Attentional Dense Convolutional Neural Network for Water Body Extraction From Sentinel-2 Images

Janak Parajuli, Ruben Fernandez-Beltran, *Senior Member, IEEE*, Jian Kang, *Member, IEEE*, and Filiberto Pla

***Abstract*—Monitoring water bodies from remote sensing data is certainly an essential task to supervise the actual conditions of the available water resources for environment conservation, sustainable development, and many other applications. Being Sentinel-2 images some of the most attractive data, existing traditional index-based and deep learning-based water extraction methods still have important limitations in effectively dealing with large heterogeneous areas since many types of water bodies with different spatial-spectral complexities are logically expected. Note that, in this scenario, optimal feature abstraction and neighborhood information may certainly vary from water to water pixel, however existing methods are generally constrained by a fix abstraction level and amount of land cover context. To address these issues, this article presents a new attentional dense convolutional neural network (AD-CNN) especially designed for water body extraction from Sentinel-2 imagery. On the one hand, the AD-CNN exploits dense connections to allow uncovering deeper features while simultaneously characterizing multiple data complexities. On the other hand, the proposed model also implements a new residual attention module to dynamically put the focus on the most relevant spatial-spectral features for classifying water pixels. To test the performance of the AD-CNN, a new water database of Nepal (WaterPAL) is also built. The conducted experiments reveal the competitive performance of the proposed architecture with respect to several traditional index-based and state-of-the-art deep learning-based water extraction models.**

***Index Terms*—Convolutional neural networks (CNNs), dense networks, residual attention networks, Sentinel-2, water bodies.**

I. Introduction

WITHOUT doubt, water is a significant part of nature with a key role in human life, environment, and climate. Being one of the most intensively exploited natural resources, the accurate and continuous knowledge on terrestrial water is fundamental in many different applications, such as precision farming, disaster management, drought detection, or Earth surface analysis [1], [2], [3], [4]. In this way, monitoring water bodies becomes an essential task to supervise the actual conditions of the available water resources along with environment conservation and sustainable development [5]. This relevance is such that even small changes in the water distribution may have a huge impact on human lives, causing soil subsidence, inland inundation, and health hazards, among other critical issues. Besides, water is also an integral part of different thematic and topographic maps used for many different purposes. Under this scenario, timely updated data are logically required to effectively monitor water bodies, which tend to change from time to time unlike other more stable structures like buildings or roads [6]. Unfortunately, this demand is difficult to cover using time consuming in situ procedures, especially in the context of developing countries [7].

With the expansion of remote sensing technologies, different satellites and constellations were designed to satisfy the regular provision of multispectral Earth observation data, which become particularly useful for water monitoring [8]. From Moderate-Resolution Imaging Spectroradiometer (MODIS) [9] and Landsat [10], to many other open and commercial satellites (e.g., Sentinel, Rapideye, ZY-3, EnviSat, Corona, radar satellite (RADARSAT), Gaofeng, etc.), multiple Earth observation data can be available for analysis [11]. Among all the available alternatives, Sentinel-2 has certainly shown to be one of the most suitable missions for the accurate detection of water bodies because of the advantages of its imaging products [12]: free availability, 13-band spectral resolution, and high spatial resolution of up to 10 m. Different water detection works published in the literature exemplify this fact, e.g., [13], [14], [15].

In general, two dominant trends can be identified when it comes to automatic water body extraction from remote sensing images [16]: traditional index-based methods and deep learning-based techniques. Despite their simplicity based on spectral indices and thresholding, traditional water extraction approaches may have important constraints in accurately distinguishing water from snow, mountains, buildings, and shadows due to the own limitations of pixel-wise computations [17], [18]. Auxiliary data like digital elevation model (DEM) may help to relieve some of these issues [19], [20]. However, how to choose the most suitable threshold value to extract even small water bodies is still a major problem [21]. As a result, traditional approaches are often not the best solution at global scales since they are unable to integrate shapes and texture information characteristic of water pixels.

In contrast, deep learning-based methods take advantage of convolutional neural networks (CNNs) to uncover more

Manuscript received 6 June 2022; revised 27 July 2022; accepted 2 August 2022. Date of publication 15 August 2022; date of current version 26 August 2022. This work was supported in part by the Ministerio de Ciencia e Innovación under Grant PID2021-128794OB-I00 and in part by the National Natural Science Foundation of China under Grant 62101371. (*Corresponding author: Jian Kang.*)

Janak Parajuli and Filiberto Pla are with the Institute of New Imaging Technologies, Department of Computer Languages and Systems, University Jaume I, E-12071 Castellón de la Plana, Spain (e-mail: al393628@uji.es; pla@uji.es).

Ruben Fernandez-Beltran is with the Department of Computer Science and Systems, University of Murcia, 30100 Murcia, Spain (e-mail: rufernan@um.es).

Jian Kang is with the School of Electronic and Information Engineering Department of Electronic Science and Technology, Soochow University, Suzhou 215006, China (e-mail: jiankang@suda.edu.cn).

The codes and data related to this article will be accessible on https://github.com/rufernan/ADCNN.

Digital Object Identifier 10.1109/JSTARS.2022.3198497

discriminating spatial-spectral features for the better identification of water bodies [22]. In this respect, different CNN technologies have been successfully exploited, being the classification scheme one of the most general mapping frameworks. For example, Pu et al. [23] propose a hierarchical CNN for water-quality classification. Analogously, Rezaee et al. [24] build a two-level network for exploiting high-level water features too. Chen et al. [25] adopt an adaptive pooling to better preserve water context and boundary information. Other works also propose different multiresolution schemes for further improving the generalization capabilities of CNN-based features for water classification [26], [27].

Despite all the conducted research, there are still important challenges in terms of the abstraction level of the uncovered water features based on the network design. Due to larger depths, deep learning models tend to suffer from the vanishing gradient problem that rapidly degrade the learning process, and hence, the quality of the results [28]. Thus, many of the existing water classification networks, e.g., [23], [24], [25], try to control the number of layers and feature maps for relieving these negative effects. Nonetheless, this strategy may often reduce the abstraction capabilities of the extracted features while limiting the resulting classification performance, especially when considering rich spatial-spectral data like in the Sentinel-2 case. In this scenario, this article proposes a new CNN-based classification model (AD-CNN) especially designed for water body extraction from Sentinel-2 imagery, based on the following key aspects: dense connectivity, residual learning, and attention. On the one hand, dense connections are used to relieve vanishing gradients as well as an excessive expansion of receptive fields at very deep layers with the objective of better preserving water local information when extracting deeper features. Besides, they also work for jointly exploiting from lower to higher level features in order to deal with the numerous spatial-spectral complexities of water pixels at large scales. On the other hand, a new residual attention module (RAM) is implemented to dynamically put the focus on the most relevant spatial-spectral features when identifying water bodies. To evaluate the performance of the proposed approach, we first create a new dataset of Nepal (WaterPAL) made of Sentinel-2 images, DEM data, and ground-truth water information. Then, we conduct multiple experiments including several state-of-the-art index-based and CNN-based water extraction methods. Summarizing, the main contributions of this work can be listed as follows:

1) We build a new database of Nepal (WaterPAL) composed by Sentinel-2 images, DEM data, and ground-truth water information.
2) We propose a novel water extraction architecture (AD-CNN) that jointly exploits dense connectivity, residual learning, and attention mechanisms to uncover more discriminating deep features from water bodies.

The remaining part of this article continues with the literature review of related works in Section II. Section III describes the geographical location of the study area and the detailed steps for the dataset preparation. Section IV delineates the workflow and structure of the proposed methodology. Section V provides details about the experimental setup and results. Finally, Section VI concludes this article.

II. Related Work

A. Traditional Index-Based Methods

In general, index-based methods focus on the spectral properties of water with the objective of defining single-band or multiband computations to isolate water pixels within a particular value range. In this way, a common practice consists in exploiting the conjugate ratio between green and red bands to segregate the spectral response of water [29]. To avoid noise from artificial constructions like buildings, this approach is often improved by using near-infrared (NIR) instead of red bands [30]. With these considerations in mind, different indices have been proposed and utilized in the literature to extract water bodies [31]. One of the most popular indices is the normalized difference water index (NDWI) [32], which is calculated using green and NIR bands as follows:

$$\mathrm{NDWI} = \frac{\rho_{\mathrm{green}} - \rho_{\mathrm{NIR}}}{\rho_{\mathrm{green}} + \rho_{\mathrm{NIR}}} \tag{1}$$

where ρ_{green} and ρ_{NIR} are green and NIR reflectance bands, respectively. The values of this index range between -1 and 1, representing positive values water bodies [33]. However, depending on the region of interest, some built-up areas could still generate noisy false positive results. Then, other authors also propose the modified NDWI (MNDWI) [34] as

$$\mathrm{MNDWI} = \frac{\rho_{\mathrm{green}} - \rho_{\mathrm{MIR}}}{\rho_{\mathrm{green}} + \rho_{\mathrm{MIR}}} \tag{2}$$

where ρ_{MIR} is the midinfrared (MIR) reflectance band. With this change, built-up areas usually become negative but some additional problems appear with mountain shadows and snow, making the MNDWI index mainly suitable for urban water extraction. Similarly, another index termed new water index (NWI) was also proposed in [35], where green and NIR bands are replaced by blue and Landsat MIR bands as follows:

$$\mathrm{NWI} = \frac{\rho_{\mathrm{blue}} - (\rho_{\mathrm{NIR}} + \rho_{\mathrm{MIR}_1} + \rho_{\mathrm{MIR}_2})}{\rho_{\mathrm{blue}} + (\rho_{\mathrm{NIR}} + \rho_{\mathrm{MIR}_1} + \rho_{\mathrm{MIR}_2})}. \tag{3}$$

In addition to these, other related indices have also shown prominent results in detecting water bodies from remote sensing data. For instance, it is the case of the normalized difference vegetation index (NDVI) [36], which employs the difference between NIR and red bands, following the same scheme as NDWI, to primarily extract vegetation while detecting water as negative values. In fact, some works in the literature show the advantages of jointly exploiting both NDWI and NDVI for water body extraction, e.g., [37]. Other authors also propose using the principal component analysis (PCA) approach to only consider the most informative image components when computing the own index, as in the case of the enhanced water index (EWI) [21]. However, the high computational cost of PCA strongly limits the applicability of this scheme over large interest regions.

B. Deep Learning-Based Methods

Despite their efficacy, traditional index-based methods usually have important limitations when working at global scales since optimal water detection ranges may often vary from local to local scenes [17]. To provide a more general solution, deep

learning methods aim at exploiting characteristic spatial-spectral information of water pixels via CNNs. In this regard, numerous approaches can be found in the related literature.

For instance, Yang et al. [38] propose using an stacked sparse autoencoder for extracting pixel-wise features that take into account neighborhood information using a feature expansion algorithm. In the case of [22], the authors opt by developing a classification CNN, named Deep-WaterMap, which is specifically trained to separate water from land, snow, ice, clouds, and shadows using Landsat images as input. Chen et al. [25], extend this concept to ZY-3 and Gaofeng satellites by adopting a self-adaptive pooling into the own network to extract water features more robust to terrain local variations.

Despite the positive results achieved by these and other relevant deep learning methods, the high spatial diversity of water bodies may lead to highly boundary-dependent features that may eventually limit the generalization power and performance of the uncovered features. To relieve these effects, different multiresolution schemes have been proposed in the literature. For example, Wang et al. [39] present a multiscale CNN for extracting urban water from Landsat imagery. Zhang et al. [26] also define a multiresolution encoder–decoder network, which is intended to characterize water pixels regardless the considered terrain conditions. Additionally, Pu et al. [23] propose a four-layer CNN with a hierarchical structure to accurately estimate nonoptically active parameters when classifying water quality levels. Following a similar scheme, Rezaee et al. [24] develop a two-level CNN for complex wetland classification from Rapideye images. Unlike these classification models that work at pixel level, other works also try to exploit different scene-based segmentation schemes to uncover water. In [40], the authors recommend using the correlations among multiresolution scales to refine the uncovered features. Xia et al. [41] take advantage of an U-shaped segmentation network (U-Net) to allow skip connections between different resolution levels. Zhang et al. [42] adopt an squeeze-and-excitation technique for the recalibration of feature channels when segmenting water. Nonetheless, these segmentation models have the disadvantage of requiring full-scene annotated data in contrast to pixel-based water classification, which become more suitable to relieve the data scarcity problem in developing countries like Nepal.

In all water extraction models, it was observed that initial layers tend to extract low-level features, like edges, whereas deeper layers are focused on higher level features, like spatial-spectral patterns and textures. In this scenario, one may think that deeper features are expected to provide more generalization capabilities for water extraction since the deeper the network the higher the abstraction level. Nevertheless, this is not always the case due to the so-called vanishing gradient problem [43]. When it comes to CNN-based methods, many of the most successful water classification networks, e.g., [23], [24], [25], need to control the number of layers and filters for avoiding a poor gradient propagation, and hence, a rapid performance saturation. Under these circumstances, the use of a reduced number of layers may certainly constrain the abstraction capabilities when characterizing water bodies, especially when dealing with rich spatial-spectral data like in the Sentinel-2 case. Although some mechanisms, such as residual [44] or dense models [45], have also been presented in the standard computer vision field to allow additional layers, how to effectively implement and exploit deeper features for outperforming state-of-the-art water classification models with remote sensing data is still an open-ended issue. Similarly, with the operational exploitation of the most recent CNN-based attention mechanisms to dynamically *pay attention* to the most relevant features [46]. Beyond existing pixel-wise water classification networks, this article pursues to design a novel CNN classification architecture especially designed for water body extraction from Sentinel-2 data by jointly exploiting the following three aspects: dense connections [45], residual learning [44], and attention [47]. Section IV will provide all the corresponding details.

III. Study Area and Dataset

The study area comprises 18 districts of the Terai region located in the southern plains of Nepal. Specifically, it occupies about 28 402.98 km^2 within 26.42° to 29.07° North latitudes and 80.47° to 87.01° East longitudes in the WGS 1984 coordinate system. The Terai is considered as the greenbelt of Nepal being covered with grasslands, tropical monsoon forests, savanna, clay, and loam soil. In terms of biodiversity, Terai is also home to 35 species of mammals, 111 of birds, 46 of herpetos, and 106 of fishes [48]. With the 55.7% of its agricultural land within an altitude range from 60 to 300 m, Terai is known as the *rice bowl* or *agricultural production house* of the country [48], [49]. Moreover, nearly a 47% of Nepal population inhabit in Terai with an increasing population density of around 350 people per km^2 [50]. Certainly, all these factors make regional water resources a major concern for the global development of the country as well as the sustainability of its agricultural sector. In this sense, Terai contains many seasonal and annual rivers mostly originated from the Siwalik hills on the northern side of the region. Besides, Terai features 163 wetlands and 4 Ramsar sites [51] that also make the automatic and remote detection of water a particularly relevant task. Fig. 1 shows the study area of this research and the corresponding Sentinel-2 tiles. Focusing on this region of interest, we build a water body extraction database (WaterPAL), made of Sentinel-2 images, DEM data, and ground-truth water information, as detailed in the following sections. The WaterPAL collection will be accessible on https://github.com/rufernan/ADCNN.

A. Sentinel-2 Images

Sentinel-2 images [52] contain 13 spectral bands with three different spatial resolutions of 10, 20, and 60 m. Blue (B), green (G), red (R), and near-infrared (NIR) bands are provided at 10 m, while the four vegetation red-edge bands and the two short wave infrared (SWIR) bands are provided at 20 m. The remaining channels, i.e., coastal aerosol, water vapor, and cirrus (SWIR), are provided at a 60-m resolution. Table I summarizes the list of bands acquired by the multi-spectral instrument carried by Sentinel-2.

Considering this data nature, a total of 11 cloud-free Level-2 A Sentinel-2 products from 2020 were downloaded to cover

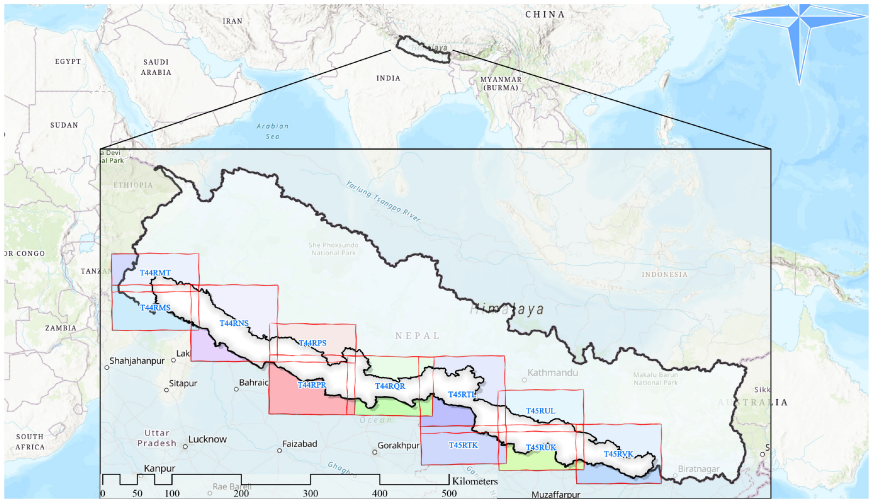


Fig. 1. Study area covering the southern region of Nepal. As it is possible to see, the considered area includes a total of 11 different Sentinel-2 tiles.

TABLE I
SENTINEL-2 BAND DESCRIPTION

Bands	Spectrum	Resolution (m)
B01	Coastal Aerosol	60
B02	Blue	10
B03	Green	10
B04	Red	10
B05	Vegetation Red Edge	20
B06	Vegetation Red Edge	20
B07	Vegetation Red Edge	20
B08	NIR	10
B8A	Vegetation Red Edge	20
B09	Water vapor	60
B10	SWIR - Cirrus	60
B11	SWIR	20
B12	SWIR	20

the whole study area. For such task, we essentially used the Copernicus Open Access Hub[1] (COAH) platform. Initially, a preliminary tile inspection regarding the amount of clouds, cirrus, number of bands, spatial coverage, etc., were performed to avoid any issue in the selected products. However, some data problems (such as, missing regions or bands) were found in some of the cloud-free tiles retrieved by COAH for the year 2020. Hence, T45RUK, T45RUL, T45RTL, and T45RVK tiles were alternatively obtained from the United States Geological Survey Earth Explorer data portal[2] (USGS-EE) to complete our dataset. For these scenes, Level-1 C data were downloaded from USGS-EE and converted into Level-2 A images by applying the Dark Object Subtraction atmospheric correction available in the Quantum Geographic Information System Desktop 3.14.15. Finally, all the downloaded products were processed by the Sentinel Application Platform to generate uniform data cubes at 10 m using a bicubic resampling kernel. Besides, B01 and B10 bands were excluded since they are only useful for atmospheric correction purposes. It is important to note that CNN-based models require spatially homogeneous input data, thus, in this work, we used the standard bicubic interpolation for up-sampling the lower resolution Sentinel-2 bands to their best resolution bands (10 m).

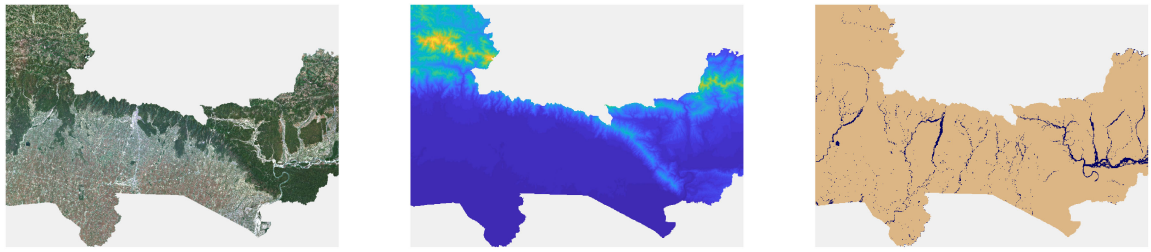
B. DEM Data

DEM raster data covering the study area were downloaded from the United Nations Office for Humanitarian Affairs Services (UN-OCHA) at the following website.[3] Specifically, these data were provided by the NASA Shuttle Radar Topographic Mission, being last updated on November 10, 2019. In more details, the downloaded data are represented at 90-m spatial resolution with geographic latitude/longitude coordinates. Since the downloaded Sentinel-2 images use the Universal Transverse Mercator projection system (with Zones 44 N and 45 N), DEM data were accordingly projected, resampled via a bicubic kernel to 10 m and converted to unsigned 16-bit integers in order to be integrated as an additional band in the corresponding Sentinel-2 products. For all these steps, we made use of ArcGIS Pro 2.6 software with its default settings.

C. Ground-Truth Information

For obtaining ground-truth information about the water bodies within the region of interest, we downloaded the River dataset from UN-OCHA website.[4] In this case, this dataset was last updated on November 24, 2015, containing different water body types in vector format. For the sake of simplicity, we binarized the available labels to *water* and *non-water* classes. Besides, we also reprojected, rasterized, and clipped the resulting data to generate a ground-truth water map for each Sentinel-2 tile.

[1][Online]. Available: https://scihub.copernicus.eu/
[2][Online]. Available: https://earthexplorer.usgs.gov/
[3][Online]. Available: https://data.humdata.org/dataset/Nepal-digital-model-elevation-dem/
[4][Online]. Available: https://data.humdata.org/dataset/Nepal-watercourses-rivers/



(a) (b) (c)

Fig. 2. Sample data tile with (a) Sentinel-2, (b) DEM, and (c) label information. Note that Sentinel-2 product is visualized in RGB, DEM data ranges from the minimum global value (blue) to the maximum global global value (yellow), and water labels are displayed in blue.

TABLE II
TOTAL NUMBER OF PATCHES EXTRACTED FROM WATERPAL

Class	Training	Validation	Test
Non-water	173 663	43 416	144 719
Water	86 831	21 708	72 360
Total	260 494	65 124	217 079

Likewise in the case of DEM data, we employed ArcGIS Pro 2.6 for all these steps. Fig. 2 shows a sample data product corresponding to the T44RQR Sentinel-2 tile. Additionally, Table II summarizes the considered number of training, validation, and test patch samples per category .

IV. METHODOLOGY

This section presents the proposed model for extracting water bodies from Sentinel-2 data. First, let us formulate the water extraction problem from a classification perspective. Let $\mathcal{I} = \{\mathbf{I}_1, \ldots, \mathbf{I}_N\}$ be a collection of Sentinel-2 images (with the possibility of including DEM data as an additional band) covering a particular region of interest with a spatial-spectral size of $(I_1 \times I_2 \times B)$. Let $\mathcal{W} = \{\mathbf{W}_1, \ldots, \mathbf{W}_N\}$ be their corresponding ground-truth water classification maps considering C classes. In this scenario, it is possible to extract M nonoverlapping patches from $\mathcal{I}$ (using a $(P \times P)$ spatial size) in order to build the following set: $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$, where $\mathbf{x}_i \in \mathbb{R}^{(P \times P \times B)} \forall i \in [1, M]$. Considering that each patch is used for representing its central pixel, i.e., $(\lfloor P/2 \rfloor, \lfloor P/2 \rfloor)$ spatial position, it is also possible to extract a label set $\mathcal{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_M\}$ with the class labels of the central pixels as one-hot-encoding vectors. Under this notation, the proposed AD-CNN architecture pursues to approximate a function $\mathcal{F} : \mathcal{X} \rightarrow \mathcal{Y}$, which essentially takes Sentinel-2 patches as input and classifies their central pixels as output. In this sense, the AD-CNN tries to relieve some limitations of current CNN-based water classification models by means of jointly exploiting two different elements: residual attention and dense connections. Now, let us describe these two components as well as the proposed network topology in details.

A. Residual Attention Module

Certainly, both residual and attentional learning paradigms have shown to be two excellent mechanisms for CNNs since they allow focusing on the most discriminating features along the learning process. On the one hand, residual blocks (RBs) [44] are able to provide better feature representations at deeper layers by using skip connections that allow the model shortcut some convolutions when convenient. In this way, over-fitting and vanishing gradient problems can be relieved since unnecessary layers may be skipped while gradients more easily restored. On the other hand, attention [47] is another important tool for allowing the network to dynamically *pay attention* to the most relevant feature maps and regions with respect to the desired output. Hence, an attention block can emphasize or suppress features with the objective of refining intermediate data representations.

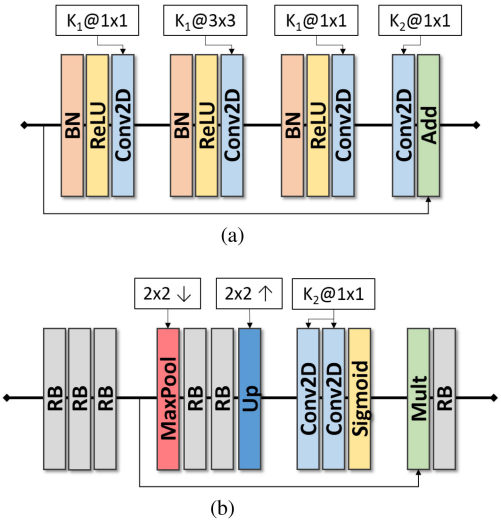


Fig. 3. Considered RB and RAM. (a) RB. (b) RAM.

Despite their potential, these two mechanisms have not yet been used in the context of extracting water bodies from RS data, e.g., [23], [24], [25]. In this scenario, the proposed approach takes advantage of residual and attentional paradigms to define an RAM especially designed to extract water features from Sentinel-2 data. Specifically, RAM is made of several RBs, which consist of batch normalization layers (BN), rectified linear activation functions (ReLU), 2-D convolutional layers (Conv2D), and residual addition layers (Add). Fig. 3(a) shows a graphical visualization of the considered residual building block. As it is possible to observe, three of the Conv2D layers use a (1×1) kernel size, whereas the other one employs (3×3)

kernels. Additionally, we set K_2 to the spectral size of the block input and $K_1 = K_2/4$ in order to compress/decompress the number of feature maps within each RB. The objective of this diabolo-shape consists in simplifying the spectral information coming from Sentinel-2 to better identify water signatures, which are typically more prominent in the visible spectrum where Sentinel-2 has only a limited number of bands.

Using our RB as basic building unit and inspired by the ideas presented in [47], we further define RAM based on additional max-pooling layers (MaxPool), up-sampling layers (Up), sigmoid activation functions (Sigmoid), and residual multiplication layers (Mult). Fig. 3(b) displays the defined RAM. In particular, MaxPool applies a maximum pooling operation with a (2×2) window size and Up does a $2\times$ up-scaling using a nearest neighbor filter. The three first RB units pursue to extract a fundamental deep representation of the input data. Then, the four following elements work for simplifying the spatial information at a higher abstraction level by down-scaling/up-scaling the corresponding feature maps. In this way, coarser texture patterns can be uncovered to better identify water pixels, which usually have rather homogeneous neighborhoods. Finally, the last RB is intended to remove some possible spectral noise that could appear after weighting the feature maps and could be rather prejudicial for water detection, given the limited spectral resolution of Sentinel-2 in visible wavelengths.

B. Dense Module

In general, increasing the number of convolutional layers in a network allows extracting higher level features that can help to achieve a better visual understanding [53]. However, standard feed-forward CNNs have two important limitations in this regard: vanishing gradients and receptive field expansion. On the one hand, the use of back-propagation requires computing the derivatives of the cost function to update the network parameters. Since the parameters of each layer logically depend on the former ones, the chain rule is used for unrolling these gradient computations. In this scenario, the deeper the network the higher the number of nested derivatives, and hence, the higher the chances of canceling the propagated gradients and network updates. On the other hand, standard CNNs process the input data layer by layer. In this way, the selected kernel sizes determine the spatial neighborhoods (or receptive fields) involved in each convolution, becoming the considered area of the input image logically bigger as more convolutional layers are sequentially stacked. As a result, very deep CNNs could also produce a degradation of the uncovered features due to an excessive increase of receptive fields.

In order to overcome these limitations when extracting water bodies from Sentinel-2 data, we design a dense convolutional module (DM) by taking advantage of the connectivity scheme presented in [45]. Specifically, our DM is made of multiple sequential blocks with the following layers: ReLU, Conv2D, and concatenation layer (Concat). Fig. 4 visualizes the defined DM. In more details, DM contains a total of D convolutional blocks with K_3 (3×3) kernels each. With this configuration, we densely propagate feature maps from shallow to deep layers in order to generate more consistent gradient computations during training while providing context information to deeper layers. In this manner, each input Sentinel-2 patch can be characterized by multiple receptive fields in order to improve its context information for a better prediction of water bodies.

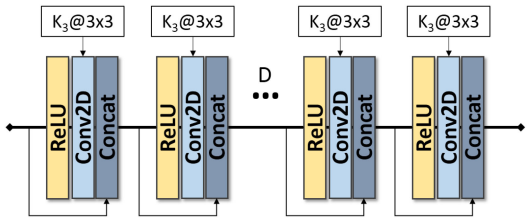


Fig. 4. Considered dense module (DM) with D convolutional blocks in total.

C. Proposed Attentional Dense CNN

In contrast to many of the existing CNN-based water extraction methods, e.g., [23], [24], [25], the proposed architecture takes advantage of the designed modules to focus on the most distinctive features of water while allowing very deep data representations. In general, it is easy to see that water bodies have particular spatial-spectral features that play a fundamental role in their recognition. From an spectral perspective, water molecules usually have spectral responses more focused on the visible and near-infrared spectrum [54]. Precisely, this is the point that classical water indices try to exploit. However, fixing the bands for such computations can often be a too rigid strategy under heterogeneous in-land scenarios, where different spectral mixtures may be expected. By contrast, existing CNN-based models take into account the whole spectral input that may eventually introduce too much noise for detecting purer water. From an spatial perspective, a similar reasoning can also be done since water bodies tend to have specific rounded and smooth shapes and textures, being other spatial information not so useful. In this sense, the proposed architecture adopts the developed RAM module to automatically pay more attention to those initial spatial-spectral features that can be more relevant to identify water pixels, but without neglecting any other input information. Moreover, the proposed network also integrates within its topology the defined DM for effectively uncovering very deep features that are able to gather multiple receptive fields that may help to decide whether a pixel is water or not at different abstraction levels.

With all these considerations in mind, we define the proposed architecture according to Fig. 5. Specifically, the AD-CNN is made of the following components: head block (HB), RAM, DM, transition block (TB), DM, TB, DM, and end block (EB). As it is possible to see in Fig. 6, HB is made of only two layers: BN and Conv2D with K_3 (3×3) kernels. Besides, TB has a total of three layers: Conv2D with K_3 (1×1) kernels, average pooling (AvgPool) with a (2×2) window, and BN. Finally, EB contains: ReLU, global average pooling (GAvgPool), a dense layer (Dense) with C units and a softmax activation function (Softmax). Let us describe the rationale behind the selected

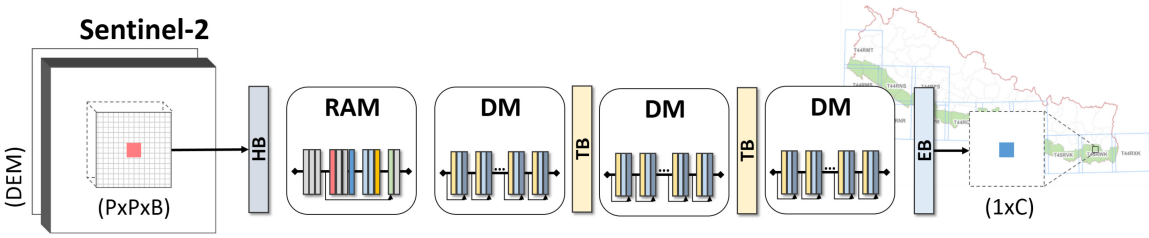


Fig. 5. Proposed AD-CNN based on different building blocks to extract water pixels from Sentinel-2 and DEM data.

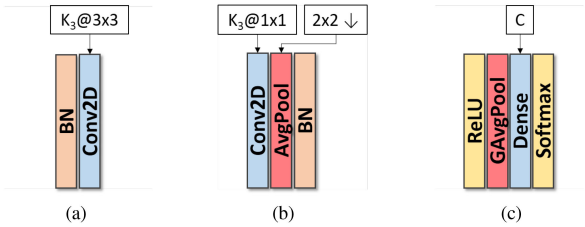


Fig. 6. Considered head (HB), transition (TB), and end block (EB). (a) HB. (b) TB. (c) EB.

components in more details. Initially, HB (1) processes the input data (i.e., a Sentinel-2 image patch with the possibility of including DEM information) to generate an initial low-level characterization of the normalized input. Then, these representations are passed through RAM (2) in order to generate a weighted version of the data, in both spatial and spectral dimensions, according to the objective task. After this process, the most relevant features for identifying water pixels are emphasized to drive the following higher level steps. Subsequently, three DMs separated by two intermediate TBs are used for extracting very deep features. In this case, transition blocks are used to reduce the data complexity since a large increase on the number of feature maps is generated within each DM. Additionally, TB also works for progressively reducing the spatial size by means of an average pooling operation. Once obtained the corresponding deep features, the objective of EB consists in projecting them to the final label space to decide whether the central pixel of the input patch is water. To further prevent overfitting, a global average pooling operation is used to summarize each feature map into a single scalar before the final fully connected classification layer.

V. Experiments

A. Experimental Settings

In order to validate the proposed architecture in the task of extracting water bodies from Sentinel-2 images, we conduct multiple classification experiments with the dataset described in Section III. For comparison purposes, we consider some of the most popular methods used for water extraction, including classical index-based and more recent CNN-based models: NDWI [32], NDVI [36], NDWI-NDVI [37], EWI [21], water quality classification CNN (WQC-CNN) [23], complex wetland classification CNN (CWC-CNN) [24] and self-adaptive pooling CNN (SAP-CNN) [25]. To complete the experimental comparison, we also test the performance of other CNNs that have not been explicitly used for water extraction but they have some connections to this work: basic CNN (base-CNN) [55], dense CNN (DenseNet) [45], and residual attention CNN (AttResNet) [47]. To validate the effectiveness of the proposed architecture, we also conduct an ablation study to compare the AD-CNN with a simplified version (named as D-CNN) that omits the RAM module. In this way, the improvements generated by the proposed dense architecture and attention mechanism can be fairly isolated. It is important to note that all CNN-based models (logically including the proposed approach) take as input an spatial-spectral patch, whereas water indices only require the spectral information of the central pixel to perform the corresponding classification.

Regarding the considered data, we selected two Sentinel-2 tiles (from the 11 tiles available in our dataset) for an external qualitative evaluation. From the remaining ones (nine tiles), we extracted their patches (i.e., $\mathcal{X}$) and labels (i.e., $\mathcal{Y}$) for training and testing the models considering different patch sizes $P = \{8, 12, 16, 20\}$. Specifically, the 60% of the data were used for training (with a 20% of it for validation) and the other 40% for testing. Since water/non-water classes may logically became highly imbalanced in inland scenarios like Nepal, we further balanced the data by means of random sampling to keep the ratio between majority (non-water) and minority (water) classes as $(2:1)$. Under this settings, we carry out the following experiments to study the performance of index-based and CNN-based methods as well as the contribution of Sentinel-2 and DEM data in the water extraction task.

1) *Experiment 1:* Index-based models using Sentinel-2 data as input. Note that index-based methods cannot be used with DEM data, hence, they are tested in isolation in this experiment.
2) *Experiment 2:* CNN-based models using only Sentinel-2 RGB channels as input data, i.e., $B = 3$.
3) *Experiment 3:* CNN-based models using all Sentinel-2 channels, i.e., $B = 11$ (note that two bands are removed by the atmospheric correction).
4) *Experiment 4:* CNN-based models using Sentinel-2 together with DEM data as input, i.e., $B = 12$ (DEM data are integrated as an additional input band).

With respect to the hyperparameters of the proposed architecture, we set $K_1 = 4$, $K_2 = 16$, $K_3 = 16$, $D = 12$, and $C = 2$ according to the information provided in Section IV. In the case of the considered competitors, we logically used the settings described in their corresponding articles. For training all CNN-based models, we made use of the standard cross-entropy

TABLE III
EXPERIMENT 1: QUANTITATIVE RESULTS FOR THE CONSIDERED INDEX-BASED METHODS IN TERMS OF THE OVERALL CLASSIFICATION ACCURACY (%) AND CLASS RECALL (%)

Methods	Patch size											
	$P = 8$			$P = 12$			$P = 16$			$P = 20$		
	Overall Accuracy	Recall		Overall Accuracy	Recall		Overall Accuracy	Recall		Overall Accuracy	Recall	
		water	no-water		water	no-water		water	no-water		water	no-water
NDWI [32]	**74**	51	97	**74**	51	97	**75**	52	97	**74**	51	97
NDVI [36]	**74**	52	97	**75**	52	97	**75**	52	97	**75**	52	97
NDWI-NDVI [37]	73	49	98	73	49	98	74	50	98	73	49	98

TABLE IV
EXPERIMENT 2: QUANTITATIVE RESULTS FOR CNN-BASED METHODS WITH SENTINEL-2 RGB DATA IN TERMS OF THE OVERALL CLASSIFICATION ACCURACY (%) AND CLASS RECALL (%)

Methods	Patch size											
	$P = 8$			$P = 12$			$P = 16$			$P = 20$		
	Overall Accuracy	Recall		Overall Accuracy	Recall		Overall Accuracy	Recall		Overall Accuracy	Recall	
		water	no-water		water	no-water		water	no-water		water	no-water
base-CNN [55]	80.83	64	89	83.45	68	91	83.98	71	91	84.80	73	90
WQC-CNN [23]	81.98	68	90	84.58	74	90	84.90	74	90	84.93	74	90
CWC-CNN [24]	N/A	N/A	N/A	81.90	66	90	82.91	68	90	83.20	71	89
SAP-CNN [25]	81.86	68	89	84.08	74	89	84.92	75	90	85.63	77	90
DenseNet [45]	82.72	71	87	84.77	75	89	85.41	75	89	85.95	78	91
AttResNet [47]	79.84	66	87	N/A	N/A	N/A	82.29	71	88	N/A	N/A	N/A
D-CNN (ours)	**83.01**	72	88	**85.11**	76	89	**85.99**	77	90	**86.22**	78	90
AD-CNN (ours)	**83.52**	73	89	**86.05**	78	90	**87.01**	80	90	**87.83**	81	91

loss with the ADAM optimizer using the following parameters: 100 epochs, $1e^{-3}$ learning rate, and 128 batch size. Additionally, we also applied a learning rate decay (0.2 factor) on each validation loss plateau after 15 epochs. All the experiments were performed on a server with an Intel(R) Core (TM) i7-6850 K processor, 64 GB of DDR4 RAM, and an NVIDIA GeForce GTX 1080 Ti. Besides, Ubuntu 20.04 $\times$64, CUDA 10.1, TensorFlow 2.1.0, Keras 2.3.1, and Python 3.6 were used as software environment. The codes of this article will be accessible on https://github.com/rufernan/ADCNN.

B. Results

Tables III–VI present the quantitative evaluation obtained for the considered experiments (i.e., Experiments 1–4, respectively). In more details, Table III contains the results of index-based methods, whereas Tables IV–VI provide the quantitative assessment of CNN-based models when considering different combinations of the input data (i.e., only Sentinel-2 RGB bands, all Sentinel-2 bands, and Sentinel-2 bands together with DEM data). As it is possible to see, all the tables are organized with the tested methods in rows and the considered patch sizes and metrics in columns. In this regard, two different quantitative classification metrics are considered: overall accuracy (%) and class recall (%). For the sake of clarity in the visualization of the tables, all recall values are rounded to integer figures. Besides, the two best accuracy values for each patch size are highlighted in bold font, being the best result displayed with gray background. Note that we use the label *N/A* to highlight that the corresponding result is not available, whether the model is unable to converge or run with the considered patch size. For conducting a qualitative evaluation of the methods, Fig. 8 also shows some of the classification maps obtained over the external tiles when focusing on the two first experiments with $P = 16$.

C. Discussion

1) Experiment 1: According to the results reported in Table III, the use of NDWI, NDVI, and NDVI_NDWI achieved a maximum overall accuracy of 75% with recall values of 52% for water and 98% for no-water classes. In general, it was found that the performance of all the considered indices were approximately similar to each other across all patch sizes. In more details, NDVI slightly edged the rest indices in terms of overall accuracy and recall metrics. Besides, NDWI was found to have exactly the same performance as NDVI at patch size 16, which reveals the affinity of both indices for the study area. Finally, the highest recall values were obtained by NDVI_NDWI for no-water classes. In contrast to the other experiments, the considered traditional indices certainly obtained the worse general performance.

2) Experiment 2: As Table IV shows, the proposed model (AD-CNN) consistently achieved the best performance through all the considered patch sizes when using Sentinel-2 RGB bands as input. In general, it is possible to observe that the larger the patch the higher the accuracy since, logically, more local information is available for consideration. In this sense, it is also important to note that the CWC-CNN was not able to converge when considering RGB bands with a small patch size of 8. Besides, AttResNet was only able to manage multiples of two as patch size due to the down-sampling operations performed inside this architecture. Regarding the other competitors, the WQC-CNN was always found to perform better than the SAP-CNN, and DenseNet was able to obtain the third best general

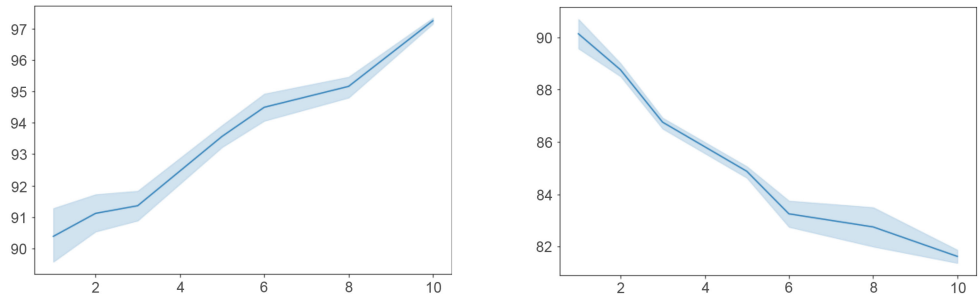


Fig. 7. (a) Average overall accuracy and (b) water recall obtained by the proposed approach when considering different class balance factors. Note that the vertical axes display the metrics in percentage, whereas the horizontal axes show the balance factors between majority (non-water) and minority (water) classes.

TABLE V
EXPERIMENT 3: QUANTITATIVE RESULTS FOR CNN-BASED METHODS WITH ALL SENTINEL-2 BANDS IN TERMS OF THE OVERALL CLASSIFICATION ACCURACY (%) AND CLASS RECALL (%)

Methods	Patch size											
	$P = 8$			$P = 12$			$P = 16$			$P = 20$		
	Overall Accuracy	Recall		Overall Accuracy	Recall		Overall Accuracy	Recall		Overall Accuracy	Recall	
		water	no-water		water	no-water		water	no-water		water	no-water
base-CNN [55]	87.14	78	92	87.84	79	92	88.47	81	92	89.52	84	92
WQC-CNN [23]	87.81	81	92	88.70	82	92	89.30	84	92	89.35	84	92
CWC-CNN [24]	85.73	72	93	86.88	78	91	87.71	79	92	87.72	81	91
SAP-CNN [25]	87.48	80	91	88.19	81	92	89.44	84	92	89.09	83	92
DenseNet [45]	88.16	81	91	89.03	83	92	89.53	84	92	89.60	84	92
AttResNet [47]	86.23	77	91	N/A	N/A	N/A	87.14	81	90	N/A	N/A	N/A
D-CNN (ours)	**88.55**	80	90	**89.65**	84	91	**89.91**	84	92	**89.98**	85	92
AD-CNN (ours)	**89.48**	82	91	**90.21**	85	92	**90.88**	86	92	**91.52**	87	93

TABLE VI
EXPERIMENT 4: QUANTITATIVE RESULTS FOR CNN-BASED METHODS WITH SENTINEL-2 AND DEM DATA IN TERMS OF THE OVERALL CLASSIFICATION ACCURACY (%) AND CLASS RECALL (%)

Methods	Patch size											
	$P = 8$			$P = 12$			$P = 16$			$P = 20$		
	Overall Accuracy	Recall		Overall Accuracy	Recall		Overall Accuracy	Recall		Overall Accuracy	Recall	
		water	no-water		water	no-water		water	no-water		water	no-water
base-CNN [55]	87.61	79	92	88.40	80	92	88.65	82	92	89.37	83	92
WQC-CNN [23]	88.42	82	92	89.26	83	92	89.34	84	93	90.23	85	93
CWC-CNN [24]	86.52	77	91	87.75	80	92	88.01	81	92	88.51	82	92
SAP-CNN [25]	88.23	81	92	89.05	83	92	89.19	83	93	89.83	85	92
DenseNet [45]	88.88	82	92	89.58	84	92	89.73	85	92	90.41	86	92
AttResNet [47]	86.90	80	91	N/A	N/A	N/A	87.67	81	92	N/A	N/A	N/A
D-CNN (ours)	**89.01**	82	92	**89.78**	83	91	**90.05**	85	92	**90.66**	86	92
AD-CNN (ours)	**89.66**	83	92	**90.22**	85	92	**90.96**	87	92	**91.34**	87	93

performance for all the considered patch sizes. In comparison to the remaining experiments, the use of RGB channels yielded the poorest results for all CNN-based models.

3) Experiment 3: In the case of Table V, it is possible to observe how all the networks were able to increase the performance around a 5% with respect to the previous experiment. Similarly, the AD-CNN provided the best results over all path sizes, being the highest overall accuracy 91.52% with a recall of 87% for water and 93% for no-water classes. Again, CWC-CNN was found to achieve the worse performances, followed by AttResNet and SAP-CNN, respectively. Besides, DenseNet and WQC-CNN obtained the most competitive results, after the proposed model ones. In general, this experiment revealed a significant performance improvement when using the complete spectral information provided by Sentinel-2 for the more accurate characterization of water.

4) Experiment 4: In Table VI, it was found that the integration of DEM data as an additional input band was only able to improve the results around a 1%. This evidence indicates that full Sentinel-2 spectral information is certainly more important than DEM data to uncover water bodies over the region of interest. Overall, the proposed model achieved again the best performance with metrics ranging from 89.66% to 91.34% of accuracy and 83 to 87 of recall for water, when

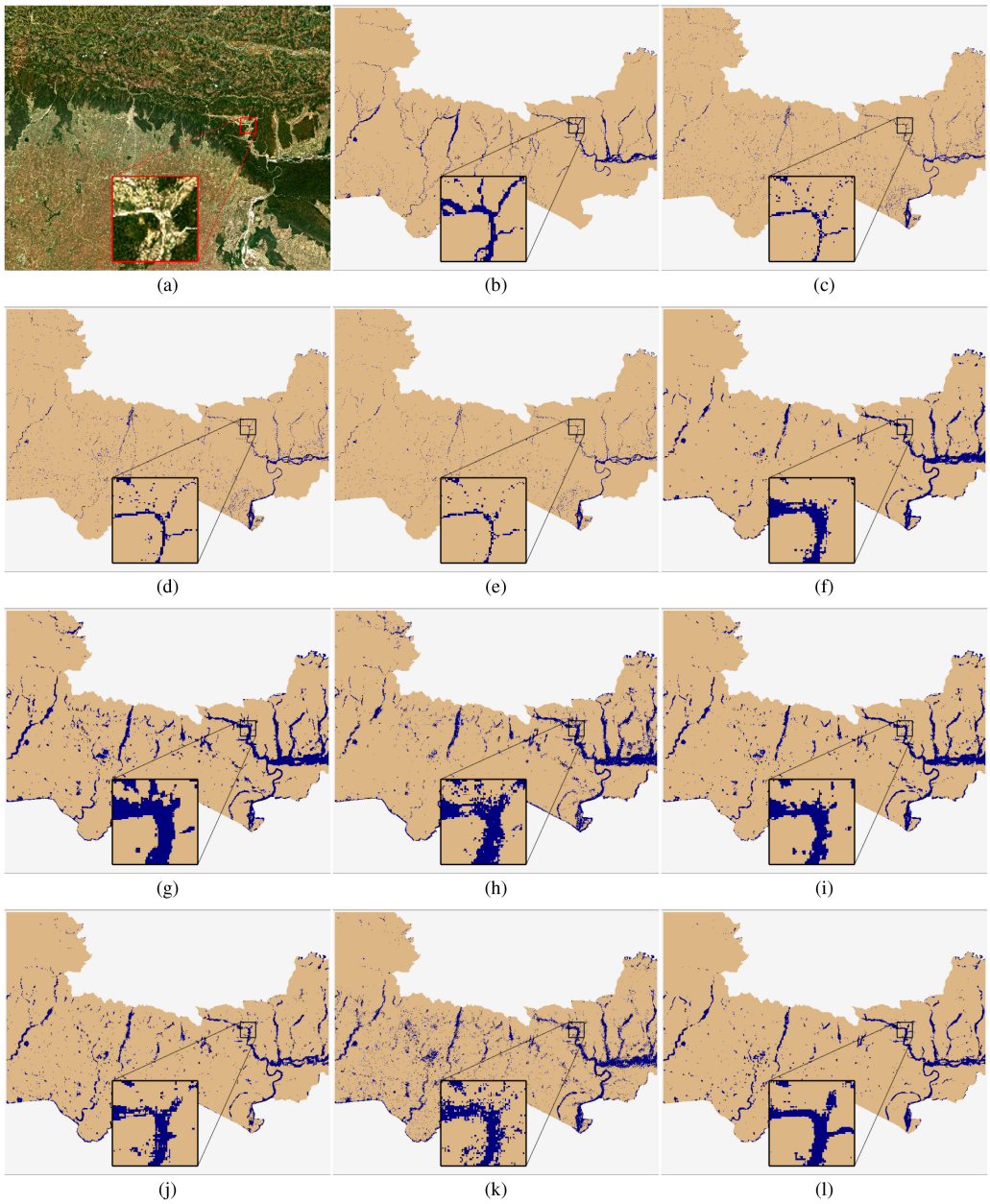


Fig. 8. Qualitative results for the T44RQR Sentinel-2 tile with $P = 16$. (a) RGB image. (b) Water ground truth. (c) NDWI. (d) NDVI. (e) NDWI-NDVI. (f) base-CNN. (g) WQC-CNN. (h) CWC-CNN. (i) SAP-CNN. (j) DenseNet. (k) AttResNet. (l) AD-CNN (ours). Note that water bodies are colored in blue, non-water in brown, and pixels outside the study area in white color.

increasing the patch size from 8 to 20. As in all the conducted experiments, the CWC-CNN was the least performing method. Moreover, DenseNet was followed by WQC-CNN and SAPCNN on the quantitative evaluation. This time the general trend showed a moderate performance increase with respect to the previous experiment due to the inclusion of DEM data.

Taking into account that the study area has an important class imbalance, Fig. 7 further analyzes the proposed approach when considering different balance factors between the majority class (non-water) and minority class (water). As it is possible to observe, the overall accuracy over the test set increases with the balance factor. However, the decreasing water recall reveals that these improvements are based on the underestimation of the minority class since less water pixels are successfully retrieved. In this way, the considered balance factor (i.e., $2:1$) shows a reasonable tradeoff between accuracy and water recall. Additionally, Table VII also provides the average number of trainable parameters and computational time (in seconds) per training/test epoch for each one of the considered CNN-based methods. As shown, the proposed approach figures are comparable to the ones of the best performing competitor (DenseNet).

From the qualitative results displayed in Fig. 8, several important observations can be made to support the conducted analysis. Regarding index-based methods, all the three considered indices (i.e., NDWI, NDVI, and NDVI_NDWI) produced similar output results where water bodies become rather underestimated. As it

TABLE VII
AVERAGE NUMBER OF PARAMETERS AND COMPUTATIONAL TIME (S) PER TRAINING/TEST EPOCH

Methods	Patch size											
	$P = 8$			$P = 12$			$P = 16$			$P = 20$		
	Num. Parameters	Time per epoch (s)		Num. Parameters	Time per epoch (s)		Num. Parameters	Time per epoch (s)		Num. Parameters	Time per epoch (s)	
		Training	Test		Training	Test		Training	Test		Training	Test
Base-CNN	284 781	12	16	366 701	13	16	481 389	13	17	628 845	14	13
WQC-CNN	7 001 986	19	19	12 244 866	31	18	19 584 898	47	20	29 022 082	69	69
CWC-CNN	5 443 410	17	21	5 443 410	18	21	5 443 410	18	22	5 443 410	19	19
SAP-CNN	1 618 007	13	16	3 715 159	13	16	6 860 887	17	16	11 055 191	23	23
DenseNet	999 810	111	91	999 810	113	97	999 810	114	96	999 810	116	103
AttResNet	33 023 543	153	158	N/A	N/A	N/A	33 023 543	240	185	N/A	N/A	N/A
AD-CNN (ours)	1 004 050	109	95	1 004 050	113	94	1 004 050	112	95	1 004 050	116	96

is possible to see, NDVI [see Fig. 8(d)] tended to obtain a slightly better estimation but, in general, many water bodies were still missed with respect to the ground-truth data [see Fig. 8(b)]. In this sense, the particularly low water recall values reported in Table III also support these observations where index-based methods tend to essentially detect the most pure water spectral signatures. When inspecting the visual results of the six considered deep learning-based competitors [see Fig. 8(f)–(k)], we can observe a different trend. Overall, all the networks seemed to extract not only pure water bodies but also other neighboring regions such as river banks or partially dried streams. In details, it was found that DenseNet [see Fig. 8(j)] was able to extract more true water pixels than the other competitors, being CWC-CNN [see Fig. 8(h)] and AttResNet [see Fig. 8(k)] the worst performing networks due to the amount of noise in their corresponding estimations. In the case of the proposed AC-CNN approach [see Fig. 8(l)], less output noise and more accurate water shapes were certainly obtained, making its estimations the most accurate results.

Although all the tested deep learning-based methods lean to overestimate water with respect to index-based ones, the need of using index thresholds according to the spectral properties of water often makes traditional indices fail to extract water bodies beyond pure water pixels. In contrast, CNNs take advantage of context information for characterizing water pixels with richer spatial-spectral features while providing a more general solution to water body extraction. Nonetheless, many of the existing water classification networks are only able to satisfactorily perform using a fix abstraction level given by a relatively small number of convolutional layers, which may eventually saturate their learning performances. Note that, when working with study areas as heterogeneous as the considered one, many types of water bodies with different complexities are naturally expected. Besides, the reasonably good spatial-spectral resolution of Sentinel-2 data is also a plus for the need of exploring deeper features. Hence, it becomes desirable to simultaneously learn from lower to higher water feature abstraction levels in order to solve from the simplest to the most challenging cases. Logically, feature abstraction and neighborhood information are important factors for identifying a pixel as water but the optimal abstraction level and amount of context may certainly vary from patch to patch. Precisely, the proposed AC-CNN model exploits this idea by implementing attentional dense connectivities that allow transferring multiple characterization levels while focusing on the most relevant features for water extraction.

VI. CONCLUSION

This article presented a new CNN classification architecture (termed AD-CNN) especially designed for water body extraction from Sentinel-2 data. Unlike other models in the remote sensing literature, the AD-CNN adopted a novel attentional dense scheme that pursues to effectively exploit deeper convolutional features for the better identification of water pixels. On the one hand, dense connections were implemented to allow extracting deeper features while characterizing multiple data complexities at once. On the other hand, a new RAM was designed to dynamically put the focus on the most relevant spatial-spectral features for classifying water pixels. In order to test the proposed model performance, a new water database of Nepal (WaterPAL) was built. The experiments, conducted on WaterPAL, revealed the competitive results achieved by the AD-CNN with respect to several traditional index-based and state-of-the-art CNN-based water extraction models.

According to the obtained results, several important conclusions can be made with regard to the use of Sentinel-2 data and the performances of the tested models. First, the most effective data configuration for water body extraction has shown to be the complete Sentinel-2 spectra together with DEM data. However, it is also important to highlight that the contribution of DEM is rather small with respect to multispectral Sentinel-2 information. Second, traditional index-based methods are generally unable to provide satisfactory results under heterogeneous large-scale scenarios since only pure water signatures are mainly detected. Third, deep learning-based methods provide more competitive results although they also tend to be more prone to overestimate water. Fourth, the proposed attentional dense scheme allows extracting deeper and more complete features for a more accurate estimation of water bodies. Although the outcomes of this work are promising, there is still room for future improvements based on extending the proposed network to different intersensor platforms, multimodal data, and multitemporal stages.

REFERENCES

[1] M. Weiss, F. Jacob, and G. Duveiller, "Remote sensing for agricultural applications: A meta-review," *Remote Sens. Environ.*, vol. 236, 2020, Art. no. 111402.
[2] G. Huang, Z. Shen, and R. Mardin, "Overview of urban planning and water-related disaster management," in *Proc. Urban Planning Water-Related Disaster Manage.*, 2019, pp. 1–10.
[3] M. Li and Z. Ma, "Soil moisture drought detection and multi-temporal variability across China," *Sci. China Earth Sci.*, vol. 58, no. 10, pp. 1798–1813, 2015.

[4] R. Fernandez-Beltran, F. Pla, and A. Plaza, "Endmember extraction from hyperspectral imagery based on probabilistic tensor moments," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 12, pp. 2120–2124, Dec. 2020.

[5] D. E. Garrick et al., "Valuing water for sustainable development," *Science*, vol. 358, no. 6366, pp. 1003–1005, 2017.

[6] J. Kang, Z. Wang, R. Zhu, X. Sun, R. Fernandez-Beltran, and A. Plaza, "PiCoCo: Pixelwise contrast and consistency learning for semisupervised building footprint segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10548–10559, 2021, doi: 10.1109/JSTARS.2021.3119286.

[7] L. Andres, K. Boateng, C. Borja-Vega, and E. Thomas, "A review of in-situ and remote sensing technologies to monitor water and sanitation interventions," *Water*, vol. 10, no. 6, 2018, Art. no. 756.

[8] C. Huang, Y. Chen, S. Zhang, and J. Wu, "Detecting, extracting, and monitoring surface water from space using optical sensors: A review," *Rev. Geophys.*, vol. 56, no. 2, pp. 333–360, 2018.

[9] C. Justice et al., "An overview of modis land data processing and product status," *Remote Sens. Environ.*, vol. 83, no. 1–2, pp. 3–15, 2002.

[10] T. R. Loveland and J. L. Dwyer, "Landsat: Building a strong future," *Remote Sens. Environ.*, vol. 122, pp. 22–29, 2012.

[11] A. S. Belward and J. O. Skøien, "Who launched what, when and why; trends in global land-cover observation capacity from civilian earth observation satellites," *ISPRS J. Photogrammetry Remote Sens.*, vol. 103, pp. 115–128, 2015.

[12] K. Toming, T. Kutser, A. Laas, M. Sepp, B. Paavel, and T. Nõges, "First experiences in mapping lake water quality parameters with Sentinel-2 MSI imagery," *Remote Sens.*, vol. 8, no. 8, 2016, Art. no. 640.

[13] Y. Du, Y. Zhang, F. Ling, Q. Wang, W. Li, and X. Li, "Water bodies' mapping from Sentinel-2 imagery with modified normalized difference water index at 10-m spatial resolution produced by sharpening the SWIR band," *Remote Sens.*, vol. 8, no. 4, 2016, Art. no. 354.

[14] X. Yang, S. Zhao, X. Qin, N. Zhao, and L. Liang, "Mapping of urban surface water bodies from Sentinel-2 MSI imagery at 10 m resolution via NDWI-based image sharpening," *Remote Sens.*, vol. 9, no. 6, 2017, Art. no. 596.

[15] Z. Wang, J. Liu, J. Li, and D. D. Zhang, "Multi-spectral water index (MuWI): A native 10-m multi-spectral water index for accurate water mapping on Sentinel-2," *Remote Sens.*, vol. 10, no. 10, 2018, Art. no. 1643.

[16] L. Dan, W. Baosheng, C. Bowei, X. Yuan, and Z. Yi, "Review of water body information extraction based on satellite remote sensing," *J. Tsinghua Univ. (Sci. Technol.)*, vol. 60, no. 2, pp. 147–161, 2020.

[17] Y. Zhou et al., "Open surface water mapping algorithms: A comparison of water-related spectral indices and sensors," *Water*, vol. 9, no. 4, p. 256, 2017.

[18] G. Kaplan and U. Avdan, "Object-based water body extraction model using Sentinel-2 satellite imagery," *Eur. J. Remote Sens.*, vol. 50, no. 1, pp. 137–143, 2017.

[19] D. Li et al., "Open-surface river extraction based on Sentinel-2 MSI imagery and DEM data: Case study of the upper yellow river," *Remote Sens.*, vol. 12, no. 17, 2020, Art. no. 2737.

[20] D. Li, G. Wang, C. Qin, and B. Wu, "River extraction under bankfull discharge conditions based on Sentinel-2 imagery and DEM data," *Remote Sens.*, vol. 13, no. 14, 2021, Art. no. 2650.

[21] J. Yang and X. Du, "An enhanced water index in extracting water bodies from Landsat TM imagery," *Ann. GIS*, vol. 23, no. 3, pp. 141–148, 2017.

[22] F. Isikdogan, A. C. Bovik, and P. Passalacqua, "Surface water mapping by deep learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 11, pp. 4909–4918, Nov. 2017.

[23] F. Pu, C. Ding, Z. Chao, Y. Yu, and X. Xu, "Water-quality classification of inland lakes using Landsat8 images by convolutional neural networks," *Remote Sens.*, vol. 11, no. 14, 2019, Art. no. 1674.

[24] M. Rezaee, M. Mahdianpari, Y. Zhang, and B. Salehi, "Deep convolutional neural network for complex wetland classification using optical remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 9, pp. 3030–3039, Sep. 2018.

[25] Y. Chen, R. Fan, X. Yang, J. Wang, and A. Latif, "Extraction of urban water bodies from high-resolution remote-sensing imagery using deep learning," *Water*, vol. 10, no. 5, 2018, Art. no. 585.

[26] P. Zhang, L. Chen, Z. Li, J. Xing, X. Xing, and Z. Yuan, "Automatic extraction of water and shadow from SAR images based on a multi-resolution dense encoder and decoder network," *Sensors*, vol. 19, no. 16, 2019, Art. no. 3576.

[27] H. Guo, G. He, W. Jiang, R. Yin, L. Yan, and W. Leng, "A multi-scale water extraction convolutional neural network (MWEN) method for GaoFen-1 remote sensing images," *ISPRS Int. J. Geo- Inf.*, vol. 9, no. 4, 2020, Art. no. 189.

[28] A. Veit, M. J. Wilber, and S. Belongie, "Residual networks behave like ensembles of relatively shallow networks," *Adv. Neural Inf. Process. Syst.*, vol. 29, pp. 550–558, 2016.

[29] T. D. Acharya, A. Subedi, and D. H. Lee, "Evaluation of water indices for surface water extraction in a Landsat 8 scene of Nepal," *Sensors*, vol. 18, no. 8, 2018, Art. no. 2580.

[30] Y. Li, X. Gong, Z. Guo, K. Xu, D. Hu, and H. Zhou, "An index and approach for water extraction using Landsat–OLI data," *Int. J. Remote Sens.*, vol. 37, no. 16, pp. 3611–3635, 2016.

[31] T. D. Acharya, A. Subedi, H. Huang, and D. H. Lee, "Application of water indices in surface water change detection using Landsat imagery in Nepal," *Sens. Mater*, vol. 31, pp. 1429–1447, 2019.

[32] B.-C. Gao, "NDWI—a normalized difference water index for remote sensing of vegetation liquid water from space," *Remote Sens. Environ.*, vol. 58, no. 3, pp. 257–266, 1996.

[33] L. Ji, L. Zhang, and B. Wylie, "Analysis of dynamic thresholds for the normalized difference water index," *Photogrammetric Eng. Remote Sens.*, vol. 75, no. 11, pp. 1307–1317, 2009.

[34] H. Xu, "Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery," *Int. J. Remote Sens.*, vol. 27, no. 14, pp. 3025–3033, 2006.

[35] D. Feng, "A new method for fast information extraction of water bodies using remotely sensed data," *Remote Sens. Technol. Appl.*, vol. 24, no. 2, pp. 167–171, 2012.

[36] C. Aguilar, J. C. Zinnert, M. J. Polo, and D. R. Young, "NDVI as an indicator for changes in water availability to woody vegetation," *Ecological Indicators*, vol. 23, pp. 290–300, 2012.

[37] K. Rokni, A. Ahmad, A. Selamat, and S. Hazini, "Water feature extraction and change detection using multitemporal Landsat imagery," *Remote Sen.*, vol. 6, no. 5, pp. 4173–4189, 2014.

[38] L. Yang, S. Tian, L. Yu, F. Ye, J. Qian, and Y. Qian, "Deep learning for extracting water body from Landsat imagery," *Int. J. Innov. Comput. Inf. Control*, vol. 11, pp. 1913–1929, 2015.

[39] Y. Wang, Z. Li, C. Zeng, G.-S. Xia, and H. Shen, "Extracting urban water by combining deep learning and Google Earth Engine," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 13, pp. 768–781, 2020.

[40] Z. Zhang, M. Lu, S. Ji, H. Yu, and C. Nie, "Rich CNN features for water-body segmentation from very high resolution aerial and satellite imagery," *Remote Sens.*, vol. 13, no. 10, 2021, Art. no. 1912.

[41] M. Xia, Y. Cui, Y. Zhang, Y. Xu, J. Liu, and Y. Xu, "DAU-Net: A novel water areas segmentation structure for remote sensing image," *Int. J. Remote Sens.*, vol. 42, no. 7, pp. 2594–2621, 2021.

[42] X. Zhang, J. Li, and Z. Hua, "MRSE-Net: Multi-scale residuals and se-attention network for water body segmentation from satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 5049–5064, 2022, doi: 10.1109/JSTARS.2022.3185245.

[43] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," *Adv. Neural Inf. Process. Syst.*, vol. 28, pp. 2377–2385, 2015.

[44] 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.

[45] G. Huang, Z. Liu, G. Pleiss, L. Van Der Maaten, and K. Weinberger, "Convolutional networks with dense connectivity," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2019.2918284.

[46] R. Hang, Z. Li, Q. Liu, P. Ghamisi, and S. S. Bhattacharyya, "Hyperspectral image classification with attention-aided CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2281–2293, Mar. 2020.

[47] F. Wang et al., "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, 2017, pp. 3156–3164.

[48] J. M. Mendez, Ed, "Nepal Biodiversity Resource Book. International Centre for Integrated Mountain Development (ICIMOD), Ministry of Environment, Science and Technology (MOEST), Government of Nepal (GoN)," 2007. [Online]. Available: https://lib.icimod.org/record/7560

[49] S. Thapa, L. Schipper, M. Shrestha, S. Nepal, and A. Taylor, "Final report: Nepal study ecosystems, development, and climate adaptation improving the knowledge base for planning, policy and management technical team," *Stockholm Environ. Institute, Tech. Rep.*, Mar. 2011.

[50] The DHS Program, 2010. [Online]. Available: https://dhsprogram.com/pubs/pdf/FR78/01Chapter01.pdf

[51] M. Siwakoti and J. B. Karki, "Conservation status of Ramsar sites of Nepal Tarai: An overview," *Botanica Orientalis, J. Plant Sci.*, vol. 6, pp. 76–84, Mar. 2010.

[52] M. Drusch et al., "Sentinel-2: ESA's optical high-resolution mission for GMES operational services," *Remote Sens. Environ.*, vol. 120, pp. 25–36, 2012.

[53] M. Liu, J. Shi, Z. Li, C. Li, J. Zhu, and S. Liu, "Towards better analysis of deep convolutional neural networks," *IEEE Trans. Visual. Comput. Graph.*, vol. 23, no. 1, pp. 91–100, 2016.
[54] Z. Kovacs et al., "Water spectral pattern as holistic marker for water quality monitoring," *Talanta*, vol. 147, pp. 598–608, 2016.
[55] M. Paoletti, J. Haut, J. Plaza, and A. Plaza, "Deep learning classifiers for hyperspectral imaging: A review," *ISPRS J. Photogramm. Remote Sens.*, vol. 158, pp. 279–317, 2019.

Janak Parajuli received the B.E. degree in geomatics engineering from Kathmandu University, Dhulikhel, Nepal, in 2014, and the M.Sc. degree in geospatial technologies from the University Jaume I, Castellón de la Plana, Spain, under Erasmus Mundus Scholarship scheme, in 2021.

He is currently serving as a Survey Officer under the Government of Nepal. His research interests include remote sensing, environment and sustainability, computer vision, and deep learning.

Ruben Fernandez-Beltran (Senior Member, IEEE) received the B.Sc. degree in computer science, the M.Sc. degree in intelligent systems, and the Ph.D. degree in computer science from the University Jaume I, Castellón de la Plana, Spain, in 2007, 2011, and 2016, respectively.

He is currently an Assistant Professor with the Department of Computer Science and Systems, University of Murcia, Murcia, Spain. He has been a visiting Researcher with the University of Bristol, U.K.; University of Cáceres, Spain; Technische Universität Berlin, Germany; and Autonomous University of Mexico State, Mexico. His research interests include multimedia retrieval, spatio-spectral image analysis, pattern recognition techniques applied to image processing, and remote sensing.

Dr. Fernandez-Beltran was the recipient of the Outstanding Ph.D. Dissertation Award at Universitat Jaume I in 2017.

Jian Kang (Member, IEEE) received the B.S. and M.E. degrees in electronic engineering from the Harbin Institute of Technology, Harbin, China, in 2013 and 2015, respectively, and the Dr.-Ing. degree in synthetic aperture radar interferometry from Signal Processing in Earth Observation (SiPEO), Technical University of Munich, Munich, Germany, in 2019 .

In August 2018, he was a Guest Researcher with the Institute of Computer Graphics and Vision (ICG), TU Graz, Graz, Austria. From 2019 to 2020, he was with the Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin (TU Berlin), Berlin, Germany. He is currently with the School of Electronic and Information Engineering, Soochow University, Suzhou, China. His research interests include signal processing and machine learning techniques, and their applications in remote sensing. In particular, he is interested in intelligent SAR/InSAR data processing, and deep learning-based techniques for remote sensing image analysis.

Dr. Kang was the recipient of the first place of the Best Student Paper Award in European Conference on Synthetic Aperture Radar 2018, Aachen, Germany. His joint work was selected as one of the ten Student Paper Competition Finalists in IEEE International Geoscience and Remote Sensing Symposium 2020. He served as a Guest Editor for IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS and REMOTE SENSING.

Filiberto Pla received the B.Sc. and Ph.D. degrees in physics from the Universitat de València, Valencia, Spain, in 1989 and 1993, respectively.

He is currently a Full Professor with the Departament de Llenguatges i Sistemes Informàtics, University Jaume I, Castellón, Spain. He has been a Visiting Scientist with Silsoe Research Institute, the University of Surrey; the University of Bristol, U.K.; CEMAGREF, France; the University of Genoa, Italy; Instituto Superior Técnico of Lisbon in Portugal; the Swiss Federal Institute of Technology ETH-Zurich; the idiap Research Institute in Switzerland; the Technical University of Delft, Netherlands; and the Mid Sweden University in Sweden. He has been the Director with the Institute of New Imaging Technologies, University Jaume I. His current research interests include color and spectral image analysis, visual motion analysis, 3-D image visualization and pattern recognition, and machine learning techniques applied to image processing.

Dr. Pla is a Member of the Spanish Association for Pattern Recognition and Image Analysis, which is a Member of the International Association for Pattern Recognition.