DAT-CNN: Dual Attention Temporal CNN for Time-Resolving Sentinel-3 Vegetation Indices

Damian Ibañez, Ruben Fernandez-Beltran, *Senior Member, IEEE*, Filiberto Pla, and Naoto Yokoya, *Senior Member, IEEE*

***Abstract*—The synergies between Sentinel-3 (S3) and the forthcoming fluorescence explorer (FLEX) mission bring us the opportunity of using S3 vegetation indices (VI) as proxies of the solar-induced chlorophyll fluorescence (SIF) that will be captured by FLEX. However, the highly dynamic nature of SIF demands a very temporally accurate monitoring of S3 VIs to become reliable proxies. In this scenario, this article proposes a novel temporal reconstruction convolutional neural network (CNN), named dual attention temporal CNN (DAT-CNN), which has been specially designed for time-resolving S3 VIs using S2 and S3 multitemporal observations. In contrast to other existing techniques, DAT-CNN implements two different branches for processing and fusing S2 and S3 multimodal data, while further exploiting intersensor synergies. Besides, DAT-CNN also incorporates a new spatial–spectral and temporal attention module to suppress uninformative spatial–spectral features, while focusing on the most relevant temporal stamps for each particular prediction. The experimental comparison, including several temporal reconstruction methods and multiple operational Sentinel data products, demonstrates the competitive advantages of the proposed model with respect to the state of the art. The codes of this article will be available at https://github.com/ibanezfd/DATCNN.**

***Index Terms*—Biophysical products, fluorescence explorer (FLEX), Sentinel-2 (S2), Sentinel-3 (S3), temporal resolution.**

I. Introduction

NOWADAYS, remote sensing (RS) data play a pivotal role in many important application fields, such as, land-cover mapping [1], [2], environmental management [3], object recognition [4], and Earth composition analysis [5] among others. In response, multiple space missions have been developed over the past years to effectively satisfy the increasing demand of RS data [6]. The Copernicus programme is one of the main projects at global level to address this demand. Managed by the European Commission in partnership with the European Space Agency (ESA), it includes several Sentinel missions to cover different spatial–spectral requirements and needs.

Inside Copernicus, there are two satellite constellations that share important synergies since both are mainly focused on multispectral imagery: Sentinel-2 (S2) [7] and Sentinel-3 (S3) [8]. On the one hand, S2 includes two satellites (S2A and S2B) that carry the multispectral instrument (MSI). In more details, MSI is able to capture 13 spectral bands (B01-B12, B8a) within the 443–2190-nm wavelength range, using a spatial resolution from 10 to 60 m. On the other hand, S3 also comprises a couple of satellites (S3A and S3B) that are equipped with the ocean and land color instrument (OLCI). In this case, OLCI is able to acquire 21 spectral bands (Oa01–Oa21) in the 390–1040-nm wavelength region, using a fix spatial resolution of 300 m. Under these settings, S2 and S3 missions are both able to provide operational data products related to vegetation, land and water, but logically with different spatial–spectral characteristics [9].

In the context of terrestrial vegetation, ESA is also developing the fluorescence explorer (FLEX) mission [10] that will launch a satellite in 2024 to work with S3 in a tandem configuration. In particular, FLEX aims at quantifying the solar-induced chlorophyll fluorescence (SIF) as an accurate measure of the vegetation photosynthetic activity [11]. To achieve this goal, FLEX will carry the fluorescence imaging spectrometer (FLORIS) which has an ultrafine spectral resolution within the 500–780-nm wavelength range and a spatial resolution of 300 m. Since FLEX will be capturing images just few seconds before one of the S3 satellites, OLCI will certainly support FLORIS for enhancing its sensing capabilities, while providing additional value in the monitoring of the vegetation status [12]. However, FLORIS has a particularly narrow field of view in contrast to OLCI. As a result, FLEX will have a lower temporal resolution of two weeks. This lower temporal resolution, together with the existing synergies between FLORIS/OLCI instruments, strongly motivate the use of S3 vegetation indices (VI) as SIF proxies for the FLEX mission [13].

Even though S3 has a relatively good temporal resolution (four days for a single-satellite and two days for the twin-satellite constellation), the highly dynamic nature of SIF requires a very accurate monitoring of S3 VIs to really serve as reliable proxies from a temporal perspective. Note that even small temporal periods may produce important vegetation changes. Then, the availability of continuous and consistent intersensor data becomes an essential issue for the early detection of changes in photosynthetic pigments across sensors [14]. In this scenario, the

Manuscript received November 19, 2021; revised February 1, 2022 and March 10, 2022; accepted March 12, 2022. Date of publication March 22, 2022; date of current version April 7, 2022. This work was supported by Ministerio de Ciencia, Innovación y Universidades under RTI2018-098651-B-C54. *(Corresponding author: Ruben Fernandez-Beltran.)*

Damian Ibañez and Filiberto Pla are with the Institute of New Imaging Technologies, University Jaume I, E-12071 Castellón de la Plana, Spain (e-mail: ibanezd@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@uji.es).

Naoto Yokoya is with the Graduate School of Frontier Sciences, The University of Tokyo, Chiba 277-8561, Japan, and also with the RIKEN Center for Advanced Intelligence Project, Tokyo 103-0027, Japan (e-mail: yokoya@k.u-tokyo.ac.jp).

Digital Object Identifier 10.1109/JSTARS.2022.3161190

open availability of Copernicus data bring us the opportunity of using intersensor S2 observations for improving S3 VI temporal resolution by means of temporal reconstruction methods [15].

In general, temporal reconstruction pursues to estimate data at unavailable temporal stamps in order to time-resolve or recover missing information. Consequently, it is a very useful process in RS, being three main trends available in the literature [15]: 1) temporal replacement, 2) temporal filtering, and 3) temporal learning models. In the case of temporal replacement, the missing information is directly (or indirectly) filled with the temporally closest available data. For this reason, this approach is mainly suited to recovering just small image regions, e.g., cloud removal [16]. Regarding temporal filtering, these methods are based on the assumption that time-series data tend to display regular fluctuations over time. In this way, it is possible to define a filtering function to interpolate missing values over a particular temporal window. For instance, it is the case of Zhao *et al.* [17] who designed a three-point changing-weight filter to reconstruct NDVI time series by considering the local maximum/minimum values for a particular growth cycle. Despite their simplicity and efficiency, both replacement and filtering approaches generally demand highly constrained scenarios, where homogeneous landscapes, temporal stability, and partially missing data are expected. In contrast, temporal learning methods provide higher generalization capabilities by means of machine learning regression models, being more suitable for time-resolving RS products. For example, Zeng *et al.* [18] presented a linear regression method for the temporal reconstruction of moderate resolution imaging spectroradiometer (MODIS) data. Zhao *et al.* [17] used a random forest for temporally modeling land surface temperature. In [19], Zhang *et al.* proposed a convolutional neural network (CNN) to reconstruct missing MODIS and Landsat-7 data. Besides, Shao *et al.* [20] also defined a generative adversarial CNN for reconstructing multisource RS data. To further exploit temporal dynamics, alternative deep learning architectures have also been proposed. For instance, Yu *et al.* [21] developed a deep recurrent neural network (DRNN) for the reconstruction of long-term MODIS data. In [22], the authors also presented a temporal CNN-based regression network for time-resolving Sentinel products.

Certainly, the latest advances on deep learning-based temporal reconstruction provide the most remarkable improvements for time-resolving VIs [21], [22]. Nonetheless, the limited operational availability of Sentinel data, together with the short lifetime of FLEX (3.5 years), may hinder the task of modeling long-term temporal dependencies that affect the performance and applicability of time-resolved S3 VIs as SIF proxies. In this sense, there are two main issues within FLEX/Sentinel context: 1) discontinuous temporal data and 2) multimodal observations. On the one hand, the operational availability of the data are logically affected by satellite orbits, cloud occlusions, data degradation, and many other factors. Therefore, the same temporal volume size may cover different temporal gaps (depending on the operational availability of the data), which can eventually make convolutional kernels unable to learn consistent temporal feature patterns over time. On the other hand, the standard scheme for time-resolving VIs typically involves a single input sensor to learn a one-to-one projection. However, the availability of multispectral data coming from different Sentinel missions is a missed opportunity for exploiting unexplored intersensor synergies. To address these challenges, this article presents a novel temporal reconstruction CNN named dual attention temporal CNN (DAT-CNN). DAT-CNN has been specially designed for time-resolving S3 VIs using S2 and S3 multitemporal observations. Fig. 1 displays a conceptualization of the proposed model objective, where intersensor S2 data together with the corresponding S3 neighboring temporal window are used to estimate S3 VIs at unavailable timestamps. In more details, DAT-CNN makes use of 4-D kernels, which slide across height, width, channel, and time dimensions, with the objective of finding temporal correlations while extracting dynamic spatial–spectral and temporal features. Additionally, DAT-CNN incorporates two different branches for processing and fusing S2 MSI and S3 OLCI modalities with the target of effectively exploiting multisensor input data. Besides, each branch also implements a newly defined spatial–spectral and temporal attention module. The developed attention modules have the objective of suppressing uninformative spatial–spectral features, while focusing on the most relevant temporal stamps for each particular prediction. Note that temporal attention is specially important in the considered discontinuous temporal data context since it allows focusing on the most informative deep features along inconsistent multimodal time intervals in order to reach a better data reconstruction. All these techniques make the proposed DAT-CNN model offer an innovative perspective on the intersensor vegetation estimation task, particularly to improve the reconstruction of vegetation indices as SIF proxies through multimodal and multitemporal data. The main contributions of this work as follows.

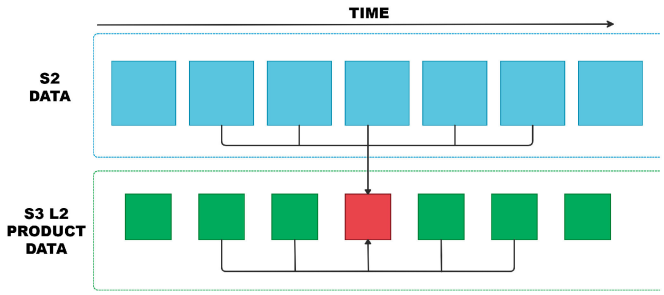

Fig. 1. Diagram of the proposed model objective. Showing S2 data in blue, S3 L2 product data in green and nonavailable S3 L2 product data in red.

1) A new deep learning architecture (DAT-CNN) is proposed for time-resolving S3 VIs using S2 and S3 multitemporal data.
2) An extended spatial–spectral and temporal attention module is defined for effectively exploiting multisensor input data.
3) The performance of several state-of-the-art temporal reconstruction methods is analyzed when resolving S3 VIs as SIF proxies.
4) The superior performance of the proposed model is demonstrated via an extensive experimentation using operational S2 and S3 data.

The rest of this article is organized as follows. Section II details some related works including their main features and limitations. Section III details the study area for the experimentation and the dataset generation process. Section IV defines the proposed architecture as well as the considered attention modules. Section V presents the experimental part of the work. Finally, Section VI concludes this article.

II. Related Work

This work is focused on filling temporal gaps in S3 VIs using multimodal S2 data and S3 information. In this context, temporal reconstruction methods take on special importance [15], being learning-based models certainly the most effective paradigms. With the development of deep learning technologies, multiple CNN models have been successfully designed to deal with multispectral RS data, while achieving positive results in land cover classification and many other downstream applications [23]. For instance, Sharma *et al.* [24] proposed a 2D-CNN, which used deep patch features to classify medium resolution Landsat-8 imagery. Analogously, the CNN-based model defined in [25] was adapted to hyperspectral imagery. In other works like [26], the authors used contextual interactions and residual information to exploit spectral–spatial features. Zhang *et al.* also showed in [27] the advantages of using a two-branch CNN architecture for classifying multisource hyperspectral and light detection and ranging (LiDAR) data.

In the literature, it is also possible to find models which are specifically used to obtain biophysical estimations through RS imagery. In [28], Pyo *et al.* were able to quantifying cyanobacteria using hyperspectral data. In [29], Aptoula *et al.* defined a CNN-based regression architecture for effectively estimating chlorophyll-a concentration from S2 data. RS data reconstruction methods have been also explored, even with different temporal methods [15]. In time-resolving problems, traditional regression models can be found as well [18]. Nevertheless, the learning-based methods are able to show further improvements in temporal reconstruction. Zhang *et al.* [19] proposed a CNN model to remove clouds, dead lines or other data anomalies in MODIS or Landsat-7 images through spatial, temporal, or spectral information. In the same line, Shao *et al.* [20] designed a contextual adversarial two-stages CNN to perform spatial reconstruction. Other authors opted for using temporal deep learning models instead. It is the case of Yu *et al.* [21] who defined the DRNN for time-resolving MODIS VIs. In [22], a temporal CNN-based regression network is also presented for the temporal reconstruction of Sentinel data.

Despite the positive results obtained by these and other relevant methods [30], many of the existing deep learning-based temporal reconstruction models still struggle at the task of managing temporal RS data from an operational perspective. In real environments, the operational data availability is not constant since there are many factors to deal with (e.g., orbit subcycles, cloud contamination, data degradation, etc.). Consequently, different temporal gaps may be considered within the same fix temporal volume, which can logically make convolutional kernels not to extract consistent temporal features over time-series data. In this sense, the possibility of extending standard attention mechanisms beyond spatial–spectral domains (e.g., [31] and [32]) certainly becomes an attractive opportunity to relieve this type of intersensor temporal limitations. Additionally, existing temporal prediction methods for VIs (e.g., [21], [22]) do not take into account the possibility of exploiting multimodal input data since they are mainly focused on learning a single domain projection. However, the open availability of multispectral Sentinel data bring us the opportunity of uncovering new intersensor synergies in the task of time-resolving S3 VIs. With these considerations in mind, the proposed model has been designed to further advance the development of time-resolved S3 VIs as SIF proxies.

Fig. 2. Study region of Extremadura, with an S2 RGB example image.

III. Dataset

This section defines the dataset created for the experiments. In Section III-A, the spatial location of the dataset and the motivation behind selecting this specific area of study are exposed. Then, Section III-B describes the process by means of S2 and S3 products were downloaded, corrected, filtered, and selected.

A. Location and Motivation

The created dataset is composed by 20 daily synchronous S2 and S3 products from the complete 2019 a. Specifically, S2 data are MSI bottom of atmosphere (BOA) reflectance products and S3 counterparts are OLCI top of atmosphere (TOA) radiance images. In this work, the normalized difference vegetation index (NDVI) was chosen as target VI, as it is a standardized VI well known in the literature. Nevertheless, any other product could be used instead, while the selected product changes along time as a result of vegetation biophysical processes. For this dataset, a highly covered vegetation area was selected. As one of the most vegetation covered regions in Spain, Extremadura was chosen with a 65% of land cover vegetation, including dehesas (an agrosilvopastoral system consisting of grassland) and oak forests (which also makes the region keep a high biodiversity. Among the natural grasslands of Extremadura, there are protected areas, such as Parque Natural de Cornalvo and Zona de Interés Regional Llanos de Cáceres y Sierra de Fuentes. The study region is located in $(-5.188652, 38.740370)$ latitudes to $(-6.498977, 39.697509)$ longitudes, which cover a total of 100 km^2, while corresponding to a complete S2 tile. Accordingly, S3 products were cropped to match and contain only this 100-km^2 area as well. An illustrative example is shown in Fig. 2.

B. Automated Data Acquisition and Processing

To automate the process of downloading the data from the open access hub Sentinel platform, a region of interest (ROI) over S2 tilling grid, i.e., R_{S2}, is initially defined in GeoJSON format. Once this region is selected, we generate a query using the Sentinel mission's name, the product type, the ROI and the temporal interval T. This process was implemented through the Sentinelsat [33] application programming interface (API) in Python. For the S2 products, the ROI ROI_{S2} is directly used, and the temporal interval T is defined. In the S3 product case, the ROI is defined as any S3 product containing the S2 ROI, $\text{ROI}_{S3} = S3_{\text{products}} \in \text{ROI}_{S2}$. The temporal interval T used is the same as for the S2 products.

When the products are already downloaded, a filtering to avoid cloud coverage is performed using the information from S2 level-2 A (L2A) products. In more details, these cloud masks are generated following S2 ground segment data pipeline through level-1 C (L1C) processing and enhanced L2A scene classification [34], [35]. After filtering the products, corrections are needed for each type of product. In the S2A case, let $X_{S2_{\text{ref}}} \in \mathbb{R}^{(M_{S2}\times W_{S2}\times H_{S2})}_{R_{S2},r_{S2}}$ be a S2 reflectance product covering the region R_{S2} in universal transverse mercator (UTM) coordinates, with M_{S2} spectral bands, W_{S2} width and H_{S2} height with a spatial resolution of r_{S2} meters per pixel. A resampling to set the spatial resolution r_{S2} equal in all the bands is needed, as some bands have different resolutions. To retain only the M_{S2} main 13 spectral bands (from B1 to B12, including B8a) a spectral subset is also performed. For the S3 products further processing is needed. For a S3 radiance L1C product $X_{S3_{\text{rad}}} \in \mathbb{R}^{(M_{S3}\times W_{S3}\times H_{S3})}_{R_{S3},r_{S3}}$ in world geodetic system 1984 (WGS84) coordinates of the region R_{S3}, with an M_{S3}, W_{S3}, H_{S3}, and r_{S2} spectral bands, width, height, and spatial resolution, respectively. First, the radiance values were transformed to reflectance using the sun zenith angle and solar flux information from the auxiliary product information. Then, a reprojection from the WGS84 system to the UTM coordinate system used in the S2 products is done to have both products, S2 and S3, in the same coordinate system. At the same time, the S3 product is trimmed to use only the intersecting area with the ROI_{S2}, and resampled to 300-m per pixel resolution. After these spatial corrections, a spectral subset is performed to maintain the main 21 reflectance bands. The product resulting after these processes is $X_{S3_{\text{ref}}} \in \mathbb{R}^{(M_{S3}\times W_{S3}\times H_{S3})}_{R_{S2},r_{S3}}$. This corrections were done using the Sentinel application platform (SNAP) in Python interface, to access to the SNAP Java API [36].

Once the corrections are applied, the S3 product is chosen and calculated. In order to illustrate the method, NDVI has been selected as target VI since it is used to estimate the quality, quantity, and development of the vegetation.

For the S3 case, before calculating the NDVI, the three red bands ($B07$, $B08$, $B09$, and $B10$) and the three near-infrared radiation (NIR) bands ($B16$, $B17$, and $B18$) are joined to one red band and one NIR band by averaging them, shown in (1) following the methodology from [37]:

$$\text{meanRED} = \frac{B07 + B08 + B09 + B10}{4} \tag{1}$$

$$\text{meanNIR} = \frac{B16 + B17 + B18}{3}. \tag{2}$$

The obtained averages are used to calculate the final NDVI value, represented in (3). Thus, the S3 level-2 (L2) NDVI product can be defined as $X_{S3_{\text{NDVI}}} \in \mathbb{R}^{(W_{S3}\times H_{S3})}_{R_{S2},r_{S3}}$

$$X_{S3_{\text{NDVI}}} = \frac{\text{meanNIR} - \text{meanRED}}{\text{meanNIR + meanRED}}. \tag{3}$$

IV. Methodology

In this section, the proposed model to perform the missing S3 data completion using multimodal S2/S3 temporal data is described. The methodology for this process has been divided in three main blocks: 1) An overview including the definition of problem and our proposed model to solve it (Section IV-A); 2) the main features of the proposed model, which allows it to deal with temporal information and improve the generation S3 enchanted products using state-of-the-art techniques and structures (Section IV-B); and 3) the dual attention temporal convolutional neural network model proposed architecture definition (Section IV-C).

A. DAT-CNN Overview

The data employed by the model is defined as follows. Let X_{S2} be an input time series of multispectral S2 products; X_{S3} another input time series, in this case of biophysical S3 products and Y_{S3}, the unavailable desired S3 L2 same biophysical output product. The ultimate goal of the DAT-CNN is to be able to generate a S3 L2 biophysical unavailable product mapping a function, such as $F : X_{S2}, X_{S3} \rightarrow Y_{S3}$. More specifically, for this process the images have been divided into patches. Therefore, each S3 L2 output pixel patch $y_{S3}(i,j) \in \mathbb{R}^{S\times S}$ at position (i,j) of the Y_{S3} target product will correspond to S2 time series image patches of pixels $Q_{S2}(m,i,j,h) \in \mathbb{R}^{(M_{S2}\times P\times P)\times T}$ from X_{S2}, where m is the spectral band, and h is the time stamp; and to a S3 time series image patches S3 $Q_{S3}(i,j,h) \in \mathbb{R}^{(S\times S)\times T}$. In this case, P is the height and width of the S2 neighboring region or image patch corresponding to each S3 L2 S height and width patch, input, and output, following the spatial resolution ratio r_{S2}:r_{S3}, and T the number of time stamps in the time interval considered.

In Fig. 3, the complete structure of the proposed DAT-CNN is shown. The architecture can be divided in three main parts as follows:1) The S2 branch, 2) the S3 branch, and 3) the tail of the network. The use of two different temporal input steams is aimed at allowing the network to learn deep features at each multitemporal modality and, then, exploiting intersensor relationships at a common resolution level embedding scheme. Each of these branches has several blocks composed by different layers. Both of the S2 and S3 branches have an attention block, helping them to focus in the most relevant information from the inconsistent time intervals between the available images. After the attention blocks, the model has head blocks and body blocks for each branch. Finally the S2 and S3 branches are connected to the tail block, which generates the time-resolved VI.

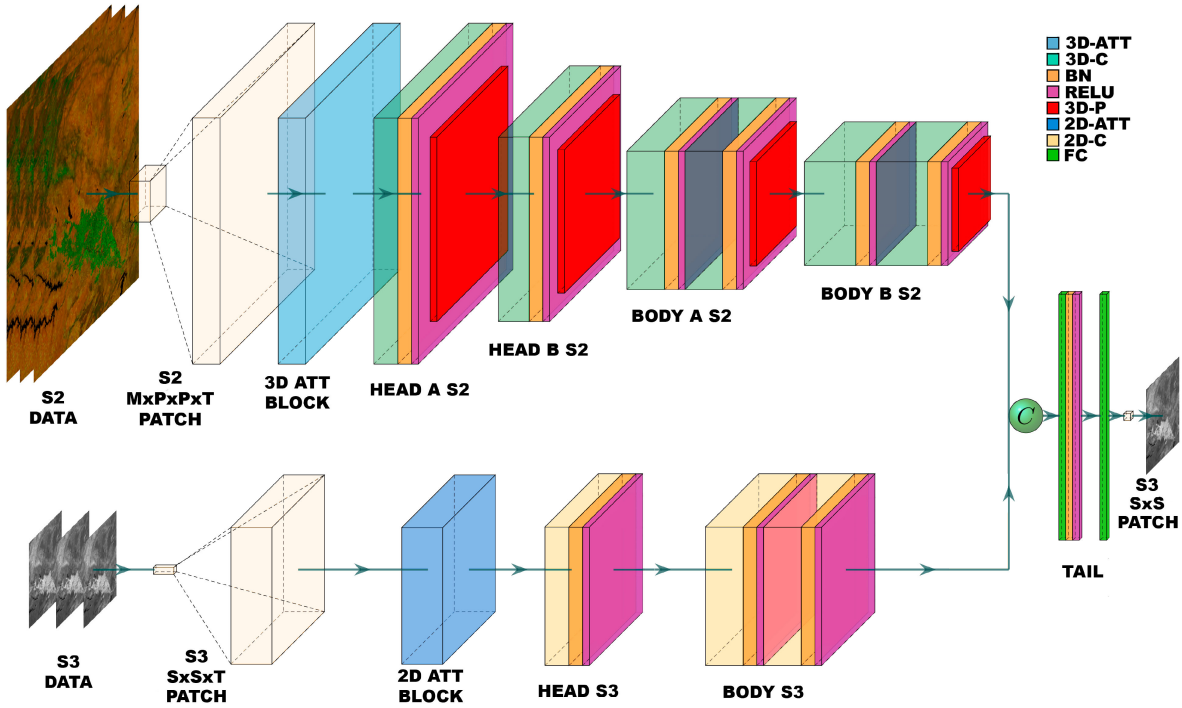

Fig. 3. Complete architecture is made of main branches, one for S2 data and another for S3 data. Each branch has several layers organized by blocks including attention blocks. After the features extraction of each branch, both set of features are concatenated and in a final block called Tail the expected S3 pixel value is generated. In the diagram the legend presents the two different attention blocks (3-D attention and 2-D attention) and the layers composing the model: 3-D convolutional layers (3D-C), batch normalization layers (BN), RELU activation layers (RELU), 3-D pooling layers (3D-P), 2-D convolutional layers (2D-C), and full connected layers (FC).

B. DAT-CNN Main Features

A CNN is a feed-forward neural network which contains convolutional layers and other kinds of filters to extract characteristics and information, specially used in the case of image data. In images, CNN models can locate lines, gradients, circles, or even more complex features. The proposed architecture has three main strengths to predict unavailable S3 L2 biophysical products. First, the use of temporal information through 3-D convolutions (Section IV-B1); second, the use of multisensor information simultaneously through two different branches (Section IV-B3); and third, an improvement of focus and representation in the feature maps applying channel and spatial attention (Section IV-B2)

$$y_k(i,j,h) = \left\{ \sum_{m=0}^{M-1} \sum_{r=0}^{N-1} \sum_{s=0}^{N-1} \sum_{t=0}^{T-1} w_k(r,s,t;m) x_m(i+r, j+s, h+t) \right\} + b_k. \tag{4}$$

1) 3-D Convolutions: The first of the main features of this model is the use of temporal information. In the literature, different temporal related tasks have been explored to solve through 3-D convolutions. For example, Diba *et al.* found the use of temporal information through 3-D convolutions to increase the accuracy for video classification. In order to manage multispectral time series data, the model contains several 3-D convolutions for the S2 data stream. For the S3 data, 3-D convolutions were not used. Instead, as the biophysical product information is concentrated in only one channel, the temporal data will be used as channels in 2-D convolutions. The 3-D convolutions perform not only in the spatial and spectral domain but also in the temporal domain, which in our case is equivalent to the image plane and channels, and the temporal interval. The 3-D convolutional kernels will expand in the fixed spectral domain, while moving through the temporal information. This allows the 3-D convolutions to find temporal correlations between samples, extracting characteristics that change dynamically in the temporal dimension. It has to be considered that vegetation indices usually change in temporal intervals of hours, days, weeks, or even longer. For example, the chlorophyll concentration changes in time periods longer than days, while other vegetation products, such as fluorescence can change drastically in a few hours. Therefore, the estimation of intersensor level-4 (L4) products can be improved with essential information from the use of the progression of the different vegetation indices through time. As the 3-D convolutional filters are allowed to move through the time dimension along with the height and width dimensions, this process creates a 3-D feature map for each spatial–spectral-temporal data input. This convolution is defined in (4). In this equation, $y_k(i,j,h)$ is the output pixel at a time h and spatial position i,j of the final reconstructed biophysical product. T is the time dimension and $N \times N$ is the spatial size of the filter and M is the set of features, in our case the number of spectral bands. In this equation, $w_k(r,s,t;m)$ are the weight values of the kth filter at the image position (r,s), time t, and m spectral band, b_k is the bias for the kth filter, and $x_m(i+r, j+s, h+t)$ is the input, for us S2 data patch

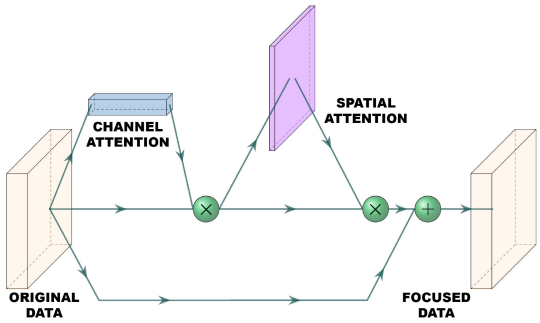

Fig. 4. Diagram of the full process of the attention block, where the original data are multiplied by the channel attention features, the spatial attention features, and finally the obtained features are added as a weight to the original data to improve the representation.

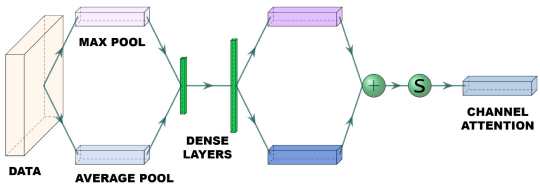

Fig. 5. Diagram of the channel attention features generation through a max and average pool, two dense layer, the addition of the resultant features, and a sigmoid activation function.

of spectral band m at image coordinates $(i + r, j + s)$ at time stamp $(h + t)$.

2) Convolutional Attention: The objective of an attention block is to suppress the unnecessary features and help the network to focus in the most significant ones. The temporal data series of S2 and S3 might be from different periods of time, with a different time interval between the samples because of missing data. Due to this reason, using an attention block to highlight the most relevant temporal features can give a great advantage to avoid outliers and temporal noisy data. This technique has been studied extensively in previous literature [38]–[41]. This block is composed by two main modules, a channel attention module and a spatial attention module, as shown in Fig. 4.

The channel attention (Fig. 5) focuses in obtaining feature maps through the interchannel possible relationships. To achieve it efficiently, the spatial dimensions are squeezed using average-pooling [42] and max-pooling features at the same time to obtain more accurate information. Once both max and average-pooling features are extracted, two shared dense layers with a reduction ratio in the first layer (to avoid overparameterization) and a second dense layer with the same number of neurons as the original number of channels are used to generate the attention channel maps. Later, both feature vectors are merged using an elementwise addition and forwarded to a sigmoid activation. In the case of the spatial attention (Fig. 6), the objective is using the interspatial relationships between the data. Max-pooling and average-pooling of the data are used again to squeeze the information. However, now the obtained features are concatenated and fed to a convolution layer to generate a single spatial feature map.

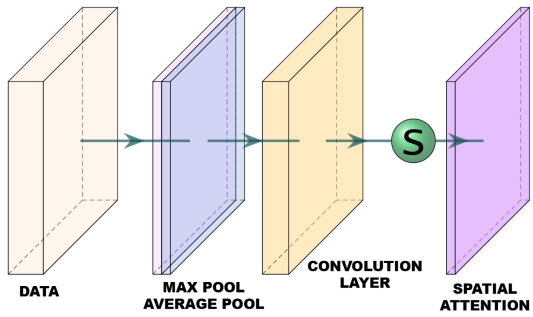

Fig. 6. Diagram of the spatial attention features generation using a max and an average pool, a convolution layer, and a sigmoid activation function.

Even though both modules can be used in a sequential or parallel order (as was shown in [43]), the sequential structure appears to give better results. Specifically, the best composition is using first the channel attention and then the spatial attention. Hence, we first calculate the channel features and multiply them to the original data to after proceed a second multiplication with the resulting values and the spatial features. These attention features are then added to the original data to increase the representation power of the network. In the DAT-CNN, an attention block was used for both S2 and S3 data, with the only difference between them being the use of 3-D pooling and 3-D convolutions for S2, and the respective 2-D layers for S3.

3) Dual Branch: For our model, the input data will be composed by two different data streams: the first one of multispectral and multitemporal image products from the S2 mission, and a second one of S3 multitemporal biophysical products. Therefore, the proposed architecture has been designed to deal with both, a sequence of S2 multispectral images and another sequence of S3 product data in order to generate an estimation of absent or inaccessible S3 L2 products at specific times. Both of the time series include previous and latter products to the objective output time. This is achieved using two different branches initially, which will eventually join the extracted features from both streams. There are examples of the use of multibranch networks using multimodal data in other fields as action recognition [44], person search [45], or even medical image categorization [46].

C. DAT-CNN Architecture

1) S2 Branch: The first block in the S2 branch is the *3-D attention block*, which was described in Section IV-B2. This block has the function to enlighten and help the network to focus in the most relevant information from the time series multispectral input patches. Due to the inconsistency in time intervals between S2 and S3 cloudless coupled data, the ability of the attention block to highlight useful temporal information in irregular periods of time has significant role. Specifically, for the S2 attention block, the number of neurons in the first dense layer of the channel attention is $N = 6$ and in the second layer equal to the number of channel, in this case, $N = 13$. The 3-D convolution performed in the spatial attention has only one filter to maintain the original spatial image size. After the

attention block an encoder structure is followed to reduce the spatial resolution, while augmenting the depth of the features, leading to the S3 biophysical product smaller patch size. This structure is composed of the other two blocks or stages of this branch, which we have called S2 head and S2 body.

The S2 head blocks consist of a sequential composite of layers. 1) A 3-D convolutional layer (3D-C), where the 3-D convolutions $K@N \times N \times T$ with K as the number of filters are done to the input patch $Q_{S2}(m, i, j, h)$ to extract temporal feature maps from the previous stage, 2) a batch normalization layer (BN) to regularize the 3-D feature maps, 3) a RELU layer (RELU) as activation function, and 4) a pooling layer (3D-P) to reduce the dimensions of the generated 3-D feature maps. The aim of the S2 head blocks is to produce initial low level 3-D feature maps from the S2 multispectral data already focused by the attention block. Two head blocks were used, a first block (S2 head A) for the complete spatial resolution features extraction, and a second block (S2 head B) to obtain a more abstract feature representation without losing much spatial resolution. Therefore, in this stage early characteristics from local information in the spectral/spatial–temporal domain with a moderated depth but in full spatio–temporal resolution feature map is obtained.

The 3D-C layers, wherein both head blocks A and B defined with a kernel size of $(3 \times 3 \times 3)$ and $(1 \times 1 \times 1)$ stride, in the S2 head A block using 64 filters and in the S2 head B the double, 128. The number of filters for each layer has been selected after experimenting with deeper and shallower convolutional layers, leading to the selected filter values. The 3D-P used a pool of $(2 \times 2 \times 1)$ and a stride in the first case of $(1 \times 1 \times 1)$ as well to highlight the stronger features without reducing its dimensionality. Meanwhile, in the second head a $(2 \times 2 \times 1)$ stride was used to obtain a higher spatial reduction maintaining the temporal information.

The S2 head blocks are followed by two *S2 body blocks* (body A S2 and body B S2), both of these blocks contain seven layers in sequential order: 1) 3D-C, 2) BN, 3) RELU, 4) 3D-C, 5) BN, 6) RELU, and 7) 3D-P. In the body blocks the previously obtained features are significantly reduced in the space domain to generate further deeper and complex maps to extract the information from the S2 data temporal series correspondent to the desired S3 biophysical product. Reducing the spatial information of the feature maps allows to reduce progressively the spatial resolution until a spatial size of 1×1. To obtain deeper feature temporal maps, the 3D-C layers in each body block are equally defined, but in the subsequent pooling layers the spatial resolution is reduced.

In the body A S2 256 filters were used in both convolutional layers with $(3 \times 3 \times 3)$ size. The 3D-P layer of this block's pool size and stride are of size $(2 \times 2 \times 1)$ as well, in order to reduce the spatial resolution without reducing the temporal deep features. Besides, the body B S2 has 512 filters and the 3D-P was designed with a pool size and stride of $(2 \times 2 \times 3)$. Is in this last part of the second body block, where the deep temporal features are also reduced.

2) S3 Branch: The S3 branch can be seen as a simplification of the S2 branch regarding to its structure, as the corresponding S3 spatial resolution to S2 is always lower. Starting by a *2-D attention block*. In this case the attention block has the purpose of highlighting the most useful samples for the desired output value. Nevertheless, as there is less spatial information, the temporal attention for irregular time intervals between samples has even higher relevance. The 2-D attention block has the same structure as the 3-D attention block with the necessary 2-D counterpart layers, as each of the temporal S3 biophysical product patches contain only one channel each. The number of neurons in the dense layers in this case is of $N = 1$ for the first layer and 4 in the second one. As in the S2 branch, in this case the 2-D convolution has one filter to maintain constant the space dimension.

Once the S3 input is focused by the attention, an *S3 head block* is used with the same purpose as the S2 head blocks, to extract superficial features. The block structure is equal to the S2 but with 2-D layers: 1) a 2-D convolutional layer (2D-C), 2) BN, and 3) RELU. From the first 2D-C after the attention block, the spatial resolution is maintained with all the filter sizes and strides of (1×1). Also, the number of filters is smaller, with only 16 filters in the 2D-C.

The *S3 body block* has the same structure as the S2 branch body blocks with 2-D layers as well: 1) 2D-C, 2) BN, 3) RELU, 4) 2D-C, 5) BN, and 6) RELU. With this block, deeper temporal characteristics are extracted from the S3 data, without reducing its spatial resolution. Again, the filter sizes and stride parameters of every layer are of size (1×1) with a number of filters of 64 and 128 for each 2D-C, layer respectively.

3) Tail: In the final segment of the architecture, which we have called *tail block*, the two branches are concatenated. The final 3-D volume containing S2 and S3 deep temporal features is then flattened. Once a single vector of features is generated, the tail block containing: 1) fully connected layer (FC), 2) BN, and 3) RELU is used. With this set of layers the spatial, temporal, and spectral features previously obtained are correlated to better approximate the desired biophysical S3 product value. The FC layer used has a total of 1024 fully connected neurons for this purpose. Finally, a last FC layer with a single output estimates S3 L2 product image patch.

V. Experiments

In this section, the experimentation conducted to validate and compare the DAT-CNN is described. In Section V-A, the experimental settings used through all the experiments presented are explained. Finally, in Section V-B, Section V-C, and Section V-D the proposed experiments are described and discussed, first an ablation study of the proposed model, second an experiment to compare its results with other state-of-the-art-methods designed to solve similar issues, and third a real case experimentation.

A. Experimental Settings

In both experiments, Section V-B and Section V-C, the experimental setting were defined equally, and executed using Python 3.6 on a Ubuntu 16.04 x64 machine with Intel(R) Core(TM) i7-6850 K processor with 110 GB RAM with a NVIDIA GeForce GTX 2080 Ti 11 GB GPU. In the experimentation, the S2 data

TABLE I
ABLATION'S STUDY QUANTITATIVE ASSESSMENT FOR NDVI PRODUCTS BASED ON RMSE ($\times 10^{-2}$), MSE ($\times 10^{-2}$), AND MAE ($\times 10^{-4}$)

VARIANTIONS	RMSE	MSE	MAE
S3 Branch	3.14±0.02	9.85±0.15	1.84±0.02
S3 Branch+Att	3.13±0.02	9.79±0.13	1.83±0.01
S2 Branch	2.07±0.01	4.27±0.05	1.25±0.02
S2 Branch+Att	2.06±0.02	4.26±0.09	1.24±0.01
S3 +S2 Branch	1.94±0.17	3.79±0.70	1.08±0.03
S3 +S2 Branch+Att (DAT-CNN)	1.88±0.03	3.53±0.14	1.09±0.05

has been reduced with a 1:3 ratio through a Lanczos interpolation. All the $X_{S2_{\text{ref}}}$ product bands have been resampled with a resolution r of 20 m per pixel, being the final S2 input products of a size of 1830×1830 pixels and 13 spectral bands. In this work, since S2 and S3 spectral ranges do not perfectly match, we decided to use all S2 bands to take full advantage of the MSI spectral range. However, other S2 bands subsets could be also considered in this regard. By the other hand, the $X_{S3_{\text{NDVI}}}$ products have a size of 366×366 pixels. The temporal windows used for in the S2 input products was of $T = 5$ with the central image of the temporal series being of the same day as the absent target S3 product. For the S3 input products a $T = 4$ interval was used, as the central image of the set was the unavailable product, selected for the training and validation as output.

For the training stage, all the methods used the same data partitions from the dataset defined in Section III-A. From this data, one complete S2 and S3 paired product was excluded to generate a real case example. The rest of the paired samples were divided in patches of $P = 5 \times 5$ for S2, and $S = 1 \times 1$ for S3 due to computational limitations, following the final spatial relation between the input and output products of $1:5$. A 40% of the data was used for training, while another 40% for test and a 20% for validation. All the networks used in the experimentation were trained using the same hyperparameters and optimization methods. We used the ADAM optimizer with a learning rate $lr = 10^{-3}$. A reduction of this learning rate with a factor of 0.5 was used when the LOSS value did not improve in 10 consecutive epochs until a minimum learning rate $lr = 10^{-7}$. The batch size and the epochs were also fixed to 128 and 150, respectively, for every training. The results for each epoch were evaluated minimizing the mse metric. To check the robustness of the methods each one was trained five different times, with different seeds used to randomize the data partition in training, test, and validation and the results shown below are the mean value of these five training.

B. Experiment 1: Ablation Study

In this first experiment, an ablation study with different variants of the proposed model have been studied. In Table I, root mean squared error (RMSE), mean squared error (mse), and mean absolute error (MAE) metrics between the objective supposed unavailable S3 NDVI product and the predicted by each model are shown. The variance of each metric in every method along the five iterations is shown as well. Note that the values displayed in Table I have been multiplied by a factor (10^{-2} in the case of RMSE and mse, and 10^{-4} for MAE) to reduce the number of uninformative digits in the results. The various model variations are explained below.

1) S3 Branch: The first of the variations proposed is as well the most simple model among them. This model consists of only the S3 branch explained in Section IV-C2 without the attention module. In this case, the input data is the temporal vector of S3 NDVI product pixels.

2) S3 Branch+Attention: In this variation an attention 2-D module has been used at the starting of the S3 branch, exactly as described in Section IV-C2. The input data are also a temporal vector of S3 NDVI product pixels.

3) S2 Branch: This model has the architecture of the S2 branch Section IV-C1 without the 3-D attention module. The input data are the multitemporal multispectral S2 image patches.

4) S2 Branch+Attention: As in the S3 case, a S2 branch including the 3-D attention module, as described in Section IV-C1, has been selected. As in the previous case, the input data are multitemporal multispectral S2 image patches.

5) S3+S2 Branch: This is the first model of the ablation study which uses both of the S2 and S3 branches. Nevertheless, neither the S2 nor the S3 branches contain any attention module. As the model contains both branches, the S3 NDVI temporal product pixels and the temporal multispectral S2 image patches window are used as input.

6) S3+S2 Branch+Attention (DAT-CNN): The DAT-CNN model described in Section IV-C containing both branches with their correspondent attention modules, using the same data has the model previously described.

According to the results shown in Table I, the multimodal S2 branch models outperform drastically the S3 simpler models. This is in part thanks to the spatial information contained in the S2 patches which the S3 pixel vector lacks, even though the S3 pixel vector consists of the same product data type as the expected output. Also, the multimodal S2 branch models containing a S2 product of the same day as the objective S3 L2 product seems to increase this advantage. In the both cases of isolated branches, the use of an attention block seems to give some stability in a few metrics, and very slight improvements in general. However, the multiple branch models show a significant improvement to the single S2 or S3 branches. In the case of the model without attention, there is an overall improvement in all the metrics, but there is a high deviation as well, we consider this as a result of the network not being able of extracting the most important features efficiently. This is solved in the last and proposed model, DAT-CNN, through the attention blocks which now show a better performance. As the number of layers and filters grow, the highlighting of the most relevant features extracted grows in significance as well.

C. Experiment 2: Comparison With Other Methods

In this second experiment, the proposed DAT-CNN is compared to other state-of-the-art methods, which were used for similar objectives and standard methods. To perform this

comparison, the dataset defined in Section III-A was used as well. For this multimodal experiment the main input data used have been the S2 image patches. The methods unable to profit from the temporal information were given only the same day S2 image patches as input data, while more information was given to the methods able to exploit temporal information or other inputs. The methods which were used in this experiment can be separated in three groups.

In the first group, four traditional regression methods were selected as representatives. This group includes the following methods: 1) The linear regression (LIN) [18]; 2) the ridge regression (RG) [47], which is an improvement of the linear regression to better model multicollinear independent variables; 3) the support vector regression (SVR) [48] using a nonlinear radial basis function kernel and a maximum number of 1000 iterations; 4) and a random forest regression (RFR) [17] with eight classification trees which results averages to improve the accuracy.

The second group contains methods based in 2-D neural networks architectures that have been used for multispectral or hyperspectral classification or temporal prediction: a multilayer perceptron (MLP) with only fully connected layers, as defined in [23], which was used in a hyperspectral image classification review; from the same review [23] a standard 2-D convolutional neural network (2D-CNN); a 2-D CNN that was used to quantifying cyanobacteria using hyperspectral imagery (PR-CNN) [28]; a deep patch-based 2-D CNN designed to classify RS land cover images (DP-CNN) [24]; an hyperspectral image classification optimized 2-D CNN (HY-CNN) [25]; a contextual CNN which uses residual information and a multiscale filter to classify multispectral images (CD-CNN) [26]; and the last of this group, a CNN used as to estimate with a regression the chlorophyll-a concentration using S2 data (CNNR) [29].

In the third group the models which take advantage and use temporal information are included: A standard 3-D convolutional neural network used in a hyperspectral image classification methods review [23]; a deep recurrent neural network with long short-term memory and gated recurrent units (DRNN) designed to predict short-term vegetation indices using temporal information from S2 or MODIS data [21]; and our proposed model DAT-CNN with two branches, 3-D convolutions, and channel and spatial attention as main features.

To analyze and compare the performance of each of the methods presented previously and the proposed DAT-CNN, three different metrics have been used. The mean of five iterations with different partitioned data and the standard deviation of RMSE, mse, and MAE for all the models was obtained. In Table II this quantitative evaluation is presented. In order to have a better visualisation the RMSE and MAE values and standard deviation were multiplied by a factor of $\times 10^{-2}$, while the mse results and deviation were multiplied by a factor of $\times 10^{-4}$.

In the first column of this table the different models are arranged in rows. In following columns, the RMSE, mse, and MAE obtained results by each method are shown. The results obtained show how the proposed DAT-CNN model outperforms every other method studied for the analyzed metrics. For the set of traditional regression methods, the method with the best performance was the RFR with a considerable improvement compared to the others, while the worst results were obtained by the SVR. In the 2D-CNN models the simple MLP and 2D-CNN had similar results to the traditional LIN or RID methods, by the other hand models like the HY-CNN or the DP-CNN obtained better results than the RFR. Among the methods unable to use temporal information, the CNNR achieved the best results. Nevertheless, all the methods which used temporal information outperformed the previous. The method with the best results behind the proposed DAT-CNN was the 3D-CNN, but still its noticeable how the difference between the DAT-CNN and the other 3-D methods is larger than the difference between the CNNR and 3D-CNN methods. The margin of improvement obtained against the following best method is comparable to the difference between using a LIN regressor and an 2D-CNN in the middle set of performances as the PR-CNN.

TABLE II
DIFFERENT METHODS' QUANTITATIVE ASSESSMENT FOR NDVI PRODUCTS BASED ON RMSE ($\times 10^{-2}$), MSE ($\times 10^{-2}$), AND MAE ($\times 10^{-4}$)

METHODS	RMSE	MSE	MAE
LIN	3.59±0.02	12.9±0.13	2.26±0.01
RID	3.59±0.02	12.9±0.13	2.26±0.01
SVR	6.93±4.40	63.6±76.85	5.71±3.82
RFR	2.48±0.01	6.14±0.04	1.63±0.01
MLP	3.59±0.02	12.9±0.13	2.26±0.01
2D-CNN	3.59±0.02	12.9±0.13	2.26±0.01
PR-CNN	3.02±0.24	9.16±1.41	2.23±0.23
DP-CNN	3.29±0.03	5.22±0.11	1.43±0.02
HY-CNN	2.39±0.02	5.74±0.06	1.54±0.01
CD-CNN	4.25±3.11	25.3±39.09	3.25±2.77
CNNR	2.27±0.02	5.13±0.07	1.41±0.01
3D-CNN	2.07±0.01	4.28±0.03	1.27±0.01
DRNN	2.24±0.02	5.02±0.09	1.34±0.01
DAT-CNN	1.88±0.03	3.53±0.14	1.09±0.05

D. Experiment 3: Real Case Experimentation

In order to test the actual ability of the proposed method to generate enchanted S3 products in a more realistic experimentation, one random complete S3 product was excluded of the previous training, testing, and validation. In this experiment, some of the models trained in the previous experiment will generate the enchanted S3 product corresponding to the excluded product. Then, their qualitative and quantitative results will be compared. For this purpose, our model and two representatives from each of the three different groups according to their performance in the previous experiment were selected: traditional (LIN and RFR), 2D-CNN (DP-CNN and CNNR), and 3D-CNN (3D-CNN and DRNN) were selected to generate S3 NDVI product predictions.

In Table III the qualitative results of this experiment are shown. Again, in the first column of the table the models are presented in rows, and in the next columns the RMSE, mse, and MAE results are shown. For this real case experimentation,

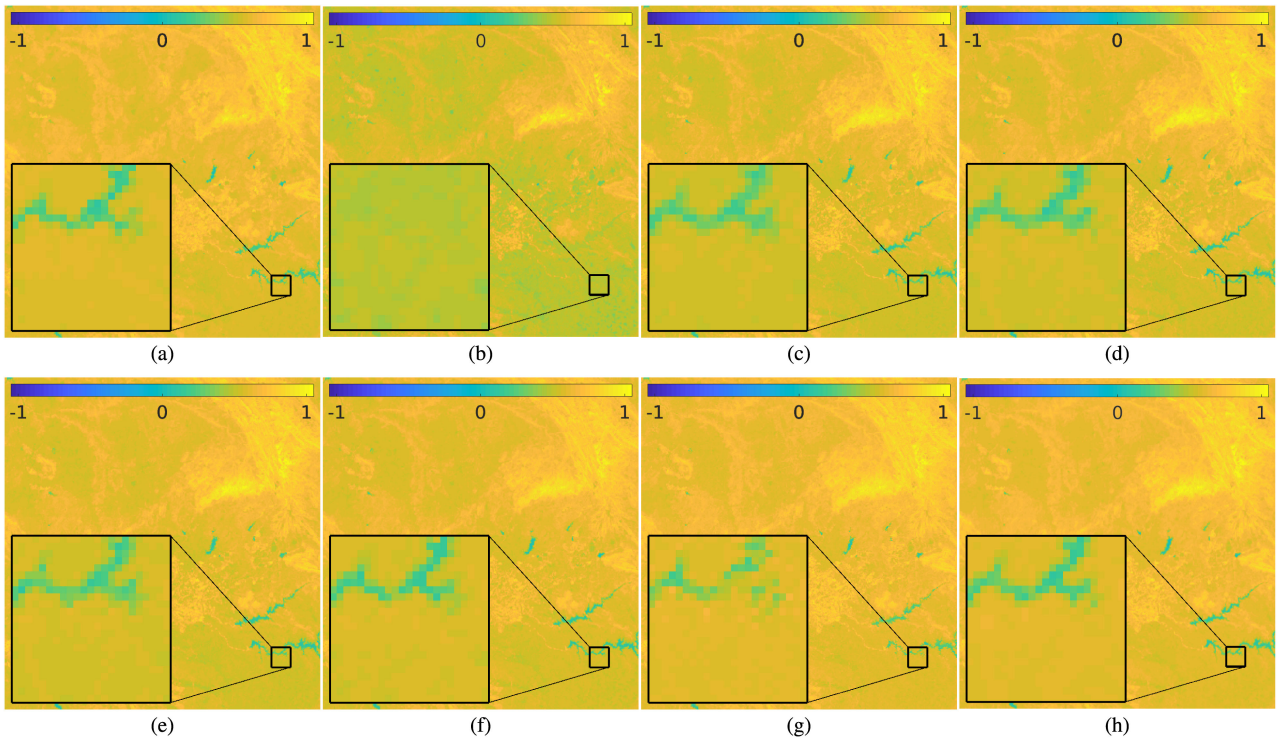

Fig. 7. NDVI qualitative assessment. (a) Ground-truth. (b) LIN. (c) RFR. (d) DP-CNN. (e) CNNR. (f) 3D-CNN. (g) DRNN. (h) DAT-CNN.

TABLE III
QUANTITATIVE ASSESSMENT FOR NDVI PRODUCTS BASED ON RMSE ($\times 10^{-2}$), MSE ($\times 10^{-2}$), AND MAE ($\times 10^{-4}$)

METHODS	RMSE	MSE	MAE
LIN	3.12	9.78	1.81
RFR	2.41	5.84	1.65
DP-CNN	2.25	5.07	1.46
CNNR	2.24	5.03	1.48
CNN3D	2.32	5.38	1.60
DRNN	2.68	7.19	1.78
DAT-CNN	2.18	4.78	1.39

the proposed DAT-CNN model is able to outperform the other methods as well. Nevertheless, for this specific case the following better performing methods in the global calculated metrics are the CNNR and the DP-CNN. Then the CNN3D, DRNN, RFR, and finally the LIN method.

Complementary to the quantitative results, a qualitative experimentation to further demonstrate the improvement obtained using the proposed DAT-CNN method in this experiment has been done. The generated NDVI predictions from each model and the ground-truth are shown in Fig. 7. To generate more intuitive visual results, a false color was used. Values from -1 to 0 represent rocks and water surfaces corresponding to blue to greenish blue, bare soil can be seen between 0.1 and 0.2 in pure green, and the vegetation is observable from 0.2 to 1 values, in greenish yellow to pure yellow.

As previously stated, the proposed model obtained the best results, and this can be easily seen visually as well through two main points. First, the overall values in the image, and second the correct features representation. In all the models (except DRNN and the proposed DAT-CNN) the images have in general lower values. These lower values appear as a greenish overall image than the ground-truth image, specially in the LIN prediction. The other visible difference can be better seen in the magnification. The shape of the river which appears in it is completely lost in the LIN prediction. In the RFR, DP-CNN, and CNNR the shape can be seen, but it has lost sharpness, appearing blurry, and bigger. This loss of sharpness can lead to a better quantitative general results when other methods are not able to correctly represent the features of the image. For the 3D-CNN, even though the reconstructed image is greenish as was discussed previously, the shape of the river is still better preserved. By the other hand, the DRNN which better represented the overall colors, appears to have difficulties with shapes as the river shown, leading to worst results than the DP-CNN or CNNR. After analyzing the different time-resolved S3 VI visual results, it can be concluded that the proposed DAT-CNN model not only obtained better quantitative results but also better shape and color representation in qualitative results according to the objective ground-truth.

From the results obtained there are four evident advantages of the proposed architecture compared to the others: 1) considering multitemporal features, 2) the adaptation to data diversity limitations, 3) the focused representation of features, and 4) the use of multimodal information of both S2 and S3 data. The use of temporal information allow the network to extract dynamic information not available in the spatial or frequency dimensions found in the temporal interval data in S2 and S3. This can be noticed in the experimental results, where the best performing

models are all 3D-CNN architectures. Besides, when working with limited data due to cloud covering (which is a common issue in RS imagery) or other limitations, avoiding overfitting is mandatory. The increment in parameters and complexity due to the 3D-convolutions has to be considered as well, possibly aggravating this issue. To decrease this possible overfitting, we reduced the number of parameters and layers, while adopting the head, body, and tail block scheme. At the same time, a it was necessary to acknowledge the necessary layers and blocks to compensate the spatial resolution difference and obtain representative features. The highlighting of the most relevant temporal features through the attention modules is essential to improve the feature representation, while the overfitting is minimized from irregular temporal period samples. Finally, the use of both S2 and S3 information, complementary to the previous mentioned reasons has a major impact in the overall better performance of the model, as has been explored in the first experiment Section V-B.

VI. Conclusion

In this article, a 3D-CNN model to generate temporal enhanced Sentinel-3/FLEX Derived L4 products has been proposed. This model takes advantage from not only spatial and spectral information, but also temporal information from previous and following days samples to obtain L4 products. The objective of those L4 products is to complete and enhance the temporal resolution of S3 or FLEX temporal series of products. To prove the accuracy of this model, it has been compared to other regression state-of-the-art methods, including other CNN models designed specifically for RS, using the RMSE, mse, and MAE metrics. The experimental results demonstrate the improvement obtained from using temporal information to generate the L4 products. Furthermore, the proposed 3D-CNN model was able to obtain better results and predictions than the other models in the same conditions, using a S2 reflectance products and S3 L2 products dataset of 20 same-day paired images from 2019. In the future, we plan to extend the proposed DAT-CNN model by exploiting early fusion interactions together with other strategies, such as increasing the temporal depth as well as the data used for training. Nevertheless, this will increment the computational costs and data volume as potential limitations.

References

[1] M. ED Chaves, M. CA Picoli, and I. D. Sanches, "Recent applications of Landsat 8/OLI and sentinel-2/MSI for land use and land cover mapping: A systematic review," *Remote Sens.*, vol. 12, no. 18, 2020, Art. no. 3062.

[2] D. Ibañez, R. Fernandez-Beltran, J. M. Sotoca, R. A. Mollineda, J. Moreno, and F. Pla, "Multitemporal mosaicing for Sentinel-3/flex derived level-2 product composites," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5439–5454, 2020.

[3] B. El Mahrad, A. Newton, J. D. Icely, I. Kacimi, S. Abalansa, and M. Snoussi, "Contribution of remote sensing technologies to a holistic coastal and marine environmental management framework: A review," *Remote Sens.*, vol. 12, no. 14, 2020, Art. no. 2313.

[4] G. Sumbul, R. G. Cinbis, and S. Aksoy, "Fine-grained object recognition and zero-shot learning in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 770–779, Feb. 2018.

[5] 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.

[6] 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. Photogramm. Remote Sens.*, vol. 103, pp. 115–128, 2015.

[7] 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.

[8] C. Donlon *et al.*, "The global monitoring for environment and security (GMES) Sentinel-3 mission," *Remote Sens. Environ.*, vol. 120, pp. 37–57, 2012.

[9] Z. Malenovskỳ *et al.*, "Sentinels for science: Potential of Sentinel-1,-2, and-3 missions for scientific observations of ocean, cryosphere, and land," *Remote Sens. Environ.*, vol. 120, pp. 91–101, 2012.

[10] J. Vicent *et al.*, "FLEX end-to-end mission performance simulator," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 7, pp. 4215–4223, Jul. 2016.

[11] G. H. Mohammed *et al.*, "Remote sensing of solar-induced chlorophyll fluorescence (SIF) in vegetation: 50 years of progress," *Remote Sens. Environ.*, vol. 231, 2019, Art. no. 111177.

[12] D. Arnas, P. Jurado, I. Barat, B. Duesmann, and R. Bock, "FLEX: A parametric study of its tandem formation with Sentinel-3," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 12, no. 7, pp. 2447–2452, Jul. 2019.

[13] S. Bandopadhyay, A. Rastogi, S. Cogliati, U. Rascher, M. Gabka, and R. Juszczak, "Can vegetation indices serve as proxies for potential sun-induced fluorescence (SIF)? A fuzzy simulation approach on airborne imaging spectroscopy data," *Remote Sens.*, vol. 13, no. 13, 2021, Art. no. 2545.

[14] R. Fernandez-Beltran, F. Pla, and A. Plaza, "Sentinel-2 and Sentinel-3 intersensor vegetation estimation via constrained topic modeling," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 10, pp. 1531–1535, Oct. 2019.

[15] H. Shen *et al.*, "Missing information reconstruction of remote sensing data: A technical review," *IEEE Geosci. Remote Sens. Mag.*, vol. 3, no. 3, pp. 61–85, Sep. 2015.

[16] C.-H. Lin, K.-H. Lai, Z.-B. Chen, and J.-Y. Chen, "Patch-based information reconstruction of cloud-contaminated multitemporal images," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 163–174, Jan. 2014.

[17] W. Zhao, H. Wu, G. Yin, and S.-B. Duan, "Normalization of the temporal effect on the MODIS land surface temperature product using random forest regression," *ISPRS J. Photogramm. Remote Sens.*, vol. 152, pp. 109–118, 2019.

[18] C. Zeng, H. Shen, M. Zhong, L. Zhang, and P. Wu, "Reconstructing MODIS LST based on multitemporal classification and robust regression," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 3, pp. 512–516, Mar. 2015.

[19] Q. Zhang, Q. Yuan, C. Zeng, X. Li, and Y. Wei, "Missing data reconstruction in remote sensing image with a unified spatial-temporal-spectral deep convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4274–4288, Aug. 2018.

[20] M. Shao, C. Wang, T. Wu, D. Meng, and J. Luo, "Context-based multiscale unified network for missing data reconstruction in remote sensing images," *IEEE Geosci. Remote Sens. Lett*., vol. 19, 2022, Art. no. 8001205.

[21] W. Yu *et al.*, "Spatial-temporal prediction of vegetation index with deep recurrent neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 2501105.

[22] R. Fernandez-Beltran, D. Ibañez, J. Kang, and F. Pla, "Time-resolved sentinel-3 vegetation indices via inter-sensor 3-D convolutional regression networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 2501505.

[23] 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.

[24] A. Sharma, X. Liu, X. Yang, and D. Shi, "A patch-based convolutional neural network for remote sensing image classification," *Neural Netw.*, vol. 95, pp. 19–28, 2017.

[25] S. Yu, S. Jia, and C. Xu, "Convolutional neural networks for hyperspectral image classification," *Neurocomputing*, vol. 219, pp. 88–98, 2017.

[26] H. Lee and H. Kwon, "Going deeper with contextual CNN for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4843–4855, Oct. 2017.

[27] M. Zhang, W. Li, R. Tao, H. Li, and Q. Du, "Information fusion for classification of hyperspectral and LiDAR data using IP-CNN," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5506812.

[28] J. Pyo *et al.*, "A convolutional neural network regression for quantifying cyanobacteria using hyperspectral imagery," *Remote Sens. Environ.*, vol. 233, 2019, Art. no. 111350.
[29] E. Aptoula and S. Ariman, "Chlorophyll—A retrieval from Sentinel-2 images using convolutional neural network regression," *IEEE Geosci. Remote Sens. Lett*., vol. 19, 2022, Art. no. 5506812.
[30] T. Kattenborn, J. Leitloff, F. Schiefer, and S. Hinz, "Review on convolutional neural networks (CNN) in vegetation remote sensing," *ISPRS J. Photogramm. Remote Sens.*, vol. 173, pp. 24–49, 2021.
[31] Y. Zeng, X. Guo, H. Wang, M. Geng, and T. Lu, "Efficient dual attention module for real-time visual tracking," in *Proc. IEEE Vis. Commun. Image Process.*, 2019, pp. 1–4.
[32] J. Fu, J. Liu, J. Jiang, Y. Li, Y. Bao, and H. Lu, "Scene segmentation with dual relation-aware attention network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 6, pp. 2547–2560, Jun. 2021.
[33] "Sentinelsat was created to search, download and retrieve the metadata of sentinel satellite images from the copernicus open access hub," Accessed: Apr. 1, 2022. [Online]. Available: https://pypi.org/project/sentinelsat/
[34] R. Richter, J. Louis, and U. Müller-Wilm, "Sentinel-2 MSI-level 2A products algorithm theoretical basis document," European Space Agency, Paris, France, S2PAD-ATBD-0001, 2012.
[35] V. Zekoll, M. Main-Knorn, J. Louis, D. Frantz, R. Richter, and B. Pflug, "Comparison of masking algorithms for Sentinel-2 imagery," *Remote Sens.*, vol. 13, no. 1, 2021, Art. no. 137.
[36] "SNAP is a common architecture for all sentinel toolboxes is being jointly developed by Brockmann consult, skywatch and C-S called the sentinel application platform (SNAP)," Accessed: Apr. 1, 2022. [Online]. Available: https://step.esa.int/main/toolboxes/snap
[37] "Algorithm theoretical basis document of the normalized difference vegetation index (NDVI) for sentinel-3 OLCI products," Accessed: Apr. 1, 2022. [Online]. Available: https://land.copernicus.eu/global/products/ndvi
[38] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," 2014, *arXiv:1406.6247*.
[39] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.
[40] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.
[41] H. Alhichri, A. S. Alswayed, Y. Bazi, N. Ammour, and N. A. Alajlan, "Classification of remote sensing images using efficientnet-B3 CNN model with attention," *IEEE Access*, vol. 9, pp. 14078–14094, 2021.
[42] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929.
[43] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
[44] Y. Zhao, K. L. Man, J. Smith, K. Siddique, and S.-U. Guan, "Improved two-stream model for human action recognition," *EURASIP J. Image Video Process.*, vol. 2020, no. 1, pp. 1–9, 2020.
[45] D. Chen, S. Zhang, W. Ouyang, J. Yang, and Y. Tai, "Person search via a mask-guided two-stream CNN model," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 734–750.
[46] S. Aslani *et al.*, "Multi-branch convolutional neural network for multiple sclerosis lesion segmentation," *NeuroImage*, vol. 196, pp. 1–15, 2019.
[47] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.
[48] Y.-S. Shiu and Y.-C. Chuang, "Yield estimation of paddy rice based on satellite imagery: Comparison of global and local regression models," *Remote Sens.*, vol. 11, no. 2, 2019, Art. no. 111.

Damian Ibañez received the B.Sc. degree in industrial electronics and automation engineering from the Polytechnic University of Valencia, Valencia, Spain, in 2019, and the M.Sc. degree in intelligent systems, in 2020, from the Universitat Jaume I, Castellón de la Plana, Spain, where he is currently working toward the Ph.D. degree in computer science, from the Universitat Jaume I.

His research interests include computer vision, machine learning and automation, with special interest in remote sensing applications.

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, Castellon de la Plana, Spain, in 2007, 2011, and 2016, respectively.

He is currently an Assisstant Professor in the Department of Computer Science and Systems, the University of Murcia, Spain, as well as collaborating member of the Institute of New Imaging Technologies, University Jaume I. He has been a visiting Researcher with the University of Bristol, Bristol, U.K., the University of Cáceres, Cáceres, Spain, Technische Universität Berlin, Berlin, Germany, and the Autonomous University of Mexico State, Mexico City, 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.

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

He is currently a Full Professor with the Departament de Llenguatges i Sistemes Informatics, University Jaume I, Castellon de la Plana, Spain. He has been a Visiting Scientist with the Silsoe Research Institute, University of Surrey, Guildford, U.K., the University of Bristol, Bristol, U.K., CEMAGREF, Montpellier, France, the University of Genoa, Italy, the Instituto Superior Tecnico, Lisbon, Portugal, the Swiss Federal Institute of Technology, ETH-Zurich, Zurich, Switzerland, the Idiap Research Institute, Martigny, Switzerland, and the Technical University of Delft, Delft, The Netherlands. He is a Faculty Member of the Institute of New Imaging Technologies, University Jaume I. His research interests include color and spectral image analysis, visual motion analysis, 3-D image capture and visualization, and pattern recognition techniques applied to image processing.

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

Naoto Yokoya (Senior Member, IEEE) received the M.Eng. and Ph.D. degrees in aerospace engineering from The University of Tokyo, Tokyo, Japan, in 2010 and 2013, respectively.

He is currently a Lecturer with The University of Tokyo and a Unit Leader with the RIKEN Center for Advanced Intelligence Project, Tokyo, Japan, where he Leads the Geoinformatics Unit.

Dr. Yokoya is currently an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. He was the Chair of the IEEE Geoscience and Remote Sensing Society Image Analysis and Data Fusion Technical Committee, from 2019 to 2021.