GENERALIZED KNOWLEDGE DISTILLATION FOR MULTI-SENSOR REMOTE SENSING CLASSIFICATION: AN APPLICATION TO LAND COVER MAPPING

Dino Ienco[1]*, Yawogan Jean Eudes Gbodjo[1], Raffaele Gaetano[2] and Roberto Interdonato[2]

[1] INRAE, UMR TETIS, LIRMM, University of Montpellier, France - dino.ienco@inrae.fr ; jean-eudes.gbodjo@inrae.fr
[2] CIRAD, UMR TETIS, Montpellier, France - raffaele.gaetano@cirad.fr ; roberto.interdonato@cirad.fr

KEY WORDS: Satellite image time series, Sentinel-1, Sentinel-2, Generalized Knowledge Distillation, Land Use Land Cover mapping

ABSTRACT:

Due to the proliferation of Earth Observation programmes, information at different spatial, spectral and temporal resolution is collected by means of various sensors (optical, radar, hyperspectral, LiDAR, etc.). Despite such abundance of information, it is not always possible to obtain a complete coverage of the same area (especially for large ones) from all the different sensors due to: (i) atmospheric conditions and/or (ii) acquisition cost. In this context of data (or modalities) misalignment, only part of the area under consideration could be covered by the different sensors (modalities). Unfortunately, standard machine learning approaches commonly employed in operational Earth monitoring systems require consistency between training and test data (i.e., they need to match the same information schema). Such a constraint limits the use of additional fruitful information, i.e., information coming from a particular sensor that may be available at training but not at test time. Recently, a framework able to manage such information misalignment between training and test information is proposed under the name of Generalized Knowledge Distillation (GKD). With the aim to provide a proof of concept of GKD in the context of multi-source Earth Observation analysis, here we provide a Generalized Knowledge Distillation framework for land use land cover mapping involving radar (Sentinel-1) and optical (Sentinel-2) satellite image time series data (SITS). Considering that part of the optical information may not be available due to bad atmospheric conditions, we make the assumption that radar SITS are always available (at both training and test time) while optical SITS are only accessible when the model is learnt (i.e., it is considered as privileged information). Evaluations are carried out on a real-world study area in the southwest of France, namely *Dordogne*, considering a mapping task involving seven different land use land cover classes. Experimental results underline how the additional (privileged) information ameliorates the results of the radar based classification with a main gain on the agricultural classes.

1. INTRODUCTION

Due to the proliferation of Earth Observation programmes, information at different spatial, spectral and temporal resolution is collected by means of various sensors (optical, radar, hyperspectral, LiDAR, etc.) (Schmitt, Zhu, 2016). These different and orthogonal sources of information can be used to characterize and monitor the evolution of Earth surfaces with the aim to better understand climate change effects (Huang et al., 2019) and provide information on the current state of agricultural or natural resources (Kolecka et al., 2018, Kussul et al., 2017, Gao et al., 2018). Considering the huge amount of multi-source data currently available, one of the main methodological research question that arises today concerns the way to take advantages from the complementarity of these information sources with the goal to improve the reliability of modern monitoring and mapping systems (Schmitt, Zhu, 2016). In spite of all this information, it is not always possible to obtain a complete coverage, in terms of different sensors, of the same area (especially for large ones) due to: (i) atmospheric conditions that may affect the quality of the signal (Zhang et al., 2019) (i.e. clouds or shadows for optical sensor) and/or (ii) the cost of certain types of information (i.e. Very High Resolution, Hyperspectral or LiDAR). In this context of data (or modalities) misalignment, only part of the area under consideration is covered by the different sensors (modalities).

However, common machine learning techniques used to perform Land Use Land Cover (LULC) mapping (e.g., Random Forest, Support Vector Machine or Convolutional Neural Networks) make a consistency hypothesis between training and test data in which the two sets need to match the same information schema (Vapnik, Izmailov, 2015) (i.e. the same number of variables with the same semantic information). Such a constraint can limit the use of additional information (i.e., that may be possible to only exploit at training time) in order to learn and calibrate LULC mapping approaches.

(Vapnik, Izmailov, 2015) introduces a machine learning setting in which a predictive model can be trained by leveraging privileged information that is not available at test time. The proposed framework is named Learning Under Privileged Information (LUPI). A generalization of such setting is proposed in (Lopez-Paz et al., 2016) under the name of Generalized Knowledge Distillation (GKD). While such setting is gaining more and more attention in the field of multi-source and multimedia signal processing (Lambert et al., 2018, Chen et al., 2017, Hoffman et al., 2016, Yao et al., 2018), surprisingly, to the best of our knowledge, the only work that makes a connection between the GKD setting and Earth Observation (EO) data (Very High Resolution satellite image) analysis is presented in (Kampffmeyer et al., 2018). In this work the authors propose to perform semantic segmentation of Very High Resolution (VHR) scenes with Digital Elevation Model as privileged source of information for urban land cover mapping. Due to the increasing availability of multi-source and multi-scale information collected by an increasing number of spatial programmes, we believe that there is large room for LUPI or GKD develop-

[*] Corresponding author

https://doi.org/10.5194/isprs-annals-V-2-2020-997-2020 | © Authors 2020. CC BY 4.0 License.

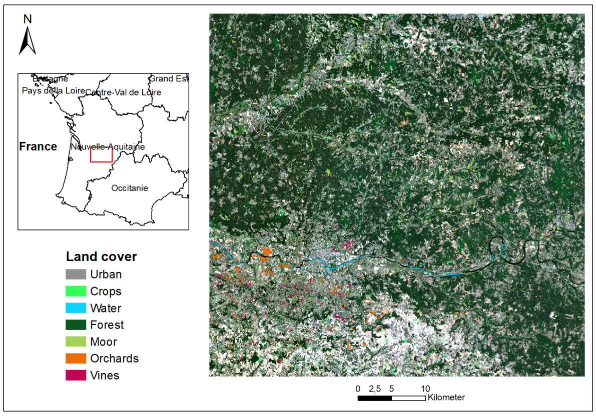


Figure 1. Location of the *Dordogne* study site.

ments in the field of EO analysis. Such methods would facilitate the integration of additional (training) information, today unexploited, to ameliorate machine learning based mapping and monitoring systems.

With the aim to provide a proof of concept regarding the benefits deriving from LUPI or GKD settings in the field of EO analysis, we here propose a Generalized Knowledge Distillation framework to deal with Satellite Image Time Series (SITS) of different nature. We focus on a LULC mapping task on a particular study site, for which we dispose of both radar and optical SITS (i.e. Sentinel-1 and Sentinel-2, respectively) at training time, while only radar SITS are accessible at inference/test. In this case, the optical/Sentinel-2 (S2) SITS data constitutes the privileged information while the radar/Sentinel-1 (S1) SITS data is available at both training and test stages. Even though such scenario focuses on a very specific case, it easily meets real world circumstances, where it may happen that a portion of the study site is constantly (or almost all the time) covered by clouds or shadows. To the sake of evaluation, with the aim to investigate the behavior of our framework in a controlled environment, we set up such a scenario starting from a real-world study site characterized by a complete information regarding both S1 and S2 SITS data. The analysis are conducted considering an OBIA (Object-Based Image analysis) process since working at object level instead of pixels has two main advantages: i) objects represent more representative and potentially feature-rich pieces of information and ii) object based approaches facilitate data analysis scale-up since, for the same area, the number of objects is usually smaller than the number of pixels by several order of magnitude (Ienco et al., 2019).

The paper is organized as follows: Section 2 introduces the *Dordogne* study site involved in the experimental study; the Generalized Knowledge Distillation framework to cope with multi-source SITS data, named S1_{S2}GKD, is described in Section 3; the experimental evaluation and the obtained findings are described in Section 4 while Section 5 concludes the work.

2. STUDY SITE

The analysis was carried out on a part of the *Dordogne* department located in the southwest of France. The considered area covers around 3 000 km^2 (5 572 × 5 390 pixels) and is characterized by an heterogeneous landscape. Figure 1 depicts the study site.

2.1 Sentinel-1 Data

The radar dataset consists of 31 Sentinel-1 (S1) SITS acquired between January and December 2016 in C-band Interferometric Wide Swath (IW) mode with dual polarization (VV+VH). All images, retrieved at level-1C Ground Range Detected (GRD) from the PEPS platform [1], are radiometrically calibrated in backscatter values (decibels, dB) using parameters included in the metadata file and then coregistered with the Sentinel-2 (see Section 2.2) grid and orthorectified at the same 10-m spatial resolution. Finally, a multitemporal filtering was applied to the time series removing artefacts resulting from speckle effect.

2.2 Sentinel-2 Data

The optical data consists of a 23 Sentinel-2 (S2) SITS also acquired between January and December 2016. All images are retrieved from the THEIA pole platform [2] at level-2A top of canopy (TOC) reflectance. Only 10-m spatial resolution bands were considered (i.e. Blue, Green, Red and Near Infrared spectrum (resp. B2, B3, B4 and B8)).

To the sake of completeness, a preprocessing step was performed over each band to replace cloudy observations as detected by the supplied cloud masks through a multi-temporal gapfilling (Inglada et al., 2017). Cloudy pixel values were linearly interpolated using the previous and following cloud-free dates. Then, the Normalized Difference Vegetation Index (NDVI) (Rouse et al., 1973) was calculated from the red (B4) and near infrared bands (B8) for each date. NDVI was chosen as supplementary optical descriptor since it describes the photosynthetic activity and the metabolism intensity of the vegetation.

2.3 Field Data and Preprocessing

The field database was built from the Registre Parcellaire Graphique (RPG) [3] for 2016 and visual interpretation of a SPOT6/7 image as well. The database (available in GIS vector format as a collection of class attributed polygons) includes 8 597 polygons distributed over 7 land cover classes. The per class details of the ground truth data are reported in Table 1. To analyse SITS data at object-level, the mean value of the pixels corresponding to each polygon was calculated over all the timestamps in the time series, resulting in 177 (31×2 for S1 $+$ 23×5 for S2) variables per object.

Class ID	Class Label	Num. of Objects
0	*Urban*	396
1	*Crops*	1457
2	*Water*	1113
3	*Forest*	2205
4	*Moor*	950
5	*Orchards*	1230
6	*Vines*	1246
TOTAL		8 597

Table 1. Characteristics of the Dordogne study site dataset.

3. GENERALIZED KNOWLEDGE DISTILLATION FOR SENTINEL-1 / SENTINEL-2 SITS DATA

In this Section we describe our framework, named S1_{S2}GKD, capable to leverage additional optical information at training

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

[2] `http://theia.cnes.fr`

[3] RPG is part of the European Land Parcel Identification System (LPIS), provided by the French Agency for services and payment

time that will not be available when inference will be performed. Our approach is based on recent developments in the field of Learning Under Privileged Information (Vapnik, Izmailov, 2015) (LUPI) as well as Knowledge Distillation (Cho, Hariharan, 2019). Recently, (Lopez-Paz et al., 2016) has showed the interplay between the two settings and the fact that they can be unified under the name of Generalized Knowledge Distillation.

In our case, we have a training dataset $X^{tr} = \{(x_i^{tr}, y_i)\}_{i=1}^n$ where each sample (x_i^{tr}) is a spatial segment with associated radar/S1 (Rad_i) and optical/S2 (Opt_i) SITS information $(x_i^{tr} = (Rad_i, Opt_i))$ on which we can learn a classification model via the corresponding class label y_i. We denote with C the number of classes. On the other hand, due to possible atmospheric (i.e. clouds or shadows) phenomena, the test dataset on which the model will be deployed is defined as $X^{te} = \{x_j^{te}\}_{i=j}^m$ where each sample (x_j^{te}) is a spatial segment with only radar/S1 (Rad_j) SITS information $(x_j^{te} = (Rad_j))$. The goal is to predict the class label for each example x_j^{te} belonging to the test set.

Standard machine learning techniques require that training and test data match the same information schema (i.e. the same number of variables with the same semantic information) and they will discard the optical information available at training time due to such limitation. Conversely, under the GKD setting (Lopez-Paz et al., 2016), the optical information takes the role of privileged information and it can be integrated in the process to regularize and guide the learning of a model with the aim to improve its generalization capabilities.

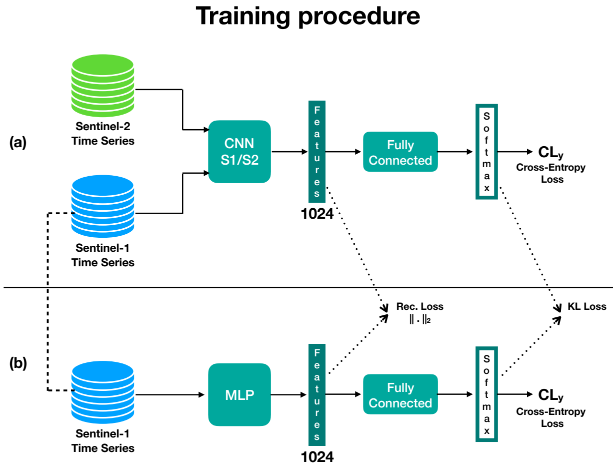


Figure 2. The overview of the training procedure of S1$_{S2}$GKD. Firstly, (a) the multi-source CNN model (teacher network) is trained leveraging both radar and optical SITS; successively (b) a Multilayer Perceptron (student network) is trained considering only radar SITS as input data and it is forced to imitate (or distill) as much as possible the behavior of the multi-source model considering both features extraction and softmax decision. At test time, the *student* model will be deployed to predict the class values of unlabeled samples defined only on radar SITS.

Figure 2 shows the training procedure of the S1$_{S2}$GKD framework. The procedure consists of two main steps: (a) a multi-source model is learnt from the whole set of data (optical and radar SITS) and, successively, (b) a mono-source model is trained considering only radar SITS as input and it is forced to imitate as much as possible the behavior of the multi-source model. The two models are learnt sequentially. Under the Generalized Knowledge Distillation setting, we name the multi-source model *teacher* while the mono-source model is named *student* since the latter is forced to imitate (or distill) the knowledge from the former (more informative) model. In our case, the imitation (or distillation) process is achieved by constraining both the extracted features and the model decisions of the student model to be as much as possible similar to the ones of the teacher model, in correspondence of the same inputs.

3.1 Teacher Model

According to the recent evaluation proposed in (Pelletier et al., 2019), we leverage one dimensional Convolutional Neural Network (CNN) to learn a multi-source discrimination model from SITS data. In (Pelletier et al., 2019), the convolution is realized on the time dimension in order to exploit the temporal information provided by the satellite images sequence.

CNN S1/S2	
Sentinel-1 Branch	**Sentinel-2 Branch**
Conv(nf=128, k=5x1, s=1, act=ReLU)	Conv(nf=128, k=5x1, s=1, act=ReLU)
Conv(nf=128, k=5x1, s=1, act=ReLU)	Conv(nf=128, k=5x1, s=1, act=ReLU)
Conv(nf=256, k=3x1, s=1, act=ReLU)	Conv(nf=256, k=3x1, s=1, act=ReLU)
Conv(nf=256, k=3x1, s=1, act=ReLU)	Conv(nf=256, k=3x1, s=1, act=ReLU)
Conv(nf=256, k=3x1, s=1, act=ReLU)	Conv(nf=256, k=3x1, s=1, act=ReLU)
Conv(nf=256, k=3x1, s=1, act=ReLU)	Conv(nf=256, k=3x1, s=1, act=ReLU)
Conv(nf=512, k=3x1, s=1, act=ReLU)	Conv(nf=512, k=3x1, s=1, act=ReLU)
Conv(nf=512, k=1x1, s=1, act=ReLU)	Conv(nf=512, k=1x1, s=1, act=ReLU)
Conv(nf=512, k=1x1, s=1, act=ReLU)	Conv(nf=512, k=1x1, s=1, act=ReLU)
GlobalAveragePooling	GlobalAveragePooling
CONCAT	

Table 2. CNN architecture of the multi-source (teacher) module.

The details of the CNN architecture for the multi-source analysis are reported in Table 2. Firstly, we can observe that the architecture has two branches: one for the Radar/S1 and one for the optical/S2 SITS. Secondly, we can note that the two branches have the same structure. Successively, at the end of each branch, the information is summarized via a Global Average Pooling (GAP) layer. The GAP layer aggregates each feature maps via the average operator producing a layer with as many neurons as the number of feature maps at the precedent step. Finally, the outputs of the GAP layers are concatenated to provide the feature extracted by the network from the multi-source SITS data. Each branch of the CNN network has nine convolutional $(Conv(\cdot))$ layers defined via the number of filters (nf), the kernel size (k), the stride size (s) and the activation function (act). In our case, we use as activation function the standard Rectifier Linear Unit (ReLU) defined as: $max(0, Z)$ where Z is the results of the convolution between the incoming information and the learnable weights. The result of each convolutional layer $(Conv(\cdot))$ is also associated to a Batch Normalization (BN) and $(DROPOUT)$ layer with drop rate equals to 0.3 $(DROPOUT(BN(Conv(\cdot)))$).

The concatenated feature set (the output of the $CONCAT$ layer) is successively used as input for a classification module (the $Fully\ Connected$ module in Figure 2). The details of such module are reported in Table 3.

Fully Connected
BN(FC(nn=512, act=ReLU))
BN(FC(nn=512, act=ReLU))
FC(nn=C, act=Linear)
Softmax

Table 3. Architectures of the Fully Connected Classification module.

The $Fully\ Connected$ module has three Fully Connected (FC) layers: the first two with a number of neurons (nn) equals to 512 and ReLU activation function while the last layer has linear activation function and as many neurons as the number of classes C. Finally, a Softmax operation is applied to obtain a kind of probability distribution over the set of possible classes. As shown in Figure 2, the multi-source (teacher) network is learnt end-to-end connecting the $CNN\ S1/S2$ module with the $Fully\ Connected$ one.

3.2 Student Model

For our *student* model, we use a simple Multilayer Perceptron (MLP). The MLP network has recently demonstrated its ability to deal with object-based SITS classification (Gbodjo et al., 2019). The MLP architecture we employ is summarized in Table 4. The MLP involves four fully connected layers (FC): three with ReLU activation function, Batch Normalization and Dropout regularisations and a fourth layer without any activation function (simple linear combination of the input with the learnable parameters). This choice is the common strategy adopted for regression tasks (Lathuilière et al., 2018) since the 1 024 features obtained from the last MLP layer will also be constrained to match the features of the *teacher* model (outputted by the concatenation layer). As depicted in Figure 2, the MLP module is connected to a $Fully\ Connected$ module in order to build the mono-source (*student*) model.

Multilayer Perceptron
DROPOUT(BN(FC(nn=2048, act=ReLU)))
DROPOUT(BN(FC(nn=2048, act=ReLU)))
DROPOUT(BN(FC(nn=2048, act=ReLU)))
FC(nn=1024, act=Linear)

Table 4. Architectures of the Multilayer Perceptron (MLP) module.

3.3 Learning and Distillation Strategy

To learn the parameters of the multi-source (*teacher*) model, we employ the Cross-Entropy loss function (Lathuilière et al., 2018) commonly employed for multi-class classification tasks and defined as follows:

$$Loss_{CE} = \sum_{i=1}^{n} \sum_{l=1}^{C} y_{il} \times log(\hat{y}_{il}) \tag{1}$$

where y_{il} is the value of the class l for the sample i (zero everywhere except for the class label associated to the sample i) and $\hat{y}_{il}$ is the value predicted by the model.

Regarding the mono-source (*student*) model, in addition to standard Cross-Entropy loss, we introduce two other loss functions that allow the *student* model to imitate the *teacher*. The first one is inspired by the work on Hallucination networks (Hoffman et al., 2016), in which the features extracted by the *student* model are constrained to be similar, as much as possible, to the features extracted by the *teacher* model. This reconstruction loss is defined as follows:

$$Loss_{REC} = ||sigmoid(f^T) - sigmoid(f^S)||_2^2 \tag{2}$$

where f^T and f^S are the *teacher* and the *student* features, respectively and $sigmoid(\cdot)$ is the sigmoid activation function. In our case, f^T is the output of the $CONCAT$ layer of the CNN $S1/S2$ module described in Table 2 while f^S is the output of the last layer of the MLP module reported in Table 4.

The second loss function devoted to support the distillation process is the Kullback-Leibler divergence between the predicted output distribution of the *teacher* and the *student* (Hinton et al., 2015) models. In (Hinton et al., 2015) the authors employ a temperature scaling factor τ to smooth the last output layer of the neural models before performing the softmax normalization. Here, we set such scaling factor equals to 1. This loss function is defined as follows:

$$LOSS_{KD} = KL(Softmax^T, Softmax^S) \tag{3}$$

where $Softmax^T$ and $Softmax^S$ are the outputs of the Softmax layer for the multi-source *teacher* and the mono-source *student* networks, respectively. The objective of the $LOSS_{KD}$ loss is to force the student model to imitate the output of the teacher model with the aim to distill the teacher behavior into the student network. The final loss function optimized by the *student* network is the follows:

$$Loss_{student} = \alpha \times Loss_{CE} + \beta \times Loss_{REC} + \gamma \times Loss_{KD} \tag{4}$$

where α, β and γ are the hyperparameters associated to each of the three basic loss functions and they are empirically set to 1.0, 1.0 and 10.0, respectively.

4. EXPERIMENTAL EVALUATION

In this section, we present and discuss the experimental results obtained on the *Dordogne* study site introduced in Section 2. We conduct several analyses in order to evaluate the benefits deriving from the use of our framework, by comparing its performances with standard competitors. The methods selected as competitors are a Random Forest (RF), a Multilayer Perceptron (MLP) and a Convolutional Neural Network (CNN) classifier. For the MLP model, we leverage the one presented in Table 4 coupled with the $FullyConnected$ block described in Table 3. For this competitor, we associate the last layer of the architecture presented in Table 4 with a ReLU activation function. For the CNN competitor, we leverage the *Sentinel-1 Branch* of the architecture presented in Table 2 in conjunction with the $FullyConnected$ block described in Table 3. Since standard classification approaches require training and test data are defined on exactly the same feature space, the three competing methods only consider Sentinel-1 SITS information during the training phase.

Firstly, we evaluate the global behavior of the different approaches. Secondly, we report per-class analysis to understand which land covers classes benefit from the use of additional optical training information and, finally, we inspect the confusion matrices produced by each approach.

4.1 Experimental Scenario

We remind the experimental scenario in which our analysis is carried out. We assume that, at training time, we dispose of both optical and radar SITS for a set of labeled samples but, at the inference/test time, only radar SITS information is available due to possible issues. Standard machine learning approaches, usually employed in the remote sensing field, require that training and test samples may be defined on exactly the same set of sources (same feature space). Here, differently from this standard assumption, we relax such constraint and we consider the scenario in which radar (Sentinel-1) SITS information is available at both training and test time while optical (Sentinel-2)

SITS information plays the role of privileged information, i.e., it is only available at training time.

4.2 Experimental Setting

To learn all the deep learning methods we use the Adam optimizer (Kingma, Ba, 2014) with a learning rate equal to 5×10^{-5}. The training process is conducted over 5 000 epochs with a batch size equals to 32. On average, each train epoch takes around a second. Considering our framework, S1$_{S2}$GKD, firstly the teacher classifier is learnt from both Sentinel-1 and Sentinel-2 data and, successively, the student model is learnt to mimic the teacher behavior considering only Sentinel-1 SITS data.

We divide the dataset into three parts: training, validation and test set with a proportion of 50%, 20% and 30% of the objects, respectively. Training data are used to learn the model. The model that achieves the best accuracy on the validation set is subsequently employed to classify the test set. For the RF models, we optimize the model via the maximum depth of each tree (in the range {20, 40, 60, 80, 100}) and the number of trees in the forest (in the set {100, 200, 300, 400, 500}). 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. The assessment of the classification performances is done considering global precision (*Accuracy*), *F-Measure* and *Kappa* measures. For each evaluation metric, we report results averaged over ten random splits performed with the previously presented strategy.

4.3 General behavior

Table 5 reports the averaged results obtained for RF, MLP, CNN and S1$_{S2}$GKD on the *Dordogne* study site. Considering the average performances, we can observe that the proposed framework outperforms all the competing approaches with a gain varying from 4 to 1.30 points of F-Measure. This result shows how integrating privileged information at the training stage is beneficial for the classification process, even if such information is not available when performing predictions.

	F-Measure	*Kappa*	*Accuracy*
RF	62.23 ± 1.20	0.5624 ± 0.0120	64.14 ± 1.09
MLP	62.90 ± 1.42	0.5620 ± 0.0153	63.75 ± 1.33
CNN	59.17 ± 1.34	0.5213 ± 0.0118	60.53 ± 1.12
S1$_{S2}$GKD	**64.27** ± 1.43	**0.5775** ± 0.0153	**65.01** ± 1.34

Table 5. F-Measure, Kappa and Accuracy considering S1$_{S2}$GKD and different competing methods. (Average over ten random splits)

In addition, the reported results are coherent with the results of (Gbodjo et al., 2019) in which the MLP approach exhibits competitive performances w.r.t. standard machine learning approaches usually employed to deal with land use land cover mapping from SITS data.

4.4 Per-class analysis

Table 6 reports the per-class F-Measure obtained by all the methods involved in the evaluation. We can observe that S1$_{S2}$GKD achieves the best performances on all the classes except the *Water* and *Forest* ones where MLP, that exploits only radar information at training time, behaves slightly better. Interestingly, we can observe that many of the land cover classes on which S1$_{S2}$GKD outperforms its competitors are related to agricultural land use: *Crops*, *Orchards* and *Vines*. Considering such classes, we can note that S1$_{S2}$GKD systematically gains more than 2 points (on average) with a major increase (around 3.5 points of F-Measure) w.r.t. the corresponding version without privileged information (MLP), on the *Crops* class.

The monitoring and detection of such agricultural land cover classes is generally performed through optical information. Passive sensors are well suited to monitor vegetation and biophysical properties on the contrary of active (radar) sensors that are more adapted to estimate structural or soil features (Marais-Sicre et al., 2020). Forcing the model trained on only radar data to mimic the behavior of a model learned from both radar and optical makes possible to guide the former to learn source correlations and it supports a better generalization on some particular classes on which it is known that optical sensor are more effective. In addition, we can also see that S1$_{S2}$GKD improves the model performances (compared to the MLP counterparts) considering *Urbanized Area* and *Moor* classes. We can suppose that also in these cases the model takes advantage of the radar/optical source correlation, and that the use of optical information at training time helps to regularise the model decision. Surprisingly, the CNN approach achieves poor results compared to all the other competing approaches. Our intuition behind this phenomena is related to the available amount of training data. Due to the limited number of labeled samples, the CNN is probably not able to learn an effective weight configuration that results in a good classification model.

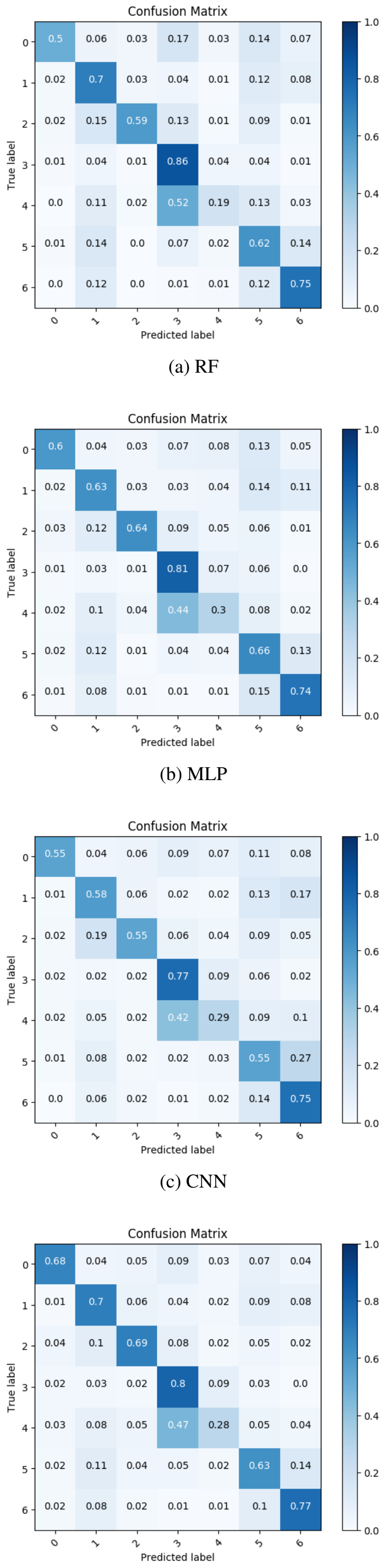
	0–Urban	*1–Crops*	*2–Water*	*3–Forest*	*4–Moor*	*5–Orchards*	*6–Vines*
RF	57.39	65.05	66.08	73.78	24.49	55.37	71.74
MLP	60.08	62.34	**66.12**	**74.2**	32.7	56.41	70.47
CNN	54.49	61.8	61.13	73.44	25.42	50.63	64.62
S1$_{S2}$GKD	**61.99**	**65.98**	66.08	73.5	**35.01**	**58.45**	**72.57**

Table 6. Per-Class *F-measure* for the *Dordogne* study site. (Average over ten random splits)

4.5 Inspection of Confusion Matrices

Figure 3 depicts the confusion matrices of the Random Forest, Multilayer Perceptron, Convolutional Neural Network and S1$_{S2}$GKD, respectively. We show the confusion matrices corresponding to one of the ten random splits of the data associated to the study site.

We can observe that the results depicted by the heat maps are coherent with the average results previously reported. S1$_{S2}$GKD benefits from the use of optical information at training time to marginally boost its performances on agricultural as well as urban land cover classes. This can be observed on the first part of the diagonal of Figure 3d, that is characterized by a darker blue color w.r.t. the ones of the competing methods. Comparing S1$_{S2}$GKD with the other competitors, we underline that our framework is able to reduce some confusions between *Crops* and *Water* classes while the main confusions that happen between *Forest* and *Moor* as well as *Orchards* and *Vines* cannot be completely recovered. Our framework can regularize/guide the model learned from radar SITS via a kind of imitation strategy w.r.t. a multi-source (teacher) model, but the final classification model is still characterized by the advantages and the limitations of the data that fed it. If the information to distinguish between some land cover classes is not present in the



(a) RF

(b) MLP

(c) CNN

(d) S1$_{S2}$GKD

Figure 3. Confusion Matrices of the land cover classification produced by (a) RF, (b) MLP, (c) CNN and (d) S1$_{S2}$GKD on the *Dordogne* study site.

radar time series, the model will not be able to discriminate among them. On the other hand, the use of optical information at training time can regularize and reduce some kind of confusion (like the ones between *Urban* vs. *Orchards* and *Orchards* vs. *Vines* that notably affect all the evaluated approaches except S1$_{S2}$GKD).

4.6 Discussion on Generalized Knowledge Distillation for Earth Observation data

In the context of modern EO analysis, due to the increasing number of spatial programmes and related orbital/aerial sensors, multi-source data are constantly acquired. Unfortunately, the acquisition of all the different programmes are not timely synchronised and this results in incomplete spatial and temporal coverage of a particular (large) area. In this context of multi-source data abundance, the Generalized Knowledge Distillation setting provides a methodological framework to effectively integrate additional information that cannot be exploited by standard machine learning approaches due to their structural intrinsic constraints. In our work, we focus on a limited but real-world scenario in which radar (Sentinel-1) and optical (Sentinel-2) are exploited together to ameliorate a predictive model learnt only on radar information. Beyond such a case that could appear restricted, many other scenarios can arise. For instance, given a study area, we can easily access to an optical (Sentinel-2) SITS but, we can hardly obtain a Very High spatial Resolution (VHR) scene that covers the whole area. In this case GKD can be deployed to integrate the partial VHR information we dispose to support and ameliorate the analysis of optical SITS data. In another scenario, the additional, expensive and limited information can be represented by hyperspectral data we use in combination with VHR image to analyze and map urban or general land use land cover classes. Finally, in the case of airborne hyperspectral acquisition, we can acquire information via a double camera system. If one of the two camera has some failures, the information collected by the two cameras will not be fully exploited due to heterogeneity in data acquisition. Such illustrations are only some exemplars scenarios in which GKD can be deployed, but many other will be soon available thanks to the ongoing trend in Earth Observation Data acquisition. We are convinced that, in the current (as well as in the near future) Earth Observation context, more and more misaligned and incomplete information will be available when a study area is analyzed and the GKD framework constitutes a practical and well founded tool to encompass the limitations of standard machine learning approaches to integrate additional (incomplete) information in their learning process.

5. CONCLUSION

In this paper we have presented a machine learning framework, based on Generalized Knowledge Distillation, to provide land use land cover mapping from radar (Sentinel-1) SITS data when additional optical (Sentinel-2) SITS information is only available at training time.

The proposed framework leverages recent advances in the domain of computer vision and signal processing to guide the model learning. More in detail, the additional information is firstly employed to learn a (teacher) multi-source model and, successively, a mono-source (student) model is trained considering only radar SITS as input data and it is forced to imitate (or distill) as much as possible the behavior of the teacher model.

The experimental evaluation, carried out on a real-world study area located in southwest of France, underlines the advantage to integrate additional information during the training process and it emphasizes that such additional knowledge is particularly useful to ameliorate the model performance on urban as well as agricultural land cover classes, in our case. As future work we plan to leverage the Generalized Knowledge Distillation idea considering other LULC mapping tasks involving multi-source Earth Observation data such as: i) optical (Sentinel-2) SITS / VHR images or ii) Hypersepctral/VHR imagery.

6. ACKNOWLEDGEMENTS

This work has been supported by the Programme National de Télédétection Spatiale (PNTS, `http://programmes.insu.cnrs.fr/pnts/`), grant n^oPNTS-2020-13 and by the French National Research Agency under the Investments for the Future Program, referred as ANR-16-CONV-0004 (DigitAg).

REFERENCES

Chen, Y., Jin, X., Feng, J., Yan, S., 2017. Training group orthogonal neural networks with privileged information. *IJCAI*, 1532–1538.

Cho, J. H., Hariharan, B., 2019. On the efficacy of knowledge distillation. *ICCV*.

Gao, Q., Zribi, M., Escorihuela, M. J., Baghdadi, N., Quintana-Seguí, P., 2018. Irrigation Mapping Using Sentinel-1 Time Series at Field Scale. *Remote Sensing*, 10(9), 1495.

Gbodjo, Y. J. E., Ienco, D., Leroux, L., Interdonato, R., Gaetano, R., Ndao, B., Dupuy, S., 2019. Object-based multi-temporal and multi-source land cover mapping leveraging hierarchical class relationships. *CoRR*, abs/1911.08815.

Hinton, G. E., Vinyals, O., Dean, J., 2015. Distilling the Knowledge in a Neural Network. *CoRR*, abs/1503.02531.

Hoffman, J., Gupta, S., Darrell, T., 2016. Learning with side information through modality hallucination. *CVPR*.

Huang, J., Hartemink, A. E., Zhang, Y., 2019. Climate and Land-Use Change Effects on Soil Carbon Stocks over 150 Years in Wisconsin, USA. *Remote Sensing*, 11(12), 1504.

Ienco, D., Gaetano, R., Interdonato, R., Ose, K., Minh, D. H. T., 2019. Combining sentinel-1 and sentinel-2 time series via RNN for object-based land cover classification. *IEEE IGARSS*, 4881–4884.

Inglada, J., Vincent, A., Arias, M., Tardy, B., Morin, D., Rodes, I., 2017. Operational High Resolution Land Cover Map Production at the Country Scale Using Satellite Image Time Series. *Remote Sensing*, 9(1), 95.

Kampffmeyer, M., Salberg, A., Jenssen, R., 2018. Urban Land Cover Classification With Missing Data Modalities Using Deep Convolutional Neural Networks. *IEEE JSTARS*, 11(6), 1758–1768.

Kingma, D. P., Ba, J., 2014. Adam: A Method for Stochastic Optimization. *CoRR*, abs/1412.6980.

Kolecka, N., Ginzler, C., Pazur, R., Price, B., Verburg, P. H., 2018. Regional Scale Mapping of Grassland Mowing Frequency with Sentinel-2 Time Series. *Remote Sensing*, 10(8), 1221.

Kussul, N., Lavreniuk, M., Skakun, S., Shelestov, A., 2017. Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data. *IEEE Geosci. Remote Sensing Lett.*, 14(5), 778–782.

Lambert, J., Sener, O., Savarese, S., 2018. Deep learning under privileged information using heteroscedastic dropout. *CVPR*, 8886–8895.

Lathuilière, S., Mesejo, P., Alameda-Pineda, X., Horaud, R., 2018. A Comprehensive Analysis of Deep Regression. *CoRR*, abs/1803.08450.

Lopez-Paz, D., Bottou, L., Schölkopf, B., Vapnik, V., 2016. Unifying distillation and privileged information. *ICLR*.

Marais-Sicre, C., Fieuzal, R., Baup, F., 2020. Contribution of multispectral (optical and radar) satellite images to the classification of agricultural surfaces. *Int. J. Applied Earth Observation and Geoinformation*, 84.

Pelletier, C., Webb, G. I., Petitjean, F., 2019. Temporal Convolutional Neural Network for the Classification of Satellite Image Time Series. *Remote Sensing*, 11(5), 523.

Rouse, J. W., Haas, R. H., Schell, J. A., Deering, D., 1973. Monitoring the vernal advancement and retrogradation (green wave effect) of natural vegetation. *Progress Report RSC 1978-1*, 112.

Schmitt, M., Zhu, X. X., 2016. Data Fusion and Remote Sensing: An ever-growing relationship. *IEEE Geoscience and Remote Sensing Magazine*, 4(4), 6–23.

Vapnik, V., Izmailov, R., 2015. Learning using privileged information: similarity control and knowledge transfer. *J. Mach. Learn. Res.*, 16, 2023–2049.

Yao, Y., Zhang, J., Shen, F., Yang, W., Hua, X., Tang, Z., 2018. Extracting privileged information from untagged corpora for classifier learning. *IJCAI*, 1085–1091.

Zhang, X., Liu, L., Chen, X., Xie, S., Lei, L., 2019. A Novel Multitemporal Cloud and Cloud Shadow Detection Method Using the Integrated Cloud Z-Scores Model. *IEEE JSTARS*, 12(1), 123–134.