IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 16, 2021, accepted January 26, 2021, date of publication January 29, 2021, date of current version February 9, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3055554

Attentive Spatial Temporal Graph CNN for Land Cover Mapping From Multi Temporal Remote Sensing Data

ALESSANDRO MICHELE CENSI[1,2], DINO IENCO[1], (Member, IEEE), YAWOGAN JEAN EUDES GBODJO[1], RUGGERO GAETANO PENSA[2], ROBERTO INTERDONATO[3], AND RAFFAELE GAETANO[3]

[1]INRAE, UMR TETIS, University of Montpellier, 34000 Montpellier, France
[2]Department of Computer Science, University of Turin, 10124 Turin, Italy
[3]CIRAD, UMR TETIS, 34090 Montpellier, France

Corresponding author: Dino Ienco (dino.ienco@inrae.fr)

This work was supported in part by the French National Research Agency through the Investments for the Future Program, under Grant ANR-16-CONV-0004 (DigitAg), in part by the GEOSUD Project under Grant ANR-10-EQPX-20, in part by the French Ministry of agriculture Agricultural and Rural Development Trust Account, and in part by the PARCELLE Project funded by the French Space Agency under Grant DAR CNES 2019.

ABSTRACT Satellite image time series (SITS) collected by modern Earth Observation (EO) systems represent a valuable source of information that supports several tasks related to the monitoring of the Earth surface dynamics over large areas. A main challenge is then to design methods able to leverage the complementarity between the temporal dynamics and the spatial patterns that characterize these data structures. Focusing on land cover classification (or mapping) tasks, the majority of approaches dealing with SITS data only considers the temporal dimension, while the integration of the spatial context is frequently neglected. In this work, we propose an attentive spatial temporal graph convolutional neural network that exploits both spatial and temporal dimensions in SITS. Despite the fact that this neural network model is well suited to deal with spatio-temporal information, this is the first work that considers it for the analysis of SITS data. Experiments are conducted on two study areas characterized by different land cover landscapes and real-world operational constraints (i.e., limited labeled data due to acquisition costs). The results show that our model consistently outperforms all the competing methods obtaining a performance gain, in terms of F-Measure, of at least 5 points with respect to the best competing approaches on both benchmarks.

INDEX TERMS Spatial temporal graph convolutional neural network, attention-based neural network, object-based image classification, satellite image time series, land cover classification, deep learning.

I. INTRODUCTION

The Food and Agriculture Organization (FAO) of the United Nations predicts that in order to meet the needs of the expected 3 billion population growth by 2050, food production has to increase by 60% [1]. Therefore, accurately mapping agricultural as well as general human activities over large areas is crucial for estimating food production across the globe and, more generally, monitoring natural resources availability in the context of climate changes [2].

Nowadays, modern Earth Observation (EO) missions allow to collect remote sensing images to support the monitoring of the Earth surface dynamics over large areas. A notorious example is the Copernicus programme with the Sentinel mission that supplies freely accessible high resolution images (up to 10m) with high revisit time period (every 5 or 6 days). This unprecedented amount of satellite imagery can be arranged as satellite image time series (SITS) and, apart from its clear value in monitoring agricultural and natural resources, it can also be employed as a powerful tool to support many other application domains like ecology [3], mobility, health, risk assessment [4] and land management planning [5].

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

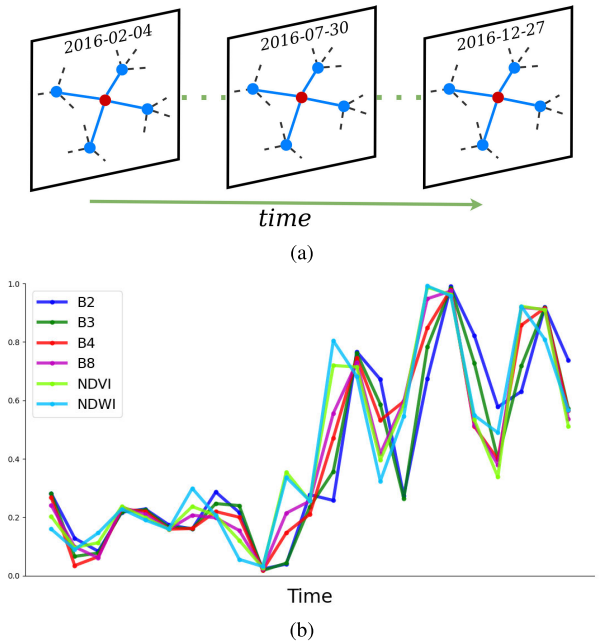


FIGURE 1. **(a) A target segment (red node) with its associated spatial neighborhood set (blue nodes) and (b) the multi-variate time series information associated to the target node. The spatial neighborhood of a generic target node is fixed over time while the target node information evolves producing a multi-variate time series.**

Due to the increasing availability of SITS information, one of the main challenges related to their exploitation today is how to simultaneously leverage the complementarity between temporal dynamics and spatial patterns characterizing such data. To this end, machine learning and, more recently, deep learning techniques are extensively adopted in the context of SITS data classification, also referred as land cover mapping [5], [6]. Nevertheless, a vast majority of existing research studies concentrate their effort to cope with the temporal dimension while the integration of the spatial context surrounding a particular location is frequently neglected.

This is especially the case when SITS analysis is conducted under the object-based image analysis (OBIA) paradigm [7]. This paradigm is widely adopted in the remote sensing community, and it is gaining increasing attention in the context of high resolution satellite images analysis [5]. Conversely to the pixel-based analysis, OBIA considers segments (objects) as working units. Segments are typically obtained via a segmentation process in which an image is partitioned into clusters of similar neighboring pixels [8] that depict visually perceptible "land units" within the image scene and that can be associated to high level semantic concepts. Due to this latter point, objects are, generally, more simpler to interpret for an expert [7] and well adapted for human level post-analysis. Moreover, from the image segmentation a Region Adjacency Graph (RAG), modeling the objects spatial interaction, can be extracted. In the RAG structure, the objects are the nodes of the graph and an edge exists between two nodes if the corresponding objects are spatially adjacent. In addition, each object is characterized by a multi-variate time series. Figure 1 depicts a node (red point), its associated spatial neighborhood (blue points) and the related multi-variate time series (Fig. 1b).

In the context of general spatio-temporal data analysis, spatial temporal graph convolutional neural networks [9] (STGCNNs) are attracting more and more attention thanks to their ability to explicitly model both dimensions at once. Despite STGCNNs being widely adopted to deal with spatio-temporal tasks, such as traffic forecasting [10], flood forecasting [11] and video activity recognition [12], surprisingly, to the best of our knowledge and according to a very recent literature survey [9], no study has been conducted yet to adopt such models in the context of satellite image time series data analysis, by leveraging RAGs in the classification process. This is probably due to the fact that an STGCNN model initially developed for a particular task can not be easily transferred to a different one. In addition, satellite image scenes cover large areas thus resulting in RAGs with (possibly) hundreds of thousands nodes. This fact limits the adoption of standard STGCNN models that are based on prior spectral graph signal processing operations, i.e., Laplacian graph extraction [9].

To deal with the SITS land cover mapping task, with the aim of explicitly integrating the segments spatial correlation in the underlying analysis, we propose an attentive Spatial Temporal Graph Convolutional Neural Network, named *STEGON*. *STEGON* leverages attention at two different stages. Firstly, the spatial neighborhood of a SITS segment is automatically aggregated weighting the contribution of each neighbor according to its importance. Secondly, an attention mechanism combines the information coming from the target SITS segment and its neighborhood.

Overall, the contributions of our work are as follows:

- We propose a novel spatial temporal graph convolutional approach to deal with SITS land cover mapping task.
- We equip *STEGON* with several attention modules, allowing the model to automatically weight the contribution of the spatial neighborhood at several stages of the processing pipeline.
- We adopt spatial graph convolutions [9] to work directly on the raw data avoiding global graph analysis like eigenvalues and/or eigenvectors computation; this allows *STEGON* to scale up on real world large study areas conversely to all the previous spatial temporal graph CNN methods [9].
- We perform extensive experiments on two different benchmarks covering two large study areas; the results confirm the quality of our model consistently with respect to all the competing methods.

To validate our proposal, we consider benchmarks representing two different study areas exhibiting contrasted land cover landscapes and state of the art machine learning and deep learning approaches commonly employed in the task of land cover mapping from SITS data.

The rest of the article is structured as follows: the literature related to our work is introduced in Section II; Section III introduces preliminary definitions about the land cover mapping task and the graph-based geographical area representation; Section IV describes the *STEGON* framework and Section V describes the data and the considered study areas. Experimental settings and results are detailed and discussed in Section VI. Finally, Section VII concludes the work.

II. RELATED WORK

A. LAND COVER MAPPING FROM MULTI-TEMPORAL SATELLITE DATA

Land cover mapping from multi-temporal satellite data constitutes a crucial task in order to monitor natural resources [13] on the Earth surfaces and human settlement evolution [14]. In [14] the authors propose an operational framework to perform large scale land cover mapping at national scale. The classification is achieved via the Random Forest classifier that, nowadays, represents the common approach for land cover mapping on multi-temporal satellite data. [15] and [16] deal with land usage and land cover (LULC) mapping via recurrent neural networks approaches. In [15], data are analyzed via Long Short Term Memory (LSTM) while [16] tackles the LULC mapping problem still considering recurrent neural network approaches but, this time, the Gated Recurrent Unit (GRU) was preferred to perform classification. Recently, [6] formalizes the use of one dimensional (temporal) Convolutional Neural Networks for satellite image time series classification. In this model, the convolution is performed on the temporal dimensions of the time series data with the aim of managing and modeling short and long time dependencies. The conducted study highlights the appropriateness of such approach with respect to the previous proposed strategies in the context of LULC mapping from multi-temporal satellite data.

B. GRAPH NEURAL NETWORKS

Recently, graph convolutional networks (GCN) have shown extraordinary performance on several graph structure tasks, such as node classification and network representation [9]. GCNs are classified into spectral [17] and spatial [18] methods. Spectral methods define convolution on the spectral domain. Many methods are derived from the work of [17]. ChebNet [19] is a powerful GCN model that uses the Chebyshev extension to reduce the complexity of Laplacians computation. GCN [20] simplifies ChebNet to a more simple form and achieves state-of-the-art performances on various tasks. Despite the value of such family of GCNs, the spectral methods require the computation of adjacency matrix eigenvectors and eigenvalues since they are based on the (normalized) Laplacian of the graph, which is challenging and prohibitive to compute when the graph structure has hundred of thousands nodes (e.g., in the case of graphs induced by the segmentation of remote sensing data). The spatial methods directly perform convolution on the graph nodes and their neighbors. The GraphSAGE model [18] generates embeddings by sampling and aggregating features from the local neighborhood of a node. Graph Attention Networks (GATs) [21] use self-attentional layers to assign different weights to different nodes in a neighborhood. PinSage [22] proposes a data efficient GCN strategy to combine random walks and graph convolutions to process large-scale graph for web recommendation.

III. PRELIMINARIES

In this section we provide the definition of the Land cover mapping task and the graph representation we adopt to model the remote sensing data.

A. LAND COVER MAPPING TASK

The Land Cover (LC) mapping task is defined as a multi-class supervised classification problem. Given satellite images (remote sensing) information covering a study area, referred as X, X can be partitioned into two sets $X = \{X_l, X_u\}$ where X_l is the portion of the study area on which the ground truth information Y_l is available and X_u is the portion without ground truth information. Usually, $|X_l| \ll |X_u|$ where the $|\cdot|$ symbol indicates the surface of the covered area. The objective of the LC mapping task is to build a classifier $CL(Y_l, X_l, X_u) \rightarrow Y_u$ which takes as input Y_l, X_l and X_u and predicts Y_u, the classification of the unlabeled portion of the study area. Finally, the LC mapping for the whole study area is provided.

B. GRAPH REPRESENTATION OF THE GEOGRAPHICAL AREA

Given a geographical area, through a segmentation process we can derive a RAG (*Region Adjacency Graph*). An example of RAG is depicted in Figure 2, the stage (b) illustrates a segmentation result while stage (c) shows the corresponding Region Adjacency Graph. More formally, considering the scenario in which satellite image time series are involved, a RAG is defined as a graph $G = (V, E, ST)$ where V is the set of segments (nodes), E is the set of edges and ST is a function that, given a segment, returns the corresponding multi-variate time series information: $ST(v_i) \in \mathbb{R}^{T \times D}$ where T is the number of timestamps of the time series and D is the number of features/dimensions on which the multi-variate time series is defined on.

The set of edges E is derived from the set of V considering spatial adjacency (Figure 2). More in detail, for each $v_i, v_j \in V$ that are spatially adjacent (segment v_i spatially touch segment v_j). Finally, we define $N(v_i)$ as the set of neighborhood segments of a node v_i, where $N(v_i) = \{v_j | \exists (v_i, v_j) \in E\}$ and $|N(v_i)|$ is the cardinality of such a set.

IV. METHOD OVERVIEW

In this section we introduce *STEGON*, our attention-based spatial convolutional graph neural network especially tailored to deal with the land cover mapping task.

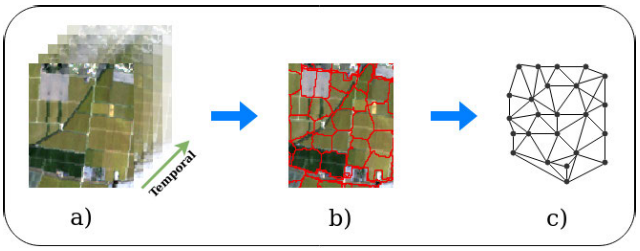


FIGURE 2. The Region Adjacency Graph extraction procedure. Among the Sentinel-2 images time series, a) an image is selected by the expert and b) the selected image is segmented via the SLIC algorithm. Finally, c) the RAG is obtained by the corresponding segmentation.

Figure 3 visually depicts the proposed framework. Given a target segment with its spatial neighborhood set as input, *STEGON* processes on one side the time series information associated to the target segment (the red node) and, on the other side the time series associated to the spatial neighborhood set (the blue nodes). In both cases, a one dimensional convolutional neural network is employed as encoder network to extract the segment embeddings. This encoder network operates on the temporal dimension of the SITS data explicitly modeling the sequential information it contains. For this encoder we use the one dimensional CNN introduced in [23] as embedding extractor for the segment SITS. Considering the neighborhood information, first the same one dimensional CNN model is applied over all the neighborhood segments then, the embeddings of the different segments are aggregated together via a graph attention mechanism [21]. To summarize, in our framework, the temporal and spatial information are not managed simultaneously but, firstly the temporal dynamics is leveraged by means of one dimensional CNN and, subsequently the spatial information is integrated by means of the graph attention mechanism.

At this point, two embeddings are available: the *target segment embedding* that contains information directly related to the time series associated to the target segment, and the *neighborhood embedding* that summarizes knowledge regarding the spatial information surrounding the target segment. Since the two embeddings constitute complementary information that permits to characterize the sample to classify, they are successively combined together by means of a self-attention mechanism [24] providing a new representation, referred to as *combined embedding*. Finally, two fully connected layers are employed on the Combined Embedding to obtain the target segment classification. An additional auxiliary classifier (white box on the right part of Figure 3) is involved in the learning procedure with the aim to directly retro-propagate the gradient error at the level of the *combined embedding* in order to improve the model behavior.

A. TARGET SEGMENT EMBEDDING AND ATTENTIVE SPATIAL NEIGHBORHOOD AGGREGATION

For the *target segment embedding*, that we name as h_{target}, the one dimensional CNN presented in [23] is employed.

As regards the *neighborhood embedding*, we remind that each target segment v_i has an associated neighborhood set $N(v_i)$ with varying size. To aggregate together such varying-size information carried out by $N(v_i)$, we adopt a graph attention mechanism [21] defined as follows:

$$h^i_{neigh} = |N(v_i)| \cdot \sum_{v_j \in N(v_i)} \alpha_{ij} \cdot h_{v_j} \tag{1}$$

where v_j is a segment in the set $N(v_i)$, h_{v_j} is the vector embedding of the segment v_j with dimension d. More precisely, the same one dimensional CNN model, based on the same set of learnable parameters, is employed over all the segments $v_j \in N(v_i)$. The segment embedding h_{v_j} is multiplied by the attention coefficient α_{ij} that weights the contribution of the segment $v_j \in N(v_i)$ in the spatial neighborhood aggregation. The aggregation is a convex combination of the contributions of each $v_j \in N(v_i)$ since $\sum_{j=1}^{|N(v_i)|} \alpha_{ij} = 1$.

The result from the aggregation is finally multiplied by the cardinality of $N(v_i)$. This is done to cope with the fact that the original graph attention mechanism fails to distinguish certain structures that can be distinguishable only by considering the cardinality of the $N(v_i)$ set. For this reason, directly taking into account this information in the analysis mitigates such phenomena and increases the discriminative power of the graph attention mechanism as discussed in [25].

Regarding the attention parameters α, following [21], we can define a generic α_{ij} as follows:

$$\alpha_{ij} = \frac{exp(LeakyReLU(a^T[Wh_{v_i}||Wh_{v_j}]))}{\sum_{v_k \in N(v_i)} exp(LeakyReLU(a^T[Wh_{v_i}||Wh_{v_k}]))} \tag{2}$$

where matrix $W \in \mathbb{R}^{d,d}$ and vectors $a \in \mathbb{R}^{2*d}$ are parameters learned during the process. *LeakyReLU* is the Leaky ReLU non-linear activation function and the $||$ symbol represents matrix concatenation. The *LeakyReLU* activation function is adopted following the original work on graph attention mechanism proposed in [21].

Here we can observe that, differently from standard attention mechanism exploited in the signal processing fields [24], the computation of α_{ij} is tightly related (or conditioned) to the embedding h_{v_i} of the target segment v_i (Equation 2). This stresses the fact that the attention computation, built upon the underlying graph structure, is contextualized with respect to the target segment information h_{v_i}. The results of this step is the spatial neighborhood embedding of the node v_i, referred as h_{neigh}. We avoid to report the superscript i to lighten the notation for the rest of the explanation. We remind that the attention weight α_{ij} is time independent since our framework, *STEGON*, firstly manages the temporal dynamics by means of one dimensional CNNs and, only subsequently, it deals with the spatial information by means of a dedicated graph attention mechanism.

B. ATTENTIVE COMBINATION OF TARGET AND NEIGHBORHOOD INFORMATION

Once the target segment embedding (h_{target}) and the spatial neighborhood embedding (h_{neigh}) are obtained, they are

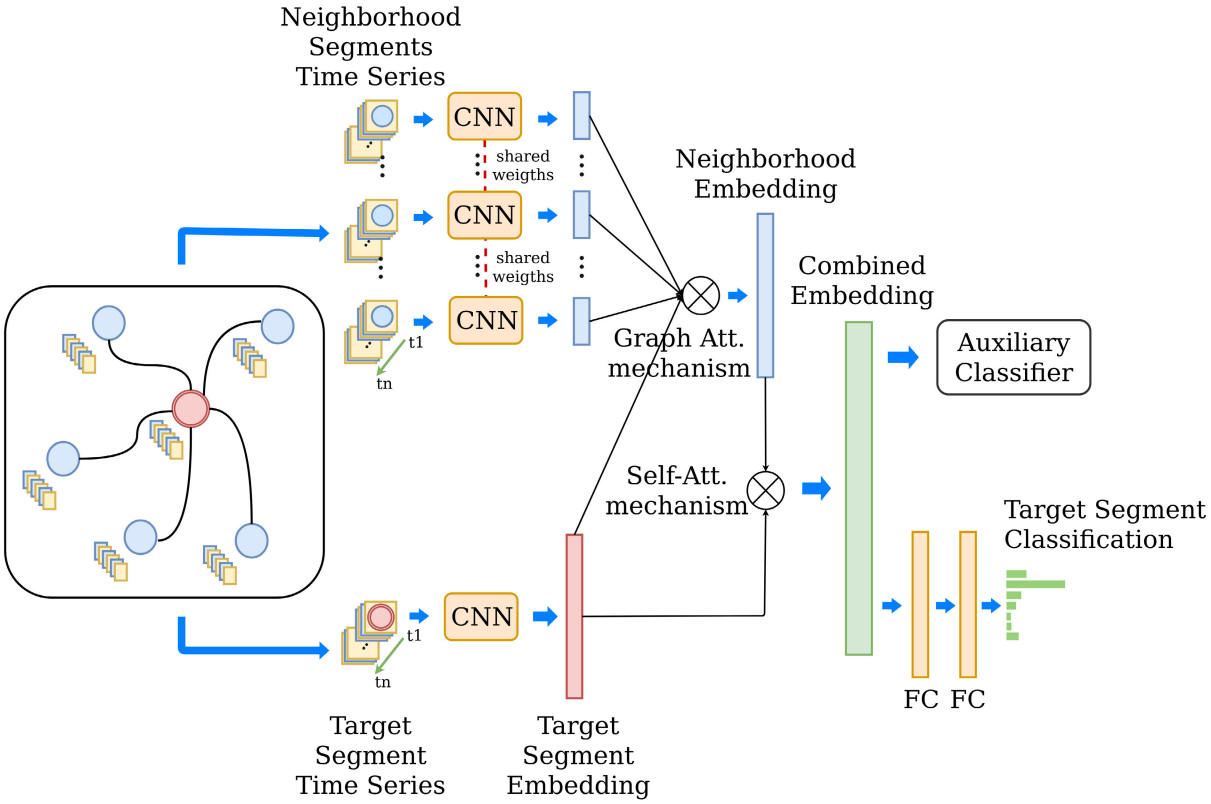


FIGURE 3. The *STEGON* architecture. The model has two branches: the top one is dedicated to analyze the spatial neighborhood information via a graph attention mechanism, while the bottom one is devoted to analyze the target segment time series. The embeddings extracted by the two branches are combined via self-attention to form the combined embedding that is successively used to perform land cover classification. An additional auxiliary classifier is employed to directly retro-propagate the error at the level of the combined embedding so as to increase the discrimination power of the learnt representation.

successively combined by means of a self-attention mechanism [24] with the goal of automatically weighting the contribution of the features extracted from the target segment as well as its spatial neighborhood. The output of this step is a representation which we refer to as $\tilde{h}$. In the case of the combination of h_{target} and h_{neigh}, the attention is not conditioned to any kind of information but it must only combine the target segment embedding and the neighborhood embedding together. To this end, we consider the attention mechanism originally introduced in [24]. Given $H = \{h_{target}, h_{neigh}\}$, we attentively combine these two embeddings as follows:

$$\tilde{h} = \sum_{l \in \{target, neigh\}} \alpha_l \cdot h_l \tag{3}$$

where α_l with $l \in \{target, neigh\}$ is defined as:

$$\alpha_l = \frac{exp(v_a^T \ tanh(W_a \, h_l + b_a))}{\sum_{l' \in \{target, neigh\}} exp(v_a^T \ tanh(W_a \, h_{l'} + b_a))} \tag{4}$$

where matrix $W_a \in \mathbb{R}^{d,d}$ and vectors $b_a, v_a \in \mathbb{R}^d$ are parameters learned during the process. These parameters allow to combine h_{target} and h_{neigh}. The purpose of this procedure is to learn weights α_{target} and α_{neigh}, and estimate the contribution of each of the embedding h_{target} and h_{neigh}. The $SoftMax(\cdot)$ function is used to normalize weights α so that their sum is equal to 1. The result of this attention-based aggregation is the final embedding $\tilde{h}$ that integrates the information related to the SITS associated to the target segment as well as the information available in the neighborhood segments SITS of $N(v_i)$ set.

C. CLASSIFICATION STEP AND TRAINING PROCEDURE

The representation $\tilde{h}$ obtained by the attentive aggregation of the target segment and its neighborhood is processed by means of two fully connected (FC) layers so as to classify the target segment. In our context we use two fully connected layers, each consisting of 512 neurons. Each FC layer is associated to a ReLU non-linearity and a batch normalization layer in order to avoid weight oscillation and ameliorate network training:

$$Cl(\tilde{h}) = SoftMax(W_3 BN(ReLU(W_2 \\ (BN(ReLU(W_1 \tilde{h} + b_1))) + b_2)) + b_3) \tag{5}$$

where W_1, W_2, W_3, b_1, b_2 and b_3 are parameters learnt by the model to process the attentive combined representation $\tilde{h}$, with $W_3 \in \mathbb{R}^{d,|Y|}$ and $b_3 \in \mathbb{R}^{|Y|}$ the parameters associated to the output layers, thus showing a dimension equal to the number of classes to predict. The model training is performed end-to-end. Due to the fact that our classification is

multi-class, we adopt standard categorical cross-entropy (CE) as cost function.

We have empirically observed that optimizing only categorical cross-entropy by considering the output of the classification layer does not allow the network to learn effective representations for the classification task, especially in the case of small size benchmark. This is due to the way in which the gradient flow back in the network and how the network parameters are updated. For this reason, we have introduced an additional auxiliary classifier to directly retropropagate error at the attentive aggregation level. Such auxiliary classifier is only considered at training time and it is defined as follows:

$$Cl^{aux}(\tilde{h}) = SoftMax(W_3'\tilde{h} + b_3') \tag{6}$$

where W_3' and b_3' are the learnt parameters that allow us to map $\tilde{h}$ to the auxiliary classification output.

The final loss function employed to learn the whole set of parameters associated to *STEGON* is defined as:

$$L = CE(Y, Cl) + \lambda CE(Y, Cl^{aux}) \tag{7}$$

where $\lambda \in [0, 1]$ is an hyper-parameter that control the importance of the auxiliary classification in the learning process. We empirically set the value of such hyper-parameter to 0.5. We remind that, at inference time, the output of the auxiliary classifier $Cl^{aux}(\tilde{h})$ is discarded and only the decision obtained via the $Cl(\tilde{h})$ classifier is considered.

D. ARCHITECTURE DETAILS OF THE ONE DIMENSIONAL CONVOLUTIONAL NEURAL NETWORK

The One Dimensional Convolutional Neural Networks we leverage in our experimental evaluation is reported in Table 1. We follow general principles applied in the design of Convolutional Neural Networks [26], where the number of filters along the network structure grows and the convolutional operations are followed by Rectifier Linear Unit (ReLU), Batch Normalization and Dropout. Our CNN1D is composed by ten blocks. The first eight blocks include parameters associated to Convolutional and Batch Normalization operations. The last two blocks do not have parameters, since they consist in a Concatenation and Global Average pooling layer, respectively. We adopt filters with a kernel size equals to 3, except for block 7 and block 8 where convolution with $k = 1$ are employed with the objective to learn per-feature combinations. The ninth block concatenates the outputs of blocks 7 and 8 along the filter dimension and the tenth block computes the Global Average Pooling with the aim to extract one value for each feature map by means of average aggregation.

TABLE 1. Architectures of the One Dimensional Convolutional Neural Network (CNN1D) where *nf* are the number of filters, *k* is the one dimensional kernel size, *s* is the value of the stride while *act* is the nonlinear activation function.

CNN1D	
Block 1	Conv(nf=256, k=3, s=1, act=ReLU) BatchNormalization() DropOut()
Block 2	Conv(nf=256, k=3, s=1, act=ReLU) BatchNormalization() DropOut()
Block 3	Conv(nf=256, k=3, s=1, act=ReLU) BatchNormalization() DropOut()
Block 4	Conv(nf=256, k=3, s=1, act=ReLU) BatchNormalization() DropOut()
Block 5	Conv(nf=512, k=3, s=2, act=ReLU) BatchNormalization() DropOut()
Block 6	Conv(nf=512, k=3, s=1, act=ReLU) BatchNormalization() DropOut()
Block 7	Conv(nf=512, k=1, s=1, act=ReLU) BatchNormalization() DropOut()
Block 8	Conv(nf=512, k=1, s=1, act=ReLU) BatchNormalization() DropOut()
Block 9	Concatenation(Block 7, Block 8)
Block 10	GlobalAveragePooling()

V. SATELLITE IMAGE TIME SERIES DATA AND GROUND TRUTH

The analysis is carried out on the *Reunion Island* dataset (a French Overseas department located in the Indian Ocean) and the *Dordogne* dataset (a French department located in the Southwest). The *Reunion Island* (resp. *Dordogne*) dataset consists of a time series of 21 (resp. 23) Sentinel-2[1] images acquired between March and December 2017 (resp. between January and December 2016). All the Sentinel-2 images we used are those provided at level 2A by the THEIA pole[2] and preprocessed in surface reflectance via the *MACCS-ATCOR Joint Algorithm* [27] developed by the National Centre for Space Studies (CNES). For all the Sentinel-2 images we only considers band at 10m: B2,B3,B4 and B8 (resp. Blue, Green, Red and Near-Infrared). A preprocessing was performed to fill cloudy observations through a linear multi-temporal interpolation over each band (cfr. *Temporal Gapfilling*, [14]). Two additional indices: NDVI[3] (Normalized Difference Vegetation Index) and NDWI[4] (Normalized Difference Water Index), are also calculated. Finally, each Sentinel-2 image has a total of six channels. The spatial extent of the *Reunion Island* dataset is 6 656 × 5 913 pixels corresponding to 3 935 Km^2 whereas the extent of the *Dordogne* site 5 578 × 5 396 pixels corresponding to 3 010 Km^2. Figure 4 and Figure 5 depicts the *Reunion Island* and *Dordogne* study site, respectively, with the associated ground truth polygons.

As regards the *Reunion Island* dataset [28], the ground truth (GT) was built from various sources: (i) the Registre Parcellaire Graphique (RPG)[5] reference data for 2014, (ii) GPS records from June 2017 and (iii) visual interpretation of very high spatial resolution (VHSR) SPOT6/7 images (1,5-m)

[1] https://en.wikipedia.org/wiki/Sentinel-2
[2] Data are available at http://theia.cnes.fr
[3] https://en.wikipedia.org/wiki/Normalized_difference_vegetation_index
[4] https://en.wikipedia.org/wiki/Normalized_difference_water_index
[5] RPG is a part of the European Land Parcel Identification System (LPIS), provided by the French Agency for services and payment

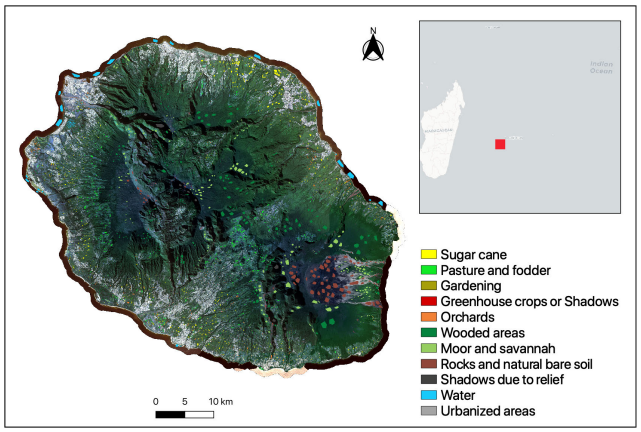


FIGURE 4. Location of the reunion island study site. The RGB composite is a SPOT6/7 image upscaled at 10-m of spatial resolution. The corresponding ground truth polygons are overlaid on the image.

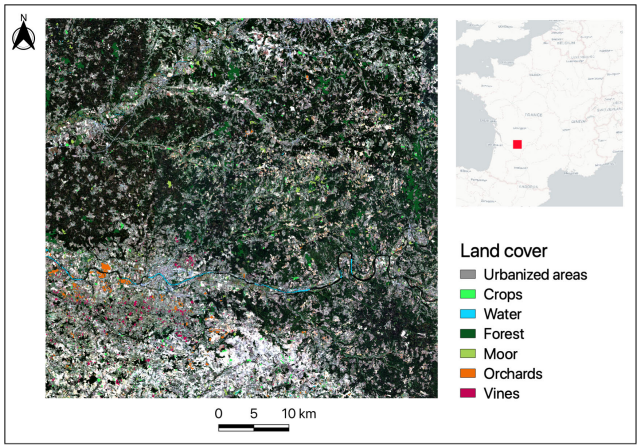


FIGURE 5. Location of the Dordogne study site. The RGB composite is a Sentinel-2 image belonging to the considered time series acquired on September 28, 2016. The ground truth polygons are overlaid on the image.

completed by a field expert with the knowledge of territory to distinguish natural and urban areas.

Regarding the *Dordogne* dataset [29], the GT was obtained via (i) the Registre Parcellaire Graphique (RPG) reference data for 2014 as the Reunion Island dataset and (ii) the Topographic database (BD-TOPO)[6] provided by the French National Geographic Institute (IGN). For both datasets, the GT comes in GIS vector file format containing a collection of polygons each attributed with a unique land cover class label.

In addition, to ensure a precise spatial matching with image data, all geometries have been suitably corrected by hand using the corresponding Sentinel-2 images as reference. Successively, the GIS vector file containing the polygon information has been converted in raster format at the Sentinel-2 spatial resolution (10m). Table 3 and Table 2 report the ground truth information of the *Reunion Island* and *Dordogne* study site, respectively,

[6]https://fr.wikipedia.org/wiki/BD_TOPO

TABLE 2. Characteristics of the Dordogne site ground truth.

Label	# Objects
Built up	849
Crops	1 554
Water	1 217
Forest	2 703
Moor	1 108
Orchards	1 099
Vines	1 389
	9 919

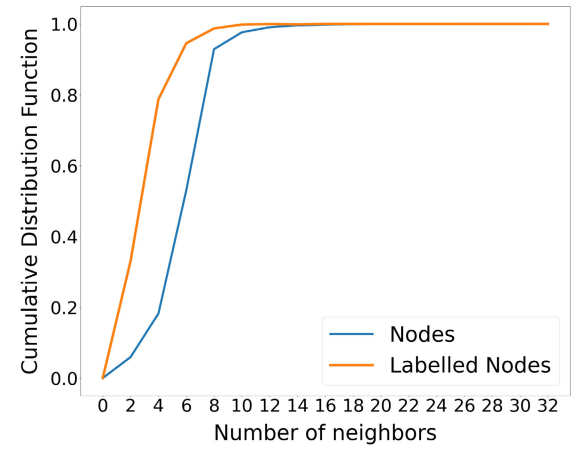
TABLE 3. Characteristics of the Reunion-Island site ground truth.

Label	# Objects
Sugar cane	2 190
Pasture and fodder	1 565
Market gardening	1 284
Greenhouse crops	339
Orchards	1 563
Wooded areas	2 741
Moor and Savannah	2 169
Rocks and bare soil	1 687
Relief shadows	560
Water	873
Urbanized areas	1 540
Total	16 511

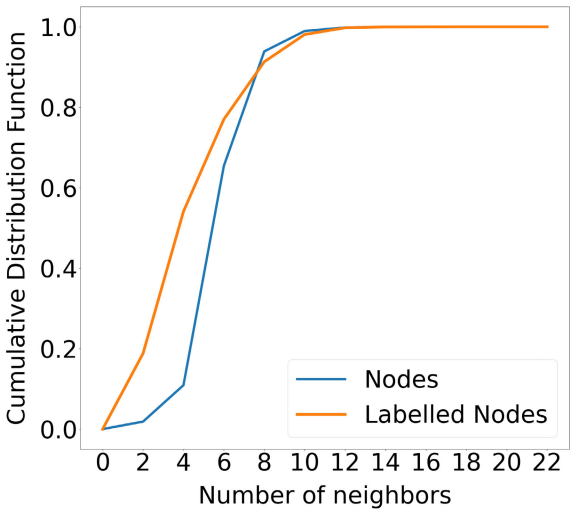
To analyze data at object level and exploit their spatial context, a segmentation was provided by field experts on each study site, according to one of the Sentinel-2 images of the time series that they considered pertinent for the adopted nomenclature. The selected image was segmented using the SLIC algorithm [8] available via the scikit-image toolkit [30]. The parameters were adjusted so that the segments obtained fit, as close as possible, the field plot boundaries. From the segmentation, a RAG (Region Adjacency Graph) is built with the aim of highlighting spatial links explicitly and identifying the spatial context (the direct neighbors) of each object clearly. The RAG extraction procedure is visually depicted in Figure 2.

A. REGION ADJACENCY GRAPH STATISTICS

The *Reunion-Island* dataset's RAG has 59 335 nodes, 186 374 edges and an average degree of 6.28, whereas the *Dordogne* dataset's RAG has 103 514 nodes, 315 027 edges and an average degree of 6.08. Moreover, considering the adjacency graph, Figures 6a and 6b show the cumulative distribution functions (in terms of number of neighbors) for the *Reunion-Island* and *Dordogne* study site, respectively, distinguishing between the whole set of segments (*Nodes*) and the set of segments with associated label information (*Labeled Nodes*). We can observe that the RAGs corresponding to the study sites have some differences, the maximum neighborhood size on the *Reunion-Island* is equal to 32 while on *Dordogne* study site the maximum neighborhood has a size equals to 21. We can also note that the majority (at least 98% of the objects considering both distributions) has a number of neighbors no greater than 8 for the



(a)



(b)

FIGURE 6. **Cumulative degree distribution of the Region Adjacency Graph induced by the segmentation of the two study area: (a) *Reunion Island* and (b) *Dordogne*. *Nodes* curve represents the whole set of segments of the RAG, while *Labeled Nodes* curve only covers the set of segments with associated ground truth information.**

Reunion-Island study site while, considering the *Dordogne* benchmark, the same value is no greater than 10.

VI. EXPERIMENTS

To assess the quality of *STEGON*, we select a panel of competitors exhibiting different and complementary characteristics:

- Random Forest (**RF** [14]). This classifier is commonly employed in the field of remote sensing for satellite image time series classification.
- Multi Layer Perceptron (**MLP**). This is a simple neural network model with two hidden fully-connected layers with 512 neurons each and ReLU activation function. Each fully-connected layer is followed by a batch normalization and a dropout layer.
- One dimensional CNN (**CNN1D**). In this strategy, the convolution operation is applied on the time dimension to model the sequential information of the satellite image time series data. This method can also be considered as an ablation of *STEGON* where the spatial neighborhood is not considered.
- Long-Short Term Memory (**LSTM** [31]). This is a deep learning method based on the Recurrent Neural Network (RNN) philosophy, where the temporal dimension is explicitly managed via an internal gated mechanism.
- Gated Recurrent Unit model (**GRU** [32]). Another RNN unit that differs from the LSTM approach due to a reduced number of parameters and competitive performances obtained in different sub-fields of signal processing.
- Hierarchical object based RNN (**Hob2srnn**) proposed in [33]. Such a method is an extension of the Recurrent Neural Network architecture, equipped with an attention mechanism, developed in the context of object-based satellite image time series classification. Originally, the approach was proposed for hierarchical classification of multi-source SITS data, here, we adapt the method to fit our scenario.
- Temporal Convolutional Neural Network (**TempCNN**) introduced in [6]. This approach was recently introduced in the field of remote sensing to deal with the task of satellite image time series classification. More in detail, it is based on temporal convolutional neural network in which the convolutional operator is deployed over the time dimension.
- Graph Attention Network (**GAT** [21]). This strategy adopts the model proposed in [21] and it only considers the neighborhood information to classify the target satellite image time series information. This method can be seen as an ablation of *STEGON* where the target segment time series branch is discarded.
- $STEGON_{noGAT}$. This strategy is an ablation of the proposed *STEGON* where the contribution of the spatial context is not aggregated via an attention mechanism but each neighbor contributes uniformly to define the neighborhood embedding.
- $STEGON_{noAux}$. This strategy is another variant of *STEGON*, where the auxiliary classifier is discarded. More in detail, concerning the loss function in Equation 7, the *STEGON* approach is trained with a value of λ equals to 0.

LSTM, **GRU**, **CNN** and **GAT** are associated to a multi layer perceptron block (like the one previously described) to perform SITS object classification; both **LSTM** and **GRU** have a dimensionality of 512 hidden units. All the competitors are evaluated under the object-based image analysis framework. For each study site, we split the corresponding data into three parts: training, validation and test set, with an object

proportion of 50%, 20% and 30% respectively. Training data are used to learn the model, while validation data are exploited for model selection. Finally, the model that achieves the best performance on the validation set is successively employed to perform the classification on the test set. For time series data, all values are normalized per band in the interval [0, 1].

The *RF* classifier is optimized by tuning two parameters: the maximum depth of each tree and the number of trees in the forest. We let the former parameter vary in the range {20,40,60,80,100}, while for the latter we take values in the set {100, 200, 300,400,500}. All the deep learning models are trained using the Adam optimizer with a learning rate equal to 1×10^{-4}. The training process, for each model, is conducted over 3 000 epochs with a batch size equals to 32. The dropout parameter, for the training stage, is equal to 0.4. In addition, in the graph attention mechanism, as previously done in [21], we use the LeakyReLU non linear activation function [34] with a slope equals to 0.3. For the *STEGON* model, we consider a maximum neighborhood size equal to 8 (resp. 10) for the *Reunion Island* (resp. *Dordogne*) study site according to the cumulative distribution function on the neighborhood size reported in Section V. When a segment has a number of neighbors bigger than the maximum neighborhood size, we pick at random 8 (resp. 10) neighbors at each training epoch for the *Reunion Island* (resp. *Dordogne*) study site. The total number of trainable parameters for our approach is 6 481 934.

The assessment of the model performances are done considering *Accuracy*, *F-Measure* and *Kappa* metrics [35]. To reduce the bias induced by the train/ validation/ test split procedure all the results are averaged over five different random splits.

Experiments are carried out on a workstation with an Intel(R) Xeon(R) W-2133 CPU@3.60GHz with 64 GB of RAM and a GTX1080ti GPU. All the deep learning methods are implemented using the Python Tensorflow library.

A. QUANTITATIVE EXPERIMENTAL RESULTS

Table 4 and Table 5 summarize the average quantitative performances obtained by the different competing approaches on the Reunion-Island and Dordogne datasets, respectively. We can note that, according to the three evaluation metrics, *STEGON* clearly outperforms all the competitors on both study sites. Regarding the best competing method (CNN), our framework achieves more than 5 points of gain in F-Measure, demonstrating the added value derived by integrating the spatial surrounding information for the land cover mapping task. In addition, the comparison between *STEGON* and its ablations demonstrates that: i) automatically weighting, by means of the attention mechanism, the importance of the spatial context is more effective than an uniform weighting of all the neighborhood objects (*STEGON* vs $STEGON_{noGAT}$) and ii) integrating the auxiliary classifier into the training stage (i.e., in order to directly retropropagate the error at the attentive aggregation level) allows to further improve the behavior of our framework (*STEGON* vs $STEGON_{noAux}$).

TABLE 4. F-Measure, Kappa and Accuracy performances of all the competing approaches on the Reunion-Island dataset.

Method	*F-Measure*	*Kappa*	*Accuracy*
RF	85.49 ± 0.18	83.98 ± 0.19	85.78 ± 0.16
MLP	78.63 ± 0.31	76.55 ± 0.34	79.19 ± 0.34
CNN	88.15 ± 0.12	85.70 ± 0.14	88.17 ± 0.12
LSTM	87.46 ± 0.24	86.02 ± 0.31	87.56 ± 0.27
GRU	87.61 ± 0.14	86.15 ± 0.14	87.68 ± 0.12
Hob2srnn	82.17 ± 0.33	80.30 ± 0.35	82.49 ± 0.31
TempCNN	83.43 ± 0.20	81.61 ± 0.22	83.67 ± 0.19
GAT	90.64 ± 0.19	89.68 ± 0.21	90.83 ± 0.19
$STEGON_{noGAT}$	91.77 ± 0.00	90.78 ± 0.00	91.80 ± 0.00
$STEGON_{noAux}$	93.08 ± 0.24	92.27 ± 0.27	93.12 ± 0.24
$STEGON$	**94.30 ± 0.16**	**93.63 ± 0.14**	**94.34 ± 0.12**

TABLE 5. F-Measure, Kappa and Accuracy performances of all the competing approaches on the Dordogne dataset.

Method	*F-Measure*	*Kappa*	*Accuracy*
RF	80.98 ± 0.30	77.41 ± 0.37	81.23 ± 0.31
MLP	85.54 ± 0.48	82.58 ± 0.65	85.54 ± 0.56
CNN	86.07 ± 0.30	83.30 ± 0.35	86.08 ± 0.29
LSTM	84.81 ± 0.57	81.73 ± 0.63	84.74 ± 0.50
GRU	85.22 ± 0.25	82.28 ± 0.31	85.23 ± 0.25
Hob2srnn	83.18 ± 0.55	79.85 ± 0.63	83.21 ± 0.52
TempCNN	85.19 ± 0.14	82.36 ± 0.15	85.32 ± 0.13
GAT	81.74 ± 0.31	78.59 ± 0.28	82.22 ± 0.22
$STEGON_{noGAT}$	88.58 ± 0.00	86.31 ± 0.00	88.58 ± 0.00
$STEGON_{noAux}$	91.22 ± 0.33	89.51 ± 0.39	91.26 ± 0.33
$STEGON$	**91.98 ± 0.33**	**90.42 ± 0.39**	**92.01 ± 0.32**

Table 6 and Table 7 report the per class F-Measure obtained by the different competing methods on the *Reunion-Island* and the *Dordogne* study site, respectively.

Regarding the *Reunion-Island* study site (Table 6), we can observe that *STEGON* consistently outperforms all the competitors on all the land cover classes. The highest gains are associated to the *Greenhouse* and *Orchards* land cover classes with an improvement of almost 30 points and 12 points, respectively, over the CNN approach. Note that the CNN approach can be seen as an ablation of our approach that does not consider spatial neighborhood information. It can also be noted how *STEGON* always improves upon the performance of its other ablations ($STEGON_{noGAT}$ and $STEGON_{noAux}$). This confirms how the combination of the graph attention mechanism and the auxiliary classifiers both significantly contribute to the performance of the proposed approach.

Concerning the *Dordogne* study site (Table 7), we can see that *STEGON* achieves the best performances on 5 over 7 classes. On the two remaining classes it is the second best method, showing comparable performances with respect to the best performing one (*MLP*). It can be noted how also in this case *STEGON* outperforms all its ablation. Similarly to what happens on *Reunion-Island*, also on *Dordogne* $STEGON_{noAux}$ always outperforms $STEGON_{noGAT}$, confirming the importance of the graph attention mechanism.

TABLE 6. Per class F-Measure performances of the different competing methods considering the ***REUNION*** study site. Best and second best performances are shown in bold face and underlined, respectively.

	Sugar Cane	Pasture	Market g.	Greenhouse	Orchards	Wooded areas	Moor	Rocks	Relief s.	Water	Urb. areas
RF	87.35	86.33	74.43	34.13	72.64	90.39	87.42	87.16	**97.95**	88.70	78.78
MLP	89.33	85.11	74.80	32.41	71.50	83.77	81.67	84.17	96.26	92.85	78.77
CNN	90.01	87.31	78.24	35.67	78.81	90.35	88.48	89.35	97.09	89.56	79.58
LSTM	90.22	85.64	76.78	33.58	78.58	88.34	87.55	88.04	92.21	91.55	78.36
GRU	90.33	86.28	76.34	33.62	78.97	89.18	88.43	87.05	92.81	88.72	78.63
Hob2srnn	87.51	79.48	67.69	28.62	69.62	88.34	85.25	91.39	97.06	93.25	76.74
TempCNN	88.33	82.79	74.05	32.34	72.74	88.42	84.56	91.43	97.49	94.74	76.21
GAT	83.70	82.80	74.75	50.35	75.57	90.16	89.26	93.25	93.00	89.37	78.82
$STEGON_{noGAT}$	93.73	91.35	81.99	57.01	86.26	96.12	95.36	97.03	98.03	95.10	85.33
$STEGON_{noAux}$	94.32	93.45	85.36	61.03	87.88	97.12	97.06	97.66	98.21	96.89	87.78
$STEGON$	**95.46**	94.45	**87.90**	**64.85**	**91.03**	**97.43**	**97.58**	**98.54**	**98.45**	**97.08**	**89.65**

TABLE 7. Per class F-Measure performances of the different competing methods considering the ***Dordogne*** study site. Best and second best performances are shown in bold face and underlined, respectively.

	Built Up	Crops	Water	Forest	Moor	Orchards	Vines
RF	84.38	79.72	86.74	87.27	65.02	68.82	85.34
MLP	**92.02**	81.48	**93.83**	88.93	70.05	78.32	90.31
CNN	87.22	84.99	91.29	90.29	74.83	79.22	88.13
LSTM	89.41	79.87	91.08	89.37	72.19	76.71	89.63
GRU	88.99	81.70	90.91	90.33	72.50	77.75	87.95
Hob2srnn	85.28	82.66	87.55	88.15	70.47	73.23	87.00
TempCNN	88.44	84.28	91.15	89.35	71.15	77.25	88.37
GAT	58.58	74.99	70.87	93.66	85.93	79.74	88.03
$STEGON_{noGAT}$	87.51	85.27	91.49	93.83	82.58	82.40	88.92
$STEGON_{noAux}$	89.62	87.81	93.21	95.98	89.16	84.75	91.75
$STEGON$	90.91	**88.19**	**93.36**	**96.31**	**90.90**	**86.02**	**92.85**

B. SENSITIVITY ANALYSIS REGARDING THE ACTIVATION FUNCTION FOR THE GRAPH ATTENTION MECHANISM AND THE HYPER-PARAMETER λ

In this section, we evaluate the sensitivity of our framework *STEGON* w.r.t. two different internal aspects: i) the activation function associated to the graph attention mechanism employed to aggregate the spatial information and ii) the impact of the λ parameter associated to the auxiliary classifier cost. For the first evaluation, beyond the *LeakyReLU* activation function (that was originally proposed to equip the graph attention mechanism [21]), we consider as additional option the widely adopted *ReLU* and *Tanh* activation function. The latter one is the activation function that is commonly used in standard attention mechanism [24]. For the second experiments, we vary the λ parameters in the set $\{0.1, 0.3, 0.5, 0.7, 0.9\}$.

Table 8 reports the performances of *STEGON* coupled with different activation functions for the graph attention mechanism. The subscript indicates which is the employed activation function. We can observe that, on both benchmarks, the *LeakyReLU* activation function achieves the best average performances in terms of F-Measure, Kappa and Accuracy. These results confirm the choice made by the authors in [21] that equip the graph attention mechanism with the *LeakyReLU* activation function.

TABLE 8. F-Measure, Kappa and Accuracy performances of ***STEGON*** equipped with different activation functions to deal with the computation of the graph attention mechanism.

	Reunion		
Method	***F-Measure***	***Kappa***	***Accuracy***
$STEGON_{ReLU}$	94.21 ± 0.14	93.55 ± 0.15	94.26 ± 0.13
$STEGON_{Tanh}$	93.68 ± 0.17	92.92 ± 0.19	93.70 ± 0.17
$STEGON_{LeakyReLU}$	$\mathbf{94.30 \pm 0.16}$	$\mathbf{93.63 \pm 0.14}$	$\mathbf{94.34 \pm 0.12}$
	Dordogne		
Method	***F-Measure***	***Kappa***	***Accuracy***
$STEGON_{ReLU}$	91.64 ± 0.30	90.01 ± 0.35	91.67 ± 0.29
$STEGON_{Tanh}$	91.21 ± 0.11	89.48 ± 0.14	91.23 ± 0.12
$STEGON_{LeakyReLU}$	$\mathbf{91.98 \pm 0.33}$	$\mathbf{90.42 \pm 0.39}$	$\mathbf{92.01 \pm 0.32}$

Figure 7 depicts the sensitivity analysis of *STEGON* w.r.t. the hyper-parameter λ (the importance related to the auxiliary classifier loss). We can observe that, generally, the proposed approach is quite stable regarding this hyper-parameter. For the case of *Reunion-Island*, the performances, in terms of F-Measure, varying between 93.95 ($\lambda = 0.1$) to 94.30 ($\lambda = 0.5$) while, on the *Dordogne* benchmark, performances varying between 91.56 ($\lambda = 0.9$) to 91.98 ($\lambda = 0.5$).

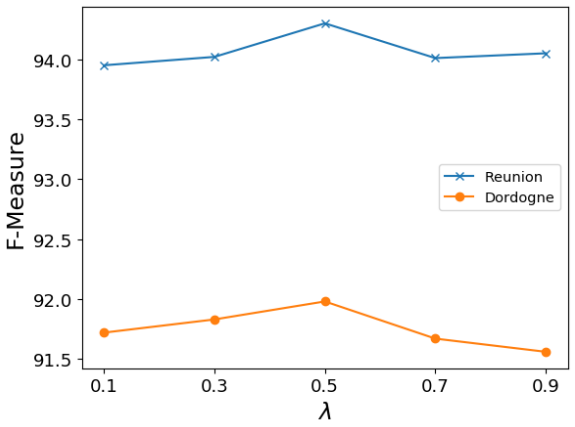


FIGURE 7. Sensitivity analysis of *STEGON* performances varying the λ parameter in the value set {0.1, 0.3, 0.5, 0.7, 0.9} in terms of F-Measure over the two considered benchmarks.

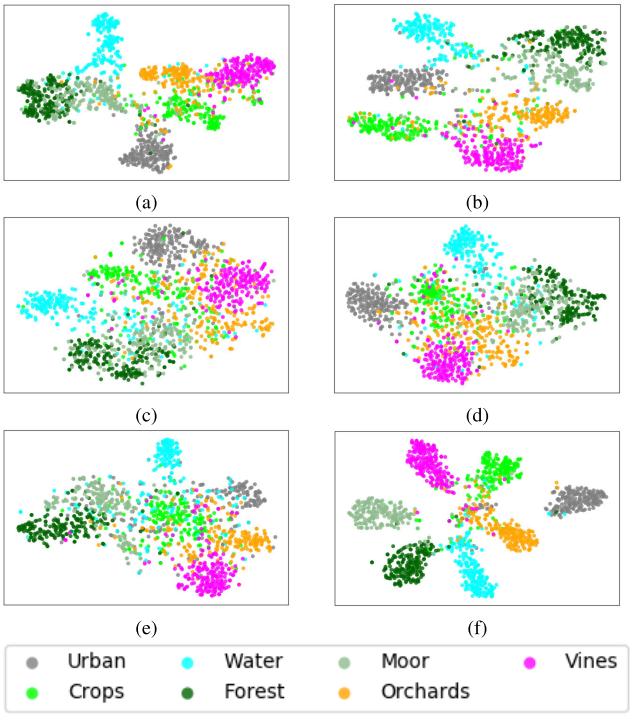


FIGURE 8. Visualization of the embeddings learnt by: (a) MLP (b) CNN (c) LSTM (d) GRU (e) GAT (f) *STEGON*. 300 examples are sampled per class from the test set and the T-SNE is used to obtain the two dimensional projection.

C. QUALITATIVE EXPERIMENTAL RESULTS

To further investigate the behavior of *STEGON*, we conduct two additional qualitative studies. First, we visualize and compare the representation learnt by *STEGON* w.r.t. the representations learnt by some of the competing deep learning methods so as to analyze their internal behavior and, second, we discuss a sample from the land cover map generated by *STEGON*. We deploy such evaluations on the Dordogne study site. Figure 8 depicts the visual projection (via T-SNE [36]) of the embeddings extracted by the competing deep learning approaches. The Dordogne dataset includes

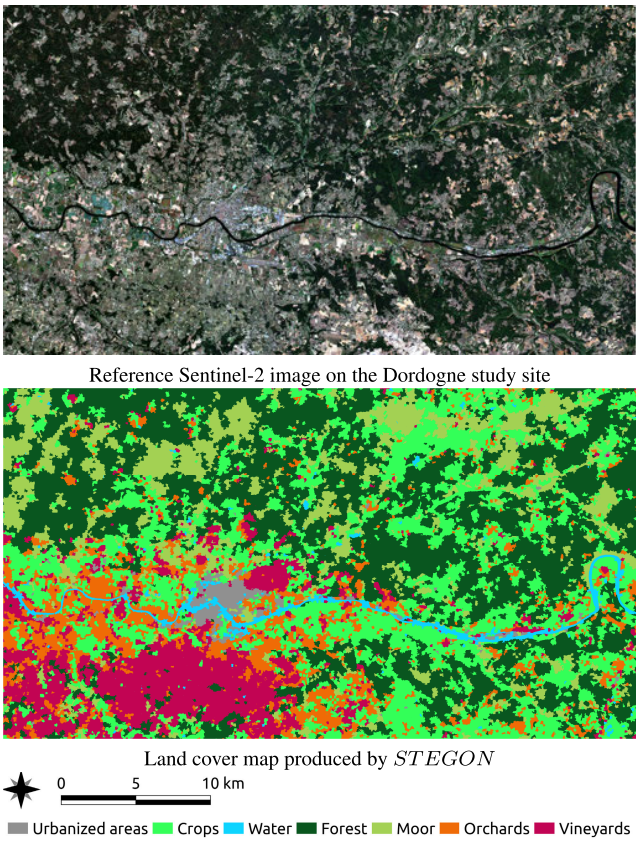


FIGURE 9. Extract of the Sentinel-2 image (top of the image) and corresponding extract of the land cover map generated by *STEGON* with the associated legend (bottom of the image) on a portion of the Dordogne study site.

seven land cover classes, that are listed in the legend of the figure with the associated color mapping. We can note that *STEGON* (Figure 8f) clearly recovers a more visible cluster structure with respect to the one exhibited by the embeddings extracted by the competitors. We can also highlight that the visual class separability (the cluster structure) is directly proportional to the quantitative performance results reported in Table 5. Similar qualitative behaviors are obtained on the Reunion Island study site (results not shown).

Figure 9 depicts the output of our model on a portion of the Dordogne study site. The top of the figure is a reference Sentinel-2 image. Only RGB bands of the image are used for visualization. The bottom of the figure reports the land cover map produced by *STEGON* (and the associated legend) corresponding to the area spanned by the Sentinel-2 extract. We clearly observe that our approach correctly recovers the river stream (blue color) crossing the image from left to right, as well as the urbanized area (gray color) corresponding to a city in the middle left part of the area. Similarly, *STEGON* accurately classifies the vineyard area (magenta color) in the bottom left part of the image.

To sum up, both quantitative and qualitative evaluations demonstrate the effectiveness of *STEGON* with respect to state of the art approaches. The obtained findings highlight the benefit of simultaneously taking into account spatial and

temporal information included in SITS data, thus improving the results on the land cover mapping task.

VII. CONCLUSION

Satellite image time series (SITS) data constitutes a valuable source of information to assess the Earth surface dynamics. Applications range from food production estimation to natural resources mapping and biodiversity monitoring. How to get the most out of such rich information source, leveraging simultaneously both spatial and temporal dimensions, is one of the main current challenges in the remote sensing community. To tackle it, in this work we have presented a novel attentive spatial temporal graph convolutional neural network to analyze SITS data in the context of land cover mapping. Our framework is equipped with a spatial attention mechanism that allows the network to automatically aggregate the spatial context information surrounding a target segment. Quantitative and qualitative evaluations demonstrate the effectiveness of our method with respect to state of the art competitors that do not integrate the spatial dimension in their analysis. We also underline that, to the best of our knowledge, this is the first spatial temporal GCNN strategy especially conceived to cope with the specific features characterizing remote sensing data (i.e. scale up to large graphs). Due to the promising performances we have obtained, we hope that this work will stimulate the scientific community to further investigate the interplay between spatial temporal GCNN models and their quality, thus dealing with the analysis of modern remote sensing data.

ACKNOWLEDGMENT

The authors would like to thank SAFER of Reunion Island, the Reunion Island Sugar Union, the DEAL of Reunion Island, the NFB, and the teams of CIRAD research units (AIDA and HortSys) for their participation in the creation of the learning database.

REFERENCES

[1] O. Dubois, J. Faurés, E. Felix, A. Flammini, J. Hoogeveen, L. Pluschke, M. Puri, and O. Ãœnver, "The water-energy-food nexus: A new approach in support of food security and sustainable agriculture," Food Agricult. Org., Rome, Italy, Tech. Rep., 2014. [Online]. Available: http://www.fao.org/policy-support/tools-and-publications/resources-details/en/c/421718/

[2] R. Albarakat, V. Lakshmi, and C. Tucker, "Using satellite remote sensing to study the impact of climate and anthropogenic changes in the mesopotamian marshlands, iraq," *Remote Sens.*, vol. 10, no. 10, p. 1524, Sep. 2018.

[3] L. Chen, Z. Jin, R. Michishita, J. Cai, T. Yue, B. Chen, and B. Xu, "Dynamic monitoring of wetland cover changes using time-series remote sensing imagery," *Ecol. Informat.*, vol. 24, pp. 17–26, Nov. 2014.

[4] S. Olen and B. Bookhagen, "Mapping damage-affected areas after natural hazard events using sentinel-1 coherence time series," *Remote Sens.*, vol. 10, no. 8, p. 1272, Aug. 2018.

[5] D. Derksen, J. Inglada, and J. Michel, "A metric for evaluating the geometric quality of land cover maps generated with contextual features from high-dimensional satellite image time series without dense reference data," *Remote Sens.*, vol. 11, no. 16, p. 1929, Aug. 2019.

[6] C. Pelletier, G. Webb, and F. Petitjean, "Temporal convolutional neural network for the classification of satellite image time series," *Remote Sens.*, vol. 11, no. 5, p. 523, Mar. 2019.

[7] G. Chen, Q. Weng, G. J. Hay, and Y. He, "Geographic object-based image analysis (GEOBIA): Emerging trends and future opportunities," *GISci. Remote Sens.*, vol. 55, no. 2, pp. 159–182, Mar. 2018.

[8] Z. Li and J. Chen, "Superpixel segmentation using linear spectral clustering," in *Proc. CVPR*, 2015, pp. 1356–1363.

[9] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, Jan. 2021.

[10] X. Kong, W. Xing, X. Wei, P. Bao, J. Zhang, and W. Lu, "STGAT: Spatial-temporal graph attention networks for traffic flow forecasting," *IEEE Access*, vol. 8, pp. 134363–134372, 2020.

[11] Y. Ding, Y. Zhu, J. Feng, P. Zhang, and Z. Cheng, "Interpretable spatio-temporal attention LSTM model for flood forecasting," *Neurocomputing*, vol. 403, pp. 348–359, Aug. 2020.

[12] J. Yang, W.-S. Zheng, Q. Yang, Y.-C. Chen, and Q. Tian, "Spatial-temporal graph convolutional network for video-based person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3286–3296.

[13] L. Morales-Barquero, M. Lyons, S. Phinn, and C. Roelfsema, "Trends in remote sensing accuracy assessment approaches in the context of natural resources," *Remote Sens.*, vol. 11, no. 19, p. 2305, Oct. 2019.

[14] J. Inglada, A. Vincent, M. Arias, B. Tardy, D. Morin, and I. Rodes, "Operational high resolution land cover map production at the country scale using satellite image time series," *Remote Sens.*, vol. 9, no. 1, p. 95, Jan. 2017.

[15] D. Ienco, R. Gaetano, C. Dupaquier, and P. Maurel, "Land cover classification via multitemporal spatial data by deep recurrent neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1685–1689, Oct. 2017.

[16] D. Ho Tong Minh, D. Ienco, R. Gaetano, N. Lalande, E. Ndikumana, F. Osman, and P. Maurel, "Deep recurrent neural networks for winter vegetation quality mapping via multitemporal SAR sentinel-1," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 3, pp. 464–468, Mar. 2018.

[17] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and locally connected networks on graphs," in *Proc. ICLR*, 2014.

[18] W. L. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proc. NIPS*, 2017, pp. 1024–1034.

[19] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Proc. NIPS*, 2016, pp. 3837–3845.

[20] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. ICLR*, 2017.

[21] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. ICLR*, 2018.

[22] R. Ying, R. He, K. Chen, P. Eksombatchai, W. L. Hamilton, and J. Leskovec, "Graph convolutional neural networks for Web-scale recommender systems," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 974–983.

[23] D. Ienco, Y. J. E. Gbodjo, R. Gaetano, and R. Interdonato, "Weakly supervised learning for land cover mapping of satellite image time series via attention-based CNN," *IEEE Access*, vol. 8, pp. 179547–179560, 2020.

[24] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. ICLR*, 2015.

[25] S. Zhang and L. Xie, "Improving attention mechanism in graph neural networks via cardinality preservation," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 1395–1402.

[26] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, "Designing network design spaces," in *Proc. CoRR*, 2020, pp. 10428–10436.

[27] O. Hagolle, M. Huc, D. Villa Pascual, and G. Dedieu, "A multi-temporal and multi-spectral method to estimate aerosol optical thickness over land, for the atmospheric correction of FormoSat-2, landSat, VENμS and sentinel-2 images," *Remote Sens.*, vol. 7, no. 3, pp. 2668–2691, Mar. 2015.

[28] S. Dupuy, R. Gaetano, and L. Le Mezo, "Mapping land cover on reunion island in 2017 using satellite imagery and geospatial ground data," *Data Brief*, vol. 28, Feb. 2020, Art. no. 104934.

[29] Y. J. Eudes Gbodjo, D. Ienco, and L. Leroux, "Toward spatio–spectral analysis of sentinel-2 time series data for land cover mapping," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 307–311, Feb. 2020.

[30] S. van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu, "Scikit-image: Image processing in Python," *PeerJ*, vol. 2, p. e453, Jun. 2014.

[31] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, Oct. 2017.
[32] K. Cho, B. van Merrienboer, C. C. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. EMNLP*, 2014, pp. 1724–1734.
[33] Y. J. E. Gbodjo, D. Ienco, L. Leroux, R. Interdonato, R. Gaetano, and B. Ndao, "Object-based multi-temporal and multi-source land cover mapping leveraging hierarchical class relationships," *Remote Sens.*, vol. 12, no. 17, p. 2814, Aug. 2020.
[34] X. Zhang, Y. Zou, and W. Shi, "Dilated convolution neural network with LeakyReLU for environmental sound classification," in *Proc. 22nd Int. Conf. Digit. Signal Process. (DSP)*, Aug. 2017, pp. 1–5.
[35] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Reading, MA, USA: Addison-Wesley, 2005.
[36] L. van der Maaten, "Accelerating t-SNE using tree-based algorithms," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3221–3245, Oct. 2014.

ALESSANDRO MICHELE CENSI received the B.Sc. degree in computer science from the University of Torino, Torino, Italy, in 2017 and the M.Sc. degree in computer science still from the University of Turin, Turin, Italy, in 2020. In 2020, he was an Intern with INRAE, Montpellier, France, working on deep learning approaches for spatio-temporal data with applications on satellite image time series data.

DINO IENCO (Member, IEEE) received the M.Sc. and Ph.D. degrees in computer science from the University of Torino, Torino, Italy, in 2006 and 2010, respectively.

He joined the TETIS Laboratory, INRAE, Montpellier, France, in 2011, as a Junior Researcher. His main research interests include machine learning, data science, graph databases, social media analysis, information retrieval, and spatio-temporal data analysis with a particular emphasis on remote sensing data and Earth Observation data fusion. He served in the program committee for many international conferences on data mining, machine learning, and database, including IEEE ICDM, ECML PKDD, ACML, and IJCAI, as well as served as a Reviewer for many international journal in the general field of data science and remote sensing.

YAWOGAN JEAN EUDES GBODJO received the M.Sc. degree in geomatics from the University of Jean Jaures, Toulouse, France, in 2018. He is currently pursuing the Ph.D. degree in computer science with the UMR TETIS Laboratory, INRAE, working on machine learning approaches devoted to manage multi-source remote sensing data for agriculture monitoring systems.

RUGGERO GAETANO PENSA received the M.Sc. degree in computer engineering from the Politecnico of Turin, Italy, in 2003, and the Ph.D. degree in computer science from INSA, Lyon, France, in 2006. He is currently an Associate Professor with the Department of Computer Science, University of Turin, Italy. His main research interests include machine learning, data science, privacy-preserving algorithms for data management and mining, social network analysis, and spatio-temporal data analysis. He served in the program committee of many international conferences on data mining and machine learning, including IEEE ICDM, ACM CIKM, SIAM SDM, IJCAI, AAAI, and ECML PKDD. He is a member of the Editorial Board of the *Data Mining and Knowledge Discovery Journal* and the *Machine Learning Journal*, and the Area Chair of ECML PKDD.

ROBERTO INTERDONATO received the Ph.D. degree in computer engineering the University of Calabria, Italy, in 2015. He was a Postdoctoral Researcher with the University of La Rochelle, France, Uppsala University, Sweden, and with the University of Calabria. His Ph.D. work focused on novel ranking problems in information networks. He is currently a Research Scientist with Cirad, UMR TETIS, Montpellier, France. His research interests include topics in data mining and machine learning applied to complex networks analysis (e.g., social media networks, trust networks, semantic networks, and bibliographic networks) and to remote sensing analysis. On these topics, he has coauthored journal articles and conference papers, organized workshops, presented tutorials at international conferences, and developed practical software tools.

RAFFAELE GAETANO received the Laurea (M.S.) degree in computer engineering and the Ph.D. degree in electronic and telecommunication engineering from the University of Naples Federico II, Naples, Italy, in 2004 and 2009, respectively.

He has been a European Research Consortium for Informatics and a Mathematics Postdoctoral Fellow of both the ARIANA Team of INRIA Sophia Antipolis and the DEVA Team of SZTAKI, Research Institute of the Hungarian Academy of Sciences. From 2010 to 2015, he conducted a Postdoctoral Research on fundamental image processing with the Multimedia Group of Telecom Paristech, Paristech, France, then with the Research Group on Image Processing, Department of Electric and Information Technology Engineering, University of Naples Federico II. Since 2015, he has been a Permanent Researcher with CIRAD, TETIS Research Unit. His current research interests include machine learning for remote sensing image analysis and processing, mainly focusing on large scale operational methods for information extraction from multisensor imagery.