This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3024133, IEEE Access

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier

Weakly Supervised learning for land cover mapping of satellite image time series via attention-based CNN

DINO IENCO[1,2], YAWOGAN JEAN EUDES GBODJO[1], RAFFAELE GAETANO[3], and ROBERTO INTERDONATO[3]

[1]INRAE, UMR TETIS, Univ. of Montpellier, France
[2]LIRMM, Univ. of Montpellier, France
[3]CIRAD, UMR TETIS, Montpellier, France

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

ABSTRACT The unprecedented possibility to acquire high resolution Satellite Image Time Series (SITS) data is opening new opportunities to monitor the different aspects of the Earth Surface but, at the same time, it is raising up new challenges in term of suitable methods to analyze and exploit such huge amount of rich image data. One of the main tasks associated to SITS data analysis is related to land cover mapping. Due to operational constraints, the collected label information is often limited in volume and obtained at coarse granularity level carrying out inexact and weak knowledge that can affect the whole process.

To cope with such issues, in the context of object-based SITS land cover mapping, we propose a new deep learning framework, named $TASSEL$ (a**T**tentive we**A**kly **S**upervised **S**atellite image time s**E**ries c**L**assifier), to deal with the weak supervision provided by the coarse granularity labels. Our framework exploits the multifaceted information conveyed by the object-based representation considering object components instead of aggregated object statistics. Furthermore, our framework also produces an additional outcome that supports the model interpretability.

Quantitative and qualitative experimental evaluations are carried out on two real-world scenarios. Results indicate that not only $TASSEL$ outperforms the competing approaches in terms of predictive performances, but it also produces valuable extra information that can be practically exploited to interpret model decisions.

INDEX TERMS Weakly Supervised Learning, Object-based image classification, Satellite Image Time Series, Land Cover classification, Deep learning

I. INTRODUCTION

Nowadays, modern Earth observation systems continuously collect massive amounts of satellite information that can be referred to as Earth Observation (EO) data. A notable example is represented by the Sentinel-2 mission[1] from the Copernicus programme, supplying optical information with a revisit time period between 5 and 10 days thanks to a constellation of two twin satellites. Due to the high revisiting period exhibited by such satellites, the acquired images can be organized in Satellite Image Time Series (SITS), which represent a practical tool to monitor a particular spatial area through time. SITS data can support a wide number of application domains like ecology [1], agriculture [2], mobility, health, risk assessment [3], land management planning [4], forest [5] and natural habitat monitoring [6] and, for this reason, they constitute a valuable source of information to follow the dynamic of the Earth Surface. The huge amount of regularly acquired SITS data opens new challenges in the field of remote sensing in relationship with the way the knowledge can be effectively extracted and how spatio-temporal interplay can be exploited to get the most out of such rich information source.

One of the main tasks related to SITS data analysis is associated to land cover mapping, where a predictive model is learnt to make the connection between satellite data (i.e., SITS) and the associated land cover classes [4]. SITS data captures the temporal dynamics exhibited by land cover classes, thus supporting a more effective discrimination among them [7].

Despite the increasing necessity to provide large scale (i.e.,

[1]https://sentinel.esa.int/web/sentinel/missions/sentinel-2

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3024133, IEEE Access

region or national) land cover maps, the amount of labeled information collected to train such models is still limited, sparse (annotated polygons are scattered all over the study site) and, most of the time, at coarser scale with respect to pixel precision. This is due to the fact that the labeling task is generally labour-intensive and time costly in order to cover a sufficient number of samples with respect to the extent of the study site.

Object Based Image Analysis (OBIA) [8] refers to a category of digital remote sensing image analysis approaches that study geographic entities, or phenomena through delineating and analyzing image-objects rather than individual pixels [9]. When dealing with supervised Land Use / Land Cover (LULC) classification, the recur to OBIA approaches is motivated by the fact that, in modern remote sensing imagery, most of the common land cover classes present an heterogeneous radiometric composition, and classical pixel-based approaches typically fail to capture such complexity. Of course, this effect is even more important when the aforementioned complexity is exhibited also in the temporal dimension, which is the case for SITS data.

To address this issue, in the OBIA framework, the main idea is to group adjacent pixels together prior to the classification process, and subsequently work on the so-obtained object layer in which segments correspond to more representative samples of such complex LULC classes (e.g. "land units") [10]. This is typically achieved by tuning the segmentation algorithms to provide object layers at an appropriate spatial scale, at which objects are generally not radiometrically homogeneous, especially on the most complex LULC classes. Matter of facts, most of the common segmentation techniques used in remote sensing allow for the parametrization of the spatial scale [11], e.g. by using an heterogeneity threshold as in [12], by defining a bandwith parameter specifically for the spatial domain as in Mean-Shift [13] or, recently, by specifying the number of required objects as in SLIC [14].

Based on these assumptions, the typical approach in the OBIA framework for automatic LULC mapping is to leverage agglomerate descriptors (i.e. object-based radiometric statistics) to build proper samples for training and classification, without explicitly managing within-object information diversity. To illustrate this point, Figure 1a depicts a segmentation result of a *Urban area*. Focusing on a single segment: this typically contains, simultaneously, sets of pixels associated to buildings, streets, gardens, and so on, which are all equivalently important in the recognition of the Urban LULC class. However, in many cases, the components of a single segment do not equally contribute to their identification as belonging to a certain land-cover class.

In another scenario, i.e. the one repoted in Figure 1b, we can have segments associated to a *Forest* land cover class that may contain only trees in the denser areas, or a mix of trees and bare soil pixels in the more open areas. Evidently, in this case the "tree" component is likely to provide the most discriminative information for classification, while the "bare soil" component may be irrelevant or even represent a source of noise, especially if it does not occur frequently in the Forest class. Our contribution is motivated by the fact that none of the recently proposed supervised classification frameworks [15], [16] relying on object-based SITS representation for land cover mapping explicitly takes into account these within-object information diversity.

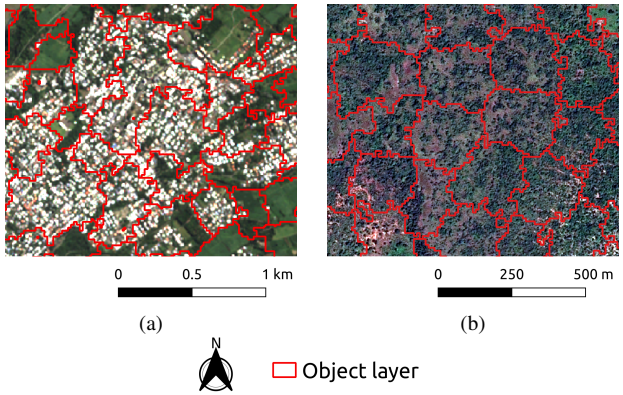


FIGURE 1: Results of a segmentation process considering: (a) an *Urban Area* and (b) a *Forest* landscape. Regarding both land cover classes, we can note that the produced segments exhibit a certain degree of within-object diversity.

We highlight that the OBIA strategy is different from recent semantic segmentation methods since the former extracts objects (segments) from the remote sensing image in a totally unsupervised fashion prior to any subsequent analysis; the latter is a supervised approach where a deep learning network is learnt over densely labeled data (which must hence be available) with the aim to provide pixel-level classification.

In this work, we propose $TASSEL$, a new deep-learning framework to deal with object-based SITS land cover mapping which can be ascribed into the weakly supervised learning (WSL) setting [17], [18]. We locate our contribution in the framework of WSL since the object-based land cover classification task exhibits label information that intrinsically brings a certain degree of approximation and inaccurate supervision to train the corresponding learning model, related to the presence of non-discriminative SITS components within a single labelled object.

Our framework includes several stages: firstly, it identifies the different multifaceted components on which an object is defined on. Secondly, a Convolutional Neural Network (CNN) extracts an internal representation from each of the different object component. Here, the CNN is especially tailored to model the temporal behavior exhibited by the object component. Then, the per component representation is aggregated together and used to provide the decision about the land cover class of the object. Beyond the pure model performance, our framework also allows to go a step further in the analysis, by providing extra information related to the contribution of each component to the final decision.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3024133, IEEE Access

Such extra information can be easily visualized in order to provide additional feedback to the end user, supporting spatial interpretability associated to the model prediction.

In order to assess the quality of $TASSEL$, we perform extensive evaluation on two real-world scenarios over large areas with contrasted land cover features and characterized by sparsely annotated ground truth data [19]. Unfortunately, such a label-specific constraint prevent us to leverage standard semantic segmentation strategies [20] as competitors. To this end, the evaluation is conducted considering state of the art land cover mapping approaches for sparsely annotated data in the OBIA framework. Finally, an in depth qualitative analysis is drawn to underline the ability of our framework to provide extra information that can be effectively leveraged to support the comprehension of the classification decision.

The main contributions of our work can be summarized as follows:

- We propose a new deep-learning framework to cope with object-based SITS classification devoted to manage the within-object information diversity exhibited in the context of land cover mapping;
- We design our framework with the goal to provide as outcomes not only the model decision but also extra information that can provide insights about (spatial) model interpretability;
- We conduct an extensive evaluation of our framework considering both quantitative and qualitative analysis on real-world benchmarks that involve ground truth data collected during field campaigns and featured by operational constraints.

The rest of the article is structured as follows: the literature related to our work is introduced in Section II; Section III introduces the Weakly Supervised Learning classification problem for object-based SITS data; Section IV describes the $TASSEL$ framework and Section V describes the data and the considered study area. Experimental settings, data and results are detailed and discussed in Section VI. Finally, Section VII concludes the work.

II. RELATED WORK

In this section we cover the literature associated to our research work. We focus on the machine learning paradigms related to the proposed framework (i.e., Weakly Supervised and Multiple Instance learning) and their connection with remote sensing analysis. Successively, we introduce recent object-based SITS classification strategies from the remote sensing literature and we conclude by highlighting the novelty of our contribution.

Weakly Supervised Learning Weakly supervised learning [18] refers to a set of approaches that have the objective to deal with weak supervision: incomplete, inexact and inaccurate. In [21], the authors introduce a convolutional neural network aimed at the joint detection and localization of objects of interest inside images. Since the only available information at training time is the presence of the object in an image, weak supervision is here used to tackle the localization problem. [22] proposes to leverage weak supervision in the context of semantic segmentation. A constrained Convolutional Neural Network is trained with labels at image level (multiple labels can be associated to an image) and the model automatically detects which part of the image is associated to the various labels. The method uses a novel loss function to optimize a set of linear constraints on the output space. Temporal action localization can also be treated using a weakly supervised approach. Authors in [23] propose a framework in which only video level labels are supplied and the deep learning system is capable to temporally localize multiple actions inside the video sequence. Also in this case the unit of analysis (the video) can be characterized by multiple actions and the multiple actions can be detected inside each video. In the remote sensing field, similarly to standard Computer Vision, weakly supervised learning frameworks are mainly devoted to deal with object localization tasks [24] or semantic segmentation [25] of high resolution (single date) satellite images.

Multiple Instance Learning Multiple Instance learning [26] (MIL) is a supervised learning paradigm in which a classification model is learnt to supply prediction for a group of instances. A bag is composed of a set of instances and the (weak) supervision is available only at bag level. Commonly, MIL approaches deal with binary classification tasks in which a negative bag is composed only by negative examples while a positive bag contains at least one positive example. Recently, [27] proposed a MIL framework based on deep learning where the decision is provided by leveraging an attention based pooling strategy. Considering the remote sensing domain, MIL frameworks have been leveraged to deal with hyper-spectral [28] and multi-spectral image classification [29] or landmine detection exploiting ground penetrating radar images [30].

Object-Based satellite image classification Object-Based image analysis [8] (OBIA) considers object instead of pixels as unit of analysis. Working at object instead of pixel granularity has several advantages: i) objects represent a more coherent piece of information since they are simpler to interpret [10], ii) label annotations can be collected with a limited human effort and iii) objects facilitate data analysis scale-up since, for the same image, the number of objects is usually smaller than the number of pixels by several orders of magnitude. The latter point is particularly important in operational remote sensing where information analysis can cover large areas (regional or national scales) involving satellite image at metric or decametric spatial resolution [16].

In [31], an object-based change detection approach of bi-temporal SITS data is introduced. The task is treated as a binary classification problem where the classification model predicts if an object changes or not between two observed time stamps. The approach is based on a supervised maximum likelihood classification. [15] tackles the problem of grasslands classification using univariate SITS data of Normalized Difference Vegetation Index (NDVI). The unit of analysis is the object but, instead of considering only the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3024133, IEEE Access

average object representation, they also retain the covariance matrix as an additional, second order, statistic characterizing the internal object distribution. Finally, a Gaussian mean kernel based on the first and second order information is developed and coupled with an SVM model in order to cope with classification. The method is especially tailored for univariate time series and its extension to multidimensional SITS is not straightforward. [32] evaluates the use of a Recurrent Neural Network (Gated Recurrent Unit) to cope with Land Use Land Cover (LULC) mapping considering both pixel-based and object-based optical (multivariate) SITS data. Object-based representation is derived via average aggregation of the pixel information belonging to the object. [33] introduces a Convolutional Neural Network (CNN) applied on the temporal domain to explicitly consider the dynamic associated to the SITS data. Despite the fact that the proposed approach is evaluated considering pixel-based time series, the same approach can be directly transposed to object-based SITS objects. The study reports an in depth evaluation of CNN models for optical SITS data and it highlights the quality of such models to manage the temporal information characterizing Earth Observation data.

In our framework we leverage weak supervision with the aim to disentangle the contributions of the different portions of the SITS object and to deal with the misalignment or approximation between the object level annotation and the object content. To this end, our aim is to deal with the multifaceted information on which the object is defined with the aim to pay more attention to useful components and, simultaneously, paying less attention to less relevant ones.

III. PROBLEM DEFINITION AND WEAKLY SUPERVISED LEARNING CHARACTERIZATION

Given a set of objects $O = \{o_i\}_{i=1}^{|O|}$ where each o_i has an associated label information $y_i \in Y$ (Y is the set of possible labels), the goal is to build a classification model $f_\Theta(o)$ parametrized with Θ to predict the label values for unlabeled objects. The parameters Θ are learnt over training information $Train = \{o_i, y_i\}$ where $y_i \in Y$ and y_i is the label information associated to object o_i. In addition, the object o_i is composed by a set of pixel time series $o_i = \{pts_{ik}\}_{k=1}^{|o_i|}$ where pts_{ik} is k-th pixel time series of the object o_i. We remind that the label y_i associated to o_i can represent a combination of object components or a portion of the object content. Such approximate or inaccurate label information can be referred as weak supervision [18].

Standard approaches in Object-Based Satellite Image Time Series Analysis [15], [32] manage the object representation via average or median aggregation over the set of pixels time series belonging to it. We can indicate the averaged information of the object o_i as $\tilde{o}_i$. In this context, the original classification problem is formulated as $y = f_\Theta(\tilde{o})$.

The aggregation procedure, that supplies the standard object characterization for satellite image time series ($\tilde{o}_i$), unfortunately, can smooth and flatten the different signal components on which the original object is defined on and it fails to deal with the weak supervision provided by object-level annotation. Moreover, it can also be sensible to outlier or anomalous signal components that can negatively influence the aggregated representation.

Differently from such standard procedure, our goal is to explicitly manage the degree of approximation and inaccurate supervision, carried out by the object-level label information, in the training of the classification process. More in detail, by leveraging the weakly supervised learning framework [18], [17], we propose to deal with the object-based classification of SITS data by means of a classification model $f_\Theta(\{pts_{ik}\}_{k=1}^{|o_i|})$ directly working on $\{pts_{ik}\}_{k=1}^{|o_i|}$, where an object o_i can be seen as a bag of pixels.

Due to the fact that object components usually involve a set of homogeneous pixels, we can consider, without loss of generality, that the pixels belonging to an object can be partitioned in a number L of components based on their radiometric similarity: $o_i = \{c_l\}_{l=1}^{L}$ and $c_l = \{pts_{ils}\}_{s=1}^{|c_l|}$ and $\forall_{c_{l1},c_{l2}} c_{l1} \cap c_{l2} = \emptyset$ and $\bigcup_{c_l} = o_i$. The set $\{c_l\}_{l=1}^{L}$ is a partition of the pixels of object o_i. In this case, an object can be seen as a bag of components. Considering object components instead of original object pixels, the classification model will be redefined as $f_\Theta(\{c_l\}_{l=1}^{|L|})$.

Problem Definition. *WSL for object-based SITS classification*
Given a set of objects $O = \{o_i\}_{i=1}^{|O|}$ with associated label information Y, each object can be represented as a partition of the pixels information belonging to it ($o_i = \{c_l\}_{l=1}^{L}$) and we refer to each c_l as a (object) component. Each object can be seen as a bag of components. The goal is to build a classification model $y, \alpha = f_\Theta(\{c_l\}_{l=1}^{L})$ parametrized with Θ to provide the class information values (y) for unlabeled objects as well as an additional extra information α that disentangles the contribution of each component c_l on which the object is defined on.

Such formulation allows to consider fine-grained information to model the classification problem, i.e., object components information instead of aggregated objects statistics. In addition, it also underlines that the outcomes of the classification process includes an extra information α, that can be leveraged to move towards the comprehension and the analysis of the decision made by the prediction model.

IV. METHOD

In this section we introduce $TASSEL$ (a**T**tentive we**A**kly **S**upervised **S**atellite image time s**E**ries c**L**assifier), a framework to deal with the object-based weakly supervised classification of SITS data following the problem definition introduced in Section III.

Figure 2 supplies a general overview of $TASSEL$. Given an object time series, firstly, the different components that constitutes the object are identified. Secondly, a Convolutional Neural Network (CNN) block is adopted to extract information from each of the different object components. The same set of weights is shared among all the CNN blocks.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3024133, IEEE Access

Then, the results of each CNN block (the component representation) is aggregated/combined via an attention mechanism [34] in which the components contribution are weighted proportionally to the information they are bringing on. After the attention combination, the new object representation is obtained and it is successively fed into the Fully Connected layers that will provide the final classification. In Figure 2 we can also observe that the outcomes of the process not only involve the model decision, but also the extra information α. Such outcome is finally leveraged to derive attention maps with the aim to analyze object contributions and, at the same time, provide qualitative information about the general model decision.

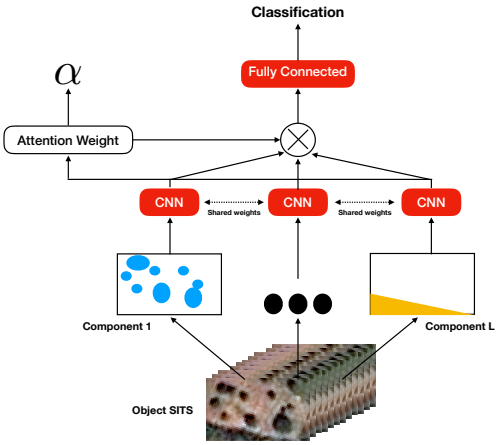


FIGURE 2: The general overview of $TASSEL$. Firstly, the different components that constitute the object are identified. Secondly, a CNN block extracts information from each of the different object components. Then, the results of each CNN block are combined via attention. Finally, the classification is performed via dedicated Fully Connected layers. The outputs of the process are the prediction for the input object SITS as well as the extra information α that provides an information related to the contribution of each object component.

A. COMPONENT PROCESSING STEP

The first step of our framework is related to the identification of components on which the SITS object is defined and the processing of such components. Firstly a fixed number of homogeneous groups, in terms of radiometric information, from each object are extracted and, successively, each component is processed by means of a Convolutional Neural Network. The output of this step is a feature representation for each of the L components. We can refer to the feature representation of component c_l with $h_l \in \mathbb{R}^d$, $H = \{h_1, \cdot, h_l\}$ the set of all the feature representations and d the dimensionality of the vector h_l.

To detect and extract such object components, we perform clustering on the pixel time series. To this aim, we use K-Means clustering with a number of clusters equals to L (the presumed number of components in an object). Once the clustering process is performed, we use the cluster prototypes (or centroids) as component information. Successively, each component information is processed by means of a Convolutional Neural Network. Due to the nature of the input signal to process (cluster prototypes of time series data), we adopt one dimensional convolutional neural network (CNN1D) where the convolution operations are applied on the time dimensions. In this way, the Convolutional Neural Network will allow to explicitly manage and exploit the temporal dimension conveyed by the time series data. Our choice is also supported by recent remote sensing literature [33] where CNN1D has recently demonstrated to be competitive and well suited to extract useful representations to support the land cover classification task. Moreover, we underline that the same CNN1D model is applied on all the different object components in order to extract an invariant per-component representation.

B. ATTENTIVE AGGREGATION STEP

The second step of our framework is devoted to the aggregation of the object components with the aim to find a global object representation. To this end, we combine all such information by means of attention [34] with the goal, in the feature aggregation, to consider the contribution of each object component differently. The outputs of this step are an object representation which we refer as $\tilde{h}$ as well as the extra information α that is related to the importance/contribution of each component on which the object is defined on.

Attention mechanisms [34] are extensively employed nowadays in standard signal processing (1D signal, language or 2D signal). At the beginning this approach was introduced to work in conjunction with recurrent neural network models, in order to combine the information extracted at different time stamps [35]. Successively, attention mechanisms were applied on 2D images [36] as well as to manage weak supervision and bag level classification [37], [27].

Given $H = \{h_1, ..., h_l\}$ the set of all the components representations, we attentively combine such information as follows:

$$\tilde{h} = \sum_{l=1}^{L} \alpha_l \cdot h_l \tag{1}$$

where each α_l is defined as:

$$\alpha_l = \frac{exp(v_a^\intercal \ tanh(W_a \, h_l + b_a))}{\sum_{l'=1}^{L} exp(v_a^\intercal \ tanh(W_a \, h_l' + b_a))} \tag{2}$$

where matrix $W_a \in \mathbb{R}^{d,d}$ and vectors $b_a, u_a \in \mathbb{R}^d$ are parameters learned during the process. These parameters allow to combine the vectors contained in matrix H. The purpose of this procedure is to learn a set of weights $(\alpha_1, \ldots, \alpha_L)$ to estimate the contribution of each component representation h_l. The $SoftMax(\cdot)$ function is used to normalize weights α so that their sum is equal to 1. In addition, the attention aggregation is a permutation-invariant operation. This means

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3024133, IEEE Access

that the results $\tilde{h}$ is invariant w.r.t. the order in which the elements of H are processed. This is a useful and important property for aggregation operation over a set of unordered elements.

C. CLASSIFICATION STEP AND TRAINING PROCEDURE

The representation $\tilde{h}$ obtained at the previous step is finally processed by means of several Fully Connected layers with the objective to provide the final classification w.r.t. the object SITS data. In our context we use two Fully Connected layers with a number of neurons equals to 512 each. Each Fully Connected layer is associated to a Rectifier Linear Unit non-linearity and followed by a Batch Normalization layer in order to avoid weight oscillation and ameliorate network training:

$$Cl(\tilde{h}) = W_3 BN(ReLU(W_2(BN(ReLU(W_1\tilde{h} + b_1))) + b_2)) + b_3 \quad (3)$$

where W_1, W_2, W_3, b_1, b_2 and b_3 are parameters learnt by the model to process the attentive aggregated 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 as cost function. The categorical cross-entropy is defined as follows:

$$CE(Y, \hat{Y}) = -\sum_{i=1}^{|O|}\sum_{j=1}^{|Y|} y_{ij} log(\hat{y}_{ij}) \quad (4)$$

where y_{i*} is the class vector (under one hot encoding) associated to object O_i and $\hat{y}_{i*}$ is the class distribution vector (after Sofmax operation) predicted by the deep learning model for the corresponding satellite image time series object o_i.

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 discriminative and effective representation 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:

$$Cl^{aux}(\tilde{h}) = W_3^{'}\tilde{h} + b_3^{'} \quad (5)$$

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

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

$$L = CE(Y, Cl) + \lambda CE(Y, Cl^{aux}) \quad (6)$$

where $\lambda \in [0, 1]$ is an hyper-parameter that control the importance of the auxiliary classification in the learning process. 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. SPATIAL INTERPRETATION VIA THE EXTRA INFORMATION α

Beyond the predictive ability of the proposed learning model, we highlight that extra information α can be leveraged to perform qualitative analysis related to the model behavior. In this direction, such extra information is exploited to interpret the internal decision of $TASSEL$ and evaluate the contribution of each component on which the object is defined on. Thanks to such information we can produce a spatial *attention* (or saliency) *map* [38] associated to each classified object SITS. More in detail, given an object o, the α information relates a weight α_l to each object component $c_l \in o$. Since each component c_l corresponds to a set of pixels, we can assign to all the pixels $p \in c_l$ the same value α_l. In this way we can visually highlight homogeneous areas (in terms of spectral evolution along the SITS) and depict their contribution to the decision process performed by $TASSEL$. An example of the outcome of this procedure is depicted in Figure 3 where the same area is replicated twice: on the left we observe the original area while on the right the attention map (blue area) is superimposed to the object extent and the degree of blue (light to dark) is proportionally related to the α values associated to the object components to which the pixel belongs to. Such tool supplies insights on the way the deep learning decision is obtained and it visually indicates which information is considered as more or less relevant by the system according to the particular land cover class. Such a stage of our framework is deeply investigated via qualitative evaluation in Section VI-B3.

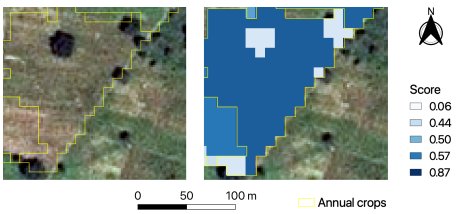


FIGURE 3: The spatial extent of an object associated to the *Annual crops* land cover class. On the left the RGB image and on the right the same image with the *attention map* superimposed to the RGB image. The yellow line represent object contours. The legend on the right of the example reports the scale (discretized considering quantiles) associated to the attention map.

E. DETAILS OF THE CNN ARCHITECTURE

The One Dimensional Convolutional Neural Networks (CNN1D) we adopt in $TASSEL$ is reported in Table 1. We follow general principles applied in the design of Convo-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3024133, IEEE Access

lutional Neural Networks [39], where the number of filters along the network structure grows and the convolutional operations are followed by non linear activation function (Rectifier Linear Unit in our case), Batch Normalization and Dropout. Our CNN1D has ten blocks where the first eight involves parameters associated to Convolutional and Batch Normalization operation. 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 aim 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 maps by means of average aggregation.

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()

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.

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 *Koumbia* dataset (a rural municipality in the province of Tuy, Burkina Faso).

The *Reunion Island* dataset consists of a time series of 21 Sentinel-2 images acquired between January and December 2017. The *Koumbia* dataset consists of a time series of 23 Sentinel-2 images acquired between January 2016 and December 2016 (see Fig 4 for acquisition date details).

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* [40] developed by the National Centre for Space Studies (CNES).

[2]Data are available via http://theia.cnes.fr

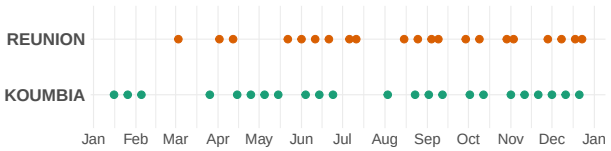


FIGURE 4: Overview of the acquisition dates of the Sentinel-2 (S2) images over the two study sites: *Reunion* and *Koumbia*. S2 acquisitions are sparse due to the ubiquitous cloudiness.

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*, [4]). Two additional indices: NDVI [3] (Normalized Difference Vegetation Index) and NDWI, defined by McFeeters [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 site is 6 656 × 5 913 pixels corresponding to 3 935 Km^2 while the extent for the *Koumbia* site is 5 253 × 4 797 pixels corresponding to 2 519 Km^2. Figure 5 depicts the study sites with the associated ground truth polygons.

Considering the Reunion island dataset [41], the ground truth (GT) was built from various sources : 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) completed by a field expert with knowledge of territory to distinguish natural and urban areas.

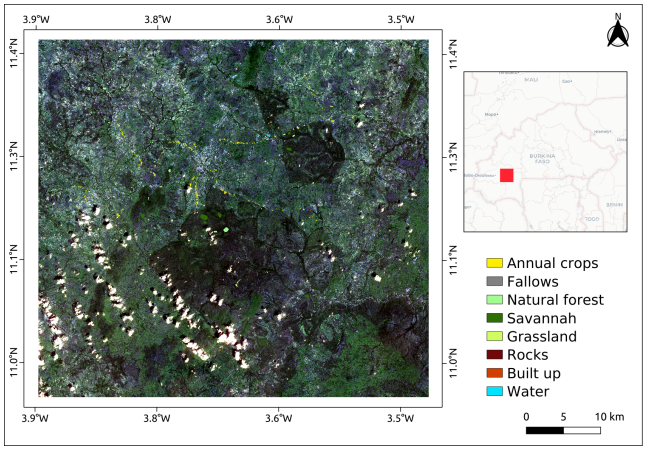
Regarding the *Koumbia* dataset [7], the reference database is a collection of (i) digitized plots from a GPS field mission performed in October 2016 and mostly covering classes within cropland and (ii) additional reference plots on non-crop classes obtained by photo-interpretation by an expert.

Class	Label	# Polygons	# Objects	# Pixels
0	*Sugar cane*	869	1 466	88 983
1	*Pasture and fodder*	582	1 042	68 069
2	*Market gardening*	758	1 038	17 574
3	*Greenhouse crops or shadows*	260	308	1 928
4	*Orchards*	767	1 174	33 694
5	*Wooded areas*	570	1 467	205 050
6	*Moor and Savannah*	506	1 172	155 229
7	*Rocks and natural bare soil*	299	845	154 283
8	*Relief shadows*	81	248	54 308
9	*Water*	177	458	82 547
10	*Urbanized areas*	1 396	1 360	19 004
Total		6 265	10 578	880 669

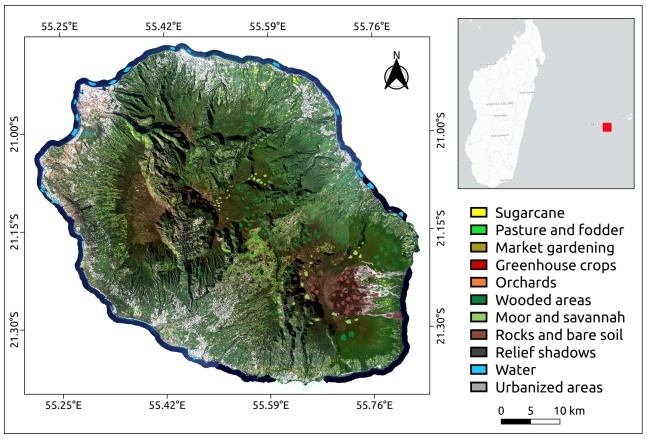
TABLE 2: Per Class ground truth statistics for the Reunion Island Dataset

[3]https://en.wikipedia.org/wiki/Normalized_difference_vegetation_index
[4]https://en.wikipedia.org/wiki/Normalized_difference_water_index
[5]RPG is part of the European Land Parcel Identification System (LPIS), provided by the French Agency for services and payment

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3024133, IEEE Access



(a) KOUMBIA Study site



(b) REUNION Study site

FIGURE 5: Location of the Koumbia (a) and Reunion (b) study sites. The RGB composite is a SPOT6/7 image upscaled at 10-m of spatial resolution. The corresponding ground truth polygons are overlaid to each image.

Class	Label	# Polygons	# Objects	# Pixels
0	*Annual Cropland*	671	481	31 075
1	*Fallows*	57	79	1 808
2	*Natural Forest*	64	174	15 843
3	*Savannah*	87	276	25 156
4	*Grassland*	142	269	12 883
5	*Rocks*	29	24	852
6	*Built up*	71	57	1 096
7	*Water*	16	19	1 410
Total		1 137	1 379	90 123

TABLE 3: Per Class ground truth statistics for the Koumbia Dataset

A. GROUND TRUTH STATISTICS AND SEGMENTATION

Considering both datasets, ground truth comes in GIS vector file format containing a collection of polygons each attributed with a unique land cover class label. 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).

The ground truth data includes 880 669 pixels (resp. 6 265 polygons) distributed over 11 classes for the *Reunion Island* dataset (Table 2) and 90 123 pixels (resp. 1 137 polygons) distributed over 8 classes for the *Koumbia* benchmark (Table 3).

To analyse data at object-level, a segmentation was provided by field experts for each study site using the VHR images (SPOT6/7 image) which have been upsampled at 10m of spatial resolution via bicubic interpolation and coregistered with the corresponding Sentinel-2 grid to ensure a precise spatial matching. The field experts adopt such a strategy since the SPOT6/7 images were acquired, on both study sites, with favorable atmospheric condition. The VHR images were segmented using the SLIC algorithm [14] available via the scikit-image toolkit [42]. The parameters were adjusted so that the obtained segments fit as closely as possible field plot boundaries. We remind that the segmentation information is an input of our process and it is not a part of our pipeline. Then, for each study site, the ground truth data were spatially intersected with the obtained segmentation finally resulting in new comparable size labeled 10 578 objects for the *Reunion Island* (resp. 1 379 segments for the *Koumbia* site).

VI. EXPERIMENTS

In this section we introduce the experimental protocol, the data on which the evaluation is carried out and the results we obtained. Firstly, we describe the real-world SITS dataset we used in our evaluation and the associated preprocessing. Seconldy, we report the experimental settings associated to the competing methods involved in the evaluation and the metrics we adopt. Successively, we report and discuss both quantitative and qualitative experiments with the aim to validate the classification performances with the former and to assess the quality of the extra information α with the latter.

A. EXPERIMENTAL SETTINGS

To assess the quality of $TASSEL$, based on recent literature, we select a panel of competitors exhibiting different and complementary characteristics:

- Random Forest (**RF**) classifiers since such general purpose machine learning approach is commonly employed to deal with the classification of SITS data [43].
- A Multi Layer Perceptron (**MLP**) model that consider the SITS data as a flat vector information. The MLP has two hidden Fully Connected Layers with 512 neurons each and ReLU activation function. Each Fully Connected layers is followed by a Batch Normalization and Dropout layers.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3024133, IEEE Access

- A Long-Short Term Memory model [44] (**LSTM**) with a recurrent unit with 512 neurons. Recurrent Neural networks are well suited to explicitly manage the temporal information that is contained in time series data. The LSTM representation is passed through a MLP block (like the one previously described) to perform SITS object classification. The model is learnt end-to-end.
- A Gated Recurrent Unit model [35] (**GRU**) with a recurrent unit with 512 neurons. GRU is another kind of Recurrent Neural networks, with a lighter architecture w.r.t. LSTM unit, that is demonstrating competitive performance considering both NLP and signal processing applications. Also in this case the GRU is stacked together with a MLP to provide the final classification. The model is learnt end-to-end.
- A one dimensional Convolutional neural network model that has the same structure of the CNN1D module employed by $TASSEL$. Also in this case the CNN is stacked together with the MLP block to provide the final classification decision. The model is learnt end-to-end. We refer to this competitor as **CNN**.
- An ablation of our framework $TASSEL$ without the auxiliary classifier $Cl^{aux}(\tilde{h})$. This ablation allows us to evaluate the effectiveness and the appropriateness to directly retropropagate the error at the attentive aggregation level. We name this competitor $\boldsymbol{TASSEL_{noAUX}}$.

All the competitors, with the exception of $TASSEL_{noAUX}$, are evaluated considering the standard average object representation.

For each study site, we split the corresponding data into three parts: training, validation and test set. Training data are used to learn the model, while validation data are exploited for model selection by varying the associated parameters. Finally, the model that achieves the best performance on the validation set is successively employed to perform the classification on the test set. The datasets were split into training, validation and test set with an object proportion of 50%, 20% and 30% respectively. The values were normalized per band (resp. indices) considering the time series, in the interval $[0, 1]$.

Considering the models leveraging the Random Forest classifier, we optimize the model via the tuning of two parameters: the maximum depth of each tree and the number of trees in the forest. For the former parameter, we vary it in the range {20,40,60,80,100} while for the latter one we take values in the set {100, 200, 300,400,500}. The weight λ is set to 0.5 for $TASSEL$.

Considering all the deep learning models, parameters learning is performed using the Adam optimizer [45] with a learning rate equal to 1×10^{-4}. The training process, for each model, is conducted over 5 000 epochs with a batch size equals to 32. For $TASSEL$ and $TASSEL_{noAUX}$, regarding the quantitative evaluation, we set the number of components equal to 6.

The assessment of the model performances are done considering *Accuracy*, *F-Measure* and *Kappa* measures. The *F-Measure* assessment criteria is particularly useful in our context since the benchmarks associated to both study sites exhibit high class unbalance. To reduce bias induced by the train/validation/test split procedure, for each benchmark and for each evaluation metric, we report results averaged over five different random splits.

Experiments are carried out on a workstation with an Intel (R) Xeon (R) CPU E5-2667 v4@3.20Ghz with 256 GB of RAM and four TITAN X GPU. All the Deep Learning methods (including $TASSEL$) are implemented using the Python Tensorflow library, while Random Forest approaches are implemented using Python scikit-learn library. The source code of $TASSEL$ is available online [6].

B. RESULTS

With the aim to assess the quality of $TASSEL$, we perform several kinds of analyses to understand the behavior of our framework. Firstly, we provide a quantitative evaluation considering metric performances of the different competing methods. During this evaluation, we report average results as well as a per-class analysis. Secondly, we conduct a sensibility analysis on the behavior of $TASSEL$ with respect to the number of components. Finally, an in-depth qualitative evaluation is carried out to investigate and exploit the extra information (α) provided by $TASSEL$ to disentangle the contribution of component objects based on the learning process.

1) Quantitative results

In this section we report the quantitative results obtained by the competing methods involved in the experimental evaluation. We consider both average and per-class analysis.

	F-Measure	Kappa	Accuracy
RF	77.51 $\pm$ 2.35	0.7259 $\pm$ 0.0273	79.23 $\pm$ 2.03
LSTM	74.10 $\pm$ 2.11	0.6784 $\pm$ 0.0282	75.26 $\pm$ 2.17
GRU	73.73 $\pm$ 1.18	0.6739 $\pm$ 0.0121	75.16 $\pm$ 0.84
MLP	74.48 $\pm$ 1.51	0.6841 $\pm$ 0.0200	75.98 $\pm$ 1.48
CNN	78.52 $\pm$ 1.99	0.7266 $\pm$ 0.0260	78.75 $\pm$ 2.06
$TASSEL_{noAUX}$	78.28 $\pm$ 2.35	0.7224 $\pm$ 0.0304	78.37 $\pm$ 2.41
$TASSEL$	**79.98** $\pm$ 2.53	**0.7476** $\pm$ 0.0308	**80.43** $\pm$ 2.42

TABLE 4: Average (and standard deviation) F-Measure, Kappa and Accuracy performances of the different competing methods considering the *KOUMBIA* study site.

Table 4 and Table 5 show the average results in terms of F-Measure, Kappa and Accuracy considering the *KOUMBIA* and the *REUNION* benchmarks, respectively. Considering the *REUNION* study site, the worst average performances are obtained by the Random Forest approach. The *CNN* strategy outperforms all the other deep learning baselines methods (*LSTM*, *GRU* and *MLP*) while the best average performances, considering all the three evaluation metrics are achieved by our proposal $TASSEL$. A bit different is

[6] https://gitlab.irstea.fr/dino.ienco/tassel.git

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3024133, IEEE Access

	F-Measure	Kappa	Accuracy
RF	81.74 ± 0.47	0.7991 ± 0.0052	82.13 ± 0.46
LSTM	82.91 ± 0.66	0.8098 ± 0.0078	83.06 ± 0.69
GRU	82.68 ± 0.98	0.8072 ± 0.0113	82.82 ± 1.00
MLP	85.81 ± 0.60	0.8423 ± 0.0074	85.94 ± 0.66
CNN	87.11 ± 0.61	0.8565 ± 0.0068	87.20 ± 0.61
$TASSEL_{noAUX}$	88.75 ± 0.70	0.8752 ± 0.0082	88.88 ± 0.72
$TASSEL$	**89.13** ± 0.62	**0.8797** ± 0.0072	**89.28** ± 0.63

TABLE 5: Average (and standard deviation) F-Measure, Kappa and Accuracy performances of the different competing methods considering the *REUNION* study site.

the situation regarding the *KOUMBIA* benchmark. On this study site, the *RF* method shows better performances than (*LSTM*, *GRU* and *MLP*) strategies but it is still outperformed by all the rest of the approaches. Also in this evaluation the best average behaviour is exhibited by $TASSEL$. On both datasets, the comparison between $TASSEL$ and its ablation variant ($TASSEL_{noAUX}$) underlines the effectiveness of the auxiliary classifier training strategy that allows to systematically increases the classification precision, this fact underlines that such component plays an important role in the training strategy. This phenomenon is particularly evident for the *KOUMBIA* benchmark that is characterized by high class imbalance and a limited number of labeled samples. Due to the reported results, we can speculate on the fact that, in presence of a limited number of labeled samples, directly inject weight updates in the middle of the network seems to facilitate the training process. Still on the *KOUMBIA* study site, we can observe that all the methods exhibit high variability (high standard deviation). This is related to the small number of samples and imbalanced class ratio such dataset exhibits. For this reason, the method performances are highly sensitive to the way the training/validation/test splits are done. Conversely, on the *Reunion* benchmark the standard deviation values are smaller but high difference (around 7 points) can be noted between the worst (*RF*) and the best ($TASSEL$) competing method.

Table 6 and Table 7 report the per class F-Measure of the different competing methods considering the *KOUMBIA* and the *REUNION* study site, respectively.

Regarding the *KOUMBIA* study site (Table 6), we can observe that $TASSEL$ achieves almost all the time the best (bold) and the second best (underlined) results considering the eight land cover classes on which this study site is defined on. The only exception is related to the *Built up* class in which $TASSEL$ achieves results that are comparable to the *CNN* method. The most notable gain, on this benchmark, can be observed for the *Fallows* class. Regarding this land cover class, $TASSEL$ achieves almost 10 points of gain w.r.t. the second direct competitor (*CNN*) and almost 20 points of F-Measure gain considering the worst competitor (*RF*). Such class constitutes a complicated land cover target since it covers heterogeneous examples that easily overlap with examples of other classes. This is also the motivation while absolute performances are quite small on such class considering all the competing methods. Nevertheless, the proposed approach is the one that better deals with the internal diversity of such heterogeneous and complicated land cover class.

Considering the *REUNION* study site (Table 7), we can note that both $TASSEL$ and $TASSEL_{noAUX}$ consistently outperform all the other competitors considering all the land cover classes with the former winning on 7 land cover classes over the total of 11 land cover classes on which the multi-class classification problem is defined. In addition, the gap between our method and its ablation are coherent with the average differences observed in Table 7. Gains between the best ($TASSEL$) and the worst (*RF*) competitors on this dataset vary from 19 points (on *Greenhouse crops*) to a couple of points (on *Relief shadows*). In the middle, we can observe notably amelioration regarding *Market gardening*, *Orchards*, *Moor*, *Pasture* and *Wooded areas* classes. All the objects of such classes, considering the landscape associated to this study site, are highly prone to contain within-object information diversity or noisy/irrelevant components conversely to class like *Relief shadows* that represents more homogeneous landscape and it mainly contains highly homogeneous information. This fact supports the ratio behind our weakly supervised learning framework and its adequateness to deal with object-based Satellite image time series classification.

2) Sensitivity analysis w.r.t. the nc parameters

Figure 6 depicts the behavior of $TASSEL$ varying the value of the nc parameters in the range $2, 4, 6, 8, 10$. In addition, the plot reports the average values (averaged over five different splits) and the associated standard deviation as error bar. We can observe that $TASSEL$ exhibits a coherent stable behaviors on both benchmarks in terms of average F-Measure performance. Considering the standard deviation, it shows a per benchmark coherence. While the *Reunion* dataset has a small standard deviation, on the *Koumbia* benchmark higher standard deviation is associated to all values of the nc parameters. For the latter study site, this is due to the reduced size of the associated dataset that can induce high performance variation depending on the specific training/validation/test split.

Generally, we can see that, considering both benchmarks, a number of object components equals to two is sufficient to achieve high level performances w.r.t. all the competitors evaluated in Section VI-B1. This is not a surprising behavior and it is in accord with the hypotheses our framework is built on. By definition, remote sensing objects represent suitable "land units" that involve multiple radiometric components but, in general, the related land cover to which the object is associated can be directly related to one of them. For this reason, a binary partition (in the majority of the cases) is sufficient to isolate relevant w.r.t. less relevant information.

3) Assessing components importance for spatial interpretation

In this section we provide a qualitative analysis related to the use of the extra information α provided by $TASSEL$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3024133, IEEE Access

	Annual Crops	Fallows	Natural Forest	Savannah	Grassland	Rocks	Built up	Water
RF	84.31	19.53	86.42	79.31	79.17	56.63	**72.58**	61.71
LSTM	80.57	15.0	82.78	76.94	75.22	54.5	62.2	84.84
GRU	81.73	13.29	80.17	75.23	74.46	58.73	63.6	84.84
MLP	82.86	17.72	80.48	75.84	75.09	53.63	63.81	78.78
CNN	83.54	29.57	**88.57**	**82.19**	77.57	**61.03**	65.94	87.51
$TASSEL_{noAUX}$	84.31	35.61	86.25	80.02	78.11	55.89	62.63	**88.0**
$TASSEL$	**85.88**	**39.12**	87.25	81.79	**79.72**	58.83	65.21	87.51

TABLE 6: Per class F-Measure performances of the different competing methods considering the *KOUMBIA* 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	88.16	83.15	74.88	34.22	71.74	89.24	84.92	89.62	95.9	87.66	78.09
LSTM	90.19	84.49	74.3	40.88	72.14	88.71	87.06	90.94	96.08	92.64	78.74
GRU	89.46	86.06	74.4	41.14	73.64	86.83	86.87	90.23	95.33	91.47	78.3
MLP	91.3	88.76	80.97	45.14	79.49	89.09	89.24	91.69	95.65	93.37	81.51
CNN	92.56	90.7	81.45	48.64	80.11	91.19	91.17	93.27	97.06	93.8	81.84
$TASSEL_{noAUX}$	93.3	90.79	84.56	**55.03**	82.42	91.44	91.72	93.28	**97.99**	**95.57**	**86.35**
$TASSEL$	**94.19**	**91.37**	**85.14**	53.21	**82.68**	**91.98**	**92.34**	**94.14**	97.6	95.37	86.14

TABLE 7: 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.

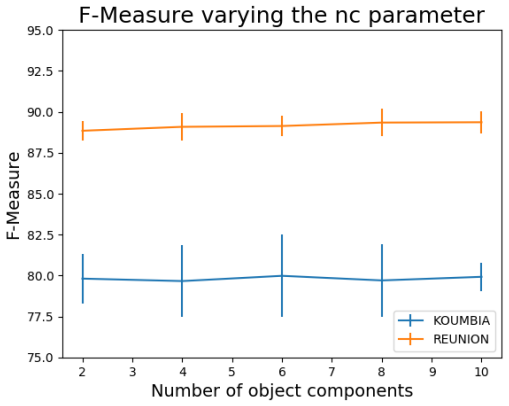


FIGURE 6: The results of the sensitivity analysis of $TASSEL$ regarding the nc parameter on the two real-world benchmarks on the *Koumbia* and *Reunion* study sites.

to interpret its internal decision and the related contribution of the object components.

With the aim to clearly highlight the internal selection process carried out by $TASSEL$, we evaluate the *attention map* derived by our framework with nc equals to 2. According to the results obtained in Section VI-B2, $TASSEL$ is stable w.r.t. such parameter and such configuration will also promote the visual investigation via higher contrasted spatial regions. The visualization we proposed is achieved considering extra images (SPOT6/7 [7] and Bing aerial view [8]) with very high spatial resolution (less than 2m). Such fine background images allow to visually depict details that are not visible for human eye at the spatial resolution of the Sentinel-2 images but, on the other hand, the pixel contours are not perfectly aligned due to the difference in spatial resolution.

Details of the *attention map* for the *Reunion* and *Koumbia* study sites are reported in Figure 7 and Figure 8, respectively. Associated to each detail a legend shows the color scale from light blue (small value of attention) to dark blue (high value of attention). With loss of generality, we can assume that the higher the attention value the more importance the model gives to a certain component.

Considering the *Reunion* study site, Figure 7a depicts an object SITS associated to the *Water* land cover class. We can clearly observe that higher importance (dark blue) is given to the component covering the dense water vegetation zone that is, probably, a confident indicator of the water class. The second detail, reported in Figure 7b, illustrates a pasture area that is recognized by $TASSEL$ thanks to the high importance supplied to the brown zone that is the direct result of animal or harvesting activities. The last detail, shown in Figure 7c, proposes a portion of the Roland Garros Reunion Airport, located in the north of the study site and classified as *Urbanized areas*. Due to the fact that this land cover class mainly includes buildings, $TASSEL$ exhibits a coherent behavior and it assigns an high attention value to the object component related to the white building (at the bottom of the detail) w.r.t. the one associated to the landing strip that cover the majority of the object extent. Such behavior pinpoints the fact that $TASSEL$ is able to recognize and leverage common (or similar) components among the examples belonging to the same coarse land cover

[7] https://en.wikipedia.org/wiki/SPOT_(satellite)#SPOT_6_and_SPOT_7
[8] https://www.bing.com/maps

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3024133, IEEE Access

class.

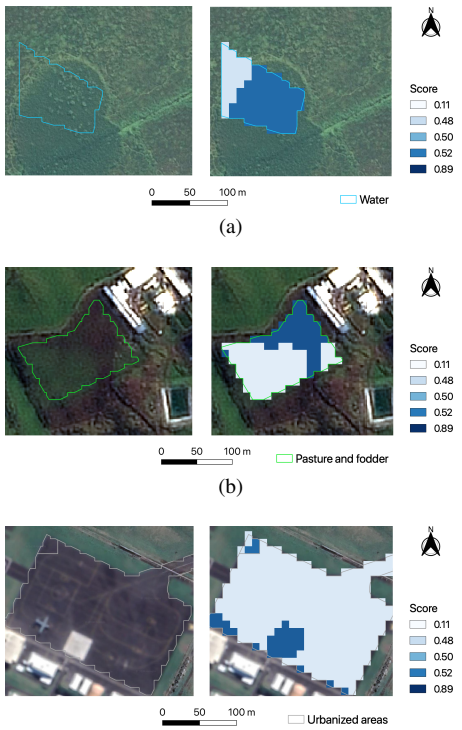


FIGURE 7: Three examples of the use of the extra information α provided by $TASSEL$ to interpret its internal decision on the *Reunion* study site. The blue, green and white lines represent object contours. Example 7a refers to the *Water* land cover class. Example 7b shows a sample related to the *Pasture and fodder* class while example 7c depicts an instance related to the *Urbanized areas* land cover class. The legend on the right of each example reports the scale (discretized considering quantiles) associated to the attention map.

Regarding the *Koumbia* study site, Figure 8a depicts an object SITS associated to the *Annual crops* land cover class. Due to the agricultural practices associated to this region of the Burkina Faso state, it is common to observe shea trees in the middle of agricultural parcels. Unfortunately, such unrelated element (with respect to the main land cover class) can negatively influence the methods leveraging the average object representation since it can inject noise in the average information. Here, we can clearly note that $TASSEL$ is able to filter out irrelevant information assigning a low attention value (light blue) to the object component associated to the shea tree. The second detail, reported in Figure 8b, illustrates an object depicting a forest area. Also in this case $TASSEL$ discriminates between relevant and irrelevant information and recover with high attention value (dark blue) the spatial extent covered by vegetation w.r.t. the spatial zone characterized by bare soil that is, clearly, unrelated to the *Forest* land cover class. The last detail, shown in Figure 8c, proposes an urban areas involving multiple objects (the red lines delimit object contours). Considering this bunch of objects, we can observe that generally, for each of them, $TASSEL$ attributes high attention score (dark blue) to built up pixels while low attention values (light blue) are related to vegetation zones coherently to the general land cover class (*built up*) to which all the objects are assigned.

To sum up, the qualitative evaluation, conducted on several details from the two study sites, has pointed out the ability of $TASSEL$ to effectively manage the multifaceted information exhibited by the object representation and, simultaneously, distinguish between relevant and irrelevant information to support and ameliorate the analysis of object SITS data for land cover mapping. Despite the fact that objects can contain high within-object information diversity, noisy signal components and, labels represent knowledge only at coarse granularity, $TASSEL$ is able to overcome such issues. More in detail, our framework is capable to learn invariant and distinctive signals with respect to a particular land cover class and, at the same time, adjust the contribution of each object components smoothing the impact of possible irrelevant information.

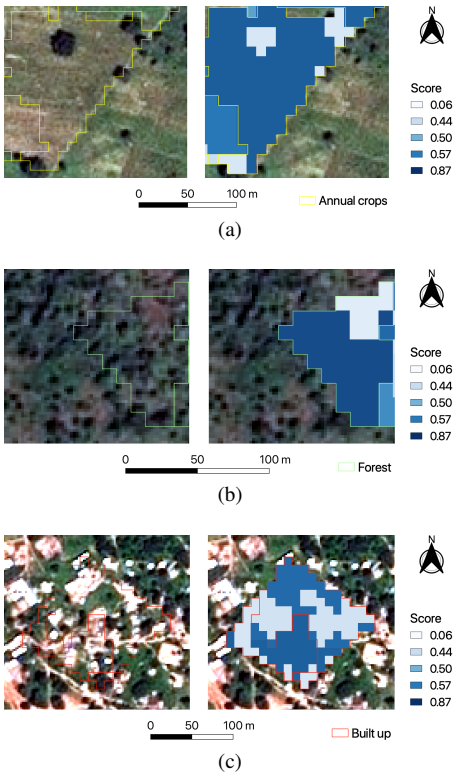


FIGURE 8: Three examples of the use of the extra information α provided by $TASSEL$ to interpret its internal decision on the *Koumbia* study site. The yellow, green and red lines represent object contours. Example 8a refers to the *Annual Crops* land cover class. Example 8b shows a sample related to the *Forest* class while example 8c depicts an instance related to the *Built up* land cover class. The legend on the right of each example reports the scale (discretized considering quantiles) associated to the attention map.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3024133, IEEE Access

VII. CONCLUSIONS

Due to the fact that object-based Satellite Image Time Series representation is characterized by high within-object information diversity, we introduce a new method, named $TASSEL$, to deal with object SITS land cover mapping under the lens of weakly supervised learning setting. Our framework, firstly, identifies the different components on which an object is defined on via cluster analysis. Secondly, a CNN block is adopted to extract an internal representation from each of the different object components. Thirdly, the results of each CNN block is aggregated via attention. Finally, the model outputs the land cover prediction associated to the object SITS as well as the extra information, referred as α, that is related to the contribution of each component to the model decision. Such outcome is directly actionable to derive attention maps with the aim to provide qualitative information about the general model behavior.

An extensive experimental evaluation on real world benchmarks underline the effectiveness of $TASSEL$, in terms of classification metrics w.r.t. state of the art competing approaches. Furthermore, the qualitative analysis pinpoints how our framework extracts knowledge that can be directly related to its decision and help the spatial interpretation of the obtained classification.

Several follows up can be drawn from the proposed work. Firstly, the CNN encoder we proposed can be ameliorated and extended according to recent research studies that investigate the use of deep learning approaches for the general analysis of time series data [46]. Secondly, other per component aggregated statistics can be considered as input for the one dimensional CNN (i.e. we can consider median value instead of mean value). Thirdly, novel strategies to adapt the number of components for each object can be inspected. Finally, the attention mechanism can be extended to also cope with the temporal dimension with the aim to discard irrelevant information and strengthen the general interpretability of our framework for spatio-temporal analysis.

VIII. ACKNOWLEDGEMENTS

This work was supported by the French National Research Agency under the Investments for the Future Program, referred as ANR-16-CONV-0004 (DigitAg), the GEOSUD project with reference ANR-10-EQPX-20 as well as from the financial contribution from the French Ministry of agriculture “Agricultural and Rural Development" trust account. It has also been conducted as a part of the PARCELLE project funded by the French Space Agency (DAR CNES 2019). We 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] L. Chen, Z. Jin, R. Michishita, J. Cai, T. Yue, B. Chen, B. Xu, Dynamic monitoring of wetland cover changes using time-series remote sensing imagery, Ecological Informatics 24 (2014) 17–26.

[2] B. Bellón, A. Bégué, D. L. Seen, C. A. de Almeida, M. Simões, A remote sensing approach for regional-scale mapping of agricultural land-use systems based on NDVI time series, Remote Sensing 9 (6) (2017) 600.

[3] S. Olen, B. Bookhagen, Mapping damage-affected areas after natural hazard events using sentinel-1 coherence time series, Remote Sensing 10 (8) (2018) 1272.

[4] J. Inglada, A. Vincent, M. Arias, B. Tardy, D. Morin, I. Rodes, Operational high resolution land cover map production at the country scale using satellite image time series, Remote Sensing 9 (1) (2017) 95.

[5] M. A. Wulder, J. G. Masek, W. B. Cohen, T. R. Loveland, C. E. Woodcock, Opening the archive: How free data has enabled the science and monitoring promise of landsat author links open overlay panel, Remote Sensing of Environment 122 (2012) 2–10.

[6] L. Khiali, D. Ienco, M. Teisseire, Object-oriented satellite image time series analysis using a graph-based representation, Ecological Informatics 43 (2018) 52–64.

[7] D. Ienco, R. Interdonato, R. Gaetano, D. H. T. Minh, Combining sentinel-1 and sentinel-2 satellite image time series for land cover mapping via a multi-source deep learning architecture, ISPRS Journal of Photogrammetry and Remote Sensing 158 (2019) 11 – 22.

[8] T. Blaschke, Object based image analysis for remote sensing, ISPRS journal of photogrammetry and remote sensing 65 (1) (2010) 2–16.

[9] G. Chen, Q. Weng, G. J. Hay, Y. He, Geographic object-based image analysis (geobia): emerging trends and future opportunities, GIScience & Remote Sensing 55 (2) (2018) 159–182.

[10] T. Lillesand, R. W. Kiefer, J. Chipman, Remote sensing and image interpretation, John Wiley & Sons, 2015.

[11] L. Ma, M. Li, X. Ma, L. Cheng, P. Du, Y. Li, A review of supervised object-based land-cover image classification, ISPRS Journal of Photogrammetry and Remote Sensing 130 (2017) 277 – 293.

[12] M. Baatz, A. Schäpe, Multiresolution segmentation: an optimization approach for high quality multi-scale image segmentation, Angewandte Geographische Informationsverarbeitung XII 58 (2000) 12–23.

[13] D. Comaniciu, P. Meer, Mean shift: A robust approach toward feature space analysis, IEEE Trans. Pattern Anal. Mach. Intell. 24 (5) (2002) 603–619.

[14] Z. Li, J. Chen, Superpixel segmentation using linear spectral clustering, in: CVPR, 2015, pp. 1356–1363.

[15] M. Lopes, M. Fauvel, S. Girard, D. Sheeren, Object-based classification of grasslands from high resolution satellite image time series using gaussian mean map kernels, Remote Sensing 9 (7) (2017) 688.

[16] D. Derksen, J. Inglada, J. Michel, Scaling up SLIC superpixels using a tile-based approach, IEEE Trans. Geoscience and Remote Sensing 57 (5) (2019) 3073–3085.

[17] L. Zhang, R. Ji, Y. Zhen, W. Lin, C. Snoek, Special issue on weakly supervised learning, J. Vis. Commun. Image Represent. 37 (2016) 1–2.

[18] Z.-H. Zhou, A brief introduction to weakly supervised learning, National Science Review 5 (1) (2017) 44–53.

[19] M. Volpi, D. Tuia, Dense semantic labeling of subdecimeter resolution images with convolutional neural networks, IEEE Trans. Geosci. Remote. Sens. 55 (2) (2017) 881–893.

[20] O. Tasar, Y. Tarabalka, P. Alliez, Incremental learning for semantic segmentation of large-scale remote sensing data, IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 12 (9) (2019) 3524–3537.

[21] H. Bilen, A. Vedaldi, Weakly supervised deep detection networks, in: CVPR, 2016, pp. 2846–2854.

[22] D. Pathak, P. Krähenbühl, T. Darrell, Constrained convolutional neural networks for weakly supervised segmentation, in: ICCV, 2015, pp. 1796–1804.

[23] P. Nguyen, T. Liu, G. Prasad, B. Han, Weakly supervised action localization by sparse temporal pooling network, in: CVPR, 2018, pp. 6752–6761.

[24] J. Han, D. Zhang, G. Cheng, L. Guo, J. Ren, Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning, IEEE Trans. Geoscience and Remote Sensing 53 (6) (2015) 3325–3337.

[25] F. Ma, F. Gao, J. Sun, H. Zhou, A. Hussain, Weakly supervised segmentation of SAR imagery using superpixel and hierarchically adversarial CRF, Remote Sensing 11 (5) (2019) 512.

[26] M. Carbonneau, V. Cheplygina, E. Granger, G. Gagnon, Multiple instance learning: A survey of problem characteristics and applications, Pattern Recognit. 77 (2018) 329–353.

[27] M. Ilse, J. M. Tomczak, M. Welling, Attention-based deep multiple instance learning, in: ICML, 2018, pp. 2132–2141.

[28] J. Bolton, P. D. Gader, Application of multiple-instance learning for hyperspectral image analysis, IEEE Geosci. Remote Sensing Lett. 8 (5) (2011) 889–893.
[29] L. Cao, F. Luo, L. Chen, Y. Sheng, H. Wang, C. Wang, R. Ji, Weakly supervised vehicle detection in satellite images via multi-instance discriminative learning, Pattern Recognit. 64 (2017) 417–424.
[30] S. E. Yuksel, J. Bolton, P. D. Gader, Multiple-instance hidden markov models with applications to landmine detection, IEEE Trans. Geoscience and Remote Sensing 53 (12) (2015) 6766–6775.
[31] V. Walter, Object-based classification of remote sensing data for change detection, ISPRS Journal of Photogrammetry and Remote Sensing 58 (3) (2004) 225 – 238.
[32] D. Ienco, R. Gaetano, C. Dupaquier, P. Maurel, Land cover classification via multitemporal spatial data by deep recurrent neural networks, IEEE Geosci. Remote Sensing Lett. 14 (10) (2017) 1685–1689.
[33] C. Pelletier, G. I. Webb, F. Petitjean, Temporal convolutional neural network for the classification of satellite image time series, Remote Sensing 11 (5) (2019) 523.
[34] D. Britz, M. Y. Guan, M. Luong, Efficient attention using a fixed-size memory representation, in: EMNLP, 2017, pp. 392–400.
[35] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, in: EMNLP, 2014, pp. 1724–1734.
[36] X. Zhu, D. Cheng, Z. Zhang, S. Lin, J. Dai, An empirical study of spatial attention mechanisms in deep networks, in: ICCV, 2019, pp. 6687–6696.
[37] B. Zhuang, L. Liu, Y. Li, C. Shen, I. D. Reid, Attend in groups: A weakly-supervised deep learning framework for learning from web data, in: CVPR, 2017, pp. 2915–2924.
[38] A. Borji, M. Cheng, Q. Hou, H. Jiang, J. Li, Salient object detection: A survey, Comput. Vis. Media 5 (2) (2019) 117–150.
[39] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, P. Dollár, Designing network design spaces, CoRR abs/14 (2020).
[40] O. Hagolle, M. Huc, D. Villa Pascual, 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 Sensing 7 (3) (2015) 2668–2691.
[41] S. Dupuy, R. Gaetano, L. L. Mezo, Mapping land cover on reunion island in 2017 using satellite imagery and geospatial ground data, Data in Brief 28 (2020).
[42] S. van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, T. Yu, the scikit-image contributors, scikit-image: image processing in Python, PeerJ 2 (2014) e453.
[43] J. Erinjery, M. Singh, R. Kent, Mapping and assessment of vegetation types in the tropical rainforests of the western ghats using multispectral sentinel-2 and sar sentinel-1 satellite imagery, Remote Sensing of Environment 216 (2018) 345–354.
[44] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, J. Schmidhuber, LSTM: A search space odyssey, IEEE Trans. Neural Networks Learn. Syst. 28 (10) (2017) 2222–2232.
[45] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, CoRR abs/1412.6980 (2014).
[46] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, P. Muller, Deep learning for time series classification: a review, Data Min. Knowl. Discov. 33 (4) (2019) 917–963.