IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 29, 2020, accepted May 10, 2020, date of publication May 19, 2020, date of current version June 3, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2995805

A Deep Multi-Attention Driven Approach for Multi-Label Remote Sensing Image Classification

GENCER SUMBUL, (Graduate Student Member, IEEE)
AND BEGÜM DEMİR, (Senior Member, IEEE)
Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin, D-10587 Berlin, Germany

Corresponding author: Begüm Demir (demir@tu-berlin.de)

This work was supported by the European Research Council (ERC) through the ERC-2017-STG BigEarth Project under Grant 759764.

ABSTRACT Deep learning (DL) based methods have been found popular in the framework of remote sensing (RS) image scene classification. Most of the existing DL based methods assume that training images are annotated by single-labels, however RS images typically contain multiple classes and thus can simultaneously be associated with multi-labels. Despite the success of existing methods in describing the information content of very high resolution aerial images with RGB bands, any direct adaptation for high-dimensional high-spatial resolution RS images falls short of accurate modeling the spectral and spatial information content. To address this problem, this paper presents a novel approach in the framework of the multi-label classification of high dimensional RS images. The proposed approach is based on three main steps. The first step describes the complex spatial and spectral content of image local areas by a novel K-Branch CNN that includes spatial resolution specific CNN branches. The second step initially characterizes the importance scores of different local areas of each image and then defines a global descriptor for each image based on these scores. This is achieved by a novel multi-attention strategy that utilizes the bidirectional long short-term memory networks. The final step achieves the classification of RS image scenes with multi-labels. Experiments carried out on BigEarthNet (which is a large-scale Sentinel-2 benchmark archive) show the effectiveness of the proposed approach in terms of multi-label classification accuracy compared to the state-of-the-art approaches. The code of the proposed approach is publicly available at https://gitlab.tubit.tu-berlin.de/rsim/MAML-RSIC.

INDEX TERMS Multi-label image classification, deep neural network, multi-attention strategy, remote sensing.

I. INTRODUCTION

Advances in satellite missions for Earth observation have led to a significant growth of remote sensing (RS) image archives. Accordingly, the development of RS image scene classification methods, which aim at automatically assigning class labels to each RS image scene in an archive, is a growing research interest in RS. In recent years, deep learning (DL) based approaches have attracted the attention of RS researchers. As an example, in [1] a gradient boosting random convolutional network is proposed as an ensemble framework to combine several deep neural networks for RS image scene classification problems. In [2] feature learning strategies defined based on different training procedures for convolutional neural networks (CNNs) are analyzed. In [3] a region attention network, which assigns attention scores to candidate regions for the expected object locations, is introduced to learn the alignment of RS image scenes. to this end, different image sources are used together for the identification of fine-grained categories. In [4] a semi-supervised approach based on a generative adversarial network is proposed for the cases that the amount of annotated training data is insufficient. In [5] an intermediate feature aggregation method that progressively combines the different level features of CNNs is proposed. In [6] a scale-free CNN that transfers the fully connected layers in a pre-trained CNN model to convolutional layers and then uses a general average pooling layer after the final convolutional layer is introduced. The above-mentioned DL based approaches in RS assume that each training image is annotated by a single (broad category) label, which is associated to the most significant content of the image. However, this assumption may not

The associate editor coordinating the review of this manuscript and approving it for publication was Qiangqiang Yuan.

be appropriate for complex scene classification applications where RS image scenes contain multiple land-cover classes and thus simultaneously associated to different class labels (i.e., multi-labels) [7].

To train DL models with training images annotated by multi-labels, few DL based multi-label scene classification methods have been recently introduced in RS. In [8] a radial basis function neural network is applied on the CNN features of aerial images as a multi-label classifier. In [9] a structured support vector machine that models the spatial contiguity is utilized based on the CNN features of the aerial images in the framework of multi-label classification. In these approaches, CNNs are used as conventional transfer learning approaches, for which pre-trained models on publicly available general purpose computer vision (CV) datasets (e.g., ImageNet) act as fixed feature extractors without changing the model parameters. However, this approach can reduce the multi-label scene classification accuracy because of the differences in image characteristics in CV and RS. In [10] a data augmentation strategy is introduced to avoid using a pre-trained network for an end-to-end training of a shallow CNN. In this approach, to adapt the standard CNN architecture in multi-label learning, the softmax function of the classification layer is changed into a sigmoid function. The direct use of standard CNNs that are actually designed for the images annotated by single-labels is a common approach in multi-label classification problems. However, it may lead to inaccurate identification of the multiple classes present in images. To overcome this limitation, integration of sequential neural network approaches into CNN architectures is introduced in RS. In [11] a class-wise attention-based recurrent neural network (RNN) is introduced to sequentially model the co-occurrence relationship of multiple classes. In this approach, class predictions are obtained one after another in the RNN sequence and each prediction is based on the decisions made until the corresponding class is reached. In [12] an attention-aware label relational reasoning network is proposed to: i) localize discriminative regions of aerial images; and ii) characterize the label relations present in the images based on the localized feature maps. In [13] an encoder-decoder neural network is introduced to characterize the aerial image features. In detail, a squeeze excitation layer is used for modeling the channel-wise interdependencies of the feature maps in the encoder, whereas a RNN based decoder is exploited as an adaptive spatial attention mechanism. The attention strategies proposed in [11]–[13] identify informative areas of images through an attention map based on the feature maps of convolutional layers. These strategies are effective for very high resolution aerial images, however they can be insufficient for accurately describing the complex content of satellite RS images with high spatial resolution (e.g., Sentinel-2 and Landsat multispectral images). Results carried out on very high resolution aerial images with only RGB bands show the success of these strategies for the description of the spatial image content. A direct adaptation of these methods for high dimensional RS images may lead to an incomplete representation of the spectral information content. These issues are critical particularly for images with several spectral bands with varying spatial resolutions acquired by the new generation satellites (e.g., Sentinel-2). Thus, methods that can efficiently and effectively describe the spatial and spectral information content of high dimensional RS images are needed in the framework of multi-label RS image scene classification.

To address this problem, we propose a DL based approach that aims at accurately describing complex spatial and spectral content of RS images in the framework of multi-label RS image scene classification. To this end, the proposed approach is based on three main steps: 1) spatial and spectral characterization of image local areas; 2) definition of a multi-attention driven global descriptor; and 3) classification of RS image scenes with multi-labels. The proposed approach assumes that RS image bands can be associated with varying spatial resolutions and a set of training images annotated with multi-labels (based on land-cover land-use classes present in the images) is available. In the first step, we introduce a novel branch-wise CNN architecture (which is called as K-Branch CNN) that efficiently describes the complex content of local areas of each image by different CNN branches specialized according to the spatial resolutions of image bands. In the second step, we present a novel multi-attention strategy in the framework of RNNs that: i) accurately identifies importance levels (i.e., scores) for different local areas; and then ii) defines a global descriptor for each image based on these scores. In the third step, multi-labels are automatically assigned to each RS image represented by the global descriptors. The main novelty of the proposed approach consists in the design and development of: i) the K-Branch CNN to efficiently model the complex information content of RS images for which the spectral bands can be associated to varying spatial resolutions; and ii) the multi-attention strategy that defines a global image descriptor based on the extraction and exploitation of importance scores of image local areas. The proposed approach has been briefly presented in [14] with limited experimental analysis. This paper extends our work introducing a detailed description of the proposed approach with a detailed ablation and comparison study. In order to evaluate the performance of the proposed approach, several experiments are carried out on the BigEarthNet since it is the only publicly available benchmark archive that includes Sentinel-2 multispectral images, each of which is annotated with multi-labels. Unlike the conventional DL based methods in RS that consider all the image bands as a single volume (after applying an interpolation method to the lower spatial resolution bands) and define a global descriptor by neglecting the importance scores of different local areas, the experimental results show the success of the proposed approach. The rest of the paper is organized as follows. Section II introduces the proposed approach for multi-label RS image scene classification, while Section III explains the BigEarthNet benchmark archive and design of experiments. Section IV provides the experimental results. Section V draws the conclusion of this work.

II. PROPOSED APPROACH

Let $\mathcal{X}=\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ be an archive that consists of M images, where $\boldsymbol{x}_i$ is the i^{th} image. We assume that a set $\mathcal{T} \subset \mathcal{X}$ of labeled images is initially available. Each image in $\mathcal{T} \subset \mathcal{X}$ is associated with multi-labels from a label set $\mathcal{L} = \{l_1, \ldots, l_S\}$, where $|\mathcal{L}| = S$. Label information of $\boldsymbol{x}_i \in \mathcal{T}$ is defined by a binary vector $\boldsymbol{y}_i \in \{0, 1\}^S$, where each element of $\boldsymbol{y}_i$ indicates the presence or absence of label $l_s \in \mathcal{L}$ in a sequence. We also assume that spectral bands of each image $\boldsymbol{x}_i$ can be associated to the K different spatial resolutions, resulting in different pixel sizes. We aim to learn $F(\boldsymbol{x}^*; \theta) = g(f(\boldsymbol{x}^*; \theta))$ that maps a new image $\boldsymbol{x}^*$ to multi-labels, where $f(\cdot)$ generates classification scores for each label l_s and $g(\cdot)$ produces $\boldsymbol{y}^*$ as a predicted label set and θ is the given set of model parameters. We propose a multi-label RS image scene classification approach made up of three main steps: 1) spatial and spectral characterization of image local areas by a novel K-Branch CNN; 2) definition of a multi-attention driven global descriptor with a novel multi-attention strategy; and 3) classification of RS image scenes with multi-labels. Fig. 1 presents the block diagram of the proposed approach and each step is explained in the following sub-sections.

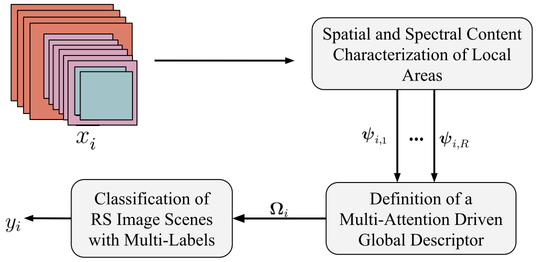

x_i
Spatial and Spectral Content Characterization of Local Areas
$\psi_{i,1}$
...
$\psi_{i,R}$
Definition of a Multi-Attention Driven Global Descriptor
Ω_i
Classification of RS Image Scenes with Multi-Labels
y_i

FIGURE 1. **Block diagram of the proposed approach for multi-label RS image scene classification.**

A. SPATIAL AND SPECTRAL CHARACTERIZATION OF LOCAL AREAS

To efficiently characterize the spatial and spectral content of image local areas, each RS image is initially divided to R non-overlapping $w \times w$ sized local areas. Let $\boldsymbol{\rho}_i^r$ be the r^{th} local area of $\boldsymbol{x}_i$. Then, for each local area, we define different sets of image bands based on their spatial resolutions. Let $\boldsymbol{\rho}_{i,k}^r$ be the k^{th} subset of the r^{th} local area for the corresponding spatial resolution, where $k \in \{1, 2, \ldots, K\}$ and $r \in \{1, 2, \ldots, R\}$. To accurately describe the local areas with varying spatial resolutions, we introduce a K-Branch CNN that utilizes separate CNNs, each of which is designed to describe the local areas of image bands with different spatial resolutions. Thus, the number K of CNN branches is selected as the total number of different spatial resolutions. If all spectral bands are associated to the same spatial resolution, the proposed K-Branch CNN turns into a single branch CNN (i.e., $K = 1$). Each $\boldsymbol{\rho}_{i,k}^r$ are fed into different branches of the K-Branch CNN. Let ϕ^k be the k^{th} branch that provides local descriptors associated with k^{th} spatial resolution by applying convolutional layers and a fully connected (FC) layer. Different local descriptors for all sets of image bands are first characterized and then concatenated into one vector for one local area. To effectively combine information from different branches, all concatenated feature vectors are fed into a new FC layer to produce the local descriptors $\boldsymbol{\psi}_{i,r}$. This step is illustrated in Fig. 2.

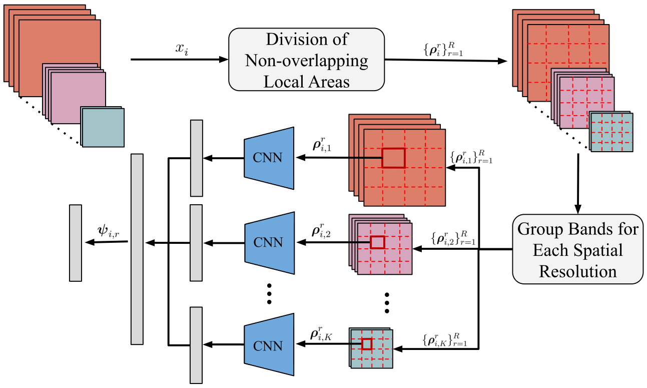

x_i
Division of Non-overlapping Local Areas
$\{\rho_i^r\}_{r=1}^R$
Group Bands for Each Spatial Resolution
$\{\rho_{i,1}^r\}_{r=1}^R$
$\{\rho_{i,2}^r\}_{r=1}^R$
$\{\rho_{i,K}^r\}_{r=1}^R$
$\rho_{i,1}^r$
$\rho_{i,2}^r$
$\rho_{i,K}^r$
CNN
CNN
CNN
$\psi_{i,r}$

FIGURE 2. **The proposed *K*-Branch CNN introduced in the first step of the proposed approach. One local area is highlighted as an example to feed into the corresponding CNN.**

The proposed K-Branch CNN describes the complex information content of image local areas through specific branches associated to different spatial resolutions. By this way, a unique CNN is used for the image bands with the same spatial resolution unlike the traditional CNN based methods in RS (which consider all the image bands as a single volume after applying interpolation to the low spatial resolution bands). On the one side, this approach leads to an accurate characterization of the content of high dimensional RS images. On the other side, due to modeling the local areas, it requires a smaller number of model parameters being estimated. Thus, the computational complexity of training phase is reduced, while the risk of over-fitting on training data with low generalizing capability is avoided (since smaller neural networks have less tendency for over-fitting).

B. DEFINITION OF A MULTI-ATTENTION DRIVEN GLOBAL DESCRIPTOR

After obtaining the local descriptors $\{\boldsymbol{\psi}_{i,r}\}_{r=1}^R$ in the first step, a global descriptor can be defined by simply stacking all local descriptors. In this way, local descriptors equally contribute to the definition of a global descriptor. However, local areas of an RS image can be subject to different levels (i.e., scores) of importance to represent the semantic content of the image. Accordingly, this step aims at accurately extracting and exploiting importance levels of local areas of each image, while defining a global image descriptor. To this end, we introduce a novel multi-attention strategy that is defined based on long short-term memory (LSTM) networks [15].

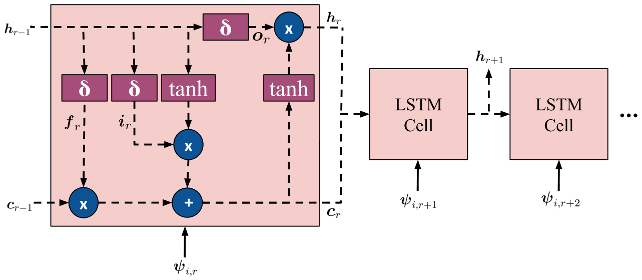


FIGURE 3. Single LSTM cell with its inputs, gates and cell state followed by two LSTM cells in a sequence. Without losing in generality, particular sequence of the LSTM network (which starts with the first local area and ends with the last local area) is chosen in the figure.

An LSTM network contains sequentially ordered LSTM nodes (i.e., cells). Each cell includes input gate ($\boldsymbol{i}$), forget gate ($\boldsymbol{f}$), output gate ($\boldsymbol{o}$) and cell state ($\boldsymbol{c}$). Cell state characterizes the knowledge of observed inputs until the corresponding cell. Different gates control how the cell state should behave according to different aims. Forget gate decides which portion of the current cell state value should be forgotten. Input gate controls which portion of the input should be read by cell state. Output gate decides which portion of the cell state should be produced as the output of the new cell state. The reader is referred to [16] for the detailed explanation. In the proposed approach, each LSTM cell takes the descriptor of r^{th} local area ($\boldsymbol{\psi}_{i,r}$) from the K-Branch CNN as input and employs the aforementioned operations as follows:

$$\begin{aligned}
\boldsymbol{f}_r &= \delta(\mathbf{W}_{f,r}\boldsymbol{\psi}_{i,r} + \mathbf{U}_{f,r}\boldsymbol{h}_\tau + \boldsymbol{b}_{f,r})\\
\boldsymbol{i}_r &= \delta(\mathbf{W}_{i,r}\boldsymbol{\psi}_{i,r} + \mathbf{U}_{i,r}\boldsymbol{h}_\tau + \boldsymbol{b}_{i,r})\\
\boldsymbol{o}_r &= \delta(\mathbf{W}_{o,r}\boldsymbol{\psi}_{i,r} + \mathbf{U}_{o,r}\boldsymbol{h}_\tau + \boldsymbol{b}_{o,r})\\
\boldsymbol{c}_r &= \boldsymbol{f}_r \odot \boldsymbol{c}_\tau + \boldsymbol{i}_r \odot \tanh(\mathbf{W}_{c,r}\boldsymbol{\psi}_{i,r} + \mathbf{U}_{c,r}\boldsymbol{h}_\tau + \boldsymbol{b}_{c,r}) \qquad (1)
\end{aligned}$$

where tanh and δ are the hyperbolic tangent and sigmoid functions, $\mathbf{W}_{\cdot,r}$ and $\boldsymbol{b}_{\cdot,r}$ are the weight and bias parameters; and the subscript of r refers to the parameters of the LSTM cell associated with r^{th} local area. All operations of one LSTM cell are illustrated in Fig. 3. Each LSTM cell produces one preliminary attention score given the sequence, $\boldsymbol{h}_{r|\tau}$, based on the cell state and the gates as follows:

$$\boldsymbol{h}_r = \boldsymbol{h}_{r|\tau} = \boldsymbol{o}_r \odot \tanh(\boldsymbol{c}_r). \qquad (2)$$

We utilize two LSTM networks in a bidirectional manner to consider the different orders of local areas and thus all LSTM cells are placed in two different sequences with different parameters. Each cell of the first LSTM network produces the preliminary attention score of one local area concerning the knowledge acquired from the attention scores of previous local areas (i.e., previous cells). Thus τ becomes $r-1$ in (1). The second LSTM network employs the same idea by considering the subsequent local areas and thus τ becomes $r+1$ in (1). In the context of bidirectional LSTM networks, forward and backward sequences can be combined by using the concatenation, the summation or the multiplication operations [17], [18]. The concatenation operation is a widely used operation in the literature. However, it requires a fully connected layer for the reduction of a vector into a single value, which can significantly increase the computational complexity of the whole approach. When multiplication operation is used, the resulting value can be dominated by one of the sequences, if the preliminary attention score is a negative value. Accordingly, we select the summation operation for combining the sequences. To this end, after obtaining two preliminary attention scores from the different orders, we apply the final attention score of the r^{th} local area $\alpha_{i,r}$ as follows:

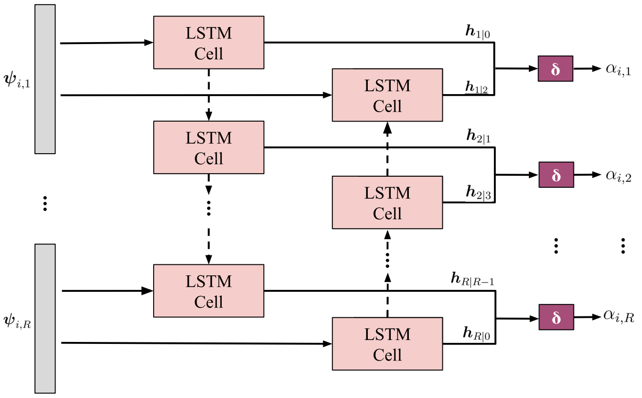


FIGURE 4. Proposed multi-attention strategy with bidirectional LSTM networks for the second step of the proposed approach.

$$\alpha_{i,r} = \delta\left(\frac{\boldsymbol{h}_{r|r-1} + \boldsymbol{h}_{r|r+1}}{2}\right). \qquad (3)$$

This produces an attention score for the r^{th} local area within the range of $[0, 1]$. For the beginning of passes ($r = 1$ or $r = R$), τ refers to an initial state of the nodes. Each attention score shows the importance level of the considered local area for the complete characterization of the whole image content. Accordingly, multi-attention scores $\{\alpha_{i,r}\}_{r=1}^{R}$ for the i^{th} image $\boldsymbol{x}_i$ show the different importance levels of the image local areas. The proposed multi-attention strategy is illustrated in Fig. 4.

Let Ω_i be the multi-attention driven global descriptor of the $\boldsymbol{x}_i$. After obtaining the multi-attention scores, the global descriptor Ω_i is defined by the concatenation of local descriptors weighted by attention scores as follows:

$$\Omega_i = [\alpha_{i,1}\boldsymbol{\psi}_{i,1}^\top, \ldots, \alpha_{i,R}\boldsymbol{\psi}_{i,R}^\top]^\top. \qquad (4)$$

Due to this step, the proposed approach extracts and exploits the importance scores of local areas of each image instead of equally considering them.

C. CLASSIFICATION OF RS IMAGE SCENES WITH MULTI-LABELS

This step aims to classify RS images into multi-labels by using the multi-attention driven global descriptor Ω_i obtained in the second step of the proposed approach. To this end, we employ a FC layer $f(\cdot)$ as a classifier that generates class scores z_{l_j} for each class label l_j in the sequence based on

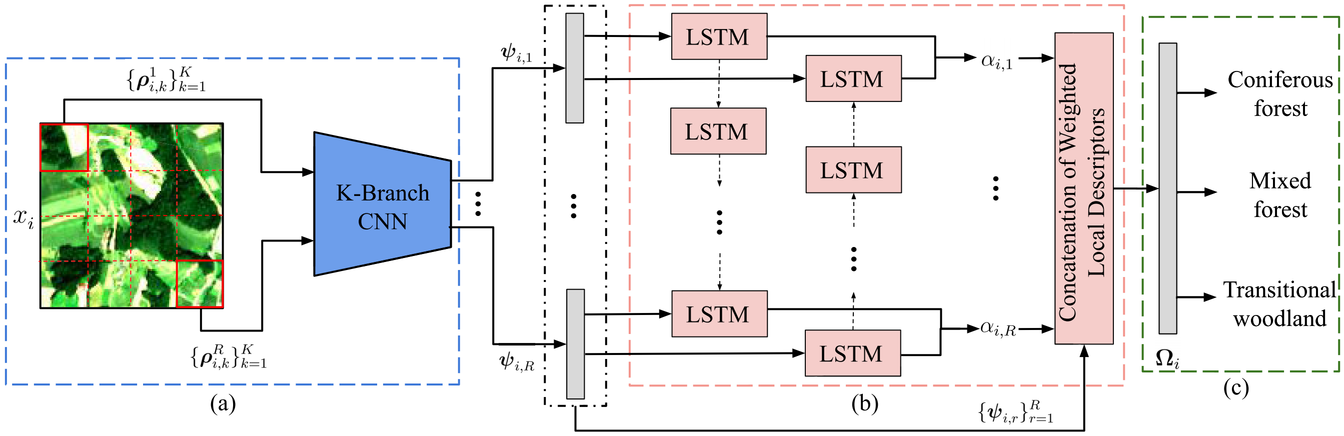


FIGURE 5. Detailed illustration of the three main steps of the proposed approach: (a) spatial and spectral characterization of local areas; (b) definition of a multi-attention driven global descriptor; (c) RS image scene classification with multi-labels.

the global descriptor Ω_i. Then, we obtain the class posterior probability of l_j for the image $\boldsymbol{x}_i$ with the *sigmoid* function as: $P(l_j|\boldsymbol{x}_i) = 1/(1 + e^{-z_{l_j}})$. After characterizing the class posterior probabilities, we define the overall loss of the approach as the cross entropy loss throughout all labels and images as follows:

$$\sum_{\boldsymbol{x}_i \in \mathcal{T}} \sum_{j=1}^{S} [l_j \in \boldsymbol{y}_i] log(P(l_j|\boldsymbol{x}_i)) + (1 - [l_j \in \boldsymbol{y}_i]) log(1 - P(l_j|\boldsymbol{x}_i)) \quad (5)$$

where $[l_j \in \boldsymbol{y}_i]$ is the Iverson bracket, which equals 1 if the l_j is one of the true multi-labels of $\boldsymbol{x}_i$, 0 otherwise. After end-to-end training of the entire neural network by minimizing the cross-entropy loss, the parameters θ of the function F (i.e., model parameters of the approach) can be learned. Accordingly, our model becomes capable of producing the posterior probabilities of multi-labels to be assigned to a new RS image scene $\boldsymbol{x}^*$. Then, the proposed approach predicts the multi-labels by thresholding the probability values.

Each step of the proposed approach is illustrated in Fig. 5.

III. DATA SET DESCRIPTION AND EXPERIMENTAL SETUP

A. DATA SET DESCRIPTION

We conducted all experiments on the BigEarthNet benchmark archive[1] [19]. BigEarthNet consists of 590, 326 Sentinel-2 images acquired between June 2017 and May 2018 over 10 European countries (Austria, Belgium, Finland, Ireland, Kosovo, Lithuania, Luxembourg, Portugal, Serbia, Switzerland). Each image in the archive is a section of: i) 120 × 120 pixels for 10m bands; ii) 60 × 60 pixels for 20m bands; and iii) 20 × 20 pixels for 60m bands and has been annotated with multi-labels among 43 land-cover classes provided from the 2018 CORINE Land Cover (CLC) database. The number of labels associated with each image is in the range of 1 and 12, whereas 5% of images have more than 5 multi-labels [19].

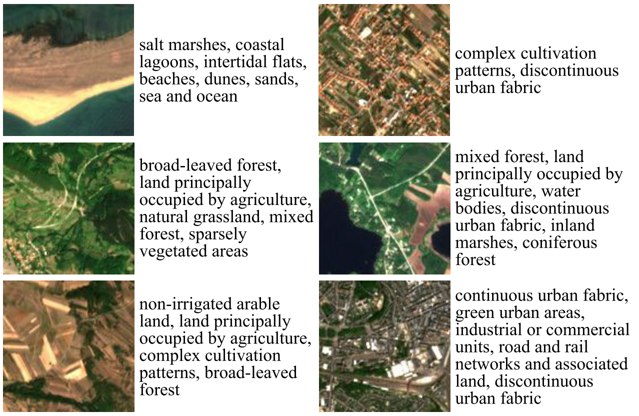

salt marshes, coastal lagoons, intertidal flats, beaches, dunes, sands, sea and ocean
complex cultivation patterns, discontinuous urban fabric
broad-leaved forest, land principally occupied by agriculture, natural grassland, mixed forest, sparsely vegetated areas
mixed forest, land principally occupied by agriculture, water bodies, discontinuous urban fabric, inland marshes, coniferous forest
non-irrigated arable land, land principally occupied by agriculture, complex cultivation patterns, broad-leaved forest
continuous urban fabric, green urban areas, industrial or commercial units, road and rail networks and associated land, discontinuous urban fabric

FIGURE 6. Example of Sentinel-2 images and their multi-labels in the BigEarthNet archive.

Fig. 6 provides an example of images with their multi-labels. Each image is atmospherically corrected. In the experiments, 70, 987 BigEarthNet images that are fully covered by seasonal snow, cloud and cloud shadow were not used[2]. According to our knowledge, BigEarthNet is the only archive in RS that includes Sentinel-2 multispectral images, each of which is annotated with multi-labels. Thus, we could only use it in the experiments in this paper. The other benchmark archives, e.g., DFC15 [11] and UC-Merced archives [20], consist of a very small number of RS images that are annotated with multi-labels and contain only RGB bands. Thus, they are not adequate to evaluate the proposed approach and are not considered in this paper.

The number of images associated with each BigEarthNet class varies significantly in the archive. To divide the BigEarthNet archive into training set (which is used for training the considered neural networks), validation set (which is used for selecting hyperparameters) and test set (which is

[1] The BigEarthNet is publicly available at http://bigearth.net.

[2] The lists of images fully covered by seasonal snow, cloud and cloud shadow are available at http://bigearth.net/#downloads.

used for accuracy assessment), one could apply random sampling. However, when images with multi-labels are considered, this approach has a risk that randomly selected images may not represent all classes present in the whole archive. There are also other approaches to divide a dataset into train, validation and test sets, however they are also designed for images annotated by single-labels and thus not suitable for multi-label applications [21]. In this paper, we develop an algorithm to represent each BigEarthNet class with a sufficient number of images in training, validation and test sets based on the label frequencies. The algorithm starts by including all images to the the training set. Let $\boldsymbol{c}_{l_m} \in \mathbb{N}$ be the number of images associated to the label l_m in the training set, where $m \in \{1, \ldots, S\}$, and thus we define the frequency γ_{l_m} of the label l_m in the training set as follows:

$$\gamma_{l_m} = \frac{\boldsymbol{c}_{l_m}}{\sum_{m=1}^{S} \boldsymbol{c}_{l_m}}. \tag{6}$$

Then, we define the cost of moving an image and its set of multi-labels from the training set to either validation or test set as follows:

$$\mathcal{C}_{\boldsymbol{x}_i,\boldsymbol{y}_i} = -\sum_{m=1}^{S} \frac{\gamma^*_{l_m} - \frac{1}{S}}{\sqrt{\gamma_{l_m}}} \tag{7}$$

where $\gamma^*_{l_m}$ indicates the new frequency of the label l_m after images are moved from the training set to the validation or test sets. The algorithm first sorts the label list in decreasing order based on the number of images associated to each class. Then, from the sorted list, the images with the decreasing cost values associated to each class are randomly selected and moved either to the validation set or to the test set. Since the algorithm starts to operate on the images associated to the majority classes, most of the images will be moved from the training set at the beginning. However, the cost value will reach the stationary point when it operates on the images associated to the minority classes. Application of this algorithm to the BigEarthNet results in a validation set of 198, 762 images, a test set of 203, 269 images, and a training set of 117, 308 images. The algorithm is summarized in Algorithm 1.

Algorithm 1 Our Algorithm for the Selection of Training, Validation and Test Sets

```
Input: $\mathcal{X}=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_M\}$, $\mathcal{L}=\{l_1,\ldots,l_S\}$, $\mathcal{Y}=\{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_M\}$
Assumption: $\mathcal{L}$ is sorted in decreasing order based the number of images associated to each class.
function LabelFreq($\mathcal{T}$, $l_m$)
    $\boldsymbol{c}_{l_m} \leftarrow |\{(\boldsymbol{x}_i,\boldsymbol{y}_i) \,|\, (\boldsymbol{x}_i,\boldsymbol{y}_i) \in \mathcal{T}, \boldsymbol{y}_{i,m} = 1\}|$
    $\gamma_{l_m} \leftarrow \boldsymbol{c}_{l_m} / (\sum_{m=1}^{S} \boldsymbol{c}_{l_m})$
    return $\gamma_{l_m}$
end function
function Cost($\mathcal{T}$, $(\boldsymbol{x}_i,\boldsymbol{y}_i)$, $S$, $\Gamma_{\mathcal{L}}$)
    $sum \leftarrow 0$
    for $m \leftarrow 1$ to $S$ do
        $\gamma_{l_m} \leftarrow \Gamma_{\mathcal{L}}$
        $\gamma^*_{l_m} \leftarrow$ LabelFreq($\mathcal{T} - (\boldsymbol{x}_i,\boldsymbol{y}_i)$, $l_m$)
        $sum \leftarrow sum - (\gamma^*_{l_m} - \frac{1}{S})/\sqrt{\gamma_{l_m}}$
    end for
    return $sum$
end function
$\mathcal{T} = \{(\boldsymbol{x}_i,\boldsymbol{y}_i) \,|\, \boldsymbol{x}_i \in \mathcal{X}, \boldsymbol{y}_i \in \mathcal{Y}\}$        ▷ Initial training set.
$\mathcal{V} = \emptyset$        ▷ Initial validation set.
$\mathcal{E} = \emptyset$        ▷ Initial test set.
$S \leftarrow |\mathcal{L}|$
$state \leftarrow$ Cost($\mathcal{T}$, $\emptyset$, $S$)
$\Gamma_{\mathcal{L}} \leftarrow \bigcup_{m=1}^{S}$LabelFreq($\mathcal{T}$, $l_m$)        ▷ Initial frequencies.
for $m \leftarrow 1$ to $S$ do
    for all $i$ such that $\boldsymbol{y}_{i,m} = 1$ do
        if Cost($\mathcal{T}$, $(\boldsymbol{x}_i,\boldsymbol{y}_i)$, $S$, $\Gamma_{\mathcal{L}}$) $< state$ then
            $\mathcal{T} \leftarrow \mathcal{T} - (\boldsymbol{x}_i,\boldsymbol{y}_i)$
            $(\mathcal{V} \leftarrow \mathcal{V} + (\boldsymbol{x}_i,\boldsymbol{y}_i)) \oplus (\mathcal{E} \leftarrow \mathcal{E} + (\boldsymbol{x}_i,\boldsymbol{y}_i))$
            $state \leftarrow$ Cost($\mathcal{T}$, $(\boldsymbol{x}_i,\boldsymbol{y}_i)$, $S$, $\Gamma_{\mathcal{L}}$)
        end if
    end for
end for
return $\mathcal{T}$, $\mathcal{V}$, $\mathcal{E}$        ▷ Resulting sets.
```

B. EXPERIMENTAL SETUP

After the selection of training, validation and test sets, we divided each image into non-overlapping local areas. Then, we employed three branch CNN (i.e., $K = 3$ for the K-Branch CNN) due to the three different spatial resolutions of Sentinel-2. Accordingly, for each local area, we split the bands into three subsets. Then, we stacked bands of each subset to obtain a single volume for each CNN branch. In detail, the bands 2 to 4 and 8 (which have 10m spatial resolution) were fed into the first branch, while the bands 5 to 7, 8A, 11 and 12 (which have 20m spatial resolution) were fed into the second branch and the third branch takes as input the remaining bands 1 and 9 (which have 60m spatial resolution). We selected the number of local areas and all other hyper-parameters with respect to the classification performance on the validation set. To select the local area size $w \times w$, w is tested within the range of $[18, 60]$ with a step size of 6. It is worth noting that, for the sizes, which are not evenly divisible by the image size (120×120 for 10m bands, 60×60 for 20m bands, 20×20 for 60m bands), we applied zero padding to the image borders. Although the same number of convolutional layers was used for all branches, the number of filters, the exploitation of pooling strategy and the filter sizes vary among branches. It is worth noting that the number of convolutional layers in all branches can be increased at a large extent to achieve deeper models. However, this would also increase the number of model parameters and thus the computational complexity. Accordingly, three convolutional layers were used for all branches. For the first branch, 32 filters with the size of 5×5, 32 filters with the size of 5×5 filters and 64 filters with the size of 3×3 filters were selected. For the second branch, the same number of filters was used, while 3×3 filters were employed in each layer. For the third branch, 32 filters with the size of 2 × 2 were used in each layer. We utilized the stride of 1 and zero padding in all convolutional layers to preserve the spatial dimensionality and not to lose information. In addition, max-pooling was utilized in the first two branches to provide partial translation invariance [22], which was not used in the last branch to avoid further decreasing the spatial resolution. For the LSTM networks, we used a 128 dimensional memory.

We jointly trained all CNN branches, FC layers and LSTM networks (i.e., an end-to-end learning of all steps was applied simultaneously). We used the Adam method [23] of Stochastic Gradient Descent with the initial learning rate of 10^{-3} to decrease the sigmoid cross entropy loss, which aims at maximizing the log-likelihood of the multi-labels in the training set. For the initialization of neural network weights, we utilized the Xavier method [24] to keep the variance of weights similar among all layers. We selected the 2×10^{-5} L2-regularization weight to layer-wise regularize the weights. 20% dropping out probability was chosen for Dropout regularization [25] to avoid the over-fitting of the proposed approach on the training set. In addition, we utilized the Batch Normalization [26] to decrease the effect of different spectral band statistics.

In the experiments, we compared the proposed approach with: 1) the Very Deep Convolutional Networks (i.e., VGG networks) [27]; 2) the Deep Residual Nets (i.e., ResNet networks) [28]; and 3) the Class-Wise Attention-Based Convolutional and Bidirectional LSTM Network [11] (denoted as CA-LSTM). For the VGG networks, we selected 16 layers (VGG16) and 19 layers (VGG19) versions. At the similar depths to the VGG networks, we selected 18 layers (ResNet18) and 34 layers (ResNet34) versions of the ResNet networks. These are widely used CNNs for the image classification problems in the CV literature. We used the same parameters presented in [27] and [28] for the VGG networks and the ResNet networks, respectively, except only the considered learning rates. CA-LSTM is one of the few DL based approaches proposed for the multi-label RS image scene classification task. For the CA-LSTM, we used the same feature extraction module (which is ResNet50 [28]), same LSTM network (bidirectional LSTM network with 2048 dimensional memory) and same parameters presented in the [11] except the learning rate.

We also evaluated the different steps at the proposed approach. To assess the effectiveness of the first step of the proposed approach (that is the K-Branch CNN), we compared it with different single branch CNN approaches. To this end, we initially applied cubic interpolation to 20m and 60m bands and stacked all bands into one volume. Then, three different approaches are considered as follows: 1) a single branch CNN that considers all the image bands as input and operates on the whole images (denoted as SiB-CNN); 2) a single branch CNN that considers all the image bands as input and operates on the local areas of images (denoted as L-SiB-CNN); and 3) a single branch CNN that considers only RGB image bands as input and operates on the whole images (denoted as SiB-CNN$_{\text{RGB}}$). For these approaches, the architecture of the first branch of the proposed K-Branch CNN is used. To evaluate the effectiveness of the second step of the proposed approach (that is the multi-attention strategy), we compared the results with those obtained without using the multi-attention strategy (i.e., only the first step is used). For all the experiments, we used the same training procedure from scratch with the same number of epochs, learning rate and the number of mini-batches to compare different approaches under the same setting. We performed our experiments on a cluster of 4 NVIDIA Tesla V100 GPUs.

Performance evaluation of any multi-label classification approach requires to analyze several factors rather than only evaluating the number of correct predictions and thus needs much more complex analysis with respect to the single-label case [29]. Accordingly, we utilized the different classification-based and ranking-based metrics with varying characteristics to accurately evaluate the accuracy of the proposed approach. Classification-based metrics consider the list of predicted classes, whereas ranking-based metrics focus on the ordered list of probabilities for all classes.

Under the category of classification-based metrics, results of experiments were provided in terms of three performance metrics: 1) Recall (R); 2) F^2-Score (F^2); and 3) Hamming loss (HL). Classification-based metrics can be calculated by: i) giving equal importance to each sample of the test set (sample averaging); ii) giving equal importance to each class (macro averaging); and iii) comparing the overall test set with the ground reference (micro averaging) regardless of giving importance to neither each sample nor each class.

Let TP_{ij}, FP_{ij}, FN_{ij} and TN_{ij} indicate the conditions of true positive, false positive, false negative and true negative, respectively, for the i^{th} image and j^{th} label (l_j), where each of them takes 0 or 1 and $TP_{ij}+FP_{ij}+FN_{ij}+TN_{ij} = 1$ holds. The recall is expressed by different averaging methods as follows:

$$R_{smpl} = \frac{1}{M}\sum_{i=1}^{M} \frac{\sum_{j=1}^{S} TP_{ij}}{\sum_{j=1}^{S} TP_{ij} + FN_{ij}} \tag{8}$$

$$R_{macr} = \frac{1}{S}\sum_{j=1}^{S} \frac{\sum_{i=1}^{M} TP_{ij}}{\sum_{i=1}^{M} TP_{ij} + FN_{ij}} \tag{9}$$

$$R_{micr} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{S} TP_{ij}}{\sum_{i=1}^{M}\sum_{j=1}^{S} TP_{ij} + FN_{ij}}. \tag{10}$$

The F^2-Score is the weighted harmonic mean of the correct prediction rates among the considered ground reference and the multi-label predictions. Thus, it is expressed by different averaging techniques as follows [30]:

$$F^2_{smpl} = \frac{1}{M}\sum_{i=1}^{M} \frac{\sum_{j=1}^{S} 5TP_{ij}}{\sum_{j=1}^{S} 5TP_{ij} + 4FN_{ij} + FP_{ij}} \tag{11}$$

$$F^2_{macr} = \frac{1}{S}\sum_{j=1}^{S} \frac{\sum_{i=1}^{M} TP_{ij}}{\sum_{i=1}^{M} 5TP_{ij} + 4FN_{ij} + FP_{ij}} \tag{12}$$

TABLE 1. Multi-label classification accuracies and the number of required model parameters (NP) when using local areas with different sizes for the proposed approach.

Local Area Size ($w \times w$)			Classification-Based Metrics (%)							Ranking-Based Metrics				NP
10m	20m	60m	R_{macr}	R_{smpl}	R_{micr}	F^2_{macr}	F^2_{smpl}	F^2_{micr}	HL	$RL(\%)$	$OE(\%)$	COV	$LRAP(\%)$	$(\times 10^6)$
18×18	9×9	3×3	51.0	68.1	61.7	46.9	68.9	64.2	4.1	2.7	6.5	5.8	85.3	**0.71**
24×24	12×12	4×4	50.0	66.2	59.2	46.0	67.4	62.2	4.1	2.8	6.8	5.9	85.1	0.93
30×30	15×15	5×5	52.8	68.5	62.3	47.2	69.4	64.7	**4.0**	**2.6**	**5.8**	**5.7**	**85.9**	1.13
36×36	18×18	6×6	54.4	70.7	65.0	47.8	71.1	66.5	4.1	2.6	6.2	5.7	85.7	1.70
42×42	21×21	7×7	53.3	70.7	64.8	46.2	71.0	66.6	4.1	2.6	6.5	5.8	85.6	2.03
48×48	24×24	8×8	**54.6**	**72.5**	67.1	48.0	**72.2**	68.2	4.1	2.6	6.5	5.8	85.3	2.81
54×54	27×27	9×9	54.2	72.3	**67.1**	**48.1**	72.2	**68.4**	4.1	2.6	6.3	5.7	85.6	3.29
60×60	30×30	10×10	54.1	72.4	66.8	46.7	71.8	67.6	4.3	2.9	6.9	6.1	84.4	4.26

$$F^2_{micr} = \frac{\sum_{i=1}^{M} \sum_{j=1}^{S} 5TP_{ij}}{\sum_{i=1}^{M} \sum_{j=1}^{S} 5TP_{ij} + 4FN_{ij} + FP_{ij}}. \tag{13}$$

The Hamming loss is the average Hamming distance between the ground reference labels and predicted multi-labels. Thus, it is defined as follows [31]:

$$HL = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{S} \sum_{j=1}^{S} [l_j \in \boldsymbol{y}_i \oplus l_j \in \boldsymbol{y}_i^*] \tag{14}$$

where $\oplus$ is the XOR logical operation.

Under the category of ranking-based metrics, results of experiments are provided in terms of four performance evaluation metrics: 1) Ranking loss (RL); 2) One error (OE); 3) Coverage (COV); and 4) Label ranking average precision ($LRAP$). All the ranking-based metrics are defined with respect to the ranking of the j^{th} label in the class probabilities result of an multi-label classification approach for the i^{th} image that is defined as $rank_{ij} = |k : P(l_k|\boldsymbol{x}_i) \geq P(l_j|\boldsymbol{x}_i)|$. Unlike the classification-based metrics, ranking-based metrics are calculated only by giving equal importance to each sample of the test set.

Accordingly, ranking loss is the rate of wrongly ordered label pairs (i.e., the probability of a label, which is irrelevant to the image, is higher than a ground reference label), and thus expressed as follows [32]:

$$RL = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{|\boldsymbol{y}_i|(S - |\boldsymbol{y}_i|)} \sum_{l_j \in \boldsymbol{y}_i} \sum_{l_k \notin \boldsymbol{y}_i} rank_{ik} \leq rank_{ij}. \tag{15}$$

The one error is the rate of test images whose predicted label having the highest ranking is not in the ground reference and thus defined as follows [29]:

$$OE = \frac{1}{M} \sum_{i=1}^{M} [\underset{j}{\text{argmax}}\, rank_{ij} \quad \notin \boldsymbol{y}_i]. \tag{16}$$

The coverage calculates the average number of labels required to be included in the prediction list of a multi-label classifier such that all ground reference labels will be predicted. Accordingly, it is defined as follows [32]:

$$COV = \frac{1}{M} \sum_{i=1}^{M} \max_{l_j \in \boldsymbol{y}_i} rank_{ij}. \tag{17}$$

For each ground reference label, the label ranking average precision calculates the rate of higher-ranked ground reference labels. This is expressed as follows [29]:

$$LRAP = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{\boldsymbol{y}_i} \sum_{l_j \in \boldsymbol{y}_i} \frac{|\{l_k : rank_{ik} \leq rank_{ij}, l_k \in \boldsymbol{y}_i\}|}{rank_{ij}}. \tag{18}$$

It is worth noting that, for any multi-label classifier, $LRAP$ provides scores strictly greater than 0 unlike the other metrics [29]. Thus, small differences in the score of this metric can be more informative compared to other metrics (e.g., recall). Smaller values of the Hamming loss, ranking loss, one error and coverage indicate better performance of an approach, whereas higher values of the recall, F^2-Score and the label ranking average precision are associated to better performance.

IV. EXPERIMENTAL RESULTS

We carried out different kinds of experiments in order to: 1) perform a sensitivity analysis with respect to different parameter settings and strategies; and 2) compare the effectiveness of the proposed approach with the widely used deep CNNs and one recent multi-label RS image scene classification approach [11].

A. SENSITIVITY ANALYSIS OF THE PROPOSED APPROACH

In this section, we performed the sensitivity analysis of the proposed approach under different parameter settings and strategies.

In the first set of trials, we analyzed the effect of utilizing local areas with different sizes in terms of the multi-label classification accuracy and computational complexity. Table 1 shows the results with the required number of parameters

TABLE 2. Results obtained by the SiB-CNN$_{\text{RGB}}$, the SiB-CNN, the L-SiB-CNN and the proposed K-branch CNN.

Method	Classification-Based Metrics (%)							Ranking-Based Metrics			
	R_{macr}	R_{smpl}	R_{micr}	F^2_{macr}	F^2_{smpl}	F^2_{micr}	HL	$RL(\%)$	$OE(\%)$	COV	$LRAP(\%)$
SiB-CNN$_{\text{RGB}}$	33.6	53.7	45.6	35.1	56.0	49.8	4.8	4.1	13.5	7.1	80.0
SiB-CNN	39.1	60.5	52.8	40.9	62.4	56.7	4.4	3.4	9.6	6.5	83.0
L-SiB-CNN	44.0	**65.7**	**58.8**	41.2	66.2	**62.9**	4.1	2.8	7.4	5.9	84.8
Proposed K-Branch CNN	**46.8**	64.7	57.7	**44.6**	**66.3**	61.0	**4.1**	**2.6**	**6.3**	**5.7**	**85.4**

TABLE 3. Multi-label classification accuracies obtained by using different steps of the proposed approach.

Steps of the Proposed Approach			Classification-Based Metrics (%)							Ranking-Based Metrics			
1st	2nd	3rd	R_{macr}	R_{smpl}	R_{micr}	F^2_{macr}	F^2_{smpl}	F^2_{micr}	HL	$RL(\%)$	$OE(\%)$	COV	$LRAP(\%)$
✓	✗	✓	46.8	64.7	57.7	44.6	66.3	61.0	4.1	2.6	6.3	5.7	85.4
✓	✓	✓	**52.8**	**68.5**	**62.3**	**47.2**	**69.4**	**64.7**	**4.0**	**2.6**	**5.8**	**5.7**	**85.9**

under different sizes of local areas. By analyzing the table, one can see that the reduction of computational complexity highly depends on the local area size $w \times w$. This is due to the fact that enlarging the local areas increases the number of parameters required to learn. As an example, using 18 × 18 sized local areas reduces the number of parameters by a half order of magnitude compared to the case for which 60 × 60 sized local areas are used. From the Table 1 one can also observe that the accuracies obtained by different sizes of local areas are similar to each other under most of the metrics. As an example, using 60 × 60 sized local areas provides almost the same F^2_{macr} score compared to the case 30 × 30 sized local area is considered. In few cases, there are noticeable differences in the results associated to metrics. As an example, using 48 × 48 sized local areas results in more than 7% higher R_{micr} compared to using 24 × 24 sized local areas. This is due to the fact that a smaller window size may reduce the capability of describing the spatial information content. All these results show that the selection of local area size in a proper range does not significantly affect the classification accuracy of the proposed approach, however considerably changes the computational complexity. Accordingly, for the rest of the experiments we used 30 × 30 sized local areas for 10m resolution bands since it provides the best values in ranking-based metrics and Hamming loss with a significantly reduced number of parameters (that is less than a half of those required for 48 × 48, 54 × 54 and 60 × 60 local area sizes).

In the second set of trials, we analyzed the effect of the first step of the proposed approach on the multi-label classification accuracy. To this end, we compare the results of the K-Branch CNN (which is introduced in the first step) with those obtained by the SiB-CNN$_{\text{RGB}}$ (which is a single branch CNN that considers RGB bands), SiB-CNN (which is a single branch CNN that considers all bands) and L-SiB-CNN (which is a single branch CNN that considers all bands and operates on the image local areas). Table 2 shows the multi-label classification accuracies under different metrics. From this table, one can observe that the proposed K-Branch CNN provides the best scores under most of the metrics. As an example, the proposed K-Branch CNN provides more than 9%, almost 4% and more than 3% higher F^2_{macr} scores compared to the SiB-CNN$_{\text{RGB}}$, SiB-CNN and L-SiB-CNN, respectively. In greater detail, the SiB-CNN provides more than 6% higher F^2_{smpl} score by achieving a reduction of about 4% in one error compared to the SiB-CNN$_{\text{RGB}}$. This shows that using spectral bands associated to 20m and 60m spatial resolutions improves the multi-label classification accuracy. Moreover, the L-SiB-CNN provides more than 4% higher F^2_{smpl} score by achieving a reduction of more than 9% in coverage compared to the SiB-CNN. This indicates that exploiting local areas of images also improves the multi-label classification accuracy. In addition, the proposed K-Branch CNN leads to a reduction of about 7% in Hamming loss and more than 7% higher R_{macr} compared to the SiB-CNN. All these results show that the K-Branch CNN much more accurately characterizes the spectral content of RS images by utilizing all spectral bands with different spatial resolutions in branch-wise CNN architecture compared to single branch CNN approaches (which require to apply interpolation to lower resolution bands).

In the third set of trials, we evaluated the effect of the second step of the proposed approach. To this end, we compared the results of proposed approach with those obtained by neglecting the multi-attention strategy (i.e., only the first step is used). When the second step is neglected, global descriptors are obtained by the concatenation of local descriptors without weighted by attention scores. Table 3 shows the multi-label classification accuracies under different metrics. From this table, one can observe that when the use of multi-attention strategy significantly improves the classification accuracy under all the metrics. As an example, the improvements are 6% in R_{macr} and more than 3% in F^2_{smpl} score. This shows the

TABLE 4. Results obtained by the ResNet18, ResNet34, VGG16, VGG19, CA-LSTM and the proposed approach together with the number of required model parameters (NP).

Method	Classification-Based Metrics (%)							Ranking-Based Metrics				NP
	R_{macr}	R_{smpl}	R_{micr}	F^2_{macr}	F^2_{smpl}	F^2_{micr}	HL	$RL(\%)$	$OE(\%)$	COV	$LRAP(\%)$	$(\times 10^6)$
ResNet18 [28]	36.1	59.9	52.6	34.8	61.1	55.5	4.9	8.1	12.8	11.4	75.5	11.2
ResNet34 [28]	37.0	64.4	57.8	35.7	64.6	59.8	4.9	6.3	12.6	9.6	77.6	21.3
VGG16 [27]	37.8	62.7	55.6	36.1	64.2	58.7	4.5	3.3	10.1	6.3	82.4	134.4
VGG19 [27]	41.5	61.5	54.2	38.1	63.0	57.4	4.6	3.5	11.1	6.4	81.5	139.8
CA-LSTM [11]	43.5	64.8	58.5	40.4	65.5	60.7	4.7	3.7	9.9	6.8	81.5	33.5
Proposed Approach	**52.8**	**68.5**	**62.3**	**47.2**	**69.4**	**64.7**	**4.0**	**2.6**	**5.8**	**5.7**	**85.9**	**1.1**

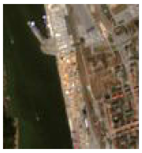
RS Images	**Multi-Labels**	**ResNet18** [27]	**ResNet34** [27]	**VGG16** [26]	**VGG19** [26]	**CA-LSTM** [11]	**Proposed Approach**
	Coniferous forest, Water bodies	Mixed forest, Water bodies	Mixed forest, Water bodies	Broad-leaved forest, Coniferous forest, Mixed forest, Water bodies	Broad-leaved forest, Coniferous forest, Mixed forest, Water bodies	Coniferous forest, Water bodies	Coniferous forest, Water bodies
	Pastures, Land principally occupied by agriculture, Coniferous forest, Transitional woodland/shrub	Pastures, Land principally occupied by agriculture, Natural grassland	Pastures, Land principally occupied by agriculture, Mixed forest	Pastures, Coniferous forest, Natural grassland, Moors and heathland, Transitional woodland/shrub	Pastures, Coniferous forest	Pastures, Land principally occupied by agriculture, Natural grassland	Pastures, Land principally occupied by agriculture, Coniferous forest, Transitional woodland/shrub
	Discontinuous urban fabric, Port areas, Pastures, Coniferous forest, Coastal lagoons	Discontinuous urban fabric, Industrial or commercial units	Discontinuous urban fabric, Port areas	Discontinuous urban fabric, Industrial or commercial units, Port areas, Green urban areas, Coniferous forest, Mixed forest, Transitional woodland/shrub, Water courses	Discontinuous urban fabric, Industrial or commercial units, Green urban areas, Coniferous forest, Mixed forest, Coastal lagoons, Sea and ocean	Discontinuous urban fabric, Pastures, Coniferous forest, Coastal lagoons	Discontinuous urban fabric, Port areas, Pastures, Coniferous forest, Coastal lagoons

FIGURE 7. An example of the BigEarthNet images with the true multi-labels and the multi-labels assigned by the ResNet18, ResNet34, VGG16, VGG19, CA-LSTM and the proposed approach.

effect of modeling the importance scores of image local areas for the characterization of a global descriptor.

B. COMPARISON AMONG THE EXISTING APPROACHES

In the fourth set of trials, we compared the effectiveness of the proposed approach with the ResNet architectures at the depths of 18 and 34 (ResNet18 and ResNet34), the VGG architectures at the depth 16 and 19 (VGG16 and VGG19) and the CA-LSTM (which is a recent multi-label RS scene classification approach). Table 4 shows the multi-label classification results of these methods under different metrics. By analyzing the table, one can observe that our proposed approach leads to the highest accuracies with the lowest number of parameters. As an example, the proposed approach provides 15% higher R_{macr}, more than 5% higher F^2_{smpl} score and a reduction of more than 21% in ranking loss compared to the VGG16 (which is one of the well known CNNs for image classification problems). Moreover, the proposed approach requires a significantly reduced number of parameters that is more than two orders of magnitude compared to the VGG16. Even with the deeper architecture (VGG19), the VGG approach is not capable of increasing the classification accuracy (while providing the lowest scores under all metrics except the R_{macr} and F^2_{macr} compared to the VGG16) and requires the highest number of parameters to learn. As an example, the VGG16 leads to a reduction of about 6% in ranking loss. This shows that increasing the depth of a CNN is not sufficient to obtain accurate multi-label RS classification results. In addition, the proposed approach leads to more than 11% higher F^2_{macr} score and more than 8% higher *LRAP* score with a reduced number of parameters that is more than an order of magnitude lower compared to the ResNet34 (which is one of the most popular CNNs due to the integration of residual connections with convolutional layers). The proposed approach provides better metric values (e.g., more than 9% higher R_{macr}, 7% higher F^2_{macr} score, 9% higher R_{macr} and a reduction of about 30% in ranking loss) also compared to the CA-LSTM. This success has been achieved with the significantly reduced number of parameters by more than an order of magnitude. All these results clearly show that the proposed approach reduces the needs for very deep CNNs to achieve a high classification accuracy. This is

TABLE 5. Table of symbols.

Symbol	Description
$\mathcal{X}$	the archive of M RS images
$\boldsymbol{x}_i$	i^{th} RS image in the archive
$\mathcal{L}$	set of class labels for the archive
S	number of classes in the archive
$\boldsymbol{y}_i$	multi-label vector of i^{th} RS image
l_s	an element of $\mathcal{L}$
$\mathcal{Y}$	set of class labels associated with all images in the archive
K	number of spatial resolutions associated with spectral bands
$\boldsymbol{x}^*$	a new RS image
$F(\boldsymbol{x}^*;\theta)$	a function which maps $\boldsymbol{x}^*$ to multi-labels
$f(\boldsymbol{x}^*;\theta)$	a function which produces classification scores of each label $l_s \in \mathcal{L}$ for $\boldsymbol{x}^*$
z_{l_j}	class score of label l_j
$\boldsymbol{y}^*$	multi-label prediction vector for $\boldsymbol{x}^*$
$g(\cdot)$	a function which produces multi-label predictions
θ	set of model parameters required to learn $F(\boldsymbol{x}^*;\theta)$
R	number of local areas
$w \times w$	local area size
$\boldsymbol{\rho}_i^r$	r^{th} local area of the i^{th} RS image
$\boldsymbol{\rho}_{i,k}^r$	k^{th} subset of the r^{th} local area with a given spatial resolution
ϕ^k	k^{th} branch of the proposed K-Branch CNN
$\boldsymbol{\psi}_{i,r}$	local descriptor for the r^{th} local area of the i^{th} RS image
$\boldsymbol{f}_r$	forget gate of the LSTM cell associated with r^{th} local area
$\boldsymbol{i}_r$	input gate of the LSTM cell associated with r^{th} local area
$\boldsymbol{o}_r$	output gate of the LSTM cell associated with r^{th} local area
$\boldsymbol{c}_r$	cell state of the LSTM cell associated with r^{th} local area
$\mathbf{W}_{\cdot,r}$	weight parameters of the LSTM cell associated with r^{th} local area
$\boldsymbol{b}_{\cdot,r}$	bias parameters of the LSTM cell associated with r^{th} local area
$\tanh$	hyperbolic tangent function
δ	sigmoid function
$\boldsymbol{h}_r$	preliminary attention score associated with r^{th} local area
$\alpha_{i,r}$	attention score associated with r^{th} local area of the i^{th} RS image
$\boldsymbol{\Omega}_i$	global descriptor of the i^{th} RS image
$\boldsymbol{c}_{l_m}$	number of images associated with label l_m in a training set
γ_{l_m}	frequency of label l_m in a training set
$\mathcal{C}_{\boldsymbol{x}_i,\boldsymbol{y}_i}$	cost of moving an image and its multi-labels $(\boldsymbol{x}_i,\boldsymbol{y}_i)$ from the training set
$\mathcal{T},\mathcal{V},\mathcal{E}$	training, validation and test sets

an important advantage, since reducing the number of model parameters to achieve promising performance is as important as the classification accuracy for DL based approaches. Figure 7 shows an example of BigEarthNet images with the true multi-labels and the multi-labels assigned by the ResNet18, ResNet34, VGG19, VGG16, CA-LSTM and the proposed approach. By analyzing the figure, one can see that our proposed approach accurately predicts all classes without predicting any wrong ones. Unlike the proposed approach, the VGG16 and VGG19 predict several unrelated classes. As an example, both of the approaches predict *broad-leaved forest* and *mixed-forest* classes for the first image, although this image does not contain these classes. ResNet18 and ResNet34 are able to accurately predict only some of the multi-labels. As an example, for the image in the center, the ResNet networks correctly predict *pastures* and *land principally occupied by agriculture* classes, however *coniferous forest* and *transitional woodland/shrub* classes are not predicted and thus missed. These results prove that the VGG and ResNet networks are less accurate in the prediction of all classes present in the images with respect to the proposed approach. From the figure, one can see that the CA-LSTM provides accurate results for the top image without any wrong classification. However, for more complex images, this approach is not capable of identifying some classes. As an example, for the image in the center, the CA-LSTM wrongly predicts *mixed forest* and *natural grassland* classes instead of *coniferous forest* and *transitional woodland/shrub* classes. However, these classes are accurately predicted by the proposed approach. As another example, for the bottom image, the CA-LSTM does not provide any correct prediction, whereas the proposed approach correctly predicts all the classes. These results, again, prove that the proposed approach more accurately describes the complex spatial and spectral content of RS images compared to the CA-LSTM.

V. CONCLUSION

In this paper, we have introduced a novel DL based approach for multi-label remote sensing image scene classification. The proposed approach is made up of three main steps. The first step achieves spatial and spectral characterization of image local areas by a novel K-Branch CNN, which includes spatial resolution specific CNN branches. The second step initially estimates the multiple attention scores to identify the importance levels (i.e., scores) of different image local areas. This is achieved by the novel bidirectional LSTM-based multi-attention strategy. Then, each image is represented by a global descriptor defined on the basis of the attention scores. In the third step, images modeled by the multi-attention driven global descriptors are classified and multi-label predictions are obtained. Experimental results obtained on the BigEarthNet (which is a large-scale Sentinel-2 benchmark archive) demonstrate that the proposed approach significantly improves the multi-label scene classification accuracy compared to the well known deep CNNs and the state-of-the-art attention driven multi-label RS image classification approach. Moreover, the proposed approach provides a computationally more efficient solution for multi-label classification problems due to the significant reduction in the number of model parameters. Decreasing the model complexity reduces the risk of over-fitting (which also contributes to the improvement in the classification accuracy). All the results confirm that the proposed approach is much more suitable to be used within the operational RS scene classification scenarios, where the images contain highly complex spatial and spectral information content. The main reasons for the success of the proposed approach are summarized as follows:

1) Due to the proposed K-Branch CNN (which includes a specialized branch in terms of the DL techniques utilized throughout layers for the set of image bands with the same spatial resolution), the proposed approach significantly improves the characterization of complex spatial and spectral content of high-dimensional RS images with high-spatial resolution. Moreover, K-Branch CNN leads to a significant reduction on the computational complexity of the entire approach by reducing the number of model parameters.
2) Due to the proposed multi-attention strategy (which efficiently exploits the bidirectional LSTM sequences on the local descriptors of each RS image to estimate the multi-attention scores), the proposed approach accurately extracts and exploits the importance levels of image local areas which are then used to define the global descriptors.

It is worth noting that although in our experiments we have used the Sentinel-2 multispectral images (which include 13 bands associated to three different spatial resolutions), the proposed approach can be used with any multispectral RS image. This can be achieved by selecting: i) the number K of branches as the total number of different spatial resolutions associated to the considered RS image bands; and ii) the proper values of the hyperparameters for each branch in the K-Branch CNN. If all the image bands are associated to the same spatial resolution value, the K-Branch CNN turns into a single branch CNN (i.e., $K = 1$). It is also important to note that when RS image bands with varying spatial resolutions are considered, the most straightforward way is to apply interpolation to the lower spatial resolution bands and then to use a single-branch CNN. However, the experimental results show that the use of interpolation may lead to a loss on the scene classification accuracy.

As a final remark, it is worth noting that to define the local areas of each image, we simply divide images into non-overlapping blocks. As a future work, we plan to apply a strategy for an adaptive definition of local areas based on the semantic content of RS images that can further improve the classification accuracy. Moreover, we also plan to develop a data summarization strategy [33] instead of stacking local descriptors in the second step of the proposed approach.

NOTATION AND SYMBOLS

A list of the notation and symbols used throughout the paper is given in Table 5.

REFERENCES

[1] F. Zhang, B. Du, and L. Zhang, "Scene classification via a gradient boosting random convolutional network framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1793–1802, Mar. 2016.

[2] K. Nogueira, O. A. B. Penatti, and J. A. dos Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recognit.*, vol. 61, pp. 539–556, Jan. 2017.

[3] G. Sumbul, R. G. Cinbis, and S. Aksoy, "Multisource region attention network for fine-grained object recognition in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4929–4937, Jul. 2019.

[4] S. Roy, E. Sangineto, N. Sebe, and B. Demir, "Semantic-fusion Gans for semi-supervised satellite image classification," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 684–688.

[5] X. Lu, H. Sun, and X. Zheng, "A feature aggregation convolutional neural network for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7894–7906, Oct. 2019.

[6] J. Xie, N. He, L. Fang, and A. Plaza, "Scale-free convolutional neural network for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6916–6928, Sep. 2019.

[7] O. E. Dai, B. Demir, B. Sankur, and L. Bruzzone, "A novel system for content-based retrieval of single and multi-label high-dimensional remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 7, pp. 2473–2490, Jul. 2018.

[8] A. Zeggada, F. Melgani, and Y. Bazi, "A deep learning approach to UAV image multilabeling," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 694–698, May 2017.

[9] S. Koda, A. Zeggada, F. Melgani, and R. Nishii, "Spatial and structured SVM for multilabel image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 5948–5960, May 2018.

[10] R. Stivaktakis, G. Tsagkatakis, and P. Tsakalides, "Deep learning for multilabel land cover scene categorization using data augmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 7, pp. 1031–1035, Jul. 2019.

[11] Y. Hua, L. Mou, and X. X. Zhu, "Recurrently exploring class-wise attention in a hybrid convolutional and bidirectional LSTM network for multi-label aerial image classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 149, pp. 188–199, Mar. 2019.

[12] Y. Hua, L. Mou, and X. X. Zhu, "Relation network for multilabel aerial image classification," *IEEE Trans. Geosci. Remote Sens.*, early access, Feb. 7, 2020.

[13] A. Alshehri, Y. Bazi, N. Ammour, H. Almubarak, and N. Alajlan, "Deep attention neural network for multi-label classification in unmanned aerial vehicle imagery," *IEEE Access*, vol. 7, pp. 119873–119880, 2019.

[14] G. Sumbul and B. Demir, "A novel multi-attention driven system for multi-label remote sensing image classification," in *Proc. IGARSS-IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 5726–5729.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[16] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural Comput.*, vol. 12, no. 10, pp. 2451–2471, Oct. 2000.

[17] S. Masum, J. P. Chiverton, Y. Liu, B. Vuksanovic, and M. Petridis, "Investigation of machine learning techniques in forecasting of blood pressure time series data," in *Proc. Int. Conf. Innov. Techn. Appl. Artif. Intell.*, 2019, pp. 269–282.

[18] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1412–1421.

[19] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "Bigearthnet: A large-scale benchmark archive for remote sensing image understanding," in *Proc. IGARSS-IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 5901–5904.

[20] B. Chaudhuri, B. Demir, S. Chaudhuri, and L. Bruzzone, "Multilabel remote sensing image retrieval using a semisupervised graph-theoretic method," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 1144–1158, Feb. 2018.

[21] Z. A. Daniels and D. N. Metaxas, "Addressing imbalance in multi-label classification using structured Hellinger forests," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 1826–1832.

[22] D. Mishkin, N. Sergievskiy, and J. Matas, "Systematic evaluation of CNN advances on the ImageNet," *Comput. Vis. Image Understand.*, vol. 161, pp. 11–19, Aug. 2017.

[23] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, Dec. 2014, pp. 1–41.

[24] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.

[25] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[26] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[29] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, Aug. 2014.

[30] R. A. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Reading, MA, USA: Addison-Wesley, 2011.

[31] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *Int. J. Data Warehousing Mining*, vol. 3, no. 3, pp. 1–13, 2007.

[32] G. Tsoumakas, I. Katakis, and I. Vlahavas, *Mining Multi-label Data*. Boston, MA, USA: Springer, 2010, pp. 667–685.

[33] M. Ahmed, "Data summarization: A survey," *Knowl. Inf. Syst.*, vol. 58, no. 2, pp. 249–273, Feb. 2019.

GENCER SUMBUL (Graduate Student Member, IEEE) received the B.S. degree in computer engineering from Bilkent University, Ankara, Turkey, in 2015, and the M.S. degree in computer engineering from Bilkent University, in 2018. He is currently pursuing the Ph.D. degree with the Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin, Germany. He is also a Research Associate with the Remote Sensing Image Analysis (RSiM) group. His research interests include computer vision and machine learning, with special interest in deep learning and remote sensing.

BEGÜM DEMİR (Senior Member, IEEE) received the B.S., M.Sc., and Ph.D. degrees in electronic and telecommunication engineering from Kocaeli University, Kocaeli, Turkey, in 2005, 2007, and 2010, respectively.

She has been a Full Professor and the Head of the Remote Sensing Image Analysis (RSiM) group, Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin, Germany, since 2018. Before starting at TU Berlin, she was an Associate Professor with the Department of Computer Science and Information Engineering, University of Trento, Italy. Her research activities lie at the intersection of machine learning, remote sensing, and signal processing. Specifically, she performs research on developing innovative methods for addressing a wide range of scientific problems in the area of remote sensing for Earth observation.

Dr. Demir is a Scientific Committee member of several international conferences and workshops, such as Conference on Content-Based Multimedia Indexing, Conference on Big Data from Space, Living Planet Symposium, International Joint Urban Remote Sensing Event, SPIE International Conference on Signal and Image Processing for Remote Sensing, and Machine Learning for Earth Observation Workshop organized within the ECML/PKDD. She was a recipient of a Starting Grant from the European Research Council (ERC) with the project BigEarth-Accurate and Scalable Processing of Big Data in Earth Observation in 2017 and the 2018 Early Career Award presented by the IEEE Geoscience and Remote Sensing Society. She is a referee for several journals such as the PROCEEDINGS OF THE IEEE, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, IEEE TRANSACTIONS ON IMAGE PROCESSING, *Pattern Recognition*, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, *International Journal of Remote Sensing*, and several international conferences. She is currently an Associate Editor of the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS and *MDPI Remote Sensing*.