mazzara, and A. Sohaib, "Hyperspectral imaging for minced meat classification using nonlinear deep features," *Appl. Sci.*, vol. 10, no. 21, 2020, Art. no. 7783.
- [9] M. Zulfiqar, M. Ahmad, A. Sohaib, M. Mazzara, and S. Distefano, "Hyperspectral imaging for bloodstain identification," *Sensors*, vol. 21, no. 9, 2021, Art. no. 3045.
- [10] H. Hussain Khan *et al.*, "Hyperspectral imaging-based unsupervised adulterated red chili content transformation for classification: Identification of red chili adulterants," *Neural Comput. Appl.*, vol. 33, pp. 14507–14521, 2021.
- [11] B. Zhang *et al.*, "Progress and challenges in intelligent remote sensing satellite systems," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1814–1822, 2022.
- [12] B. Rasti *et al.*, "Feature extraction for hyperspectral imagery: The evolution from shallow to deep: Overview and toolbox," *IEEE Geosci. Remote Sens. Mag.*, vol. 8, no. 4, pp. 60–88, Apr. 2020.
- [13] T. V. Bandos, L. Bruzzone, and G. Camps-Valls, "Classification of hyperspectral images with regularized linear discriminant analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 3, pp. 862–873, Mar. 2009.
- [14] D. Hong *et al.*, "Endmember-guided unmixing network (EGU-net): A general deep learning framework for self-supervised hyperspectral unmixing," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: [10.1109/TNNLS.2021.3082289](https://doi.org/10.1109/TNNLS.2021.3082289).
- [15] S. Liu, H. Zhao, Q. Du, L. Bruzzone, A. Samat, and X. Tong, "Novel cross-resolution feature-level fusion for joint classification of multispectral and panchromatic remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5619314.
- [16] M. Ahmad, S. Protasov, and A. M. Khan, "Hyperspectral band selection using unsupervised non-linear deep auto encoder to train external classifiers," 2017, *arXiv:1705.06920*.
- [17] D. Hong, N. Yokoya, J. Chanussot, J. Xu, and X. X. Zhu, "Learning to propagate labels on graphs: An iterative multitask regression framework for semi-supervised hyperspectral dimensionality reduction," *ISPRS J. Photogrammetry Remote Sens.*, vol. 158, pp. 35–49, 2019.
- [18] M. Ahmad, S. Shabbir, R. A. Raza, M. Mazzara, S. Distefano, and A. M. Khan, "Hyperspectral image classification: Artifacts of dimension reduction on hybrid CNN," 2021, *arXiv:2101.10532*.
- [19] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.
- [20] S. Shabbir and M. Ahmad, "Hyperspectral image classification-traditional to deep models: A survey for future prospects," 2021, *arXiv:2101.06116*.
- [21] M. Ahmad, A. M. Khan, M. Mazzara, S. Distefano, M. Ali, and M. S. Sarfraz, "A fast and compact 3-D CNN for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2020, Art. no. 5502205.
- [22] D. Hong, N. Yokoya, N. Ge, J. Chanussot, and X. X. Zhu, "Learnable manifold alignment (LeMA): A semi-supervised cross-modality learning framework for land cover and land use classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 147, pp. 193–205, 2019.
- [23] J. Yao, D. Hong, L. Xu, D. Meng, J. Chanussot, and Z. Xu, "Sparsity-enhanced convolutional decomposition: A novel tensor-based paradigm for blind hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5505014.
- [24] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
- [25] D. Hong, N. Yokoya, G.-S. Xia, J. Chanussot, and X. X. Zhu, "X-modalnet: A semi-supervised deep cross-modal network for classification of remote sensing data," *ISPRS J. Photogrammetry Remote Sens.*, vol. 167, pp. 12–23, 2020.
- [26] B. Zhang, S. Li, X. Jia, L. Gao, and M. Peng, "Adaptive Markov random field approach for classification of hyperspectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 5, pp. 973–977, May 2011.
- [27] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2321–2325, Nov. 2015.
- [28] D. Hong, J. Hu, J. Yao, J. Chanussot, and X. X. Zhu, "Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model," *ISPRS J. Photogrammetry Remote Sens.*, vol. 178, pp. 68–80, 2021.
- [29] B. Zhang, W. Yang, L. Gao, and D. Chen, "Real-time target detection in hyperspectral images based on spatial-spectral information extraction," *EURASIP J. Adv. Signal Process.*, vol. 2012, 2012, Art. no. 142.
- [30] D. Nyasaka, J. Wang, and H. Tinaga, "Learning hyperspectral feature extraction and classification with resnext network," 2002, *arXiv:2002.02585*.
- [31] M. Ahmad, A. Khan, and R. Hussain, "Graph-based spatial-spectral feature learning for hyperspectral image classification," *IET Image Process.*, vol. 11, no. 12, pp. 1310–1316, 2017.
- [32] L. Gómez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, "Multimodal classification of remote sensing images: A review and future directions," *Proc. IEEE*, vol. 103, no. 9, pp. 1560–1584, Sep. 2015.
- [33] S. K. Roy, P. Kar, D. Hong, X. Wu, A. Plaza, and J. Chanussot, "Revisiting deep hyperspectral feature extraction networks via gradient centralized convolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5516619.
- [34] H. Gao, D. Yao, Y. Yang, C. Li, H. Liu, and Z. Hua, "Multiscale 3-D-CNN based on spatial-spectral joint feature extraction for hyperspectral remote sensing images classification," *J. Electron. Imag.*, vol. 29, no. 1, 2020, Art. no. 013007.
- [35] Z. Zhong, J. Li, Z. Luo, and M. Chapman, "Spectral-spatial residual network for hyperspectral image classification: A 3-D deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 847–858, Feb. 2018.
- [36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [37] S. K. Roy, R. Mondal, M. E. Paoletti, J. M. Haut, and A. Plaza, "Morphological convolutional neural networks for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 8689–8702, 2021.
- [38] L. He, J. Li, C. Liu, and S. Li, "Recent advances on spectral-spatial hyperspectral image classification: An overview and new guidelines," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, Mar. 2017, Art. no. 1579–1597.
- [39] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "HybridSN: Exploring 3-D-2-D CNN feature hierarchy for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 277–281, Feb. 2020.
- [40] M. Ahmad, M. Mazzara, and S. Distefano, "3D/2D regularized CNN feature hierarchy for hyperspectral image classification," 2021, *arXiv:2104.12136*.
- [41] B.-C. Kuo, C.-H. Li, and J.-M. Yang, "Kernel nonparametric weighted feature extraction for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 4, pp. 1139–1155, Apr. 2009.
- [42] D. Hong, L. Gao, J. Yao, B. Zhang, P. Antonio, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5966–5978, Jul. 2021.
- [43] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
- [44] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
- [45] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.
- [46] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. Int. Conf. Comput. Vision Workshops*, 2019, pp. 1971–1980.
- [47] D. Hong *et al.*, "Spectralformer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5518615.
- [48] S. K. Roy, A. Deria, D. Hong, B. Rasti, A. Plaza, and J. Chanussot, "Multimodal fusion transformer for remote sensing image classification," 2022, *arXiv:2203.16952*.
- [49] B. Fang, Y. Li, H. Zhang, and J. C.-W. Chan, "Hyperspectral images classification based on dense convolutional networks with spectral-wise attention mechanism," *Remote Sens.*, vol. 11, no. 2, 2019, Art. no. 159.
- [50] W. Ma, Q. Yang, Y. Wu, W. Zhao, and X. Zhang, "Double-branch multi-attention mechanism network for hyperspectral image classification," *Remote Sens.*, vol. 11, no. 11, 2019, Art. no. 1307.
- [51] E. Pan *et al.*, "Spectral-spatial classification of hyperspectral image based on a joint attention network," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 413–416.
- [52] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11531–11539.
- [53] Z. Zheng and Y. Zhong, "S3NET: Towards real-time hyperspectral imagery classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 3293–3296.

- [54] Y. Li, H. Zhang, and Q. Shen, "Spectral-spatial classification of hyperspectral imagery with 3D convolutional neural network," *Remote Sens.*, vol. 9, 2017, Art. no 67.
- [55] D. Hong *et al.*, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021.
- [56] M. F. Baumgardner, L. L. Biehl, and D. A. Landgrebe, "220 band AVIRIS hyperspectral image data set: Jun. 12, 1992 indian pine test site 3," Sep. 2015. [Online]. Available: <http://pur.purdue.edu/publications/1947/1>
- [57] "Hyperspectral datasets description," 2021. Accessed: Jan. 1, 2021. [Online]. Available: http://www.ehu.es/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes
- [58] D. Hong, N. Yokoya, J. Chanussot, J. Xu, and X. X. Zhu, "Joint and progressive subspace analysis (JPSA) with spatial-spectral manifold alignment for semisupervised hyperspectral dimensionality reduction," *IEEE Trans. Cybern.*, vol. 51, no. 7, pp. 3602–3615, Jul. 2021.



Muhammad Ahmad received the M.S. degree in electronics engineering from International Islamic University, Islamabad, Pakistan, in 2011, the Ph.D. degree in computer science and engineering from Innopolis University, Innopolis, Russia, in 2019, and the other Ph.D. degree in cyber-physical systems from the University of Messina, Messina, Italy, in 2021.

He is currently with the National University of Computer and Emerging Sciences (FAST-NUCES), Islamabad. He has also served as an Assistant Professor, Lecturer, Instructor, Research Fellow, Research Associate, and Research Assistant for a number of international/national universities. He was also a Radio Access Network (RAN) Supervisor worked with Ericsson (Mobilink Project). He is supervising/co-supervising several graduates (M.S. and Ph.D.). He has authored and coauthored more than 70 scientific contributions to international journals, conferences, and books. His research interests include hyperspectral imaging, remote sensing, machine learning, computer vision, and wearable computing.

Dr. Ahmad has/is served/serving as a Academic/Associate and Lead/Guest Editor on several special issues for SCI/E, JCR journals. He has delivered a number of invited and keynote talks and reviewed (reviewing) the technology-related articles for journals.

Dr. Ahmad has/is served/serving as a Academic/Associate and Lead/Guest Editor on several special issues for SCI/E, JCR journals. He has delivered a number of invited and keynote talks and reviewed (reviewing) the technology-related articles for journals.



Adil Mehmood Khan (Member, IEEE) received the B.S. degree in information technology from the National University of Sciences and Technology, Islamabad, Pakistan, in 2005, and the M.Sc. and Ph.D. degrees in computer engineering from Kyung Hee University, Seoul, South Korea, in 2011.

He is currently a Professor with the Institute of Artificial Intelligence and Data Science, Innopolis University, Innopolis, Russia. His research interests include machine learning and deep learning.



Manuel Mazzara received the Ph.D. degree in computing science from the University of Bologna, Bologna, Italy, in 2006.

He is a Professor of computer science with Innopolis University, Innopolis, Russia, with a research background in software engineering, service-oriented architectures and programming, concurrency theory, formal methods, software verification, and artificial intelligence. He cooperated with European and US industry, plus governmental and intergovernmental organizations, such as the United Nations, always at

the edge between science and software production. His research interests include the development of theories, methods, tools and programs covering the two major aspects of software engineering, and artificial intelligence: the process side, describing how we develop software, and the product side, describing the results of this process.



Salvatore Distefano received the Ph.D. degree in computer engineering from the University of Messina, Italy, in 2006. He is currently an Associate Professor with the University of Messina, University of Messina, Messina, Italy. He is the coordinator of the Italian CINI Working Group on System and Service Quality. He is also one of the cofounders of the SmartMe.io start-up, a University of Messina spin-off, established in 2017. He was a Visiting Scholar and Professor with different universities and research centers, such as University of Massachusetts Dartmouth, Dartmouth, MA, USA, UCLA, Los Angeles, CA, USA, Duke University, Durham, NC, USA, Innopolis University, Innopolis, Russia, and Kazan Federal University, Kazan, Russia, collaborating with top scientists. He took part in several national and international projects, such as Reservoir, Vision (EU FP7), SSMCOM (EU FP7 ERC Advanced Grant), Beacon, IoT-Open.EU (EU H2020). He authored and coauthored more than 250 scientific papers and contributions to international journals, conferences, and books. During his research activity, he contributed to the development of several tools, such as WebSPN, ArgoPerformance, GS3, and Stack4Things. His research interests include non-Markovian modeling; quality of service/experience; parallel and distributed computing, grid, cloud, autonomic, volunteer, crowd, edge, fog computing; internet of things; cyber-physical social systems; smart cities; intelligent transportation systems; big data, stream processing; software-defined and virtualized ecosystems; hyper spectral imaging; and machine learning.

Dr. Distefano is a member of international conference committees and he is on the editorial boards for IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, *Journal of Cloud Computing*, and *International Journal of Big Data*.



Swalpa Kumar Roy (Student Member, IEEE) received the bachelor's degree in computer science and engineering from the West Bengal University of Technology, Kolkata, West Bengal, India, in 2012, the master's degree in computer science from the Indian Institute of Engineering Science and Technology, Shibpur, Howrah, West Bengal, in 2015, and the Ph.D. degree in computer science and engineering from the University of Calcutta, Kolkata, in 2021.

From July 2015 to March 2016, he was a Project Linked Person with the Optical Character Recognition Laboratory, Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata. He is currently an Assistant Professor with the Department of Computer Science and Engineering, Jalpaiguri Government Engineering College, Jalpaiguri, West Bengal. His research interests include computer vision, deep learning, and remote sensing.

Dr. Roy was nominated for the Indian National Academy of Engineering engineering teachers mentoring fellowship program by INAE Fellows, in 2021. He was the recipient of the Outstanding Paper Award in second Hyperspectral Sensing Meets Machine Learning and Pattern Analysis (HyperMLPA) at the Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS), in 2021. He has served as a Reviewer for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and *IEEE Geoscience and Remote Sensing Letters*.



Xin Wu (Member, IEEE) received the M.Sc. degree in computer science and technology from the College of Information Engineering, Qingdao University, Qingdao, China, in 2014, and the Ph.D. degree in information and communication engineering from the School of Information and Electronics, Beijing Institute of Technology (BIT), Beijing, China, in 2020.

In 2018, she was a Visiting Student with the Photogrammetry and Image Analysis Department of the Remote Sensing Technology Institute, German Aerospace Center, Oberpfaffenhofen, Germany. She is currently a Postdoctoral Researcher with the School of Information and Electronics, BIT. Her research interests include signal/image processing, fractional Fourier transform, and deep learning and their applications in biometrics and geospatial object detection.

Dr. Wu was the recipient of the Best Reviewer Award of the IEEE JSTARS, in 2022, and the Jose Bioucas Dias award for recognizing the outstanding paper at the Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS), in 2021.