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/JSTARS.2021.3119191, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

Multi-sensor land cover classification with sparsely annotated data based on Convolutional Neural Networks and Self-Distillation

Yawogan Jean Eudes Gbodjo, Olivier Montet, Dino Ienco, Raffaele Gaetano and Stephane Dupuy

***Abstract*— Extensive research studies have been conducted in recent years to exploit the complementarity among multi-sensor (or multi-modal) remote sensing data for prominent applications such as land cover mapping. In order to make a step further with respect to previous studies which investigate multi-temporal SAR and optical data or multi-temporal/multi-scale optical combinations, here we propose a deep learning framework that simultaneously integrates all these input sources, specifically multi-temporal SAR/optical data and fine scale optical information at their native temporal and spatial resolutions. Our proposal relies on a patch-based multi-branch convolutional neural network (CNN) that exploits different per source encoders to deal with the specificity of the input signals. In addition, we introduce a new self-distillation strategy to boost the per source analyses and exploit the interplay among the different input sources. This new strategy leverages the final prediction of the multi-source framework to guide the learning of the per source CNN encoders supporting the network to learn from itself.**
Experiments are carried out on two real world benchmarks, namely the *Reunion island* (a french overseas department) and the *Dordogne* study site (a southwest department in France) where the annotated reference data were collected under operational constraints (sparsely annotated ground truth data). Obtained results, providing an overall classification accuracy of about 94% (resp. 88%) on the *Reunion island* (resp. the *Dordogne*) study site highlight the effectiveness of our framework based on CNNs and self-distillation to combine heterogeneous multi-sensor remote sensing data and confirm the benefit of multi-modal analysis for downstream tasks such as land cover mapping.

***Index Terms*—Multi-sensor, multi-temporal and multi-scale remote sensing, convolutional neural networks (CNNs), self-distillation, land use and land cover mapping, sparsely annotated data.**

I. Introduction

NOWADAYS, a plethora of satellite missions continuously provides remotely sensed images of the Earth surface via various modalities (e.g. SAR or optical) and at different spatial and temporal scales. Therefore, the same study area can be effectively covered by rich, multi-faceted and diverse information. In particular, with the advent of the European Space Agency's Sentinel missions [1], a set of quasi-synchronous SAR and optical data is systematically made available over any area of the planet's continental surface at high spatial (order of 10m) and temporal (an acquisition up to every five/six days) resolution. The remote sensing community has been focusing its efforts for a while now to demonstrate the benefit to combine the multi-modal information provided by such sensors [2].

With particular emphasis on land use and land cover (LULC) mapping, recently, the community is investigating the potential of deep learning (DL) approaches to integrate complementary sensor acquisitions available on the same study area [3] with the aim to leverage as much as possible the interplay between input sources exhibiting different spectral as well as spatial content to ameliorate the underlying mapping result.

Differently from standard and/or legacy approaches devoted to remote sensing data fusion [2], [4] where, firstly each source is processed independently to extract additional information (i.e. indices in the context of optical data), secondly a machine learning approach is still deployed (independently) for each source and, finally, a voting schema is applied on the output of each source-specific method in order to get the final prediction; deep learning methods have the ability to directly work with raw signal data avoiding intermediate steps (i.e. data harmonization or spatial/temporal resampling) and automatically deal with the process of source combination in an end-to-end manner.

In the works presented in [5], [6], panchromatic and multispectral bands at different spatial resolutions are directly combined to provide LULC mapping at the finest resolution. Recently, Hong et al. [7] proposes to fuse together multispectral LIDAR with hyperspectral optical information for urban land use and land cover classification.

Considering multi-modal remote sensing classification, when at least one of the sources depicts a satellite image time series (SITS), Kussul et al. [8] and Ienco et al. [9] combine together SAR and optical SITS with the aim to leverage the complementarity between active and passive sensors. Moreover, Benedetti et al. [10] and Gadiraju et al. [11] propose to combine multi-temporal and single date very high spatial resolution (VHSR) optical data with the objective to jointly exploit multi-temporal and multi-scale information.

The majority of DL-based multi-modal approaches proposed in remote sensing literature mainly involves two different sources as input. This is especially the case when SITS

Y. J.E. Gbodjo is with INRAE, UMR TETIS, University of Montpellier, Montpellier, France (email: jean-eudes.gbodjo@inrae.fr)
O. Montet is with INRAE, UMR TETIS, University of Montpellier, Montpellier, France (email: olivier.montet@inrae.fr)
D. Ienco is with INRAE, UMR TETIS, LIRMM, University of Montpellier, Montpellier, France (email: dino.ienco@inrae.fr)
R. Gaetano is with Cirad, UMR TETIS, Montpellier, France (email: raffaele.gaetano@cirad.fr)
S. Dupuy is with Cirad, UMR TETIS, Montpellier, France (email: stephane.dupuy@cirad.fr)

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/JSTARS.2021.3119191, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

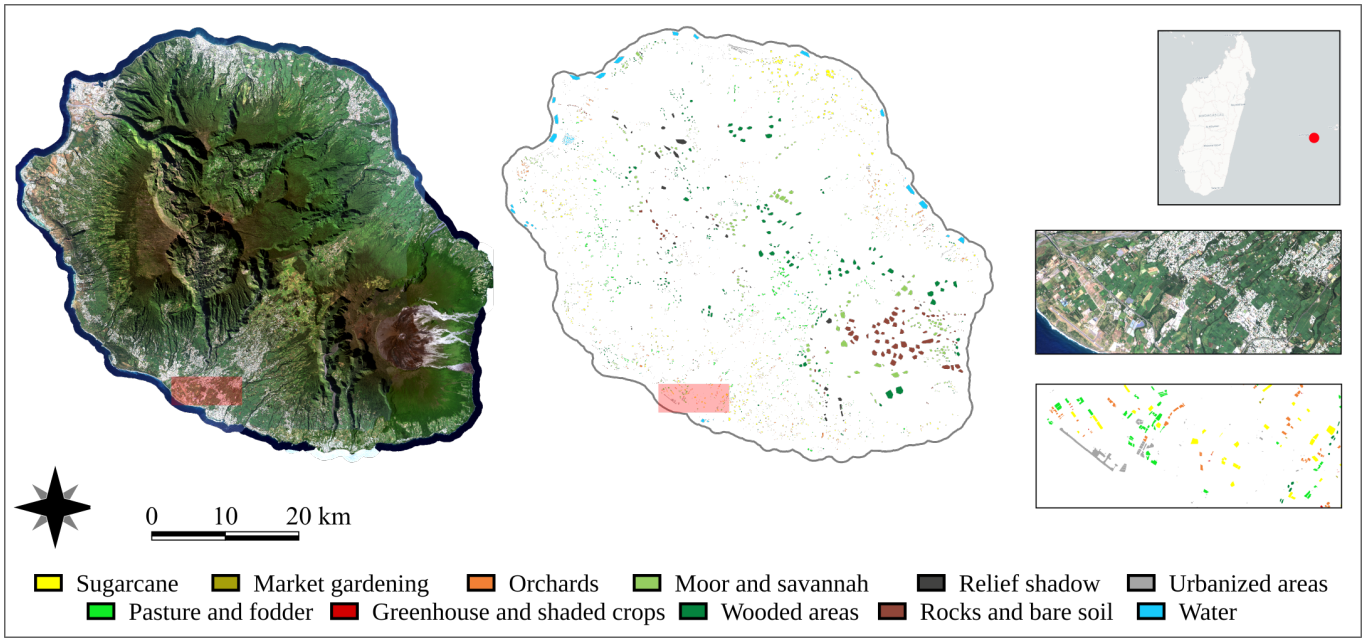

Fig. 1. Location of the *Reunion island* study site. The RGB composite is the VHSR SPOT-6 image. The corresponding ground truth is shown on the right.

data are leveraged in the analysis (SAR with optical and optical multi-temporal/multi-scale).

Here propose a patch-based Convolutional Neural Network (CNN) framework to cope with the combination of SAR and optical SITS data as well as Very High Spatial Resolution (VHSR) optical imagery to support real-world operational LULC mapping under sparsely annotated ground truth data scenario where three different input sources as combined together to ameliorate the underlying land cover mapping process.

The goal is to produce the mapping of a study area from a limited set of per LULC class samples on the same area [12], [13], [14]. Furthermore, in order to get the most out of the interplay among multi-modal information, we design a self-distillation strategy [15], [16] in which per source encoders are optimized considering the final multi-modal classification output. In this way, we allow the DL model to learn from itself. More in detail, we enable the network architecture to distill knowledge from deeper layers (the output of the model) to shallow layers (the per source encoders) with the aim to steer the learning process associated to lower levels of the network. While this process has recently getting attention in computer vision to strength the performance of standard CNN frameworks [17] for mono-source analysis, it is still unexplored in the context of multi-modal (or multi-source) image classification. To assess the effectiveness of the proposed framework, we consider two real-world benchmarks, namely the *Reunion island* (a French overseas department located in Indian Ocean) and the *Dordogne* study site (a southwest department in France) both involving highly sparse ground truth data obtained by means of field campaigns and institutional surveys (see Fig.s 1 and 2). Our framework adopts CNNs as per-source encoders since they are consolidated strategies to deal with VHSR image and, recent studies (e.g. [8], [18], [9], [19]) have highlighted that such models are even competitive for multi-temporal information such as SITS data.

When dealing with real-world LULC mapping in an operational setting, the collected GT is generally sparse due to human-effort and cost constraints [20], [21], [22]. This means that a limited number of polygons (in terms of surface with respect to the study site) is annotated by field experts with the aim to have samples covering the whole study area without taking care to highlight possible spatial correlations among classes (class polygons are far away from each other). For instance, Fig. 1 depicts a study area characterized by sparse GT data. In the extract to the right of the figure we clearly observe that only a small portion of the area is labelled and polygons are spatially sparse. Matter of fact, the most common GT data collection protocol in operational settings prevents the use of standard semantic segmentation approaches [23], [24], [25] widely adopted in the computer vision community, since semantic segmentation strategies require densely annotated patches on which the model is trained on (each pixel should be associated to a label information). For this reason, when sparsely annotated data are considered, patch-based approaches are usually preferred [26], [9], [13]. For more details about patch-based and semantic segmentation approaches, the interested reader can refer to [27].

To summarize, the contributions of our work are:

- A patch-based multi-branch CNN framework to deal with multi-modal remote sensing land cover mapping considering simultaneously three different input sources: SAR/optical SITS and VHSR optical imagery;
- A new self-distillation strategy to transfer knowledge from deeper layers (the output of our model) to shallow ones (the per source encoder layers) with the aim to boost the final classification performances of our multi-modal framework;

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/JSTARS.2021.3119191, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

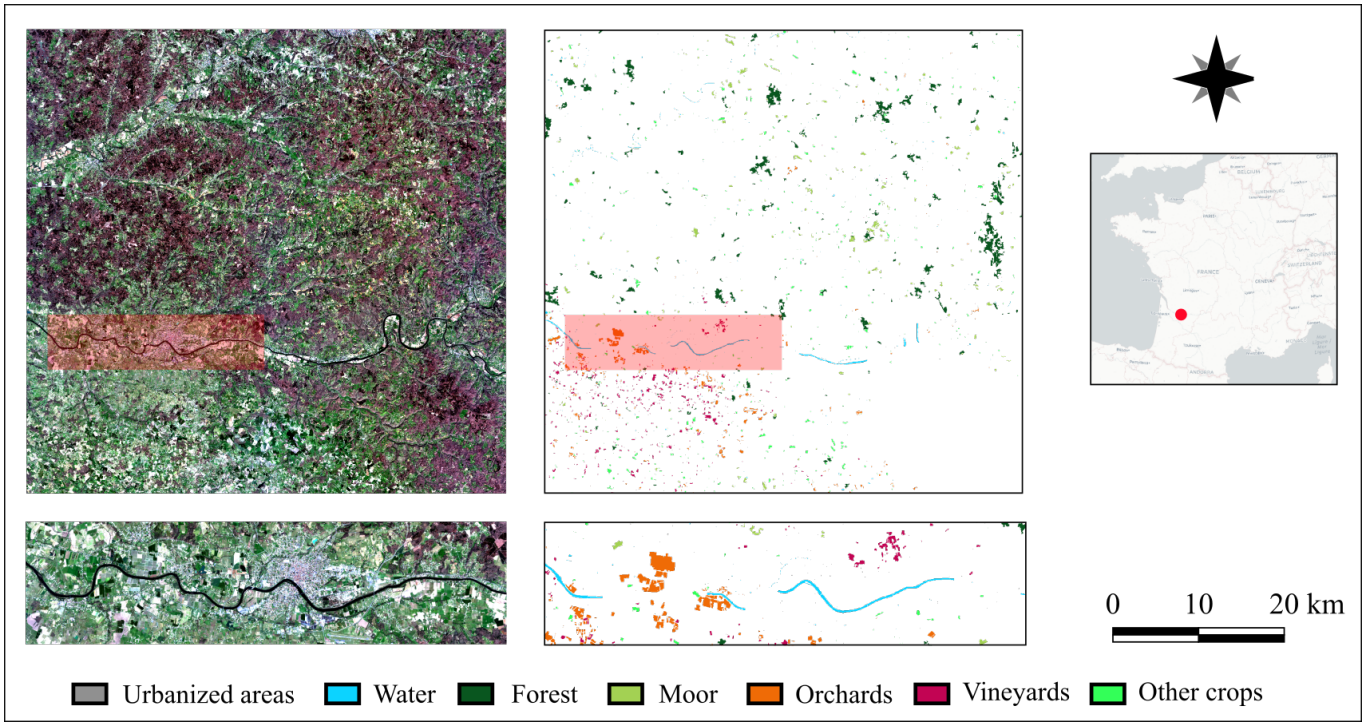

Fig. 2. Location of the *Dordogne* study site. The RGB composite is the VHSR SPOT-6 image. The corresponding ground truth is shown on the right.

- An in-depth experimental study to characterize the interplay among the different input sources. The same study also underlines that the proposed framework is capable to take the most out of the multi-modal information associated to the study sites.

The remainder of this work is structured as follows: first, Section II introduces the data associated to the two study sites; then Section III describes the proposed framework while the experimental settings and the results are reported and discussed in Section IV. Finally, Section V draws the conclusion and possible follows up.

II. Data

The study was carried out on the *Reunion island*, a French overseas department located in Indian Ocean (Fig. 1), and on a part of the Dordogne department located in the southwest of France (Fig. 2). Satellite data on the *Reunion island* consists of a Sentinel-1 (S1) and Sentinel-2 (S2) time series of 26 and 21 images, respectively, acquired over the year 2017, as well as a VHSR SPOT-6 image. The latter was obtained via a radiometrically harmonized mosaic [28] of 4 images acquired respectively on December 26, 2016 and on May 10, June 11, and November 20, 2017 in order to ensure a cloud-free coverage of the whole study area. The *Dordogne* study site dataset includes respectively time series of 31 S1 and 23 S2 images, both acquired in 2016, and a cloud free VHSR SPOT-6 image dated March 3, 2016.

S1 data was acquired in C-band with co- and cross-polarization (VH and VV) and in ascending orbit. The data was downloaded from the PEPS platform [1] in the *Ground Range Detected* format and *Interferometric Wideswath* mode [2] with a pixel spacing of 10×10-m. The S1 images were first radiometrically calibrated in back-scatter values, then orthorectified and finally a multi-temporal filtering [29] was performed over the time series in order to reduce speckle. The S2 images were downloaded from the THEIA pole platform [3] at level-2A (top of canopy reflectance values) and were provided with cloud masks. Only 10-m spatial resolution bands (Blue, Green, Red and Near infrared spectrum) were considered in this analysis. A preprocessing was performed over each band to fill cloudy pixel values as detected by the supplied cloud masks through a linear multi-temporal interpolation (cf. temporal gap-filling [12]). In addition, two spectral indices were then extracted and involved in the analysis: the NDVI [30] and the NDWI [31] leading to a total of six channels describing each Sentinel-2 image. The SPOT-6 images consist of one panchromatic and four multispectral bands (Blue, Green, Red and Near infrared spectrum) at 1.5-m and 6-m spatial resolution respectively, which have been preprocessed in top of atmosphere reflectance.

The GT data for the *Reunion island* was built from various sources: the *Registre Parcellaire Graphique* (RPG) reference data for 2016 (the French land parcel identification system), Global Positioning System records from June 2017 and the visual interpretation of a SPOT image completed by a field expert with knowledge of territory. The *Reunion island* dataset is publicly available[4] [32]. Similarly for the *Dordogne* site[5], the GT was built from RPG reference data for 2016 and

[1]https://peps.cnes.fr/

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

[3]http://theia.cnes.fr

[4]https://doi.org/10.18167/DVN1/TOARDNandadditionalinformationcanbefoundin

[5]Currently available upon request

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/JSTARS.2021.3119191, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

TABLE I
CHARACTERISTICS OF THE REUNION ISLAND GROUND TRUTH

Class	Label	Polygons	Pixels
1	*Sugarcane*	869	88 962
2	*Pasture and fodder*	581	68 098
3	*Market gardening*	758	17 488
4	*Greenhouse and shaded crops*	249	1 908
5	*Orchards*	767	33 721
6	*Wooded areas*	570	205 023
7	*Moor and Savannah*	506	155 231
8	*Rocks and natural bare soil*	299	154 343
9	*Relief shadow*	81	54 301
10	*Water*	177	82 592
11	*Urbanized areas*	1 126	19 056
Total		5 983	880 723

TABLE II
CHARACTERISTICS OF THE DORDOGNE SITE GROUND TRUTH

Class	Label	Polygons	Pixels
1	*Urbanized areas*	253	2 002
2	*Water*	679	50 471
3	*Forest*	199	378 969
4	*Moor*	184	99 627
5	*Orchards*	608	97 546
6	*Vineyards*	593	92 259
7	*Other crops*	584	93 562
Total		3 100	814 436

the visual interpretation of a SPOT image as well. For both study sites, the GT was assembled in Geographic Information System (GIS) vector file, containing a collection of polygons, each attributed with a land cover category (See Tables I and II).

Finally, the polygons have been rasterized at the Sentinel spatial resolution (10-m), obtaining 880 723 labeled pixels for the *Reunion island* (respectively 814 436 labeled pixels for the *Dordogne* site). Owing to the fact that the GT is sparsely annotated, as can be observed (Fig.s 1 and 2), we focus our efforts on patch-based multi-modal remote sensing classification strategies instead of semantic segmentation ones since the latter requires densely labeled GT data conversely to the ones we dispose in our context.

III. FRAMEWORK

In this section we introduce our framework, named *MMCNN*$_{SD}$ (Multi-Modal CNN with per source Self-Distillation). Firstly, we supply an overview of the general multi-modal architecture , then, we describe the new self-distillation strategy we have introduced and finally, we introduce the per source components we have adopted to manage the different remote sensing data sources.

A. Multi-modal patch-based CNN

Fig. 3 depicts the proposed framework, *MMCNN*$_{SD}$. In our scenario, each geospatial location is described by means of different and complementary information, each of them coming from a different sensor.

The model has three branches, one for each of the input sources: S1 SITS, S2 SITS and VHSR SPOT imagery. Each branch is associated to an encoder network that extracts a source specific representation: R_{S1}, R_{S2} and R_{SPOT}. Successively, the different per source representations are aggregated together considering a late fusion schema [33] by summing together the three per source representations with the aim to obtain a multi-sensor representation (R_M) of the specific geospatial location. Finally, the multi-sensor representation is fed through two fully connected layers and an output layer with the goal to obtain the final classification decision for the considered geospatial location.

MMCNN$_{SD}$ leverages a self-distillation component [15], [16] that supports the network to learn from itself. More precisely, for each per source encoder we add an output layer (auxiliary classifier) with the aim to forcing the extraction of complementary and discriminative information from each of the input modality. The per source output layers are trained to mime the behavior of the final multi-modal classification as showed in Fig. 3 with the goal to distill knowledge from deeper layers (the output of our model) to shallow ones (the per source encoder layers). While classical knowledge distillation [16] is based on a teacher-student framework where the objective is to distill/transfer the dark knowledge of the teacher model to the student one, self-distillation [17] does not require a pair (or a set) of distinct models since a model tries to distill/transfer knowledge from itself, autonomously. To make a connection with standard teacher-student frameworks, in our case, the output of *MMCNN*$_{SD}$ (the final multi-modal classification) can be considered as the teacher output while the per source encoders represent the students models that have the goal to mime the teacher behaviour. Here we introduce such a strategy in the context of multi-modal remote sensing analysis. To the best of our literature review [15], [16], this is the first time that such kind of strategy is employed in a multi-source scenario for image analysis and classification.

We formally define the loss of *MMCNN*$_{SD}$ as follows:

$$L = CE(Y, CL(R_M)) + \lambda \sum_{s \in \{S1,S2,SPOT\}} CE(CL(R_M), OUT(R_s)) \tag{1}$$

where Y is the supervision provided by the labeled information, $CE(\cdot,\cdot)$ is the standard Cross-Entropy loss function, $CL(\cdot)$ is a neural network with two fully connected layers with ReLU activation function and Batch Normalization followed by an output layer with SoftMax activation and $OUT(\cdot)$ is a fully connected output layer with SoftMax activation. The λ hyper-parameter controls the trade off between the cost involving the multi-sensor representation and the costs concerning the self-distillation associated to the per source output layers. While the model training involves both the main classifier and the auxiliary classifiers associated to the self-distillation strategy, at inference stage, only the decision provided by the main classifier $CL(R_M)$ is considered. The set of parameters associated to the entire framework (per source feature encoders, prediction and auxiliary classifiers) are learnt end-to-end.

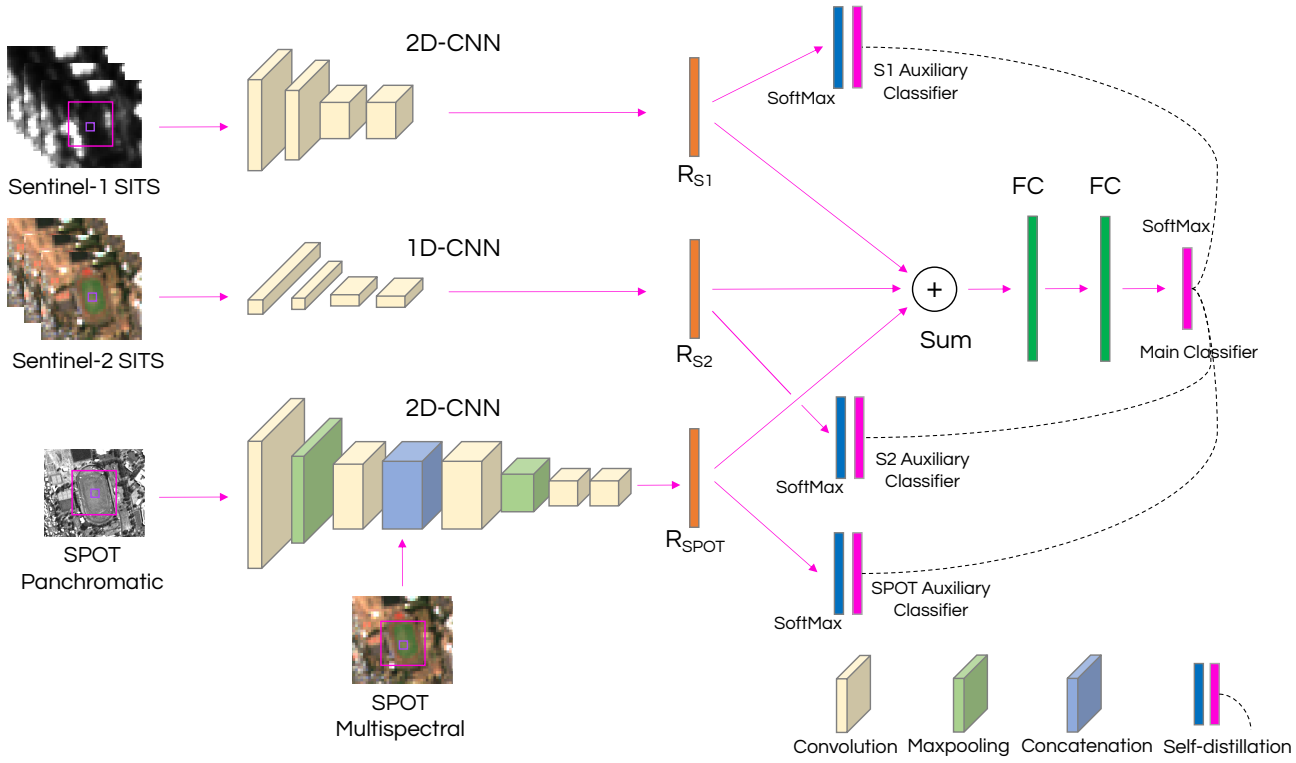

Fig. 3. Overview of $MMCNN_{SD}$ framework. The architecture has three branches, each of them dedicated to an input source. Sentinel-1 SITS and SPOT data are processed by means of 2D-CNN encoders while Sentinel-2 SITS is analyzed through a 1D-CNN encoder. Then, the per-source feature representations are aggregated by the means of the sum operation in order to perform the final land cover classification. To this end, a main classifier associated to the aggregated features and per-source auxiliary classifiers, supervised from the distillation of the main classifier, are employed.

B. Per Source CNN encoders

Due to the fact that the different sensors contain diverse and complementary information, we design specific CNN encoders for each of them.

For the *S1* SITS data we consider a two dimensional convolutional neural network (2D-CNN) with the goal to alleviate possible issues induced by spatial speckle phenomena that usually affects SAR signal [34]. To this end, the S1 SITS described in Section II is organized as a stacked image with as many bands as the number of timestamps times 2 since S1 data have backscatter values with two polarizations: VV and VH. Patches extracted from the stacked image are then concatenated and constitute the input information for the Sentinel-1 encoder branch.

For the *S2* SITS data, according to recent literature on land cover mapping [19], [35], we adopt a one dimensional convolutional neural network (1D-CNN). Such model explicitly manages the sequential information of the SITS since it performs multi-dimensional convolutions on the temporal dimension. Here, only pixel time series information is considered.

For the *VHSR SPOT* image, we still consider a 2D-CNN model with the aim to exploit as much as possible the available fine scale spatial information. In addition, the SPOT image has Panchromatic (PAN) and Multi-Spectral (MS) bands with a resolution of 1.5 and 6 meters, respectively. With the aim to manage such data at their native resolution avoiding as much as possible intermediate resampling steps (e.g. pan-sharpening), the 2D-CNN model for the SPOT image starts processing the PAN information and, once feature maps at the same resolution of the MS information are produced, the MS bands are integrated in the analysis by concatenation. In addition, manage PAN and MS at their original spatial resolution allows to reduce the computational burden that can be introduced if the MS bands are resampled at the same resolution of the PAN information [6].

TABLE III
ARCHITECTURE OF THE MULTI-MODAL CNN ENCODERS. THE PER SENSOR FEATURE REPRESENTATIONS ARE SUCCESSIVELY AGGREGATED TOGETHER BY MEANS OF THE SUM OPERATION AND PROCESSED BY FULLY CONNECTED LAYERS TO PERFORM THE FINAL CLASSIFICATION. (FOR THE SAKE OF READABILITY AUXILIARY CLASSIFIERS ARE OMITTED).

Sentinel-1	**Sentinel-2**	**SPOT**
		7×7 Conv2D (128) on PAN
		MaxPooling2D 3×3
3×3 Conv2D (128)	5×1 Conv1D (128)	3×3 Conv2D (256)
3×3 Conv2D (128)	3×1 Conv1D (128)	Concatenation with MS
3×3 Conv2D (256)	3×1 Conv1D (256)	3×3 Conv2D (256)
1×1 Conv2D (256)	1×1 Conv1D (256)	MaxPooling2D 3×3
GlobAvgPooling2D	GlobAvgPooling1D	3×3 Conv2D (256)
		1×1 Conv2D (256)
		GlobAvgPooling2D
Sum of feature representations		
Fully Connected (512) + ReLU + Batch Normalization		
Fully Connected (512) + ReLU + Batch Normalization		
Fully Connected Output Layer with SoftMax		

To summarize, Table III reports the whole architecture associated to the proposed framework. Conv1D and Conv2D

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/JSTARS.2021.3119191, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

represents one dimensional and two dimensional convolutions, respectively. The associated value (128, 256, 512) is the number of filters. Each convolutional layer is followed by a ReLU activation function, a Batch Normalization and a Dropout layer.

The top of the table (including the Global Average Pooling layers) describes the per source encoders according to the choice we have discussed above. Successively, the per source representations produced by the pooling layers are aggregated together by means of the sum operation and exploited to provide the final land cover prediction. For the sake of clarity and readability, in Table III we have voluntarily omitted to report the auxiliary classifiers associated to the self-distillation strategy. We remind that our framework manages the different sensor information at their original spatial resolutions and, therefore, it explicitly deals with the fusion of multi-scale sensor information.

IV. Experiments

In this section, we present the experimental settings and discuss the results obtained on the datasets previously introduced.

A. Experimental settings

First of all, we validated the architectural choices related to our framework by assessing the behaviour of each sensor encoder. For this evaluation, S1 and S2 SITS are analyzed considering 1D, 2D and 3D-CNN. The 1D and 2D-CNN are the same as in the proposed architecture (See Table III). As concerns the 3D-CNN, it has the same number of convolutional layers and filters as 1D and 2D-CNN. A kernel size of $(3 \times 3 \times 3)$ was employed for the first 3 convolutional layers, as suggested in [18] which found it suitable for SITS data, while the last layer is set up with a kernel size of $(1 \times 1 \times 1)$ similarly to 1D and 2D-CNN encoders. In addition, we used a stride of 2 in the timestamp axis, i.e. $(1 \times 1 \times 2)$, for the second and third convolutional layers with the aim to further explore the temporal signal. Lastly, a global average pooling layer was employed to extract the feature representation before classification.

Then, we evaluate the integration of multi-modal data via the proposed framework. We also consider as competitor for this evaluation an extension of the model introduced in [10], named $M^3Fusion$. The $M^3Fusion$ approach was originally designed to perform land cover classification from S2 SITS and a VHSR SPOT image. It processes input data through dedicated streams (encoders) based on a Recurrent Neural Network (RNN) block to manage S2 SITS and a 2D-CNN branch for the SPOT image. In order to make a fair comparison considering our setting, we have equipped this model with an additional RNN stream especially dedicated to process S1 SITS.

To further assess the behaviour of the proposed framework, we also perform ablation studies to disentangle the interplay among the different input sources (the variants are named *MMCNN*$^{S1+S2}_{SD}$ and *MMCNN*$^{S2+SPOT}_{SD}$, respectively) as well as the contribution of the per source auxiliary classifiers that support the self-distillation strategy (this variant is named *MMCNN*$_{noSD}$). This latter can be assimilated to a standard late fusion procedure as reported in [3]. Additionally, we consider two other baselines: the first one is a variant of the proposed framework named *MMCNN*$_{HardLabels}$ that follows studies on multi-source land cover mapping as [9], [10], in which per source auxiliary classifiers are supervised from the original (hard) labels; the second one named *MMCNN*$^{10}_{SD}$ is a version of our framework which treats all input sources at the same spatial resolution i.e. 10-m. Finally, we gauge the effect of varying in our framework, the per source features dimensionality and the λ hyper-parameter that controls the self-distillation process.

As regards sensor input data, we extracted image patches to describe each specific geospatial location. The Sentinel (S1 and S2) patch size was fixed to 9×9 while similarly to [6], SPOT MS and PAN patch size were set to 8×8 and 32×32, respectively. To fit the input requirements of the $M^3Fusion$ competitor, the VHSR SPOT images were pansharpened on both study sites and multi-spectral image patches of size 32×32 at the highest spatial resolution i.e 1.5-m were extracted. For the *MMCNN*$^{10}_{SD}$ baseline, the pansharpened images were resampled to 10-m spatial resolution using the nearest neighbor method and finally multi-spectral image patches of size 5×5 (covering approximately the same spatial extent as the native resolution image patches) were extracted. Note that we have considered the 2D-CNN designed for the Sentinel data in order to process the SPOT patches at 10-m spatial resolution. Nonetheless, for compatibility purposes, a zero padding was set up for the first convolutional layer.

The values of the dataset were normalized per band in the interval $[0, 1]$, considering the time series and the VHSR, pansharpened and resampled images. The datasets were split into training, validation and test set with a proportion of 50%, 20% and 30% of samples respectively. We imposed that pixels belonging to the same ground truth polygon were assigned exclusively to one of the data partition (training, validation or test) with the aim to avoid possible spatial bias in the evaluation procedure. The evaluated models were optimized via training/validation procedure [36]. Their hyper-parameter settings are reported in Table IV. For the settings of the $M^3Fusion$ model, we adopted the same hyper-parameter values as reported in [10].

TABLE IV
Hyper-parameter settings of the evaluated approaches

Hyper-parameter	Setting or Value
Epochs	300
Learning rate	10^{-4}
Optimizer	Adam [37]
Dropout rate	0.4
Batch size	256
λ (for all the multi-modal approaches)	0.3

The assessment of the model performances was done considering test set and the following metrics: *Accuracy* (global precision), *F1 score* (harmonic mean of precision and recall) and Cohen's *Kappa* (level of agreement between two raters relative to chance). Since the model performances may vary depending on the split of the data due to simpler or more

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/JSTARS.2021.3119191, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

complex samples involved in the different partitions, all metrics were averaged over five random splits of the dataset following the strategy mentioned above. Experiments were carried out on a workstation with an AMD Ryzen 7 3700X CPU, 64 GB of RAM and RTX 2080 NVIDIA GPU. The number of trainable parameters of the evaluated models and the associated time costs are reported in Table V. The different architectures were implemented using the Python Tensorflow library. The code implementation of *MMCNN*$_{SD}$ is available at https://github.com/eudesyawog/S1S2VHSR.

TABLE V
TRAINABLE PARAMETERS OF THE DIFFERENT MODELS AND ASSOCIATED TIME COSTS OVER THE 300 TRAINING EPOCHS

Sensor		Trainable parameters		Training time	
		Reunion	*Dordogne*	*Reunion*	*Dordogne*
S1	1D-CNN	0.62 M	0.62 M	0.37 h	0.40 h
	2D-CNN	0.97 M	0.97 M	0.61 h	0.58 h
	3D-CNN	1.80 M	1.80 M	7.54 h	8.40 h
S2	1D-CNN	0.62 M	0.62 M	0.38 h	0.37 h
	2D-CNN	1.05 M	1.07 M	0.84 h	0.81 h
	3D-CNN	1.82 M	1.81 M	6.35 h	6.48 h
SPOT		2.48 M	2.48 M	2.32 h	2.19 h
$M^3Fusion$		12.6 M	12.58 M	15.37 h	15.80 h
MMCNN$_{SD}^{S1+S2}$		1.20 M	1.20 M	0.96 h	0.93 h
MMCNN$_{SD}^{S2+SPOT}$		2.71 M	2.70 M	2.71 h	2.53 h
MMCNN$_{SD}$		3.28 M	3.29 M	3.39h	3.18 h
MMCNN$_{SD}^{10}$		1.71 M	1.72 M	1.27 h	1.22 h

B. Per Sensor encoder assessment

The performances of the per sensor encoders at the two study sites are reported in Table VI and Table VII, respectively. As regards average results, we note first that leveraging temporal or spatial dependencies for S1 and S2 exhibits different behaviours. Employing 2D convolutions in the CNN instead of 1D convolutions is clearly more effective for S1 while obtained results are comparable for S2. This specific behaviour comes from the fact that 2D convolutions reduces in turn the spatial speckle noise [34] in the S1 data exploiting the spatial context information available when input patches are used. About the 3D-CNN, it achieves overall slightly lower (e.g. for S1) or similar results (e.g. for S2) than the 2D-CNN encoder. Only average results for S2 in the case of *Reunion island* are slightly better than those of the 2D-CNN. Then, the benefit here of leveraging simultaneously convolutions in both spatial and temporal domains via the 3D-CNN is minimal, especially regarding trainable parameters and training time costs (See Table V). For the rest, SAR data (S1) is less effective than optical ones (S2 or SPOT) for the land cover mapping tasks. However, note the significance of the fine scale spatial information provided by the VHSR SPOT data on the *Reunion island*, which gives competitive performances than those of S2 data, with respect to the *Dordogne* site. Overall, the validation of per source CNN encoders suggests that the 2D-CNN model is the most effective to deal with S1 SITS while the 1D-CNN seems more appropriate to manage S2 SITS owing to a cheaper cost in terms of computational training time. Hereafter, S1 and S2 refer to the single-modality models with 2D and 1D-CNN, respectively.

TABLE VI
AVERAGE LAND COVER CLASSIFICATION PERFORMANCES CONSIDERING THE PER SENSOR CNN ENCODERS ON THE REUNION ISLAND

Sensor		F1 Score	Kappa	Accuracy
S1	1D-CNN	64.82 ± 1.32	0.587 ± 0.018	65.63 ± 1.64
	2D-CNN	**73.09** ± 2.62	**0.684** ± 0.030	**73.39** ± 2.66
	3D-CNN	72.35 ± 2.94	0.673 ± 0.036	72.63 ± 3.16
S2	1D-CNN	87.98 ± 1.12	0.859 ± 0.017	88.09 ± 1.06
	2D-CNN	87.41 ± 1.61	0.851 ± 0.021	87.41 ± 1.66
	3D-CNN	**88.62** ± 1.45	**0.866** ± 0.017	**88.66** ± 1.36
SPOT		88.35 ± 1.33	0.862 ± 0.017	88.35 ± 1.39

TABLE VII
AVERAGE LAND COVER CLASSIFICATION PERFORMANCES CONSIDERING THE PER SENSOR CNN ENCODERS ON THE DORDOGNE SITE

Sensor		F1 Score	Kappa	Accuracy
S1	1D-CNN	73.54 ± 2.96	0.644 ± 0.028	75.15 ± 2.76
	2D-CNN	**80.50** ± 2.17	**0.730** ± 0.024	**80.73** ± 2.21
	3D-CNN	78.87 ± 3.12	0.709 ± 0.034	79.43 ± 2.88
S2	1D-CNN	**85.97** ± 2.15	**0.806** ± 0.025	86.04 ± 2.01
	2D-CNN	85.90 ± 1.92	**0.806** ± 0.018	**86.05** ± 1.66
	3D-CNN	85.29 ± 2.35	0.793 ± 0.024	84.88 ± 2.46
SPOT		81.75 ± 2.53	0.745 ± 0.028	81.39 ± 2.62

C. Multi-modal patch-based CNN assessment

The performances of the multi-modal models at the two study sites are reported in Table VIII and Table IX, respectively. Following average behaviour, we first note that combining complementary sensor information systematically ameliorates the land cover classification with respect to per sensor performances. The integration of all available modality via the proposed framework is the most efficient. Our framework achieved the best performances on both study sites, more than 94% (resp. 88%) of accuracy on the *Reunion island* (resp. on the *Dordogne* site) and it also demonstrates its effectiveness considering the $M^3Fusion$ competitor.

TABLE VIII
AVERAGE LAND COVER CLASSIFICATION PERFORMANCES CONSIDERING THE MULTI-MODAL COMBINATION ON THE REUNION ISLAND

Sensor	F1 Score	Kappa	Accuracy
$M^3Fusion$	92.58 ± 0.51	0.912 ± 0.006	92.59 ± 0.50
MMCNN$_{SD}$	**94.34** ± 0.49	**0.934** ± 0.006	**94.38** ± 0.49
MMCNN$_{SD}^{S1+S2}$	91.99 ± 0.42	0.906 ± 0.004	92.05 ± 0.30
MMCNN$_{SD}^{S2+SPOT}$	93.07 ± 1.18	0.918 ± 0.014	93.12 ± 1.16
MMCNN$_{noSD}$	93.21 ± 0.79	0.920 ± 0.009	93.25 ± 0.77
MMCNN$_{HardLabels}$	93.74 ± 0.94	0.926 ± 0.011	93.77 ± 0.96
MMCNN$_{SD}^{10}$	93.87 ± 0.68	0.928 ± 0.008	93.91 ± 0.64

As regards the ablation study on the efficiency of the self-distillation strategy (i.e. *MMCNN*$_{noSD}$ vs *MMCNN*$_{HardLabels}$ vs *MMCNN*$_{SD}$), we note that this architectural component contributes to the final land cover classification performances. Firstly, we observe that the models with the auxiliary classifiers (*MMCNN*$_{SD}$ and *MMCNN*$_{HardLabels}$) achieve better classification results than the baseline model that does not adopt such architectural component (*MMCNN*$_{noSD}$). In order

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/JSTARS.2021.3119191, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

TABLE IX
AVERAGE LAND COVER CLASSIFICATION PERFORMANCES CONSIDERING THE MULTI-MODAL COMBINATION ON THE DORDOGNE SITE

Sensor	F1 Score	Kappa	Accuracy
$M^3Fusion$	87.16 ± 1.47	0.825 ± 0.017	87.48 ± 1.51
$MMCNN_{SD}$	**88.73** ± 1.80	**0.845** ± 0.021	**88.90** ± 1.68
$MMCNN_{SD}^{S1+S2}$	87.09 ± 1.86	0.823 ± 0.020	87.33 ± 1.78
$MMCNN_{SD}^{S2+SPOT}$	88.36 ± 1.70	0.840 ± 0.020	88.56 ± 1.62
$MMCNN_{noSD}$	87.87 ± 1.73	0.832 ± 0.020	87.94 ± 1.54
$MMCNN_{HardLabels}$	88.20 ± 1.72	0.836 ± 0.021	88.18 ± 1.69
$MMCNN_{SD}^{10}$	88.07 ± 1.73	0.837 ± 0.018	88.31 ± 1.70

to further investigate such a phenomena, with a major emphasis on the proposed framework, we depict in Figure 4 the behaviors of $MMCNN_{SD}$ and $MMCNN_{noSD}$ over the established number of training epochs considering their performances on both training and validation sets. As can be noted, while both models clearly fit the training set, the proposed approach ($MMCNN_{SD}$) exhibits superior performances on the validation set underlying that the use of self-distillation strategy clearly supports the model to better generalize on previously unseen data.

Secondly, regarding the direct comparison between our framework ($MMCNN_{SD}$) and the strategy that uses the original (hard) labels to supervise per source auxiliary classifiers ($MMCNN_{HardLabels}$), we can see that the use of self-distillation systematically ameliorates, in terms of evaluation metrics, the joint exploitation of multi-modal sources. This behavior is inline with recent studies on knowledge distillation [15], [16] where it is observed that the soft labels produced by the teacher model (in our case the fused classifier) carry on more useful and easy to exploit information for the student network (in our case the auxiliary classifiers) than the original (hard) label information thus, facilitating the student network to mime the behavior of the teacher model.

Finally, by comparing the $MMCNN_{SD}^{10}$ baseline to the proposed framework, we also notice on both study sites the helpfulness of the fine scale information provided by the VHSR data as well as the significance of integrating multi-scale data at their native spatial resolution for the land cover classification task.

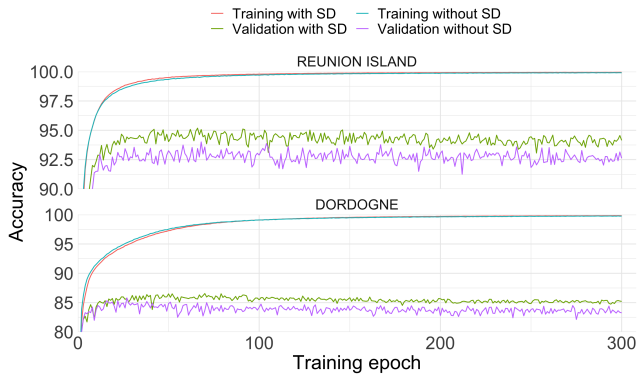

Fig. 4. Learning history, considering Accuracy on training and validation sets, of the proposed framework with and without self-distillation strategy. The latter refers to the behaviour of the method named ($MMCNN_{noSD}$).

D. Effect of varying the framework hyper-parameters

In this evaluation, we analyse two main hyper-parameters associated to the proposed framework. We evaluate how: (i) the dimensionality of per source features extracted by the CNN encoders and (ii) the λ hyper-parameter controlling the self-distillation strategy influence the behaviour of the proposed framework. We vary the former hyper-parameter considering the set of values {64, 128, 256, 512} while the latter one is evaluated according to the following values: {0.1, 0.2, 0.3, 0.4, 0.5}. Results are summarized in Fig.s 5 and 6, respectively.

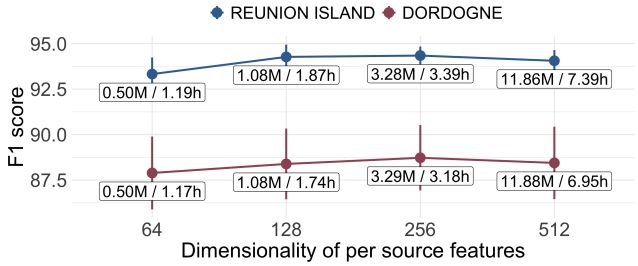

Fig. 5. Land cover classification performances varying the dimensionality of the per source features. Standard deviation is displayed as error bar. Trainable parameters and time costs are shown beside.

The analysis on the dimensionality of per source features shows that 256 features seem suitable for the proposed framework on both study sites and the performance is relatively stable (between 93% and 94% of F1 score on the *Reunion island* and around 88% on the *Dordogne* site) with respect to the considered range. Particularly, it is noteworthy that the model can already generalize well with only 64 features, which could reduce the number of trainable parameters and the associated computational cost related to the training stage.

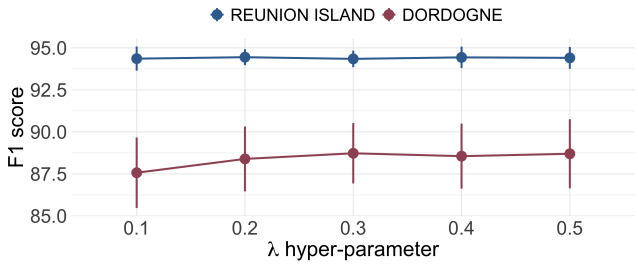

Fig. 6. Land cover classification performances varying the λ hyper-parameter that controls the cost involving the self-distillation strategy. Standard deviation is displayed as error bar.

As regards the assessment on the λ hyper-parameter, here also, we note relatively stable performances on the two study sites for values equal or greater than 0.2. This result underlines that such hyper-parameter does not influence the behaviour of $MMCNN_{SD}$ when it is varied among the considered range.

E. Per class analysis

The per class F1 score at the two study sites are shown in Fig.s 7 and 8, respectively. In this analysis, we note that leveraging complementary sources of information is fully beneficial for almost all the land cover classes, particularly when

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/JSTARS.2021.3119191, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

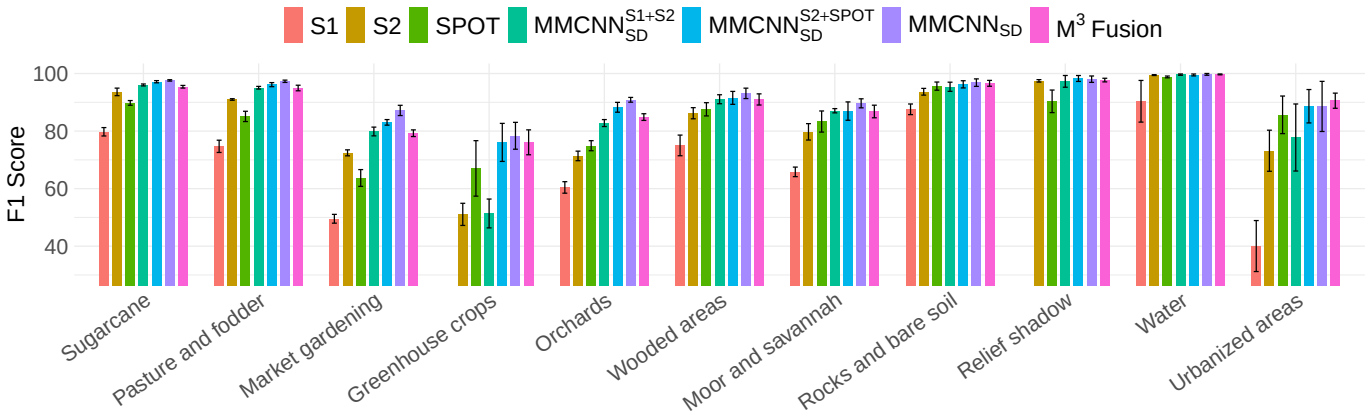

Fig. 7. Average per land cover class F1 score (standard deviation as error bar) considering the various combinations of the multi-modal data (i.e. S1, S2, SPOT, $MMCNN_{SD}^{S1+S2}$, $MMCNN_{SD}^{S2+SPOT}$, $MMCNN_{SD}$).

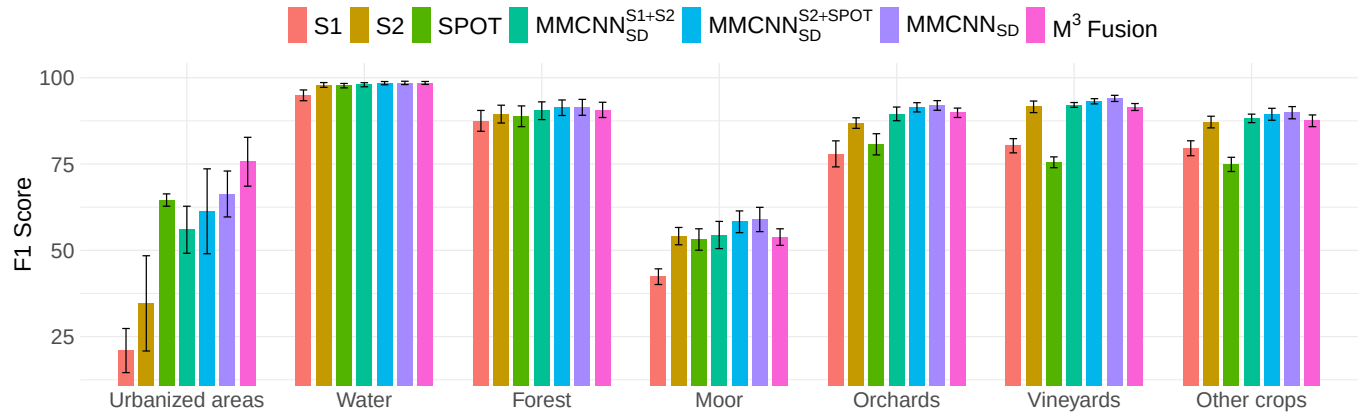

Fig. 8. Average per land cover class F1 score (standard deviation as error bar) considering the various combinations of the multi-modal data (i.e. S1, S2, SPOT, $MMCNN_{SD}^{S1+S2}$, $MMCNN_{SD}^{S2+SPOT}$, $MMCNN_{SD}$).

all modalities are combined. Salient examples on the *Reunion island* are the *Greenhouse and shaded crops*, *Market gardening*, *Orchards* or *Urbanized areas* land cover classes. The F1 score of *Greenhouse and shaded crops*, for instance, improved from 50% (with S2) to 75% (with $MMCNN_{SD}$). Such land cover especially benefits from the fine resolution information provided by SPOT data (67% of F1 score). The benefit is similar for *Urbanized areas* and *Orchards* classes which are better distinguished with fine scale spatial information. On the *Dordogne* site, *Urbanized areas* and crop classes especially profit from the multi-modal combination. To go further with the per land cover class analysis, we supply in Fig.s 9 and 10 the confusion matrices for both study sites. The trend observed in the per class score analysis is confirmed by the confusion matrices. The more complementary sources are combined, the less confusions remain between land cover classes. Only some minor misclassifications remain on the *Reunion island* with the proposed framework, especially between *Greenhouse and shaded crops* and *Urbanized areas*. On the *Dordogne* site, the major confusions between *Moor* and *Forest* classes are also alleviated. Overall, the simultaneous combination of multi-sensor, multi-temporal and multi-scale information was valuable for characterizing land cover classes carrying out not only temporal dependencies, such as the ones related to crops or natural vegetation, but also spatial patterns as evidenced by the performance improvement associated to the *Urbanized areas* land cover class.

F. Qualitative investigation of land cover maps

In Figure 11, we report some extracts from the land cover maps produced on the *Reunion island*. We focused only on this study site since it exhibits a more heterogeneous and challenging landscape in terms of land cover classes than the *Dordogne* site. We recall that all land cover maps were generated at Sentinel spatial resolution (10-m). In addition, owing to the fact that the models are patch-based, the border pixels of the maps (i.e. 4 pixels in each direction since considered Sentinel patch size is 9×9) remain unlabeled. For the sake of clarity, we only considered extracts of the maps produced by considering $MMCNN_{SD}^{S1+S2}$, $MMCNN_{SD}^{S2+SPOT}$ and $MMCNN_{SD}$. The extracts were selected following discussions we had with field experts and with the aim to be representative of observations made in the per land cover class analysis.

The first extract (Fig. 11a–d) depicts a part of Saint-Pierre, a coastal urban area with sugarcane and orchards plantations.

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/JSTARS.2021.3119191, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

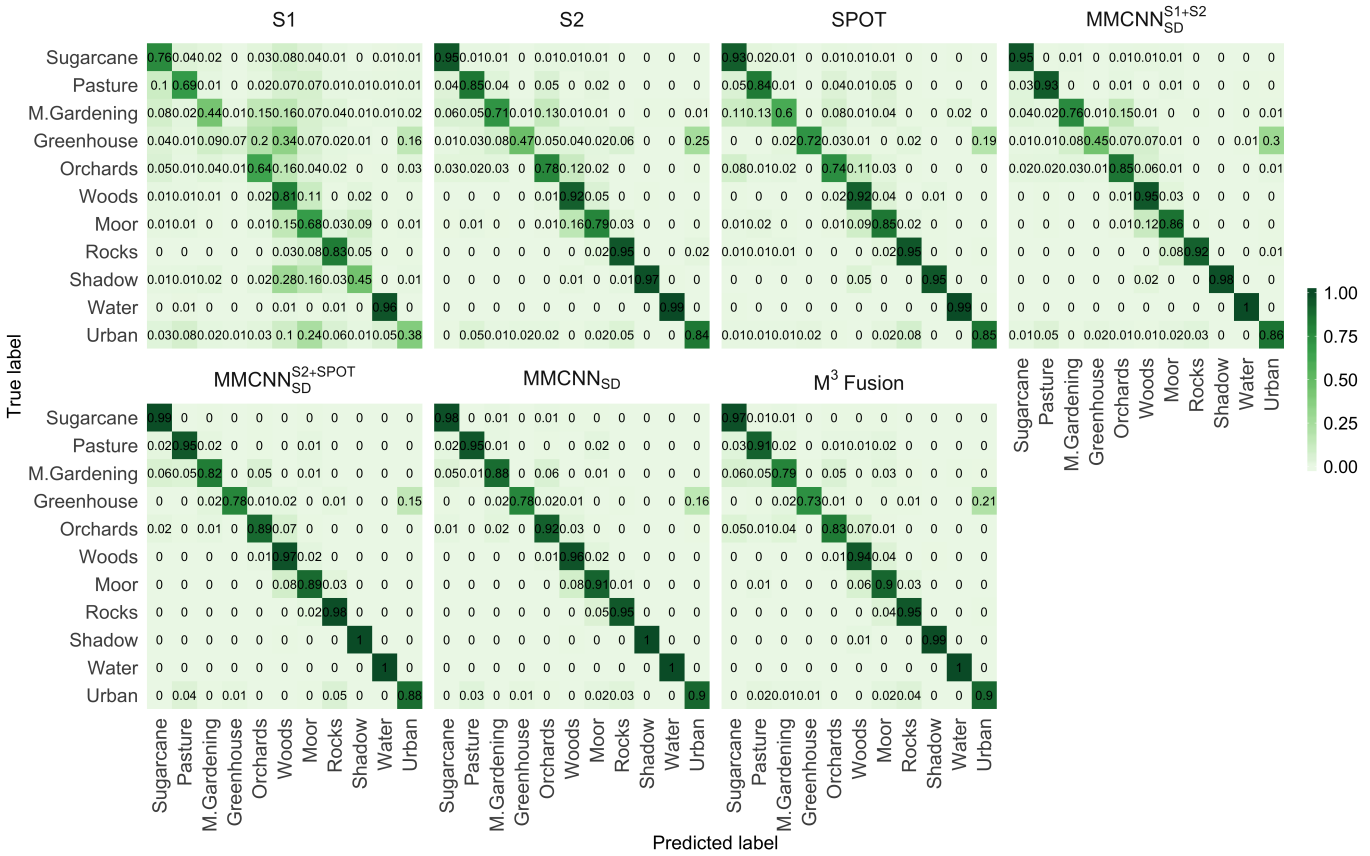

Fig. 9. Confusion matrices of the land cover classification considering the various combinations of the multi-modal data (i.e. S1, S2, SPOT, *MMCNN*$_{SD}^{S1+S2}$, *MMCNN*$_{SD}^{S2+SPOT}$, *MMCNN*$_{SD}$, $M^3 Fusion$).

Misclassifications between *Urbanized areas* and *Greenhouse and shaded crops* can be highlighted in *MMCNN*$_{SD}^{S1+S2}$ extract while the introduction of fine scale spatial information (cf. *MMCNN*$_{SD}^{S2+SPOT}$ and *MMCNN*$_{SD}$ extracts) significantly reduced this issue. The second extract (Fig. 11e–h) is located within the Cilaos cirque, a landscape consisting of hamlets with some market gardening activities surrounding. Here, the *MMCNN*$_{SD}^{S1+S2}$ map exhibits major misclassifications between *Rocks and natural bare soil* class and *Urbanized areas*. This artifact is still slightly noticeable in the *MMCNN*$_{SD}$ classification, while S2 and SPOT combination (i.e. *MMCNN*$_{SD}^{S2+SPOT}$) better deals with the *Rocks and natural bare soil* class. The third extract (Fig. 11i–l) shows an area around Le Tampon, a mixed urban and pasture landscape with some market gardening. Beyond the confusions exhibited by *MMCNN*$_{SD}^{S1+S2}$ between *Urbanized areas* and *Greenhouse and shaded crops*, we note a general overestimation of *Orchards* plantations although minimized by *MMCNN*$_{SD}^{S2+SPOT}$ and *MMCNN*$_{SD}$. The fourth extract (Fig. 11m–p) depicts the Belouve forest which consists of a primary growth forest and forest plantations. There is some minor inaccuracies in the forest detection, misclassified with *Orchards* and *Moor and savannah* classes, which are suppressed in the *MMCNN*$_{SD}$ map. Finally, the fifth and last extract (Fig. 11q–t) focused on the Saint-Gilles les Bains area. The landscape consists of orchards, savannah, some sugarcane plantations as well as built-up. According to field experts, there is a general underestimation of *Moor and savannah* class which is classified as *Wooded areas*, although *MMCNN*$_{SD}^{S2+SPOT}$ combination, slightly alleviate this issue. To wrap up, this qualitative investigation also validate the benefit to combine multi-modal remote sensing data for land cover mapping. Overall, *MMCNN*$_{SD}^{S2+SPOT}$ and *MMCNN*$_{SD}$ land cover maps are of a satisfying quality while *MMCNN*$_{SD}^{S1+S2}$ exhibits extensive errors. This fact is probably due to the noise remaining in SAR data, which sometimes leads to inaccuracies such as the overestimation of orchards areas, and the precious information provided by the SPOT image that is especially pertinent for the considered study area.

G. Visualisation of internal feature representations

In this last stage of our experimental results, we supply a visualisation of the internal feature representation learned by considering the various combinations of the multi-modal data at the two study sites. To this end, we randomly chose 300 samples per land cover class in the test set and we extracted their feature representation. Subsequently, we applied t-SNE [38] and reduced the feature dimensionality to 2 for visualisation purposes. Results are displayed in Fig.s 12 and 13, respectively. On both study sites, we can observe an improved separability of the per land cover class representations as additional and complementary sensors information are combined. As underlined before, S1 is less discriminative than

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/JSTARS.2021.3119191, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

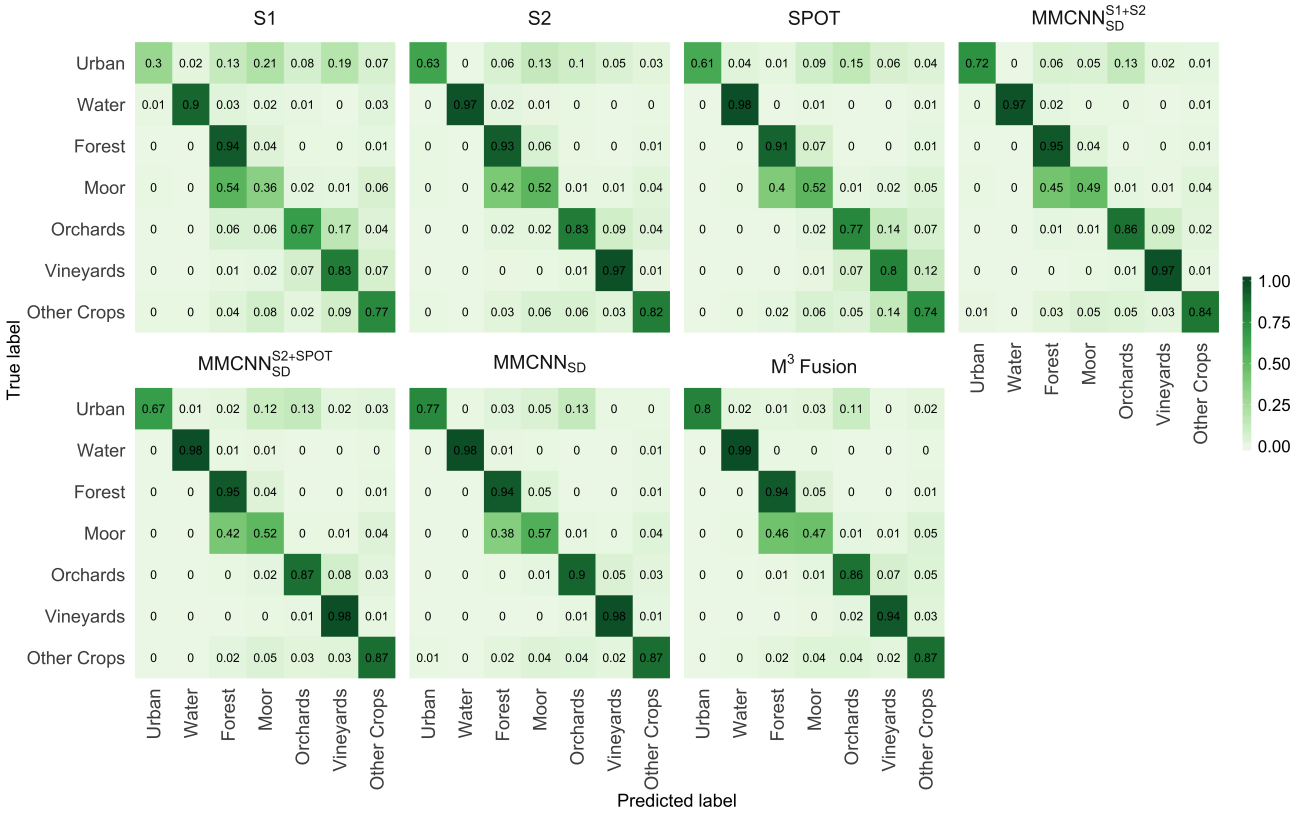

Fig. 10. Confusion matrices of the land cover classification considering the various combinations of the multi-modal data (i.e. S1, S2, SPOT, $MMCNN_{SD}^{S1+S2}$, $MMCNN_{SD}^{S2+SPOT}$, $MMCNN_{SD}$, $M^3 Fusion$).

optical sensors (i.e. S2 or SPOT) while the fine scale spatial information carried out by SPOT is particularly relevant to disentangle the per class feature visualisation on the *Reunion island*. However, some land cover class representations are still barely separable with single-modality data especially *Orchards* and *Wooded areas* or *Pasture and fodder* and *Market gardening* on the *Reunion island* (respectively *Moor* and *Forest* or *Orchards*, *Vineyards* and *Other crops* on the *Dordogne* site). Such ambiguities are successively alleviated by the combination of the multi-modal data, especially $MMCNN_{SD}^{S2+SPOT}$ and $MMCNN_{SD}$ which separate in a similar way the land cover classes, while $MMCNN_{SD}^{S1+S2}$ notably on the *Reunion island* site is still affected by these confusions. Overall, the visualisation of internal features representation is coherent with the quantitative as well as qualitative findings we previously discussed.

V. Conclusion

In this work, we have presented a framework, named $MMCNN_{SD}$, to deal with the task of multi-modal land cover mapping. More specifically, $MMCNN_{SD}$ exploits, simultaneously, multi-temporal and multi-scale remote sensing data, namely Sentinel-1, Sentinel-2 SITS and SPOT VHSR image, for land cover mapping through a three branch patch-based Convolutional Neural Network model that integrates a new self-distillation strategy especially tailored for multi-source analysis. The new knowledge distillation component allows to effectively transfer knowledge from the final prediction to the per source CNN encoders supporting the network to learn from itself. All the process is performed end-to-end.

The results obtained on two real-world benchmarks, the *Reunion island* and the *Dordogne* study sites, have highlighted the quality of the proposed framework regarding both quantitative and qualitative analysis. Furthermore, the obtained results have also validated the importance to boost the representation extracted by per source encoders combining auxiliary classifiers with self-distillation. To sum up, all the experimental findings clearly support the hypothesis that complementary sensor information are definitively valuable for downstream tasks such as land cover mapping.

Possible future work can be related to extend our approach to deal with possible temporal as well as spatial transfer. As of now, our framework deals with a standard land cover mapping setting where a map of a particular study site is derived by learning a classification model from some per-class samples that belongs to the same area. How to transfer a model learnt on a particular area (resp. period of time) to another different area (resp. period of time) is an active domain of research considering multi-temporal mono-source strategies [39], [40] while it is still more challenging and open to investigation when multi-source data are involved.

The proposed framework can also be extended going further with the exploitation of Sentinel-1 and Sentinel-2 data integrating for the former sensor, data coming from both ascending

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/JSTARS.2021.3119191, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

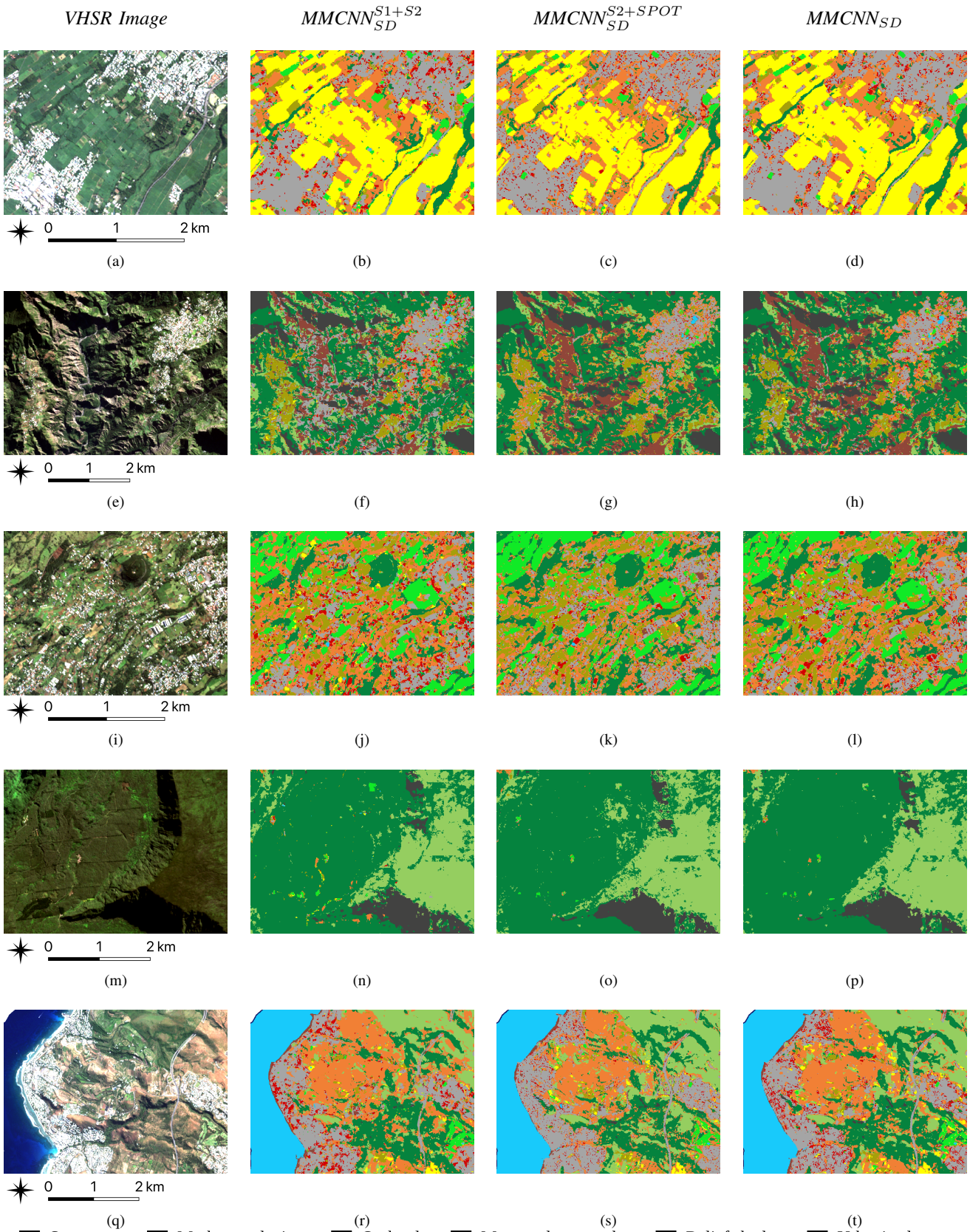

Fig. 11. Qualitative investigation of land cover maps produced by considering $MMCNN_{SD}^{S1+S2}$, $MMCNN_{SD}^{S2+SPOT}$ and $MMCNN_{SD}$. The VHSR SPOT image is supplied as reference. Five areas are detailed, from top to bottom: Saint-Pierre, the Cilaos cirque, Le Tampon, the Belouve forest and Saint-Gilles les Bains.

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/JSTARS.2021.3119191, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

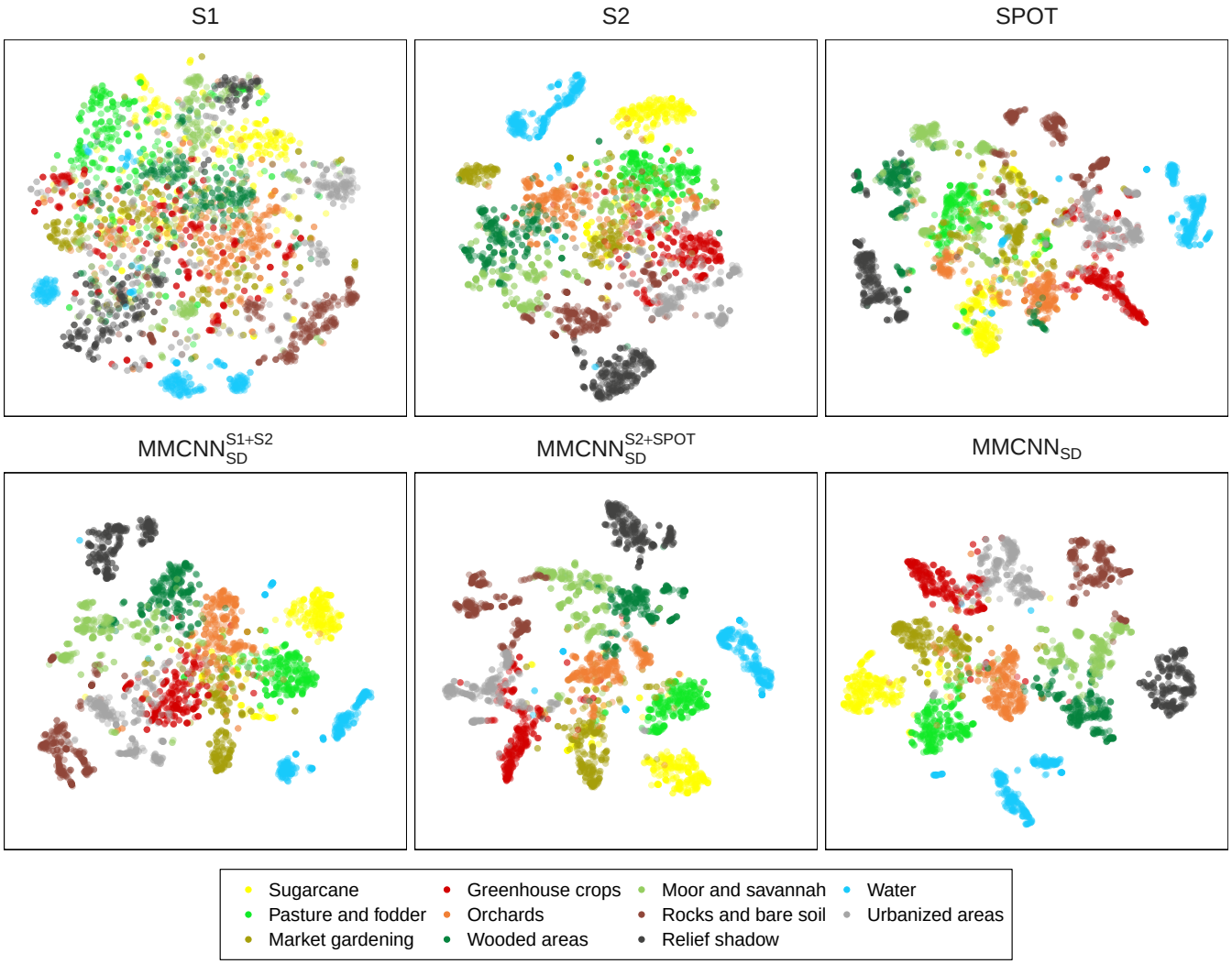

Fig. 12. t-SNE visualisation of internal feature representation learned by considering the various combinations of the multi-modal data (i.e. S1, S2, SPOT, $MMCNN_{SD}^{S1+S2}$, $MMCNN_{SD}^{S2+SPOT}$, $MMCNN_{SD}$) on the *Reunion island* site.

and descending orbits and for the latter sensor, the rest of Sentinel-2 bands, following a schema like the one we have used for the PAN and MS bands of the SPOT image.

FUNDING

This work was supported by the French National Research Agency under the Investments for the Future Program, referred as ANR-16-CONV-0004 (DigitAg) and the Programme National de Télédétection Spatiale (PNTS, https://programmes.insu.cnrs.fr/pnts/), grant no PNTS-2020-13.

REFERENCES

[1] M. Berger, J. Moreno, J. A. Johannessen, P. F. Levelt, and R. F. Hanssen, "Esa's sentinel missions in support of earth system science," *Remote Sensing of Environment*, vol. 120, pp. 84 – 90, 2012.

[2] M. Schmitt and X. X. Zhu, "Data fusion and remote sensing: An ever-growing relationship," *IEEE Geosc. and Rem. Sens. Mag.*, vol. 4, no. 4, pp. 6–23, 2016.

[3] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, and B. Zhang, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–15, 2020.

[4] S. Valero, L. Arnaud, M. Planells, E. Ceschia, and G. Dedieu, "Sentinel's classifier fusion system for seasonal crop mapping," in *IGARSS*. IEEE, 2019, pp. 6243–6246.

[5] X. Liu, L. Jiao, J. Zhao, J. Zhao, D. Zhang, F. Liu, S. Yang, and X. Tang, "Deep multiple instance learning-based spatial-spectral classification for PAN and MS imagery," *IEEE Trans. Geoscience and Remote Sensing*, vol. 56, no. 1, pp. 461–473, 2018.

[6] R. Gaetano, D. Ienco, K. Ose, and R. Cresson, "A two-branch CNN architecture for land cover classification of PAN and MS imagery," *Remote. Sens.*, vol. 10, no. 11, p. 1746, 2018.

[7] D. Hong, J. Chanussot, N. Yokoya, J. Kang, and X. X. Zhu, "Learning-shared cross-modality representation using multispectral-lidar and hyperspectral data," *IEEE Geosci. Remote. Sens. Lett.*, vol. 17, no. 8, pp. 1470–1474, 2020.

[8] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geosci. Remote Sensing Lett.*, vol. 14, no. 5, pp. 778–782, 2017.

[9] D. Ienco, R. Interdonato, R. Gaetano, and 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*, vol. 158, pp. 11 – 22, 2019.

[10] P. Benedetti, D. Ienco, R. Gaetano, K. Ose, R. G. Pensa, and S. Dupuy, "M3 fusion: A deep learning architecture for multiscale multimodal multitemporal satellite data fusion," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 12, pp. 4939–4949, 2018.

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/JSTARS.2021.3119191, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

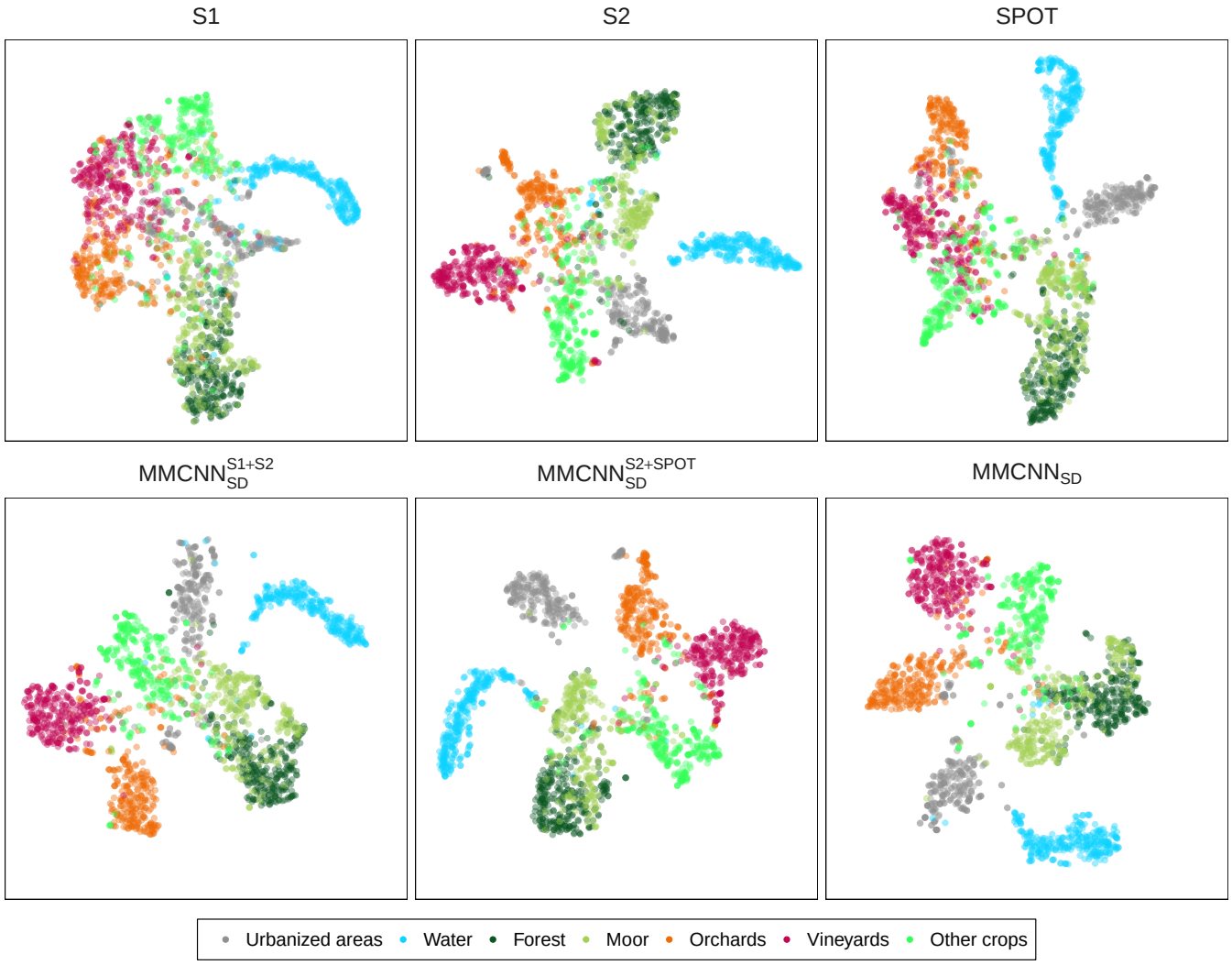

Fig. 13. t-SNE visualisation of internal feature representation learned by considering the various combinations of the multi-modal data (i.e. S1, S2, SPOT, *MMCNN*$^{S1+S2}_{SD}$, *MMCNN*$^{S2+SPOT}_{SD}$, *MMCNN*$_{SD}$) on the *Dordogne* site.

[11] K. K. Gadiraju, B. Ramachandra, Z. Chen, and R. R. Vatsavai, "Multimodal deep learning based crop classification using multispectral and multitemporal satellite imagery," in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2020, pp. 3234–3242.

[12] 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 Sensing*, vol. 9, no. 1, p. 95, 2017.

[13] R. Interdonato, D. Ienco, R. Gaetano, and K. Ose, "Duplo: A dual view point deep learning architecture for time series classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 149, pp. 91 – 104, 2019.

[14] Y. Yuan and L. Lin, "Self-supervised pretraining of transformers for satellite image time series classification," *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, vol. 14, pp. 474–487, 2021.

[15] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *CoRR*, vol. abs/2006.05525, 2020.

[16] L. Wang and K. J. Yoon, "Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021.

[17] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, and K. Ma, "Be your own teacher: Improve the performance of convolutional neural networks via self distillation," in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 3712–3721.

[18] S. Ji, C. Zhang, A. Xu, Y. Shi, and Y. Duan, "3d convolutional neural networks for crop classification with multi-temporal remote sensing images," *Remote Sensing*, vol. 10, no. 2, p. 75, Jan 2018.

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

[20] S. Liu, Q. Shi, and L. Zhang, "Few-shot hyperspectral image classification with unknown classes using multitask deep learning," *IEEE Trans. Geosci. Remote. Sens.*, vol. 59, no. 6, pp. 5085–5102, 2021.

[21] Y. Dong, T. Liang, Y. Zhang, and B. Du, "Spectral-spatial weighted kernel manifold embedded distribution alignment for remote sensing image classification," *IEEE Trans. Cybern.*, vol. 51, no. 6, pp. 3185–3197, 2021.

[22] D. He, Y. Zhong, X. Wang, and L. Zhang, "Deep convolutional neural network framework for subpixel mapping," *IEEE Trans. on Geosc. and Rem. Sens.*, vol. -, no. -, pp. 1–22, 2020.

[23] N. Audebert, B. Le Saux, and S. Lefèvre, "Semantic segmentation of earth observation data using multimodal and multi-scale deep networks," in *Computer Vision – ACCV 2016*. Springer International Publishing, 2017, pp. 180–196.

[24] M. Volpi and D. Tuia, "Dense semantic labeling of subdecimeter resolution images with convolutional neural networks," *IEEE Trans. Geoscience and Remote Sensing*, vol. 55, no. 2, pp. 881–893, 2017.

[25] H. Guo, Q. Shi, B. Du, L. Zhang, D. Wang, and H. Ding, "Scene-driven multitask parallel attention network for building extraction in

high-resolution remote sensing images," *IEEE Trans. Geosci. Remote. Sens.*, vol. 59, no. 5, pp. 4287–4306, 2021.

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

[27] S. Liu and Q. Shi, "Local climate zone mapping as remote sensing scene classification using deep learning: A case study of metropolitan china," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 164, pp. 229–242, 2020.

[28] R. Cresson and N. Saint-Geours, "Natural color satellite image mosaicking using quadratic programming in decorrelated color space," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 8, pp. 4151–4162, 2015.

[29] S. Quegan and J. J. Yu, "Filtering of multichannel sar images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 11, pp. 2373–2379, 2001.

[30] J. W. Rouse, R. H. Hass, J. Schell, and D. Deering, "Monitoring vegetation systems in the great plains with ERTS," *Third Earth Resources Technology Satellite (ERTS) symposium*, vol. 1, pp. 309–317, 1973.

[31] B. cai Gao, "Ndwi—a normalized difference water index for remote sensing of vegetation liquid water from space," *Remote Sensing of Environment*, vol. 58, no. 3, pp. 257 – 266, 1996.

[32] S. Dupuy, R. Gaetano, and L. L. Mézo, "Mapping land cover on reunion island in 2017 using satellite imagery and geospatial ground data," *Data in Brief*, vol. 28, p. 104934, 2020.

[33] Y. Hu, A. Soltoggio, R. Lock, and S. Carter, "A fully convolutional two-stream fusion network for interactive image segmentation," *Neural Networks*, vol. 109, pp. 31–42, 2019.

[34] P. Wang, H. Zhang, and V. M. Patel, "Sar image despeckling using a convolutional neural network," *IEEE Signal Processing Letters*, vol. 24, no. 12, pp. 1763–1767, 2017.

[35] P. Tang, P. Du, J. Xia, P. Zhang, and W. Zhang, "Channel attention-based temporal convolutional network for satellite image time series classification," *IEEE Geosci. Remote Sensing Lett.*, vol. -, no. -, pp. 1–5, 2021.

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

[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[38] L. van der Maaten, "Accelerating t-sne using tree-based algorithms," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3221–3245, 2014.

[39] B. Lucas, C. Pelletier, D. F. Schmidt, G. I. Webb, and F. Petitjean, "Unsupervised domain adaptation techniques for classification of satellite image time series," in *IGARSS*. IEEE, 2020, pp. 1074–1077.

[40] B. Lucas, C. Pelletier, D. Schmidt, G. I. Webb, and F. Petitjean, "A Bayesian-inspired, deep learning-based, semi-supervised domain adaptation technique for land cover mapping," *Machine Learning*, Mar. 2021.

Yawogan Jean Eudes received the M.Sc. degree in Geomatics from the University of Jean Jaures, Toulouse, France, in 2018. He is currently working toward his Ph.D. in computer science at the UMR TETIS laboratory, INRAE, Montpellier working on machine learning approaches devoted to manage multi-source remote sensing data for agriculture monitoring systems.

Olivier Montet received the M.Sc. degree in computer science from the University of Montpellier, Montpellier, France in 2020. From February to September 2020, he was an Intern with UMR TETIS, INRAE, Montpellier working on deep learning approaches to manage multisource and multiscale remote sensing data.

Dino Ienco received the M.Sc. and Ph.D. degrees in computer science both from the University of Torino, Torino, Italy, in 2006 and 2010, respectively. He joined the TETIS Laboratory, IRSTEA, 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. Dr. Ienco served in the program committee of many international conferences on data mining, machine learning, and database including IEEE ICDM, ECML PKDD, ACML, IJCAI as well as served as a Reviewer for many international journal in the general field of data science and remote sensing.

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. He has been a European Research Consortium for Informatics and 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 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 multi-sensor imagery.

Stephane Dupuy was born in France in 1972. He received the M.Sc. degree in geography and remote sensing from the Montpellier University, Montpellier, France, in 2011. From 1994 to 2007, he was with CS Company, Toulouse, France. He joined the Cirad and the TETIS Research Unit in 2007. From 2007 to 2015, he was with Montpellier University. Since 2015, he has been with the CIRAD, UMR TETIS, Runion Island (Indian Ocean), France.