Temporal Domain Adaptation for Satellite Image Time Series Land Cover Mapping with Adversarial Learning and spatially-aware Self-Training

Emmanuel Capliez, Dino Ienco, Raffaele Gaetano, Nicolas Baghdadi and Adrien Hadj Salah

***Abstract*—Nowadays, Satellite Image Time Series (SITS) are commonly employed to derive land cover maps (LCM) to support decision makers in a variety of land management applications. In the most general workflow, the production of LCM strongly relies on available ground truth data to train supervised machine learning models. Unfortunately, this data is not always available due to time-consuming and costly field campaigns.**

In this scenario, the possibility to transfer a model learnt on a particular year (*source domain*) to a successive period of time (*target domain*), over the same study area, can save time and money. Such a kind of model transfer is challenging due to different acquisition conditions affecting each time period thus, resulting in possible distribution shifts between *source* and *target* domains. In the general field of machine learning, Unsupervised Domain Adaptation (UDA) approaches are well suited to cope with the learning of models under distribution shifts between *source* and *target* domains. While widely explored in the general computer vision field, they are still under investigated for SITS-based land cover mapping, especially for the temporal transfer scenario. With the aim to cope with this scenario in the context of SITS-based land cover mapping, here we propose *SpADANN* (Spatially Aligned Domain-Adversarial Neural Network), a framework that combines both adversarial learning and self-training to transfer a classification model from a time period (year) to a successive one on a specific study area. Experimental assessment on a study area located in Burkina Faso characterized by challenging operational constraints demonstrates the significance of our proposal. The obtained results have shown that our proposal outperforms all the UDA competing methods by 7 to 12 points of F1-score across three different transfer tasks.

***Index Terms*—Land Cover Mapping, Satellite Image Time Series, Temporal Domain Adaptation, Deep Learning**

I. Introduction

Today, satellite imagery represents a fundamental source of information to monitor the dynamic of the Earth surface providing valuable knowledge to support decision makers in several application domains [1]. Recent spatial programmes (i.e. the European Union's Copernicus programme and its Sentinel missions) provide open access satellite imagery with both high spatial resolution as well as high revisit frequency. They capture Satellite Image Time Series (SITS) data that can be leveraged to monitor phenomena in a variety of different domains, such as ecology [2], agriculture [3], forestry [4] and natural habitat monitoring [5].

SITS data, conversely to mono-date imagery, contains signal information about the evolution of the Earth surface allowing, for instance, to distinguish vegetated land covers that evolve differently over a yearly cycle of seasons (e.g. in agriculture, different cropping practices exhibit a different dynamic in their radiometric signal over a growing season).

Among the possible use of SITS data, the production of Land Cover Maps (LCM) over a specific region [6] is of paramount importance. The increasing availability of SITS data along with advances in machine learning [7], more precisely deep learning [8], has led to land cover mapping systems that take largely profit of the information carried out by time series of remote sensing imagery.

Nonetheless, supervised machine learning methods require large amount of reference (or Ground Truth (GT)) data to be trained, hence posing serious challenges to their use in situations characterized by a reduced amount of, or unavailable, reference data. For instance, when LCM have to be updated from previous years, costs or restrictions related to new field campaigns can prevent the possibility to collect new reference data thus, hindering to learn an up-to-date classification model [9].

An ideal solution would be to reuse already available data on a study site, for instance collected in previous field campaigns or shared in the last years by some public/government agency, to save time and money for the production or update of land cover maps. This option can, on one hand, take advantages of the efforts previously done and, on the other hand, limit the needs of fresh reference data on a study area whose accessibility may be reduced or compromised.

However, in the specific, yet common case in which significant land cover changes occur over a certain reference period, i.e. for agricultural landscapes with year-to-year crop type changes among fields, simply training a new model using up-to-date images and legacy reference data is not a solution, and the use of transfer learning strategies becomes urgent.

We here start from the observation that directly transfer a model trained on a particular year (the source domain) to a successive period of time (the target domain) can be challenging since the two time periods can be affected by different environmental, weather or climate conditions [10], [11]. This results in differences or shifts in the distributions

E. Capliez is with INRAE, UMR TETIS, University of Montpellier, Montpellier, France (email: emmanuel.capliez@inrae.fr) and also with Airbus Defence and Space, Toulouse, France.

D. Ienco is with INRAE, UMR TETIS, University of Montpellier, Montpellier, France (email: dino.ienco@inrae.fr).

R. Gaetano is with CIRAD, UMR TETIS, University of Montpellier, Montpellier, France (email: raffaele.gaetano@cirad.fr).

N. Baghdadi is with INRAE, UMR TETIS, University of Montpellier, Montpellier, France (email: nicolas.baghdadi@inrae.fr).

A. Hadj Salah is with Airbus Defence and Space, Toulouse, France (email: adrien.hadjsalah@airbus.com).

of the acquired yearly remote sensing data.

Addressing the distribution shift problem to adapt a model trained on a source domain to an unlabelled target domain is known as Unsupervised Domain Adaptation (UDA) [12] in the general field of machine learning. The UDA approach has the objective to provide methods and strategies to cope with distribution shifts between the data on which the model is trained (*source domain*) and the data on which the model is deployed (*target domain*) [13].

Here, we consider the temporal UDA (or tUDA) problem where data is satellite image time series and the task is to provide yearly land cover mapping. The goal is to train a classification model capable to provide a reliable LCM using an image time series on a given year for which no specific ground truth is provided (*target domain*) as well as both SITS and sparsely annotated GT data from a previous year (*source domain*).

When dealing with real-world land cover mapping the collected GT is generally sparse due to the operational constraints related to time and efforts associated to field campaigns [14], [15]. 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. Matter of fact, the common operational GT data collection protocol prevents the use of standard semantic segmentation approaches due to the fact that the latter requires densely annotated GT data as underlined in [16], [17] forcing the conceived land cover mapping solution to work at the pixel [18] or at the parcel [19] granularity.

To cope with the tUDA challenging setting affecting SITS-based land cover mapping, in this paper, we propose *SpADANN* (Spatially Aligned Domain-Adversarial Neural Network), a framework that combines both adversarial learning and self-training for temporal unsupervised domain adaptation for SITS-based land cover mapping under sparsely annotated ground truth data. More precisely, *SpADANN* leverages adversarial learning with the aim to extract domain-invariant features and it progressively transfers the underlying classification model from source to target domain via self-training. With the aim to leverage the peculiarity of remote sensing data, the self-training process generates pseudo-labels on the target domain identifying stable spatial areas between the two considered years (domains) and use such spatial areas (anchor points) to further alleviate the distribution shift between domains. In addition, with the goal to explicitly cope with the temporal dimension charaterizing SITS data we leverage one dimensional convolutional neural networks as backbone of our framework. Extensive experimental evaluations are carried out to assess the behavior of *SpADANN* considering state-of-the-art UDA approaches and assessing both quantitative and qualitative aspects on a rural study site located in Burkina Faso, referred as *Koumbia* site and characterized by a mostly agricultural land cover nomenclature (crop types as well as natural and built-up classes). The associated GT data is highly sparse due to operational constraints related to labor-intensive and costly field campaigns spanning the year 2018, 2020 and 2021.

The rest of this manuscript is organized as follows. Section II presents the related literature in SITS-based land-cover mapping, self-training and domain adaptation. Section III describes the temporal UDA problem setting and introduces the proposed *SpADANN* framework to cope with tUDA for SITS-based land cover mapping. Section IV presents the study site and the associated data while the experimental evaluation is reported in Section V. Section VI discusses the obtained results and short-term follow-ups. Section VII draws final conclusions on the work.

II. Related works

A. SITS based land-cover mapping under sparsely annotated GT data

Land cover mapping from satellite image time series data is of paramount importance to monitor and characterize spatio-temporal phenomena occurring on the Earth surface, i.e. quantify natural resources [20], estimate agricultural surfaces [21] or assess human settlement evolution [7]. In [7] the authors propose an operational framework to perform large scale land cover mapping at national scale from time series data. The classification is achieved via the Random Forest (RF) classifier that, still today, represents a well-established approach for land cover mapping based from SITS data. [22], [23] and [24] deal with land use and land cover (LULC) mapping via recurrent neural network approaches. In both [23] and [22], SITS data is managed via Long Short Term Memory (LSTM) while [24] deals with land use land cover mapping still considering recurrent neural network strategies but, this time, the performances of the Gated Recurrent Unit (GRU) were inspected to perform classification. [3] proposes the use of one dimensional (temporal) Convolutional Neural Networks for SITS based land-cover mapping, referred as TempCNN. In this model, the convolutional operator is performed on the temporal dimension of the SITS data with the purpose to manage and model short and long time correlations. The conducted study highlights the appropriateness of such approach w.r.t. previous proposed strategies in the context of general LULC mapping from satellite image time series data. Furthermore, [25] provides a comparison of both recurrent and convolutional neural network for the classification of summer crops highlighting that the latter approach achieves the best performances in their study case. More recently, [26] proposes the pixel set encoder temporal attention encoder (PSE-TAE), a transformer based strategy equipped with a pixel-set encoder and a self-attention module for agricultural parcels classification.

Despite the recent progress in the field of SITS based land-cover mapping, the proposed algorithms still struggle to manage data coming from different temporal periods thus limiting their applicability in a temporal transfer scenario.

B. Self-training methods

Self-training [27] can be seen as a particular case of semi-supervised learning [28] where a machine learning model is trained using a reference data set composed of a small set of labelled samples and a big amount of unlabelled ones. More precisely, in the self-training setting, a model is trained iteratively by assigning pseudo-labels to the set of

unlabeled training samples and, successively, enriching the current labelled training set with pseudo-labeled samples on which the model exhibited a high confidence. Co-training [29] is one of the earliest and widely popular techniques that have been proposed in the context of self-training learning. In co-training, examples are defined by two views that are decorrelated to each other. The goal of learning is to train a classifier on each view by first initializing it with the available labeled training data. Then, one of the classifiers assigns pseudo-labels to unlabeled data, which the other one will use to learn. Following training, the classifiers switch roles, with the learned classifier assigning pseudo-labels to unlabeled examples, which will then be used to train the first classifier. This procedure continues until there are no more unlabeled instances to be pseudo-labeled. Tri-training [30] is a direct extension of the co-training approach in which three classifiers from the original labeled set are generated. These classifiers are then refined using unlabeled examples in a tri-training process. In detail, in each round of tri-training, an unlabeled example is labeled for a classifier if the other two classifiers agree on the pseudo-labels. Another popular self-training techniques is Mean-Teacher [31]. This method employs two Neural Networks (NNs) as supervised classifiers, one of the models is named teacher, while the other is called student. These two models are structurally identical, and their weights are related in that the teacher's weights are an exponential moving average of the student' weights. In this scenario, the student model is the only one that is trained over the labeled training set and, a consistency loss is computed between the teacher's probability distribution prediction and the student's one.

In [32] the authors proposed a clustering based approach to perform land use/land cover mapping from multi-spectral satellite image data in an unsupervised fashion. More in detail, firstly pixels are clustered together, then a clustering label procedure is employed to obtain initial labelled samples and finally the machine learning classifier is iteratively trained via self-training. More recently, [33] proposes to cope with hyperspectral image classification under the lens of self-training. The authors exploit self-training to alleviate issues related to tedious and time-consuming process of data annotation. In addition to only use classifier confidence to select pseudo-labels, the proposed approach leverages spatial consistency (in terms of spatial neighborhood) to correct possible mistakes in the training enrichment step. [34] introduces a framework that combines self-training (referred as self-pace learning) and active learning in order to iteratively enrich an initially small labelled training set with informative samples for land use/land cover classification of SITS data via Support Vector Machines on the Google Earth Engine platform.

Self-training methodologies are receiving more and more attention due to their ability to train machine learning models in a data paucity scenario. While many frameworks have already been proposed for image or scene classification [35], only few research studies have leveraged self-training for time series analysis [36] or SITS-based land cover mapping [34]. Furthermore, all the research studies associated to time series analysis work in a classical context where no domain shift exists between training and target data.

C. Unsupervised domain adaptation

Unsupervised domain adaption [12] (UDA) methods belong to the family of transfer learning approaches [37] which has the main objective to transfer a model trained on a labelled source domain to an unlabelled target domain. Recent advances in unsupervised domain adaptation focus their efforts to extract domain-invariant features by either align domains through data transformation or perform adversarial training with the aim to reduce the distribution gap between the source and the target domain [12]. Regarding the first category of methods, the one that aligns domains through data transformation, [38] proposed a geodesic flow kernel-based strategy (GFK) to align source and target data distributions. The method allows to project both source and target data into a shared, low-dimensional, space in which the distribution shift between the two domains should be reduced. Since GFK only provides the low-dimensional data projection, a standard supervised model needs to be successively trained to perform the final classification on target data. Concerning domain-invariant approaches based on adversarial training, [39] defines the Adversarial Discriminative Domain Adaptation (ADDA) method. Inspired by the concept of generative adversarial network (GAN), this approach set up a two-player learning game where a discriminator network tries to distinguish between source and target sample representations derived by the generator while the generator tries to fool the discriminator network. Currently, adversarial learning is one of the main trends when it comes to unsupervised domain adaptation.

Still based on the adversarial training principle, [40] introduces the Domain-Adversarial Neural Network (DANN) model where a standard neural network model is augmented with a domain classifier that may distinguish between source and target samples in a multi-task learning setting. The domain classifier is associated with a gradient reversal layer (GRL) that enforces the features extracted by the encoder to be invariant w.r.t. the domains. The CDAN+E approaches [41] extends the DANN framework conditioning the discriminator on the prediction of the classification network for source and target data and it introduces an entropy regularization to prioritize the transfer of easy-to-transfer samples. This should, in theory, focus the source-target matching of instances belonging to the same class. The GRL principle introduced in the DANN framework is also the core of more recent unsupervised domain adaptation approaches as the Margin Disparity Discrepancy (MDD) [42] and the Adversarial-Learned Loss for Domain Adaptation ALDA [43] frameworks.

Concerning the remote sensing field, early research focused on proposing UDA strategies for high spatial resolution images [44], while only recently some strategies are emerging in the context of satellite image time series [11]. More generally, in this context distribution shifts between training (source) and test (target) data can be induced by different factors, and among others, differences in sensor acquisitions and environmental conditions are the most recurrent ones. However, such differences can be related to either the geographical shift from

a study site to another one [45] or the temporal delay among acquisitions, for data covering the same area in two different periods [46].

Regarding differences in sensor acquisitions, [47] proposes a cross-sensor UDA framework to cope with spatial and spectral distribution shifts between airbone and spaceborne very high spatial resolution (VHR) imagery with a specific focus on urban land cover map. The domain adaptation process leverages a self-training approach to transfer a classification model from the source to the target domain. Concerning differences in environmental condition, [45] proposes an adversarial based strategy to adapt a semantic segmentation model to be transferred from a spatial location to a different one. The method is conceived, also in this case, to cope with mono-date VHR imagery mainly considering urban land cover mapping.

Related to SITS-based land cover mapping, very few domain adaptation frameworks exist. Focusing on spatial transfer, [48] proposes a framework based on recurrent neural network and Maximum Mean Discrepancy (MMD) principle in order to embed both source and target SITS pixels in a common shared space. This is achieved by using an encoder per source and the MMD strategy to align the two domains. In [49] the combination of a transformer encoder-based classifier and the Domain Adversarial Neural Network (DANN) strategy with gradient reversal layer is evaluated to cope with spatial transfer learning in the context of land cover mapping from satellite image time series data. More recently, [11] proposes a framework to cope with agricultural parcel mapping under the objective to achieve spatial transferability. The approach combines together a module to align time series information, based on the estimation of time shift between SITS coming from the source and the target domain, and a self-training strategy in order to adapt the model to samples coming from the target domain. For the case of temporal transferability, [46] performs preliminaries investigation via optimal transport baselines for the case of temporal unsupervised domain adaptation from multiple source domain (multiple annual SITS) to a specific target domain (annual SITS). The obtained findings reveal that the use of optimal transport baselines results in a low level accuracy with respect to the use of a direct transfer of a supervised classifier from the source to the target domain thus, underlying that the problem of temporal transfer is quite complex and advanced methods are needed.

The extensive literature review we have performed clearly underlines that recent UDA approaches, especially the ones based on deep learning strategies, are still unexplored and under-exploited in the context of unsupervised domain adaptation for satellite image time series analysis. More in detail, a major lack is related to frameworks and methodologies addressing the important challenge related to temporal unsupervised domain adaptation, on which we set the focus of this work.

III. SpADANN

In this section, we introduce our proposed framework *SpADANN* (Spatially Aligned Domain-Adversarial Neural Networks with self-training) to deal with temporal UDA for SITS-based land cover mapping. We firstly provide the problem setting, then we give an overview of *SpADANN*. Successively, we supply the details of the different components on which *SpADANN* is built on.

A. Problem setting

In this work, we consider the problem of temporal UDA. We are giving a source domain $\mathcal{D}^s = \{(x_i^s, y_i^s)\}_{i=1}^{n^s}$ and a target domain $\mathcal{D}^t = \{x_i^t\}_{i=1}^{n^t}$ with n^s and n^t the number of samples for the source and target domain, respectively. We indicate with X^s, Y^s and X^t the set of source samples, source labels and target samples, respectively $\mathcal{D}^s = \{X^s, Y^s\}$ and $\mathcal{D}^t = \{X^t\}$. Each sample $x_i \in \mathbb{R}^{T \times B}$ is a satellite image time series pixel defined over T timestamps and characterized by B spectral bands. The land cover information (y_i^s) is only available for the source domain and $y_i^s \in \{1, \cdots, K\}$ can take one value between 1 and K, with K the number of land cover classes on which the multi-class classification problem is defined.

The set of SITS pixels belonging to the two domains cover exactly the same spatial area but at different periods of time (i.e. different years). This means that the same spatial location is covered by a SITS pixel coming from the source domain as well as one coming from the target domain. Thus, n^s is equal to n^t and $location(x_i^s)$ is equal to $location(x_i^t)$ where $location(\cdot)$ is a function providing the spatial location of a SITS pixel in terms of geographical coordinates. Due to differences in environmental, weather or climate acquisition conditions between the pixel SITS belonging to the source ($\mathcal{D}^s$) and the target ($\mathcal{D}^t$) domain, distribution shifts can affect the two sets of data thus, impacting the performances of standard inductive supervised classification approaches [12], [10], [11].

Here, the goal is to train a robust (in terms of data distribution shifts) SITS-based land cover mapping model that exploits both source ($\mathcal{D}^s$) and target ($\mathcal{D}^t$) domain information with the aim to predict, for a given pixel x_i^t belonging to the target domain ($\mathcal{D}^t$) the corresponding land cover class y_i^t. We remind that the set of land cover classes of the target domain spans exactly the same set of land cover classes of the source domain, as in a general closed-set scenario [50].

B. SpADANN overview

Hereafter, we provide a general overview of our framework, with the aim to supply a picture of how *SpADANN* behaves as well as describe the general principles behind it. Figure 1 visually sketches the *SpADANN* framework.

SpADANN combines both adversarial learning and self-training with the aim to: i) learn an invariant representation space (features) with respect to possible distribution shifts between source and target domains (in our case pixel SITS coming from two time periods - years - covering exactly the same geographical area) and ii) progressively transfer the underlying classification model from the source to the target domain via self-training/pseudo-labelling on the target domain.

While the adversarial learning strategy is based on the model proposed in [40], that we adapt for the special case

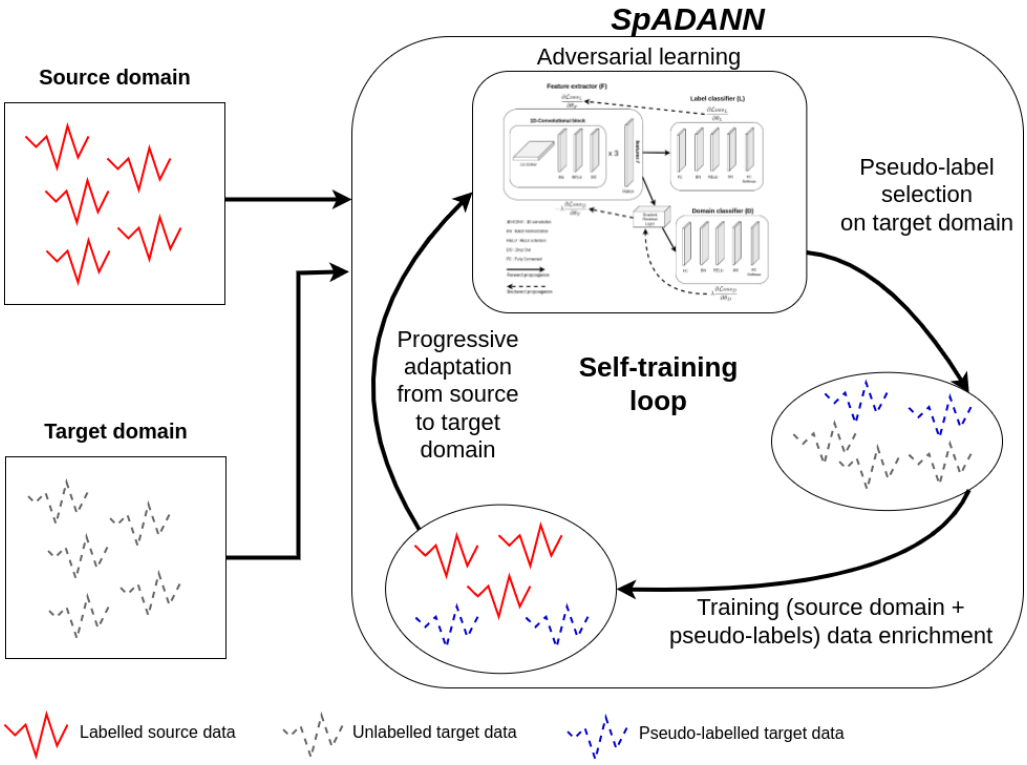


Fig. 1. A graphical overview of *SpADANN*. Our framework takes as input source ($\mathcal{D}^s$) SITS and ground truth data and target ($\mathcal{D}^s$) SITS data with the aim to train a classification model to predict land cover labels for target SITS data. During the iterative training process (self-training loop), it combines both adversarial learning and pseudo-labelling with the aim to progressively adapt the underlying classification model from the source to the target SITS data to cope with the temporal UDA problem.

of time series data, the pseudo-labelling procedure deeply exploits the features characterizing the tUDA problem. The pseudo-label selection is based on the fact that the source and the target domains are spatially aligned (i.e. they cover exactly the same geographical area). More precisely, given two spatially aligned pixels SITS ($location(x_i^s) = location(x_i^t)$) where the ($x_i^s$) comes from the source domain and the (x_i^t) comes from the target domain, if the land cover classifier provides the same decision for both pixels ($Cl(x_i^s|\Theta_F, \Theta_L) = Cl(x_i^t|\Theta_F, \Theta_L)$) with $Cl(\cdot)$ the prediction of the land cover classifier $\mathbf{L}$ and the predicted class for the source pixel SITS (x_i^s) is the correct one, then the target pixel SITS (x_i^t) is associated with the pseudo-label generated by the land cover classifier. Finally, as the iterative training procedure goes on, pseudo-label information gets more importance with the aim to progressively transfer the underlying classification model from the source to the target data.

C. Adversarial learning

With the aim to extract SITS pixel representations that are invariant to the particular domain they come from (source or target), we adapt the strategy proposed in [40], namely DANN (Domain-Adversarial Neural Network), as backbone block in the *SpADANN* framework. Figure 2 depicts the architecture of the proposed backbone network.

The network architecture has three components, an encoder network $\mathbf{F}$ relying on the Θ_F parameters, a land cover classifier network $\mathbf{L}$ with parameters Θ_L, and a domain classifier network $\mathbf{D}$ with parameters Θ_D. Due to the fact that we are dealing with satellite image time series pixels, we adopt an encoder model especially tailored for such kind of data, namely the TempCNN model [3], due to its confirmed ability to cope with the task of SITS-based land cover mapping in a standard in-domain setting through 1-D convolution on the time dimension.

The *SpADANN* backbone is a multi-output network that has the objective to generate a new data representation via the encoder $\mathbf{F}$ ensuring high land cover classification accuracy and, simultaneously, making difficult to distinguish between the domain each SITS pixel comes from.

The DANN loss function is defined as follows:

$$L_{DANN}(X^s, Y^s, X^t|\Theta_F, \Theta_L, \Theta_D) = L_c(X^s, Y^s|\Theta_F, \Theta_L) - \lambda L_{Adv}(X^s, X^t|\Theta_F, \Theta_D) \quad (1)$$

where $L_c(X^s, Y^s|\Theta_F, \Theta_L)$ is the loss associated to the land cover classification problem modeled with standard Categorical Cross-Entropy function [40] while $L_{Adv}(X^s, X^t|\Theta_F, \Theta_D)$ is the loss related to the domain classifier modeling a binary classification problem in which class label represents the possibility to belong exclusively to the source or the target domain. Also in this case, the Categorical Cross-Entropy function is employed. Finally, the hyper-parameter λ controls the influence of the domain classifier loss on the learnt features.

In order to leverage standard stochastic gradient descent to optimize the L_{DANN} loss function, the $L_c(\cdot|\cdot)$ loss is opti-

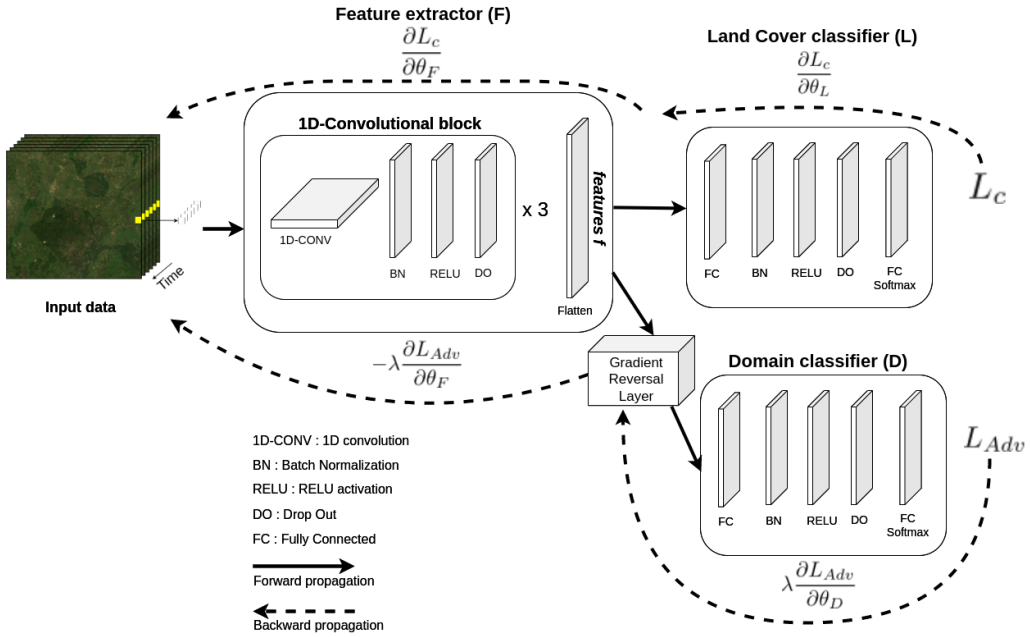


Fig. 2. The internal classification architecture of *SpADANN*. It is based on the DANN adversarial learning strategy [40] coupled with the TempCNN [3] encoder to customize the architecture for the special case of SITS data. The model has three components, an encoder network **F** (the TempcNN model), a land cover classifier network **L** and a domain classifier **D**. The multi-task network has the objective to generate a new data representation via the encoder **F** that has high land cover classification accuracy (maximizing the **L** performances) and, simultaneously, make difficult to distinguish between the domains the SITS pixels come from (confusing the **D** component).

mized as commonly done for general neural network models while for the $L_{Adv}(\cdot|\cdot)$ loss we employ the Gradient Reversal Layer (GRL) trick [40]. More in detail, the GRL acts as the identity transform during the forward propagation pass while it reverses the gradient (the gradient is multiplied by -1) during the backward propagation pass when the gradient is exploited for the update of the encoder **F** weights. In this way, the GRL trick allows to implement the adversarial training strategy with a standard backpropagation of the gradients without adding any extra parameters to the model. More precisely, the domain classifier parameters are updated in a standard way with the aim to support the model to distinguish between source and target samples while, the reversed gradient applied to the encoder network forces the model to generate domain-invariant features with the goal to fool the domain classifier [49].

Figure 2 visually highlights the difference between forward propagation (solid line) and backward propagation (dashed line) passes in the neural network backbone model of the *SpADANN* framework.

D. Spatial consistent pseudo-labelling

In order to further adapt the SITS classification model, presented in Section III-C, to effectively classify pixels coming from the target domain, we leverage a self-training strategy that allows to associate pseudo-labels to a subset of data coming from $\mathcal{D}^t$. This is done with the aim to inject, in the training process, pseudo supervision on the target domain permitting to the land cover classifier sub-network **L** to tackle the classification of SITS pixels coming from $\mathcal{D}^t$.

In a standard self-training pipeline [51], given a set of unlabelled samples, the model output distribution is employed to select a subset on which the model has high confidence. Successively, such pseudo-labelled samples are used to enrich the current training set as the training process proceeds. More precisely, this mechanism is implemented by defining a threshold on the model output *softmax* and, subsequently, choose all the samples on which the value of the most probable prediction is greater than the defined threshold. This widely adopted process suffers from the fact that a threshold needs to be defined and the way this hyper-parameter is set can drastically affects the performance of the underlying sampling process [52].

In our case, we leverage the specificity of the land cover mapping tUDA problem conceiving a process based on the spatial consistency between the two SITS pixels x_i^s and x_i^t sharing the same spatial location ($location(x_i^t) = location(x_i^s)$). Such a strategy provides a solution to the pseudo-labelling selection process that avoids the definition of any kind of threshold thus reducing possible hyper-parameter tuning associated to our framework. More precisely, the set of target pixels to which pseudo-labels will be associated are chosen based on two criteria that need to be met simultaneously. The first criteria is based on spatial consistency as described below: $Cl(x_i^s|\Theta_F, \Theta_L) = Cl(x_i^t|\Theta_F, \Theta_L)$ (given that $location(x_i^t) = location(x_i^s)$) and the second criteria requires that the land cover classifier **L** supplies the correct prediction for the source sample ($Cl(x_i^s|\Theta_F, \Theta_L) = y_i^s$). The idea behind this selection process is to choose target samples that remain stable, in terms of model output prediction, w.r.t. the corresponding source pixel in terms of spatial location, and simultaneously we enforce the fact that the model predicts the correct land cover class on the source sample x_i^s. In this way,

the procedure allows to select pseudo-labelled samples that act as anchor points between the source and the target domains exploiting the model output stability and, at the same time, leveraging target samples that are in principle characterized by a small distribution gap, thus more effective to support the classification model transfer from the source to the target domain.

More formally, we can define the loss associated to the pseudo-labelled samples as follow:

$$L_p(X^s, X^t, Y^s, \hat{Y}^t|\Theta_F, \Theta_L) = \sum_{x_i^t \in X^t} \mathbb{1}_{\{Cl(x_i^s)=Cl(x_i^t) \text{ and } Cl(x_i^s)=y_i^s)\}} H(\hat{y}_i^t, Cl_{prob}(x_i^t)) \quad (2)$$

where $\mathbb{1}_{cond}$ is an indicator function that returns 1 if the condition $cond$ is verified and 0 otherwise, $Cl_{prob}(\cdot)$ provides the model output distribution over the possible land cover set, $H(\cdot,\cdot)$ is the classical Categorical Cross-Entropy function, $\hat{Y}^t$ is the whole set of possible pseudo-label for the target domain and $\hat{y}_i^t$ is the pseudo-label land cover class with the highest model output probability w.r.t. $Cl_{prob}(x_i^t)$ for the pixel x_i^t coming from the target domain.

E. SpADANN training procedure

In this section we introduce the general training procedure we have used to optimize the parameters of the *SpADANN* framework. Algorithm 1 briefly summarizes the pseudo-code of the training procedure. The inputs of the procedure are constituted by the data coming from the source ($\mathcal{D}^s = (X^s, Y^s)$) and the target ($\mathcal{D}^t = (X^t)$) domains, the hyper-parameter β associated to the progressive transfer strategy and N_e, the number of epochs associated to the learning procedure.

Algorithm 1 *SpADANN* Training procedure

Require: X^s (the source SITS pixels), Y^s (the source labels), X^t (the target SITS pixels), β (the progressive transfer hyper-parameter), N_e (the number of epochs).
Ensure: Θ_F (param. of the encoder), Θ_L (param. of the land cover classifier).

1: e = 0
2: **while** e $< N_e$ **do**
3: $\quad \hat{Y}^t = Cl_{prob}(X^t)$
4: $\quad \alpha = \beta \times \frac{e}{N_e}$
5: $\quad L_{TOT}$ = (1-α) $\times$ $L_{DANN}(X^s, X^t, Y^s|\Theta_F, \Theta_L, \Theta_D)$ + $\alpha \times L_p(X^s, X^t, Y^s, \hat{Y}^t|\Theta_F, \Theta_L)$
6: $\quad \nabla_{\Theta_F, \Theta_L, \Theta_D} L_{TOT}$ with mini-batch SGD
7: $\quad$ e = e + 1
8: **end while**
9: **return** Θ_F, Θ_L

Firstly, the current classification model is applied to the target data X^t in order to obtain the set of possible pseudo-labels $\hat{Y}^t$ (line 3). Then, we compute the trade off value α as a linear function of the current epoch (e), the total number of epochs (N_e) and the input hyper-parameter β (line 4). The α value is subsequently employed to weight the contribution of the L_{DANN} and the L_p losses in a convex combination of the two terms with the aim to vary their importance during the learning procedure (line 5). More in detail, at the beginning of the procedure, the α value starts from zero and linearly increases with the objective to, progressively, give more importance to the L_p term.

This is done since at the early iterations of the procedure we want that the model exploits as much information as possible from the labelled source domain while learning SITS pixels' representations that are invariant w.r.t. the specific domain. The reason is that at the first iterations the trained model is not yet effective, so that the prediction on the target data could be highly biased. As the learning procedure goes on, the α value increases, hence decreasing the importance of the first term while increasing the weight of the second one $L_p(X^s, X^t, Y^s, \hat{Y}^t|\Theta_F, \Theta_L)$. This mechanism implements a kind of progressive transfer from the first to the second term during the learning procedure, allowing the underlying classification model to smoothly focus on the specificity of the target SITS pixels via the use of the pseudo-labels selected as described in Section III-D. The hyper-parameter β controls the range of the α trade-off value with the aim to avoid the latter to get extreme values that can completely move the learning process towards the target domain, resulting in a degeneration of the behaviour of *SpADANN*. For this reason, β is supposed to range between 0.5 and 1. After that the current loss L_{TOT} is computed, the network weights Θ_F, Θ_L and Θ_D are updated by mini-batch stochastic gradient descent (line 6). At the end of the training procedure (line 9), the network weights Θ_F and Θ_L associated to the encoder **F** and the land cover classifier **L** are returned as output of the training process associated to our framework. This set of parameters represents the classification model that will be finally employed to provide the land cover mapping predictions on the SITS pixels coming from the target domain.

IV. Data

The study site covers an area around the town of *Koumbia*, in the Province of Tuy, *Hauts-Bassins* region, in the south-west of Burkina Faso. This area has a surface of about 2 338 km^2, and is situated in the sub-humid sudanian zone. The surface is covered mainly by natural savannah (herbaceous and shrubby) and forests, interleaved with a large portion of land (around 35%) used for rainfed agricultural production (mostly smallholder farming). The main crops are cereals (maize, sorghum and millet) and cotton, followed by oleaginous and leguminous. Several temporary watercourses constitute the hydrographic network around the city of Koumbia.

Figure 3 presents the study site with the 2018 reference data (ground truth) superposed on a Sentinel-2 image of September 12, 2018. A more detailed view corresponding to the red box in the overview is also depicted on the bottom right of the figure. A specific analysis of the ground truth is provided in the Section IV-B.

A. Satellite Image Time Series

We collected satellite image time series of Sentinel-2 imagery spanning the years 2018, 2020 and 2021, amounting for

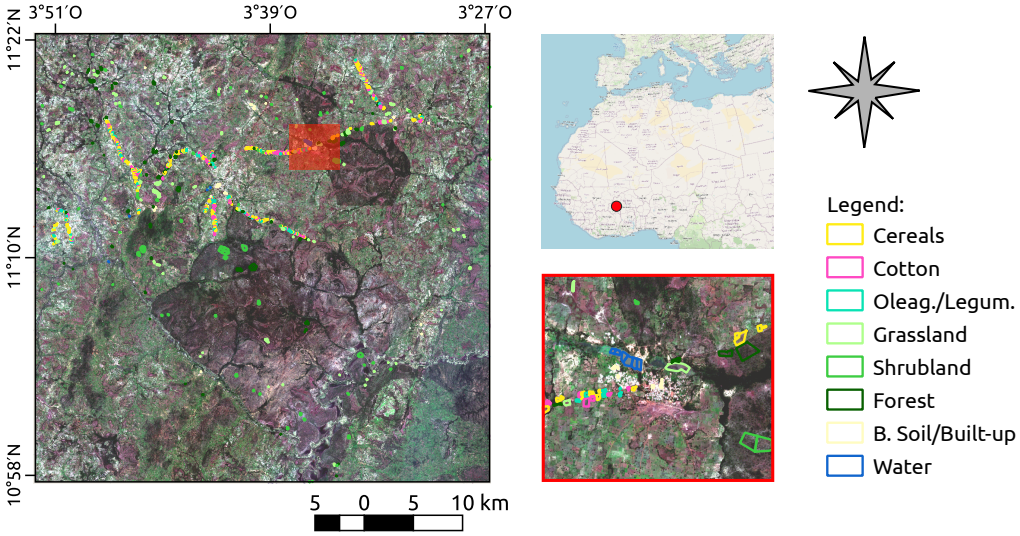


Fig. 3. View and location of Koumbia study site. The ground truth data coming from the 2018 year is superposed to a Sentinel-2 image covering the whole area. In the red box (bottom right) a more detailed view of the study site is depicted.

a total of respectively 35, 41 and 39 available scenes. Based on the available acquisitions, we conducted a visual analysis and we select 24 images for each year. Acquisitions are selected in order to account for an uniform temporal distribution among the three years. The main selection criteria used were (i) filtering out images that were visually impacted by cloud coverage and (ii) keep a sufficient amount of acquisitions over the rainy (cropping) season, occurring between May and October. Figure 4 depicts the acquisition dates of the three Sentinel-2 satellite image time series.

All images were provided by the THEIA Pole platform[1] at level-2A, which consist in atmospherically corrected surface reflectances (cf. MAJA processing chain [53]) and relative cloud/shadow masks. Only 10-m spatial resolution bands (Blue, Green, Red and Near infrared spectrum) were considered in this analysis in order to limit the computational burden related to the experimental assessment. A standard pre-processing was performed over each band to replace cloudy pixel values as detected by the available cloud masks based on the method proposed in [54]. In this pre-processing, the value of a cloudy pixel (w.r.t cloud/shadow mask and a threshold of 0) is linearly interpolated considering precedent and posterior acquisitions.

B. Ground truth data

Ground truth data for 2018, 2020 and 2021 has been derived from a large agricultural land cover data set available online [55], mainly consisting of field data collected by local experts on several sites all over the tropics. For the Koumbia site, these field surveys were conducted yearly around the growing peak of the cropping season from 2013 to 2021. GPS waypoints were gathered following an opportunistic sampling approach along the roads or tracks according to their accessibility, while ensuring the best representativity of the existing cropping practices in place. Records were also provided on different types of non-crop classes (e.g. natural vegetation, settlement areas, water bodies) to allow differentiating crop and non-crop classes. Moreover, some additional non-crop reference polygons are also provided, obtained by photo-interpretation of very high resolution (SPOT 6/7 and PLEIADES) optical satellite images.

Our final ground truth has been assembled in a Geographic Information System (GIS) vector file, containing a collection of polygons, each attributed with a land cover category based on information reported in the original database. Statistics about the yearly reference data sets used here are reported in Table I. In order to ensure consistency with the proposed method, we kept the exact same surface for the three reference years by performing a year by year intersection of the polygons of the original database.

This also allows measuring the changes occurring in the ground truth from one year to another, which are obviously more important on crop classes due to the presence of cropping cycles in this type of agricultural system. It is important to note that for year 2018 the surface of $cotton$ crop is about two times that of $oleaginous/leguminous$ when this ratio is balanced for years 2020 and 2021.

Figure 5 quantifies these changes in terms of land cover classes between each couple of reference years. They indeed highlight the presence of $cereals$, $cotton$ and $oleaginous/leguminous$ crops on the same agricultural parcels over the years. Conversely, $baresoil/builtup$ and $water$ classes remain unchanged, and few changes occur on non-crop classes, mainly due to occasional shifts in the density of natural vegetation or conversion to active cropland (e.g., 16% of $grassland$ in 2018 became $shrubland$ - 10% - or was converted to cereal crops - 6%).

Figure 6 shows the NDVI profiles over 2018, 2020 and 2021 for the three agricultural classes: $cereals$, $cotton$ and $oleaginous/leguminous$. We can observe that profiles over the years are similar. In the case of $cereals$ and $cotton$ classes, 2018 profile presents its value peak about ten days earlier than

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

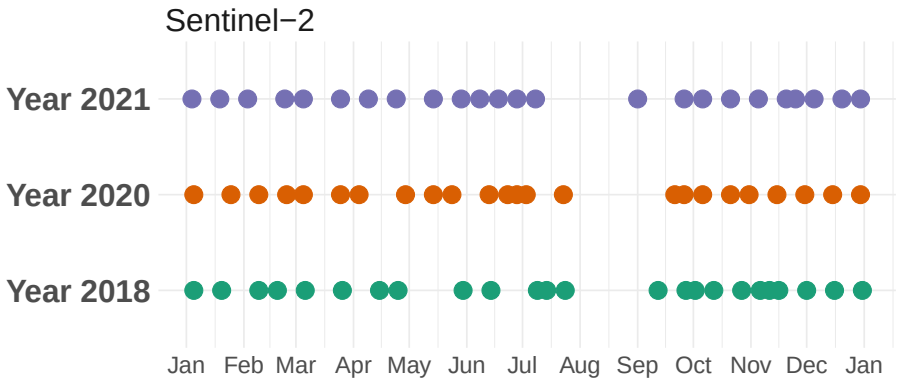
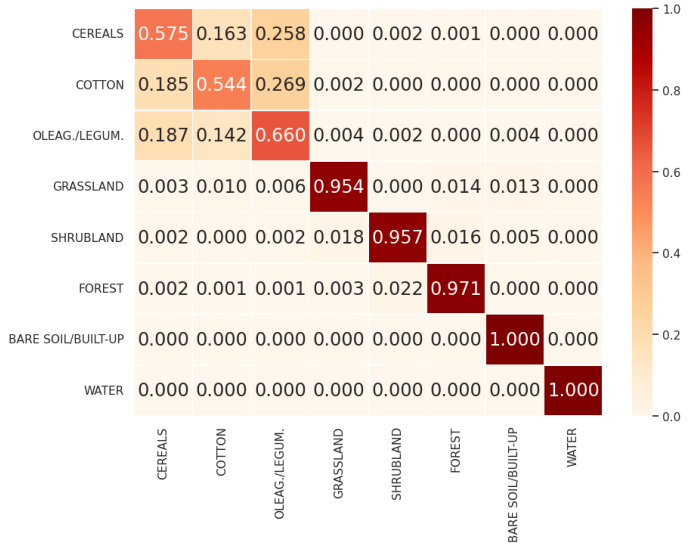
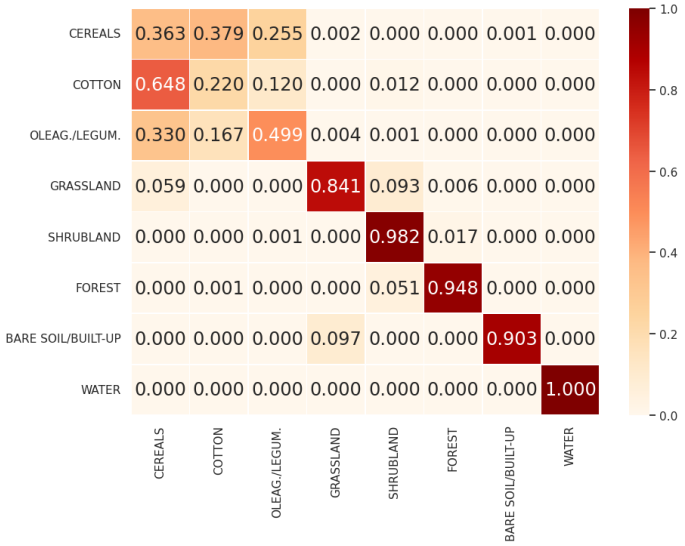


Fig. 4. Acquisition dates of Sentinel-2 Satellite Image Time Series.



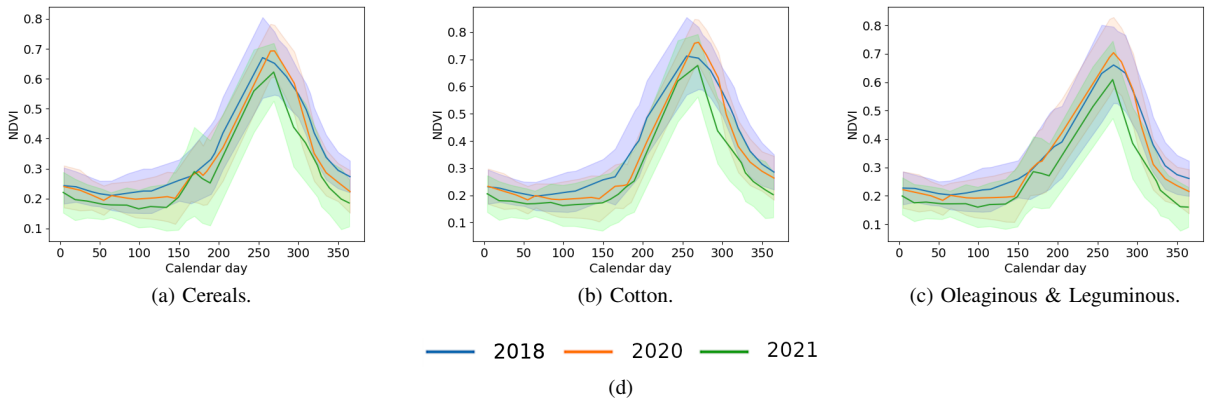
(a) From 2018 (rows) to 2020 (columns)

(b) From 2018 (rows) to 2021 (columns)



(c) From 2020 (rows) to 2021 (columns)

Fig. 5. Ground truth class transition between each pair of considered years.



(a) Cereals. (b) Cotton. (c) Oleaginous & Leguminous.

(d)

Fig. 6. Normalized Difference Vegetation Index (NDVI) profiles for a representative parcel of a) Cereal b) Cotton and c) Oleaginous & Leguminous land cover classes over the years 2018, 2020 and 2021.

TABLE I
GROUND TRUTH STATISTICS FOR YEAR 2018, 2020 AND 2021.

Class Name	Class ID.	# Polygons	# Pixels	% Pixels
CEREALS	1	330	13056	16.3
COTTON	2	153	7672	9.6
OLEAGINOUS/LEGUMINOUS	3	161	3595	4.5
GRASSLAND	4	123	13108	16.4
SHRUBLAND	5	87	23121	28.9
FOREST	6	88	17369	21.7
BARE SOIL/BUILT-UP	7	46	835	1.0
WATER	8	10	1205	1.5
	Total	998	79961	

Class Name	Class ID.	# Polygons	# Pixels	% Pixels
CEREALS	1	230	9731	12.2
COTTON	2	139	6971	8.7
OLEAGINOUS/LEGUMINOUS	3	281	7950	9.9
GRASSLAND	4	122	12998	16.3
SHRUBLAND	5	83	22546	28.2
FOREST	6	82	17435	21.8
BARE SOIL/BUILT-UP	7	51	1125	1.4
WATER	8	10	1205	1.5
	Total	998	79961	

Class Name	Class ID.	# Polygons	# Pixels	% Pixels
CEREALS	1	268	11435	14.3
COTTON	2	121	6575	8.2
OLEAGINOUS/LEGUMINOUS	3	263	7316	9.1
GRASSLAND	4	113	11100	13.9
SHRUBLAND	5	90	24324	30.4
FOREST	6	82	16984	21.2
BARE SOIL/BUILT-UP	7	51	1022	1.3
WATER	8	10	1205	1.5
	Total	998	79961	

for 2020 and 2021.

Finally, Table II reports the average (and standard deviation) percentage associated to how many times a labelled pixel is covered by clouds in the whole time series as well as considering the portion of the time series covering the growing and harvesting stages (between days 150 and 300). This latter period is of particular interest in order to distinguish between the crops classes. Inspecting Table II, we can see that the 2018 SITS data exhibits the highest percentage of cloudiness when statistics are computed considering the whole time series of 24 dates. Although, when only the period covering both growing and harvesting stages is considered, we can clearly observe that both 2018 and 2021 SITS data are more affected by cloud coverage than the SITS data from 2020. The cloudiness diversity is probably due to the differences in the environmental and climatic conditions that have affected the study areas in three considered periods. More precisely, the three considered years are affected by non homogeneous weather conditions that result in a heterogeneous level of non-cloudiness per time series.

These preliminaries analysis clearly indicate the presence of inter-annual differences in environmental, weather or climate conditions that can challenge the "naive" transfer of supervised machine learning models from one year to another one.

TABLE II
AVERAGE PERCENTAGE OF LABELLED PIXELS ASSOCIATED FOR YEARS 2018, 2020 AND 2021 COVERED BY CLOUDS. THE AVERAGE AND STANDARD DEVIATION ARE COMPUTED CONSIDERING BOTH THE WHOLE TIME SERIES (*Full year*) AS WELL AS THE PORTION OF THE TIME SERIES COVERING THE GROWING AND HARVESTING STAGES (*Days 150 to 300*).

Year	**Full year**	**Days 150 to 300**
2018	24.2 $\pm$ 32.9	42.8 $\pm$ 31.7
2020	15.5 $\pm$ 23.0	23.1 $\pm$ 20.0
2021	15.7 $\pm$ 28.3	37.5 $\pm$ 37.9

V. EXPERIMENTS

In this section we describe and discuss the experimental results obtained on the study site introduced in Section IV. We carried out several experiments with the aim to provide an extensive analysis of the performance of *SpADANN*. We investigate different aspects: (i) we perform an in-depth analysis of the performance of *SpADANN* with respect to competing methods and (ii) we provide a qualitative evaluation through the visualisation of the internal representation learnt by our framework and the exploration of the land cover maps.

A. Competing methods

With the aim to compare the performance of *SpADANN* to state-of-the-art UDA strategies we consider the following competitors:

- The Geodesic Flow Kernel (GFK) approach introduced in [38]. This approach leverages a kernel-based method that projects both source and target information in a low-dimensional manifold. Since GFK only provides the low-dimensional data projection, a standard supervised model needs to be successively trained to perform the final classification on target data. To this end, we couple the GFK with a Multi-Layer Perceptron as well as a Random Forest classifier. We indicate the former approach with GFK-MLP and the latter with GFK-RF;
- The Adversarial Discriminative Domain Adaptation (ADDA) method proposed in [39]. This approach employs adversarial learning with a two players game (discriminator and generator) in order to learn invariant representations w.r.t. the domain. Due to the fact that we are dealing with multi-variate time series analysis, we use as backbone the TempCNN network [3] that was especially designed to perform land cover mapping from SITS data;
- The Domain Adversarial Neural Network (DANN) method originally introduced in [40]. This is a standard UDA approach that exploits gradient reversal layer in order to obtain data representations that are invariant to the particular domain they come from. Also in this case, we use as backbone model the TempCNN network. This competitor can indeed be considered as an ablation of our proposed framework.
- The Conditional Adversarial Domain Adaptation with Entropy Conditioning (CDAN+E) approach proposed in [41] upgrades DANN by conditioning the domain discriminator on the classification output and minimizing an entropy loss on target data. We use as backbone the TempCNN network.
- The Margin Disparity Discrepancy (MDD) method introduced in [42]. This theory-inspired technique is designed to measure the distribution discrepancy in domain adaptation and it is built on top of the DANN approach. We use as backbone the TempCNN network.
- The FixMatch method proposed in [35]. This competitor is a state-of-the-art semi-supervised learning approach that exploits consistency regularization between a weak and a strong augmentation of the unlabelled data. We include this competitor in order to further highlight the need for domain adaptation in the temporal transfer task. To adapt FixMatch to time series data, we follow what proposed in [11] where identity function (resp. random time steps selection) corresponds to weak (resp. strong) data augmentation. We use as backbone the TempCNN network.
- The adversarial-learned loss for domain adaptation (ALDA) method presented in [43]. ALDA combines self-training and domain-adversarial learning to reduce the gap and align the feature distributions by means of a noise-correcting domain discriminator. We use as backbone the TempCNN network.

Moreover, we also consider three baseline strategies in which: i) a supervised classification model is trained with only source data and directly deployed on target data, referred as "only $\mathcal{D}_s$"; ii) a supervised classification model is trained on labelled target data and deployed on the rest of the target examples referred as "only $\mathcal{D}_t$" and iii) a supervised classification model is trained on the union of the source data and a portion of the target data and deployed on the rest of the target examples referred as "$\mathcal{D}_s + \mathcal{D}_t$". The first constitutes a straightforward baseline that does not take into account the necessity to deal with temporal distribution shifts. The second represents the performances we can (theoretically) achieve if we have knowledge about the labels associated to the target domain $\mathcal{D}_t$. The third provides a baseline that directly combines all the source domain data with some labelled samples from the target domain in order to assess the possibility to combine data from different domains.

For all the baseline strategies we consider, as supervised classification methods, both the Random Forest (RF) and the TempCNN [3] models. These two models are chosen due to the fact that they are standard and widely-adopted methodologies for land cover mapping from satellite image time series data. More in detail, the former has an established popularity in the remote sensing community due to the accuracy of its classifications [56] while the latter approach is representative of the recent deep learning methods that explicitly manage the temporal dimension that heavily characterizes SITS data [57].

B. Experimental settings

Evaluation of baseline strategies: Concerning the first baseline strategy (only $\mathcal{D}_s$), the supervised classification model is trained over all the source data and then deployed on the target data. Regarding the second baseline strategy (only $\mathcal{D}_t$), solely the target data are exploited. More in detail, target data is split into three parts: training, validation and test sets following a proportion of 70%, 10% and 20% of the original target data set, respectively. Furthermore, with the aim to avoid possible spatial bias in the evaluation procedure [58], we impose that all the pixels belonging to the same object will be exclusively associated to one of the data partition (training, validation or test). The splitting procedure is repeated ten times and the average results are reported. To what concern the third baseline strategy ($\mathcal{D}_s + \mathcal{D}_t$), the complete set of data from the source domain is combined with 80% of the target data. The amount of target samples corresponds to the union of training and validation set for the baseline strategy (only $\mathcal{D}_t$). Successively, the learnt classifier is deployed on the rest of the target data. The procedure is repeated ten times (fixing the source data and varying the selected labelled target data) and the average results are reported.

Evaluation of UDA competing methods: All the UDA models are trained exploiting the whole set of source and target samples with the sole access to label information coming from the source domain.

Concerning the evaluation tasks, according to the data presented in Section IV, we set up three temporal transfer tasks ($\mathcal{D}_s \rightarrow \mathcal{D}_t$) where the right arrow indicates the transfer direction from the source ($\mathcal{D}_s$) to the target ($\mathcal{D}_t$) domain: (2018 $\rightarrow$ 2020), (2018 $\rightarrow$ 2021) and (2020 $\rightarrow$ 2021).

To evaluate the different methodologies, once the models are trained, we consider two different scenarios referring to three different tasks. Regarding the evaluation scenarios, we distinguish between the follows:

1 We use the same test set as the one employed for the second baseline strategies (only $\mathcal{D}_t$). Following this evaluation, we compare all the approaches to each other. We name such context *Subset* $\mathcal{D}_t$;
2 We use the whole target data $\mathcal{D}_t$ as test set (this is possible for all the baseline and UDA methods except for the second baseline strategy). We name such a context *Full* $\mathcal{D}_t$.

The values of the three satellite image time series benchmarks (2018, 2020 and 2021) were normalized per band in the interval $[0, 1]$, with Min-Max method. The assessment of the model performances was done considering the following metrics: *accuracy* (global precision), *weighted F1-score* and *Cohen's Kappa* (level of agreement between two raters relative to chance).

Implementation details: For the neural network approaches, the training stage has been conducted for 300 epochs, with a learning rate of 10^{-4} and a batch size of 32. Batch normalization layers has been inserted after each fully connected or convolutional layer (except for the classification layer). The drop out value is set to 50%. For *SpADANN* we set the value of β equal to 0.8 and the value of the hyper-parameter λ as suggested by [40]. In addition, similar to [11], *SpADANN* implements domain-specific batch normalization [59] by processing the source and target mini-batches separately. This ensures that batch normalization [60] statistics are computed separately for each domain.

Considering Random Forest classifiers, we optimize the model via the tuning of one parameter: the number of trees in the forest. We vary this parameter in the range {100, 200, 300, 400, 500}. The Multi-Layer Perceptron classifier coupled with the GFK approach has two fully connected layers both with ReLU activation function, each one with 512 neurons and followed by batch normalization and drop out layers. A final output layer, with softmax activation function, is employed to perform classification.

Regarding ALDA, and according to recent literature on pseudo-labeling in the context of SITS based land cover mapping [11], we set the threshold of pseudo-labels to 0.9. The same value of pseudo-labels threshold is also used for FixMatch. For this latter method, we set the relative weight of the unlabelled loss (λ_u) to 2, the strong data augmentation is implemented by means of the Python *TSAUG* library[2] via the "dropout" function with parameters $p = 0.05$ (probability to drop a timestamps).

Experiments are carried out on a workstation with a dual Intel (R) Xeon (R) CPU E5-2667v4 (@3.20GHz) with 256 GB of RAM and four TITAN X (Pascal) GPU. All the deep learning methods are implemented using the Python *TensorFlow* library except ALDA that was implemented in Pytorch based on the original open source implementation[3]. The MDD and CDAN+E competitors are implemented via the Python *ADAPT* library [61]. All the models run on a single GPU. The Random Forest is implemented using the Python *Scikit-learn* library. The code implementation of *SpADANN* is available at this link[4].

C. Quantitative analysis

The results, in terms of F1-score, Accuracy and Kappa, are reported in Table III, IV and V for the (2018 $\rightarrow$ 2020), (2018 $\rightarrow$ 2021) and (2020 $\rightarrow$ 2021) transfer tasks, respectively.

Firstly, we can notice that, whatever the transfer task, a direct application of a model learnt on the source domain to data coming from the target domain results in poor performances as expected. This is evident when we compare, for each table, the metric values achieved by the *Only* $\mathcal{D}_s$ and *Only* $\mathcal{D}_t$ strategies. We can also observe that supervised learning models trained under *Only* $\mathcal{D}_t$ and $\mathcal{D}_s + \mathcal{D}_t$ strategies achieve very similar performances thus, providing empirical evidences that, when the training set contains samples coming from different distributions, increasing the amount of training data does not result in an increasing of classification performances. All this points clearly indicate that a serious distribution shift exists between two years of SITS data on the considered study site. The highest gap is shown by the task (2018 $\rightarrow$ 2021) where, the best supervised machine learning method (RF) degrades its performances of around 18 points of F1-score.

Secondly, we can see that, generally, UDA strategies allow to reduce the performances gap induced by the distribution shifts between the source ($\mathcal{D}_s$) and the target ($\mathcal{D}_t$) domains. While this improvement is evident for the deep learning based techniques, it is not so explicit for the GFK approach. This is probably due to the fact that the GFK approach aligns domains independently from the underlying classification task while all the other approaches perform an end-to-end process that optimizes together the data distribution alignment and the classification process.

Thirdly, we can observe than *SpADANN* always obtains the best scores among the UDA methods. The gains compared to DANN and CDAN+E, which are the two most competitive approaches for all the transfer tasks, varies between 1.5 and 8.5 points of F1-score. Regarding the transfer task (2018 $\rightarrow$ 2020), the gap induced by the data distribution shift is largely reduced, especially in terms of accuracy and Kappa score.

Concerning the other two transfer tasks (2018 $\rightarrow$ 2021) and (2020 $\rightarrow$ 2021), *SpADANN* outperforms the supervised classifier approaches when they are trained on the target data ($\mathcal{D}_t$). This unexpected result is tightly related to several factors associated to the SITS data covering the 2021 year that describes the target domain ($\mathcal{D}_t$) in both transfer tasks. More precisely, as highlighted by the statistics reported in Table II, the cloud cover associated to the 2021 SITS data (regarding the ground truth pixels) is quite high and it affects periods of the year (growing and harvesting stages) that are crucial for the monitoring of the agricultural classes involved in the study site. Due to the gap filling process we have used to obtain complete time series data, the standard supervised

[2]https://tsaug.readthedocs.io/en/stable/
[3]https://github.com/ZJULearning/ALDA
[4]https://github.com/ecapliez/SpADANN

TABLE III
WEIGHTED F1-SCORE, ACCURACY AND KAPPA SCORE OF THE COMPETING APPROACHES FOR THE TRANSFER TASK (2018 → 2020). THE BEST SCORE, FOR UDA METHODS, IS HIGHLIGHTED IN BOLD.

		Subset $\mathcal{D}_t$			*Full* $\mathcal{D}_t$		
Strategy	**Method**	**F1-score**	**Accuracy**	**Kappa**	**F1-score**	**Accuracy**	**Kappa**
Only $\mathcal{D}_s$	TempCNN	60.1 ± 0.5	62.1 ± 0.5	53.7 ± 0.6	58.8	60.7	52.1
	RF	63.0 ± 0.4	66.0 ± 0.3	57.9 ± 0.4	63.0	65.7	57.7
SSL	FixMatch	32.3 ± 1.6	46.2 ± 1.6	35.7 ± 1.5	34.6	48.6	38.0
w/ UDA	GFK+MLP	56.7 ± 1.2	57.8 ± 1.4	49.2 ± 1.6	56.2	57.4	48.6
	ALDA	57.0 ± 1.4	62.9 ± 1.3	53.4 ± 1.5	59.0	64.1	55.3
	GFK+RF	63.0 ± 1.2	66.0 ± 1.0	58.1 ± 1.2	63.5	66.2	58.4
	MDD	60.3 ± 1.6	65.6 ± 1.4	56.7 ± 1.6	63.6	67.4	59.4
	ADDA	65.5 ± 1.1	69.0 ± 1.0	61.5 ± 1.2	66.3	69.3	62.2
	DANN	69.2 ± 0.8	72.1 ± 0.7	65.6 ± 0.8	69.2	71.9	65.6
	CDAN+E	70.8 ± 1.0	73.0 ± 1.0	66.6 ± 1.2	72.1	73.6	67.7
	SpADANN	**72.1 ± 1.1**	**75.1 ± 1.0**	**69.3 ± 1.1**	**73.8**	**76.5**	**71.1**
Only $\mathcal{D}_t$	TempCNN	77.6 ± 0.8	77.8 ± 0.8	72.5 ± 1.1			
	RF	78.0 ± 0.8	78.0 ± 0.9	72.9 ± 1.1			
$\mathcal{D}_s + \mathcal{D}_t$	TempCNN	78.3 ± 1.2	78.5 ± 1.2	73.5 ± 1.4			
	RF	78.2 ± 0.8	78.3 ± 0.8	73.2 ± 1.0			

TABLE IV
WEIGHTED F1-SCORE, ACCURACY AND KAPPA SCORE OF THE COMPETING APPROACHES FOR THE TRANSFER TASK (2018 → 2021). THE BEST SCORE, FOR UDA METHODS, IS HIGHLIGHTED IN BOLD.

		Subset $\mathcal{D}_t$			*Full* $\mathcal{D}_t$		
Strategy	**Method**	**F1-score**	**Accuracy**	**Kappa**	**F1-score**	**Accuracy**	**Kappa**
Only $\mathcal{D}_s$	TempCNN	48.3 ± 0.7	51.4 ± 0.6	41.1 ± 0.7	48.9	52.0	41.9
	RF	56.4 ± 0.6	58.6 ± 0.6	49.0 ± 0.7	57.4	59.6	50.5
SSL	FixMatch	36.0 ± 2.0	49.3 ± 1.8	38.3 ± 1.7	33.3	47.0	36.1
w/ UDA	GFK+MLP	50.0 ± 1.9	52.7 ± 1.9	42.8 ± 2.1	49.5	52.6	43.0
	ALDA	50.9 ± 1.6	55.7 ± 1.6	44.7 ± 1.6	54.6	59.2	49.1
	GFK+RF	57.5 ± 2.1	59.9 ± 1.9	50.4 ± 2.2	58.4	61.0	51.9
	MDD	59.7 ± 1.6	62.8 ± 1.5	53.6 ± 1.6	62.5	64.4	56.0
	ADDA	62.1 ± 1.6	63.3 ± 1.5	54.9 ± 1.7	64.0	65.1	57.4
	DANN	68.5 ± 1.5	69.1 ± 1.4	61.7 ± 1.7	70.2	70.7	63.9
	CDAN+E	65.2 ± 2.2	65.9 ± 2.0	57.7 ± 2.3	73.0	73.3	67.1
	SpADANN	**78.4 ± 1.1**	**79.9 ± 1.1**	**75.0 ± 1.2**	**78.9**	**80.5**	**75.9**
Only $\mathcal{D}_t$	TempCNN	72.0 ± 2.4	72.0 ± 2.3	65.2 ± 2.7			
	RF	74.3 ± 2.4	74.2 ± 2.3	67.9 ± 2.7			
$\mathcal{D}_s + \mathcal{D}_t$	TempCNN	72.7 ± 2.4	72.9 ± 2.2	66.2 ± 2.6			
	RF	74.4 ± 2.2	74.2 ± 2.1	68.0 ± 2.5			

machine learning methods could be biased by such synthetic information thus, leveraging the gap filled information in order to derive their decision boundary. Conversely, *SpADANN* is based on a domain alignment process, implemented via adversarial learning, that forces the whole pipeline to extract invariant features w.r.t. the two domains ($\mathcal{D}_s$ and $\mathcal{D}_t$). Such quest for invariant characteristics allows our framework to focus on common information, hence reasonably discarding specific per-year information that can be related to local (in terms of domain) behaviors or artifacts.

Finally, we can observe that in all transfer task the performances obtained on the *Full* $\mathcal{D}_t$ scenario are similar with those obtained by evaluating the method on subsets of the target domain *Subset* $\mathcal{D}_t$. This fact pinpoints that the test subsets extracted from the whole target domain are well representative of the whole target distribution.

1) Per-class analysis: In this section we report and discuss per-class analysis regarding the competing methods on the three transfer tasks we have considered. We firstly report per-class F1-score and, successively, we examine the different confusion matrices to understand possible inter-class mistakes. For this analysis, we focus our attention on the supervised methods (*Only* $\mathcal{D}_s$ and *Only* $\mathcal{D}_t$) as well as *SpADANN* and its direct ablation DANN.

The per-class F1-score are depicted in Figure 7, 8 and 9 for the (2018 → 2020), (2018 → 2021) and (2020 → 2021) transfer tasks, respectively.

We can clearly note that our framework achieves superior performances on the majority of the land cover classes ($grassland$, $shrubland$, $forest$, $baresoil/built-up$ and $water$). Such land cover classes show a more stable pattern among the years, hence exhibiting a smaller gap in terms of distribution shifts to fill between source and target domain.

TABLE V
WEIGHTED F1-SCORE, ACCURACY AND KAPPA SCORE OF THE COMPETING APPROACHES FOR THE TRANSFER TASK (2020 → 2021). THE BEST SCORE, FOR UDA METHODS, IS HIGHLIGHTED IN BOLD.

		Subset $\mathcal{D}_t$			*Full* $\mathcal{D}_t$		
Strategy	**Method**	**F1-score**	**Accuracy**	**Kappa**	**F1-score**	**Accuracy**	**Kappa**
Only $\mathcal{D}_s$	TempCNN	56.4 ± 0.8	57.9 ± 0.7	48.3 ± 0.8	56.6	57.9	48.6
	RF	65.2 ± 0.5	65.6 ± 0.5	57.6 ± 0.6	66.4	66.6	59.0
SSL	FixMatch	36.1 ± 2.0	49.2 ± 1.9	38.1 ± 1.8	37.6	49.4	38.8
w/ UDA	GFK+MLP	55.2 ± 1.8	54.8 ± 1.7	45.3 ± 1.9	52.5	52.2	42.6
	ALDA	60.5 ± 1.3	61.9 ± 1.4	52.1 ± 1.3	63.8	65.1	56.4
	GFK+RF	67.9 ± 1.7	67.8 ± 1.7	60.3 ± 1.9	68.7	68.5	61.4
	MDD	65.7 ± 1.6	66.4 ± 1.5	57.9 ± 1.6	66.9	67.4	59.2
	ADDA	69.3 ± 1.6	68.9 ± 1.5	61.6 ± 1.6	70.2	69.7	62.7
	DANN	72.2 ± 1.5	71.8 ± 1.5	65.0 ± 1.6	72.8	72.4	65.9
	CDAN+E	72.0 ± 1.4	71.5 ± 1.4	64.7 ± 1.5	72.8	72.2	65.8
	SpADANN	**81.5 ± 1.4**	**81.1 ± 1.4**	**76.7 ± 1.6**	**81.6**	**81.1**	**76.8**
Only $\mathcal{D}_t$	TempCNN	72.0 ± 2.4	72.0 ± 2.3	65.2 ± 2.7			
	RF	74.3 ± 2.4	74.2 ± 2.3	67.9 ± 2.7			
$\mathcal{D}_s + \mathcal{D}_t$	TempCNN	71.9 ± 2.7	72.0 ± 2.6	65.2 ± 3.1			
	RF	75.1 ± 2.0	74.8 ± 2.0	68.8 ± 2.3			

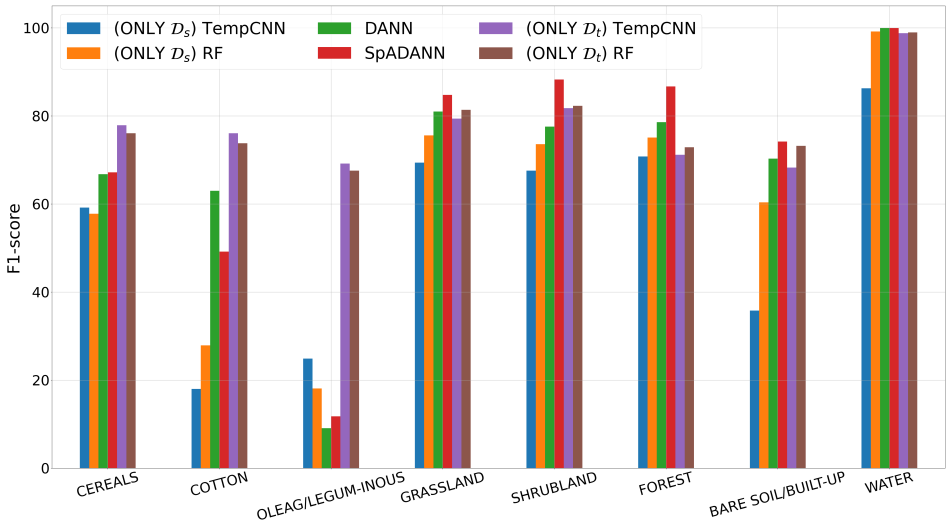


Fig. 7. Per land cover class F1-score considering different competing approaches considering the direct transfer strategy (only $\mathcal{D}_s$), the model trained on the target domain (only $\mathcal{D}_t$) as well as *SpADANN* and the best UDA competitor on the transfer task (2018 → 2020).

Interestingly, on such classes, *SpADANN* always achieves better performances than a supervised machine learning model directly learnt on the considered target domain.

Concerning the remaining classes (*cereals*, *cotton* and *oleaginous*) a different pattern is exhibited. When the 2018 year is considered as source domain ($\mathcal{D}_s$), the transfer on the agricultural classes has some issues to achieve performances on pair with the supervised methods trained on the target domain with the (2018 → 2021) task showing better transferability behaviour, using *SpADANN*, than the (2018 → 2020) one. Another interesting point is related to the poor performances that all the UDA methods exhibit for the *oleaginous* class. This is probably due to the fact that between 2018 and the other two subsequently years, the distribution of the ground truth data on such class drastically changes (see Table I). More precisely, in 2018 the *oleaginous* class cover an area of 350 000 m^2 while in 2020 and 2021 this surface doubles attaining a surface bigger than 730 000 m^2. This indicate that, in 2018, the *oleaginous* surfaces are under represented w.r.t. the other land cover classes thus, producing a dataset featured by high unbalancedness. Matter of facts, this shift in such agricultural practice significantly affects the capacity of all the model to generalize on the *oleaginous* land cover class when 2018 is considered as the source domain. In addition, due to the pseudo-labelling procedure associated to our framework, if a class in the source domain is highly under represented, the same class in the target domain will inherited this feature, with all the possible issues related to learning classification models under imbalance scenarios.

Regarding the (2020 → 2021) transfer task, here *SpADANN* effectively shows transfer capabilities also on the agricultural classes. The same behavior can be observed for all the other competing methods. These results can be explained, also in this case, by the fact that all the land cover classes are

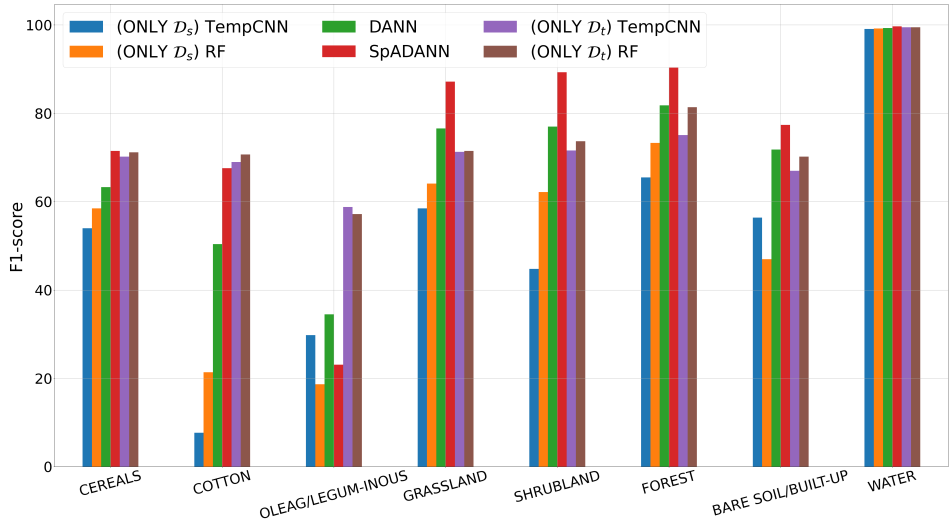


Fig. 8. Per land cover class F1-score considering different competing approaches considering the direct transfer strategy (only $\mathcal{D}_s$), the model trained on the target domain (only $\mathcal{D}_t$) as well as *SpADANN* and the best UDA competitor on the transfer task (2018 $\rightarrow$ 2021).

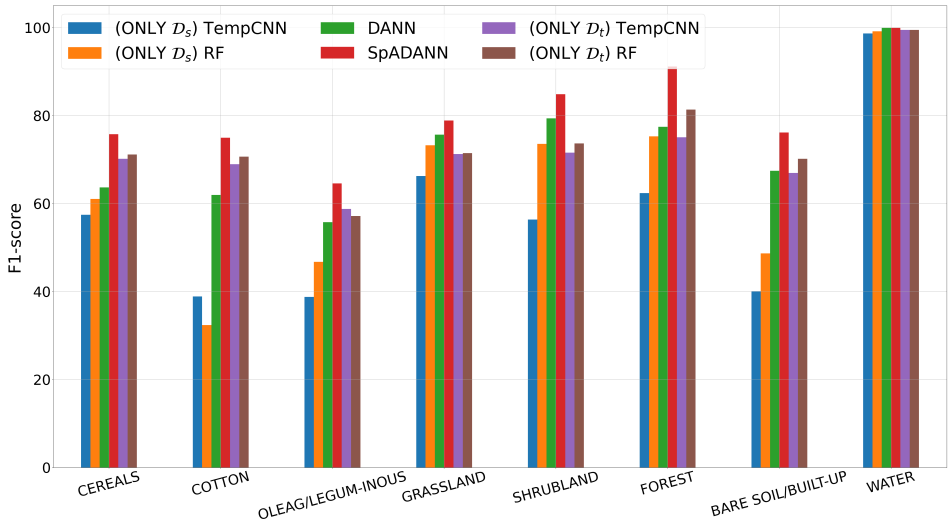


Fig. 9. Per land cover class F1-score considering different competing approaches considering the direct transfer strategy (only $\mathcal{D}_s$), the model trained on the target domain (only $\mathcal{D}_t$) as well as *SpADANN* and the best UDA competitor on the transfer task (2020 $\rightarrow$ 2021).

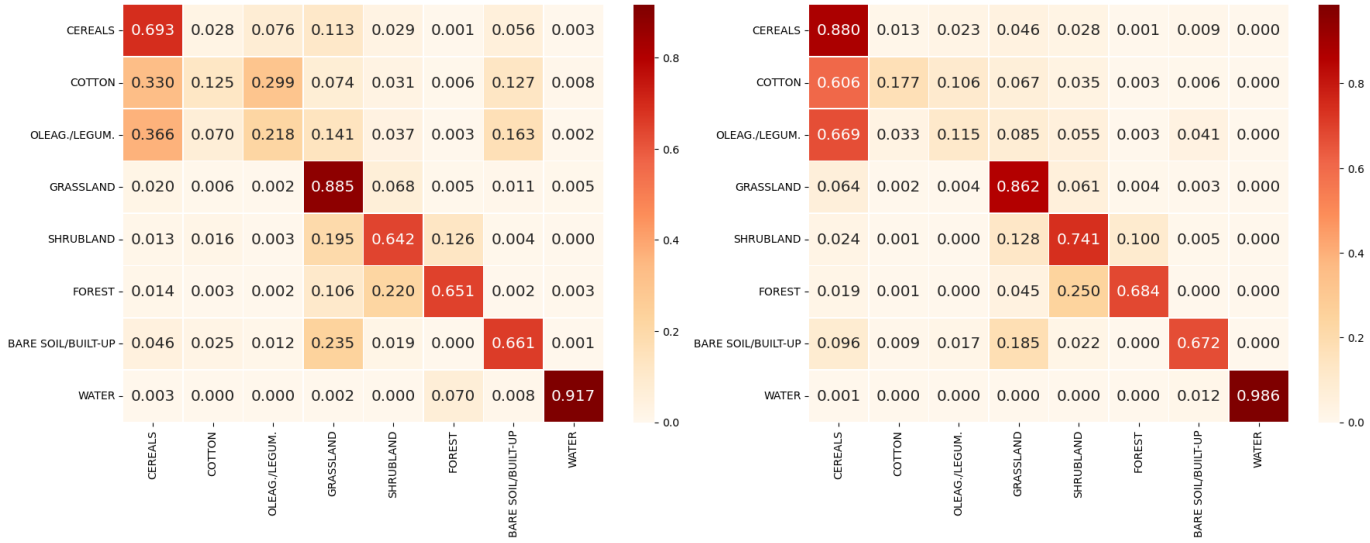
sufficiently represented in the source domain with a more balanced representation while class distributions in the source and target domain are more similar to each other (see Table I).

Figure 10, 11 and 12 depict confusion matrices for the (2018 $\rightarrow$ 2020), (2018 $\rightarrow$ 2021) and (2020 $\rightarrow$ 2021) transfer tasks, respectively.

Globally, the confusion matrices confirm the trend observed in the per-class F1-score analysis. All the methods have some troubles in discriminating among the different agricultural classes. As discussed before in Section V-C1, the UDA approaches suffer from class imbalance related to transfer tasks like (2018 $\rightarrow$ 2020) and (2018 $\rightarrow$ 2021). We can also note that some other coherent confusions arise between $grassland$ and $shrubland$, $shrubland$ and $forest$ land cover classes. This is expected since these three classes refer to three different degrees of density of woody vegetation in natural areas, which vary in a continuous way over the site, making a neat discrimination more challenging.

Despite this fact, *SpADANN* provides a more visible diagonal structure (the dark red blocks concentrated on the diagonal) than the second best competing UDA method alleviating some of the major confusions exhibited by the competing approaches. Same goes for the (2020 $\rightarrow$ 2021) transfer task, where *SpADANN* clearly outperforms the supervised approaches trained on the target domain.

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3263755

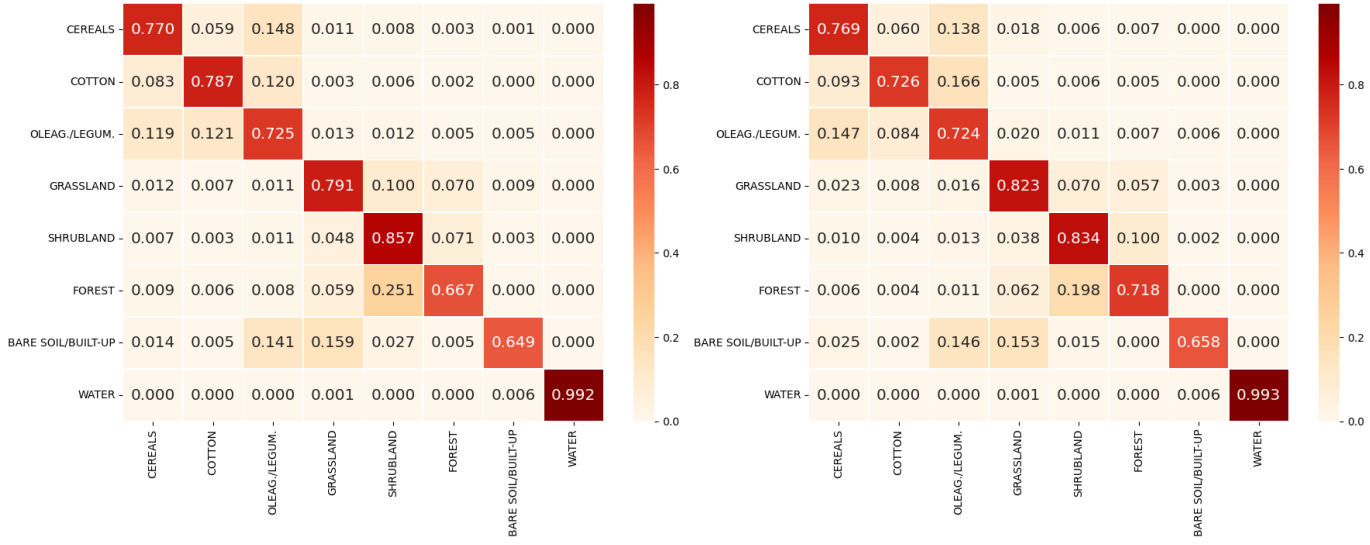


(a) (ONLY $\mathcal{D}_s$) TempCNN.

(b) (ONLY $\mathcal{D}_s$) RF.

(c) DANN.

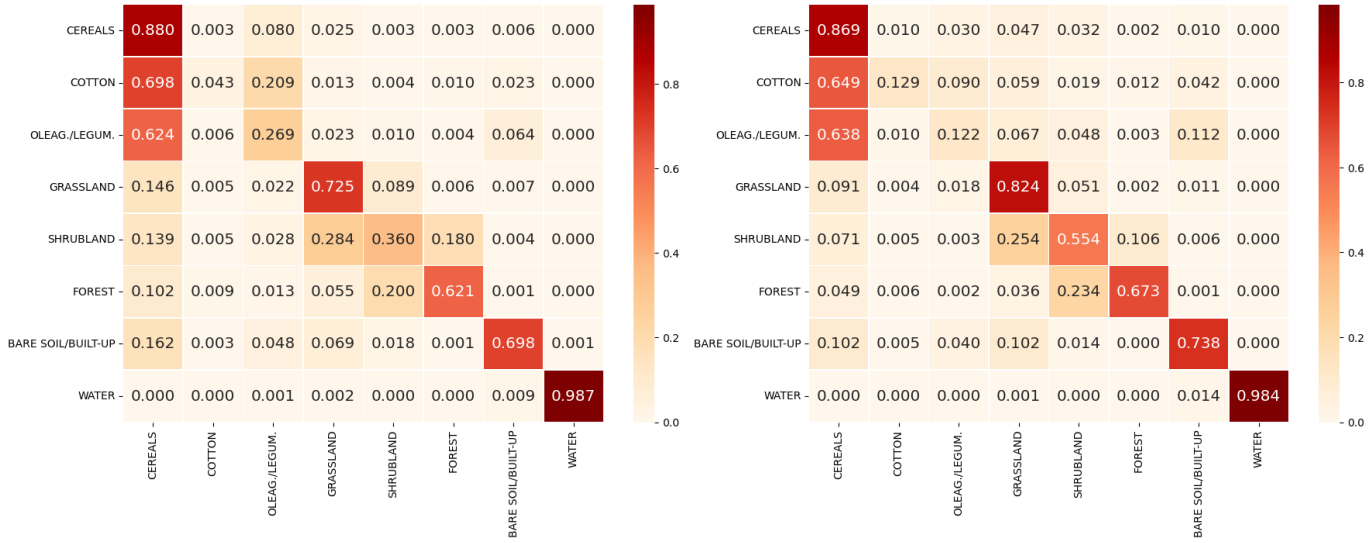
(d) *SpADANN*.



(e) (ONLY $\mathcal{D}_t$) TempCNN.

(f) (ONLY $\mathcal{D}_t$) RF.

Fig. 10. Confusion matrices of the land cover classification for the transfer task (2018 $\rightarrow$ 2020); True class (rows), Predicted class (columns).

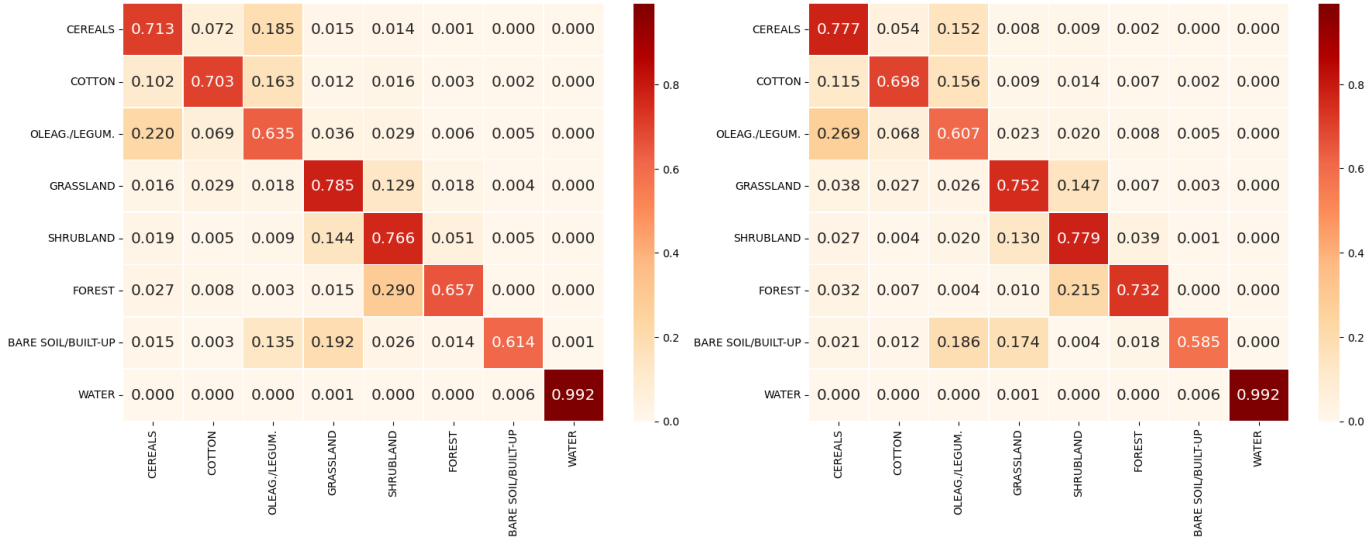


(a) (ONLY $\mathcal{D}_s$) TempCNN.

(b) (ONLY $\mathcal{D}_s$) RF.

(c) DANN.

(d) *SpADANN*.

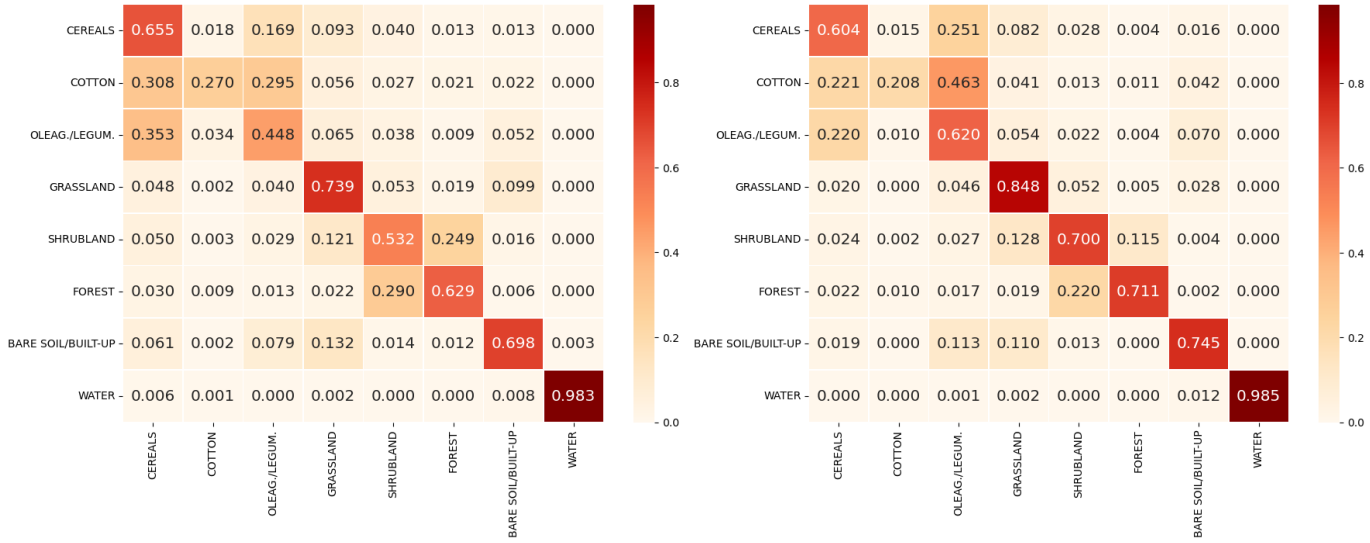


(e) (ONLY $\mathcal{D}_t$) TempCNN.

(f) (ONLY $\mathcal{D}_t$) RF.

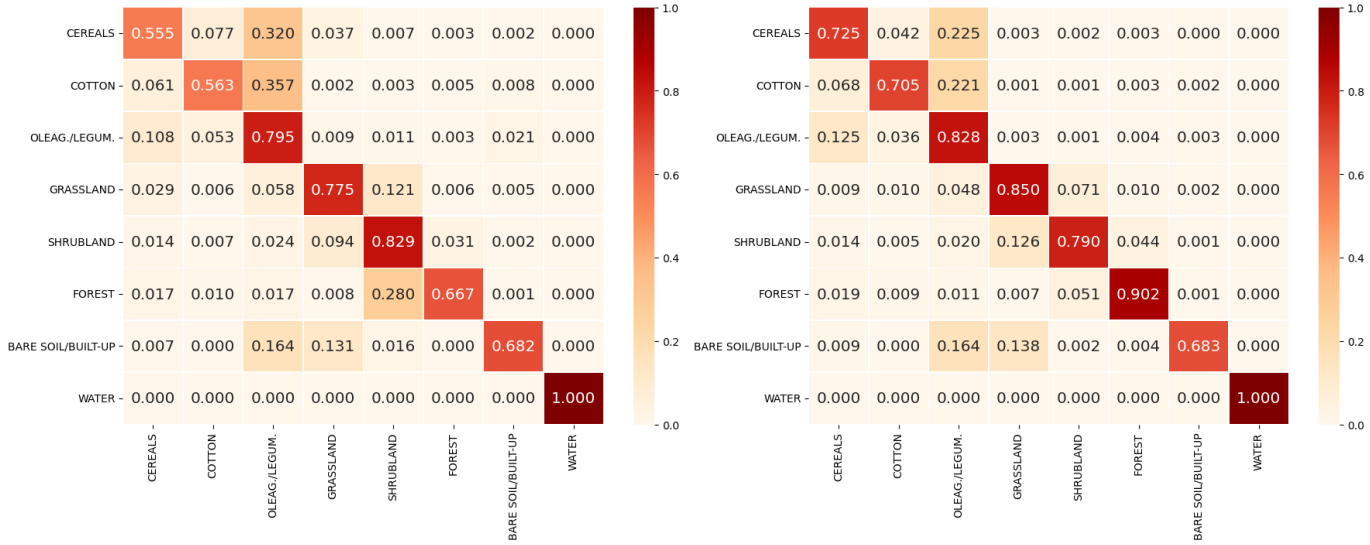
Fig. 11. Confusion matrices of the land cover classification for the transfer task (2018 $\rightarrow$ 2021); True class (rows), Predicted class (columns).

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3263755



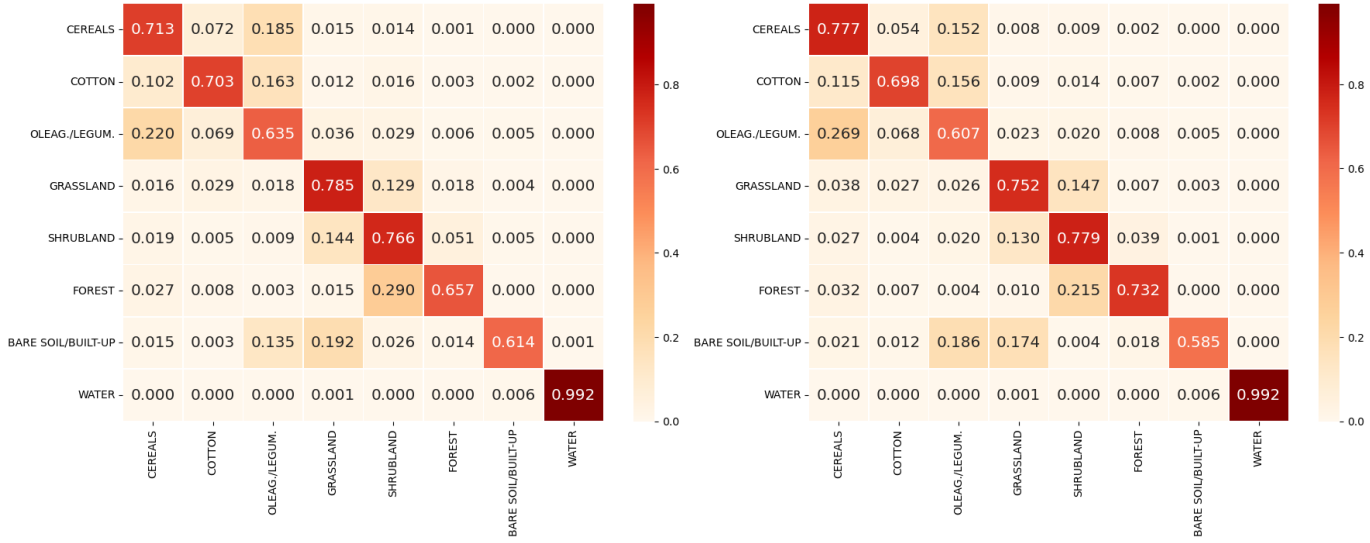
(a) (ONLY $\mathcal{D}_s$) TempCNN.

(b) (ONLY $\mathcal{D}_s$) RF.



(c) DANN.

(d) *SpADANN*.



(e) (ONLY $\mathcal{D}_t$) TempCNN.

(f) (ONLY $\mathcal{D}_t$) RF.

Fig. 12. Confusion matrices of the land cover classification for the transfer task (2020 $\rightarrow$ 2021); True class (rows), Predicted class (columns).

TABLE VI
TIME PERFORMANCES OF THE DIFFERENT UDA METHODS REGARDING THE TRANSFER TASK (2018 → 2020).

Method	Training Time
GFK + MLP	147 mins.
ALDA	298 mins.
GFK + RF	005 mins.
MDD	144 mins.
ADDA	885 mins.
DANN	357 mins.
CDAN+E	178 mins.
SpADANN	454 mins.

2) Running time of the UDA competing methods: Table VI summarizes the training time of the different unsupervised domain adaptation methods involved in the experimental evaluation. Beyond the GFK+RF strategy that requires only few minutes to learn its classification model, all the other methods require between 2 and 15 hours with *SpADANN* demanding around 7.5 hours in order to learn its internal parameters. Due to the fact that, in our situation, a LULC classification model demands to be trained once per season (or year), all the exhibited times remain more than reasonable with respect to the constraints associated to the downstream task.

3) Ablation analysis: In this section we disentangle the added value of the different components on which *SpADANN* relies. Table VII summarizes the behavior of the different ablations of *SpADANN* for the (2018 → 2020) task. In particular, we make reference to three specific ablations of *SpADANN* :

- *SpADANN*$_{noST}$, an ablation of the proposed method without the self-training step in the overall training stage. In this ablation for each source pixel in a batch, the corresponding target pixel (in terms of spatial location) is present in the same batch. In this way the adversarial learning stage is constrained to extract domain-invariant features for spatially correspondent source and target pixels conversely to what is done during the training of the DANN method where source and target pixels, in a batch, are selected completely at random. Here $L_{TOT} = L_{DANN}$;
- *SpADANN*$_{Th}$, an ablation of the proposed method where the selection of pseudo-labelled samples is achieved by the traditional thresholding approach [62]. More precisely, during the iterative process, samples from the target domain are associated to pseudo-label if the most confident class predicted by the land cover classifier has an associated confidence bigger than a specific threshold θ. We set θ equals to 0.9 similarly to what done for the ALDA and FixMatch approaches.
- *SpADANN*$_{onlyC1}$, an ablation of the proposed method that removes the pseudo-label condition requiring that the predicted class is equal to the true class ($Cl(x_i^s) = y_i^s$) for the self-training stage. Here, the L_p loss is redefined as:

$$L_p(X^s, X^t, Y^s, \hat{Y}^t|\Theta_F, \Theta_L) = \sum_{x_i^t \in X^t} \mathbb{1}_{\{Cl(x_i^s)=Cl(x_i^t)\}} H(\hat{y}_i^t, Cl_{prob}(x_i^t)) \quad (3)$$

We can first note that *SpADANN* provides by far better behaviors than DANN. This latter can be seen as a baseline ablation of our framework. Secondly, we can see that no real difference exists between DANN and *SpADANN*$_{noST}$. This underlines that the spatial alignment between source and target training batches, alone, does not provide any added value. Moreover, we observe that choosing pseudo-labels based on a traditional thresholding mechanism (*SpADANN*$_{Th}$) or only based on the spatial consistency of the model output classification (*SpADANN*$_{onlyC1}$) degrades the performances. This is probably due to the fact that the condition ($Cl(x_i^s) = y_i^s$), in conjunction with the condition ($Cl(x_i^s) = Cl(x_i^t)$), allows to filter out spurious information, consequently providing more guarantees on the quality of the pseudo-labels selected (from the target domain) to enrich the current training set. Finally, we observe that *SpADANN* always outperforms all its ablations underlying that the interplay among the different components on which it is built eventually provides a robust strategy for the temporal unsupervised domain adaptation problem from SITS data.

TABLE VII
F1-SCORE, ACCURACY AND KAPPA SCORE OF THE *SpADANN* ABLATIONS FOR THE TRANSFER TASK (2018 → 2020).

Method	F1-score	Accuracy	Kappa
DANN	69.2	71.9	65.6
SpADANN$_{noST}$	69.3	71.8	65.3
SpADANN$_{Th}$	68.4	71.5	64.9
SpADANN$_{onlyC1}$	69.1	73.2	67.0
SpADANN	**73.8**	**76.5**	**71.1**

4) Sensitivity to the β hyper-parameter: In this section we test the sensitivity of *SpADANN* to the value of the β hyper-parameter. Figure 13 summarizes the behavior of *SpADANN*, in terms of accuracy, on the three transfer tasks when the value of β varies between 0.5 and 1.0.

For the transfer tasks (2018 → 2021) and (2020 → 2021) we can note that, generally, as the value of the β hyper-parameter increases, the performances of our framework increases as well. The only exception is represented by the transfer task (2018 → 2021) in which a value of β equal to 1 (only consider pseudo-label extracted from the target domain at the end of the training process) degrades the final performances. This is probably due to the fact that, as highlighted by the previous results, a serious distribution shift exists between SITS data coming from 2018 and 2021, so that forcing the learning process to make a complete transfer from the source to the target domain results in a less appropriate classification model. Regarding the transfer task (2018 → 2020), *SpADANN* exhibits a slightly fluctuating behaviour with a variation of less than a point around an Accuracy value of 76%.

As empirical rule, we can state that considering values of the β hyper-parameter between 0.7 and 0.9 is the most

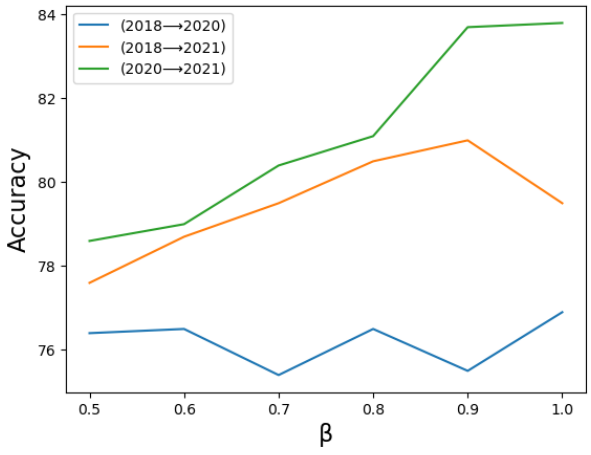


Fig. 13. Influence of β hyper-parameter on *SpADANN* performance, in terms of accuracy, for the three transfer tasks (2018 $\rightarrow$ 2020), (2018 $\rightarrow$ 2021)and (2020 $\rightarrow$ 2021).

appropriate choice since this setting can prevent the model to suddenly degenerate due to a complete transition from a source to a target domain characterized by very different data distributions.

D. Visual analysis

In this part of the experimental evaluation we conduct some qualitative analysis to assess further the behaviour of *SpADANN* in the case of transfer task (2018 $\rightarrow$ 2020), other transfer tasks are evaluated in appendix A. More precisely, we firstly investigate some extracts related to the land cover maps provided by *SpADANN* and some of the competing approaches and, successively, we visually investigate the internal representations learnt by the involved deep learning models.

1) Land cover maps: In Figure 14(b-e), maps corresponding to the 2018 to 2020 transfer task are compared, referred to the scene subset depicted in Figure 14(a). Maps shown here are, respectively, the one obtained using the RF classifier trained on the target domain $\mathcal{D}_t$ followed by the one obtained through a "naive" transfer (direct transfer without UDA of the RF model trained on source domain $\mathcal{D}_s$), and the two maps obtained through the DANN and *SpADANN* domain adaptation methods.

Accordingly to what reported in the quantitative analysis of Figure 10, the visual analysis confirms that transferring knowledge from 2018 to 2020 is a challenging task, probably due to longer term changes in seasonal vegetation dynamics that appear after a 2 year delay, as well as a redistribution of the proportions among the different crop classes. The main difference concerns the strong under-estimation of the $oleaginous/leguminous$ and $cotton$ classes in almost all maps using a transfer approach (Figure 14(c-e)), to the benefit of the $cereal$ class, with the direct transfer method being particularly destructive. However, if both UDA methods seem to effectively restore the extent of the $cotton$ class, it appears quite evidently that the *SpADANN* map is less noisy w.r.t. the DANN map, once again confirming a better potential in recovering spatial structures than its competitor.

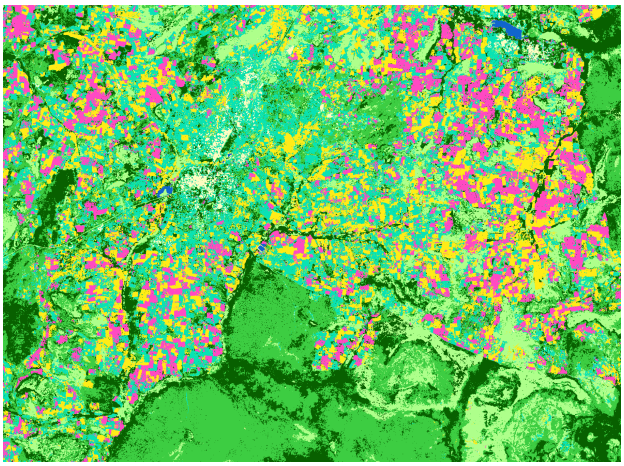
To better appreciate the spatial precision of the *SpADANN* maps w.r.t. its direct competitor, we also report some zoomed-in areas in Figure 15 and 16. In both cases, spatial details better emerge in the maps provided by *SpADANN*, both over agricultural fields, with more structured and less noisy plots (in terms of salt and pepper error) especially over the $cotton$ and $cereals$ classes, and over natural spaces.

The under-estimation phenomena related to the $oleaginous/leguminous$ and $cotton$ classes can be related to the crop class unbalancedness that features the 2018 GT data. More precisely, as shown in Section IV, the collected reference data for 2018 for both $oleaginous/leguminous$ and $cotton$, in terms of surfaces, is much lesser than the one collected for the $cereal$ class. This evident unbalancedness among crop classes affects both the direct transfer as well as the domain adaptation strategies thus, bringing distribution bias related to the source domain to the target one.

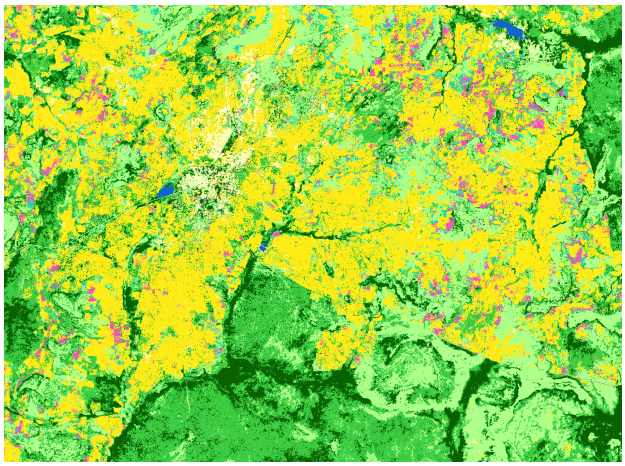
Such effects can be once again better observed in the zoomed-in areas of Figure 15 and Figure 16. This last observation seems to be in an opposite direction to the quantitative results previously reported on the same land cover classes. This is probably due to the fact that the GT data, that we use to both train and validate the different models, can only partially represent the study area, in terms of land cover class distributions. This fact underlines that, when the GT data collection is affected by operational constraints associated to costly and labour-intensive field campaigns, the investigation of the produced land cover maps is encouraged to evaluate the behaviour of the land cover classifiers. Only rely on quantitative analysis, via standard classification metrics, can provide a limited comprehension of the methods behaviour regarding the whole study area.

2) Visualisation of internal feature representations: In this last stage of our experimental evaluation, we provide a visual inspection of the internal feature representation learned by GFK, ADDA, DANN and *SpADANN* on the transfer task (2018 $\rightarrow$ 2020). To this end, we randomly chose 300 samples per land cover class from the target domain and we extracted the corresponding feature representation per method. Subsequently, we have applied t-SNE [63] to reduce the feature dimensionality for visualisation purposes. Results are depicted in Figure 17. We can note that all the methods well separate samples coming from the $water$ and $baresoil/built-up$ classes from the rest of the data. While GFK and ADDA clearly mix samples from all the other land cover classes together, DANN and *SpADANN* partially alleviate clutter issues on the remaining classes with the latter providing a slightly better visual behaviour in terms of cluster structure, on the considered subset of target data, than the former. This can be noted, for instance, regarding both the $grassland$, $shrubland$ and the $forest$ classes. Overall, the visualisation of internal features representation is coherent with the quantitative as well as qualitative findings we previously discussed.

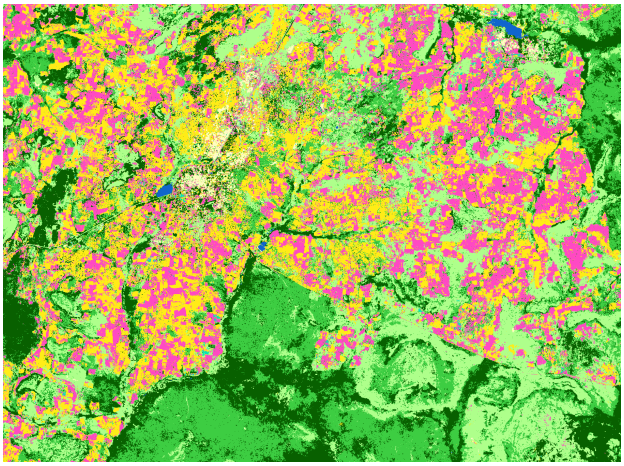
(a) Sentinel-2 image acquired on September 21, 2020.



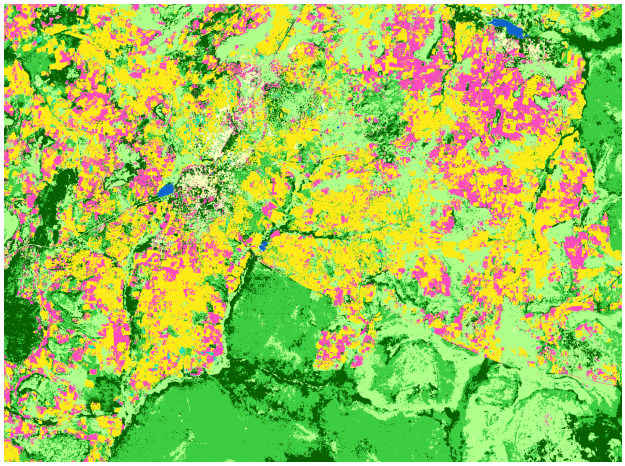
(b) (ONLY $\mathcal{D}_t$) RF.



(c) (ONLY $\mathcal{D}_s$) RF.



(d) DANN.

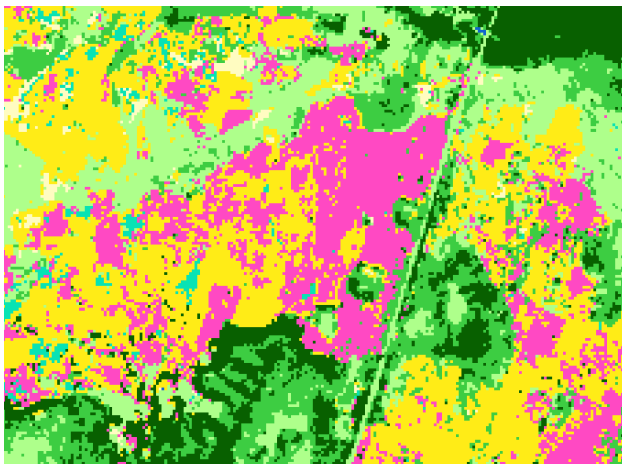


(e) *SpADANN*.

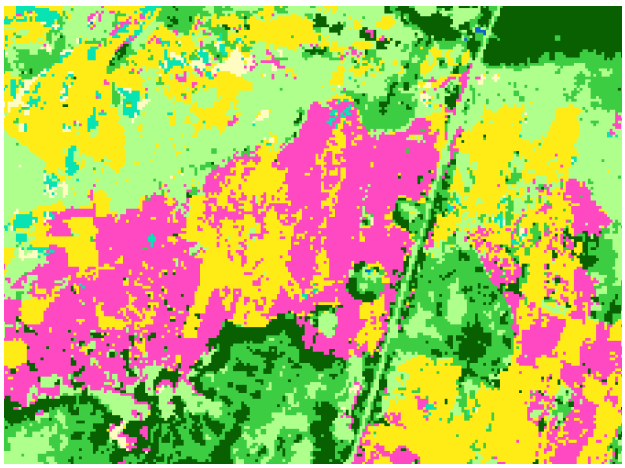
(f)

Fig. 14. Qualitative investigation of land cover maps produced by the RF methods - (ONLY $\mathcal{D}_t$) and (ONLY $\mathcal{D}_s$), DANN and *SpADANN* for the transfer task (2018 $\rightarrow$ 2020): zoom on the Koumbia city.

(a) Sentinel-2 image acquired on September 21, 2020.



(b) DANN.



(c) *SpADANN*.

(d)

Fig. 15. Qualitative investigation of land cover maps produced by DANN and *SpADANN* for the transfer task (2018 → 2020): zoom on Area n°1.

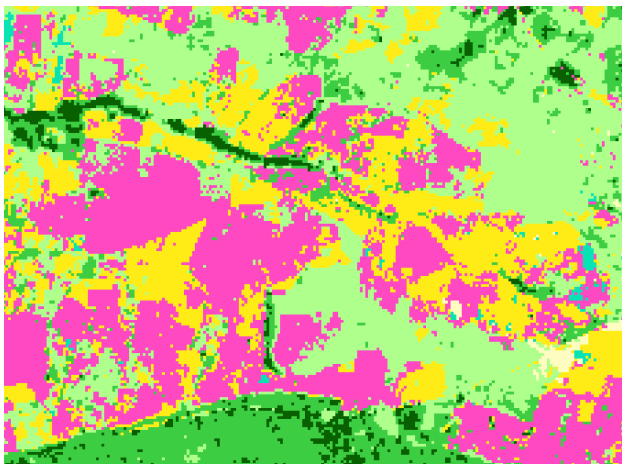
VI. Discussion

To summarize, our research study proposes a novel framework to perform temporal unsupervised domain adaptation (tUDA) for land cover mapping from SITS data. It couples together adversarial and self-training learning with the aim to cope with the distribution shifts affecting data coming from different years and hindering the transfer of standard machine learning models. In addition, to the best of our literature survey, this is the first time that recent deep learning methods are leveraged in the context of temporal transfer of LULC models from satellite image time series data.

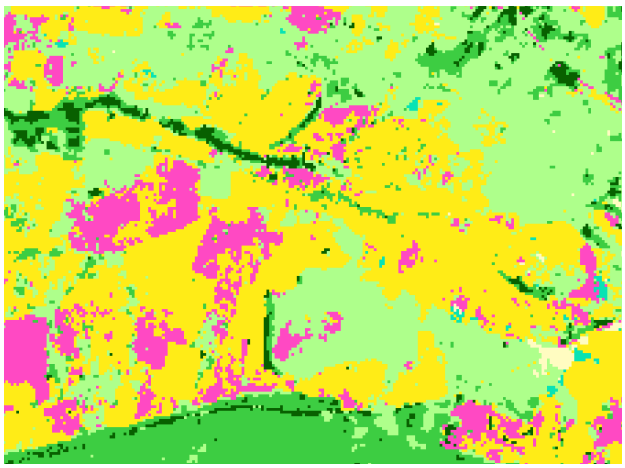
Firstly, we underline that our framework exploits the spatio-temporal information carried out by remote sensing data in order to temporally transfer the final LULC classification model. It explicitly leverages spatial information in order to transfer the model from one year (the source domain) to another year (the target domain) via self-training. The spatial alignment facilitates the identification of stable regions that act as tie points between the two domains while the self-training strategy allows the model to learn from its predictions. As underlined by the ablation analysis, such components are fundamental to support the behaviour of *SpADANN* in order to achieve its final goal.

Secondly, we have observed that, in general, performances vary from one transfer task to another. This is well-known in the general field of domain adaption since not all transfer tasks are equal [64]. More precisely, in our experimental evaluation we have noted that class imbalance in the source domain (i.e. $\mathcal{D}_s$ = 2018) can negatively influence the transfer from one year to another one as well as major changes in class distributions (i.e. the underlying cropping practices). These points

(a) Sentinel-2 image acquired on September 21, 2020.



(b) DANN.



(c) *SpADANN*.

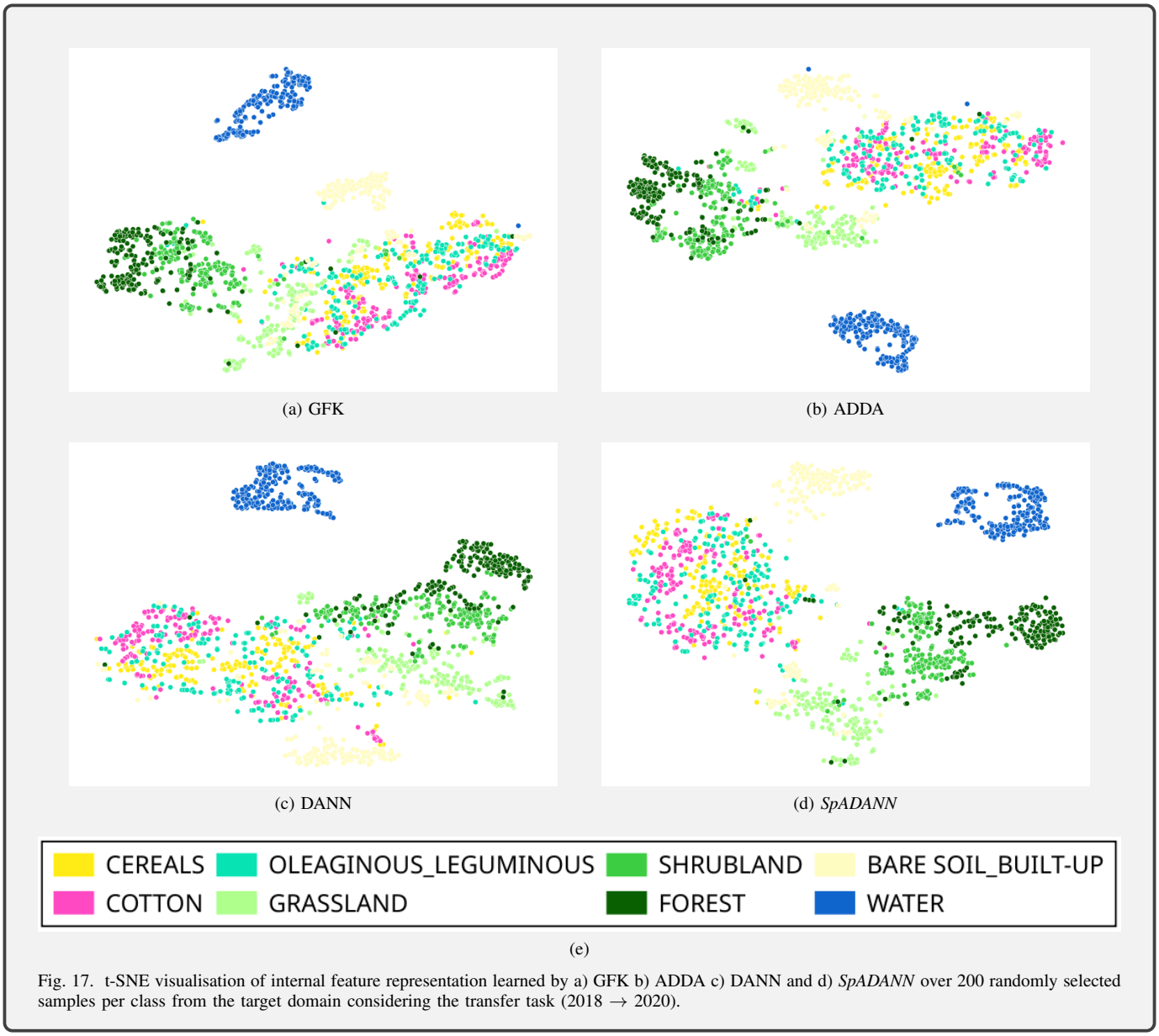
(d)

Fig. 16. Qualitative investigation of land cover maps produced by DANN and *SpADANN* for the transfer task (2018 $\rightarrow$ 2020): zoom on Area n°2.

suggest that *SpADANN* can be deployed in situations where no dramatic changes in the underlying landscape happen, thus potentially limiting costs and human efforts associated to field campaigns while reducing their frequency, for instance, from annual to every two or three years.

Thirdly, the use of domain adaptation for temporal LULC transfer opens new room for investigation in order to re-use already acquired data (on the same study site) with the aim to increase the return of investment on field campaigns and efforts done in the past. We have observed that, in some cases, combining together two years of satellite image time series data and previously acquired reference data can ameliorate the classification performances on the target domain (i.e. transfer task 2020 $\rightarrow$ 2021) due to the fact that *SpADANN* is steered to extract invariant representations w.r.t. a specific domain thus, alleviating year/domain specific issues (i.e. complex and unfavourable acquisition conditions). In addition, here we have focused our attention on the mono-source (mono-year) setting in which only a specific year is used as source domain while, in different real world LULC applications, we could access reference and satellite data spanning several previous years thus allowing the process to exploit multi-year information under the lens of unsupervised multi-source domain adaptation [12], [65], a recent family of techniques that extends standard unsupervised domain adaptation to consider multiple (related) source domains with the aim to generalize on the unlabelled target domain.

Fourthly, connections between recent spatial [11], [48] and temporal UDA approaches for SITS land cover mapping can be drawn. Both families of methods have the objective to cope with possible distribution shifts between source and target SITS data thus, coping with intrinsic domain shifts



(a) GFK (b) ADDA (c) DANN (d) *SpADANN* (e)

Fig. 17. t-SNE visualisation of internal feature representation learned by a) GFK b) ADDA c) DANN and d) *SpADANN* over 200 randomly selected samples per class from the target domain considering the transfer task (2018 $\rightarrow$ 2020).

that can be caused by different environmental, weather or climate conditions of acquisition. In our framework, in order to cope with the temporal UDA scenario, the spatial alignment between source and target data is explicitly exploited with the aim to alleviate distribution shifts while, this characteristic cannot be leveraged in the context of spatial UDA since the two domains are spatially unrelated. This fact prevents the use of *SpADANN*, as it is, for the spatial UDA scenario while the contrary should be possible. Nevertheless, as shown in the ablation study section, the use of spatial alignment derived information constitutes a crucial asset that effectively guides the self-training process. For this reason, methods that will not integrate such knowledge will probably fail to provide an effective solution for the temporal unsupervised domain adaptation scenario.

Finally, we remind that our task is characterized by operational/realistic constraints, implying a limited and sparse amount of reference data from which the relationships between remote sensing data and the fine land cover classes is learnt. This is why, in our case study, the TempCNN deep learning approach does not exhibit competitive behavior compared to standard machine learning techniques such as Random Forest, regarding intra-domain classification. Conversely, the use of both source and target domains, simultaneously, together with the self-training strategy we have proposed, permits to increase the amount of data labels the model can access to learn its internal classification function. This means that the proposed framework can be deployed in situations characterized by moderate data labels availability due to its capacity to progressively and incrementally exploit knowledge coming from the two domains in a complementary manner.

The conducted research opens the way to several future works. As of now, *SpADANN* works in a standard (mono-source) UDA setting where only a single labelled source

domain is considered. Due to the fact that previous field campaigns can span multiple years, a possible research direction is the extension of *SpADANN* to a multi-source unsupervised domain adaptation setting where multiple labelled source domains can be exploited in order to further improve the temporal transferability performances. Another possible follow-up is related to extend our framework to a multiple modality scenario where the study area is described by multi-sensor remote sensing data like, for instance, satellite image time series coming from both Synthetic Aperture Radar (SAR) and optical sensors (e.g., Sentinel-1 and Sentinel-2). While the majority of UDA approaches consider a mono modality setting where domains are described by only one modality, very few research studies exist for the unsupervised domain adaptation under the multi-modality scenario, even in the general field of computer vision and signal processing.

VII. Conclusion

In this work we have presented *SpADANN*, a new framework to cope with temporal unsupervised domain adaptation for land cover mapping from Satellite Image Time Series (SITS) data. Our approach combines adversarial learning and self-training with the aim to progressively transfer/adapt a neural network model from a source domain (a specific year featured by ground truth data) to a target domain (a successive year where no label information is available) in order to provide land cover classification on the latter. While the adversarial learning strategy is implemented by means of gradient reversal layer, in order to extract domain-invariant features, the self-training stage selects pseudo-labels on the target domain leveraging spatial consistency between domains.

The obtained results on the *Koumbia* study site have highlighted the quality of our framework regarding both quantitative and qualitative analyses with respect all the UDA competitors. This is tightly related to the fact that *SpADANN* explicitly takes advantage of the spatio-temporal features that highly characterize SITS data. In addition, we could also show that, when the general land cover distribution does not exhibit drastic changes between source and target domain, the proposed method is highly competitive compared to a model directly trained on the target domain. This last point can be explained by the fact that our framework focuses its attention on domain-invariant characteristics thus, probably, discarding specific per-year information that can be related to local (in terms of domain) behaviors or artifacts and leveraging more training data due to the new self-training strategy we have proposed.

Acknowledgements

This work was supported by the French National Association of Research and Technology through the Convention Industrielle de Formation par la REcherche (CIFRE refered as 2019/1810) Ph.D. grant.

References

[1] C. Toth and G. Jóźków, "Remote sensing platforms and sensors: A survey," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 115, pp. 22–36, 2016.

[2] N. Kolecka, C. Ginzler, R. Pazur, B. Price, and P. H. Verburg, "Regional scale mapping of grassland mowing frequency with sentinel-2 time series," *Remote Sensing*, vol. 10, no. 8, p. 1221, 2018.

[3] C. Pelletier, G. I. Webb, and F. Petitjean, "Temporal convolutional neural network for the classification of satellite image time series," *Rem. Sens.*, vol. 11, no. 5, 2019.

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

[5] F. Guttler, D. Ienco, J. Nin, M. Teisseire, and P. Poncelet, "A graph-based approach to detect spatiotemporal dynamics in satellite image time series," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 130, pp. 92–107, 2017.

[6] C. Gómez, J. C. White, and M. A. Wulder, "Optical remotely sensed time series data for land cover classification: A review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 116, pp. 55–72, 2016.

[7] 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.

[8] E. Cherif, M. Hell, and M. Brandmeier, "Deepforest: Novel deep learning models for land use and land cover classification using multi-temporal and -modal sentinel data of the amazon basin," *Remote Sensing*, vol. 14, no. 19, 2022. [Online]. Available: https://www.mdpi.com/2072-4292/14/19/5000

[9] B. Tardy, J. Inglada, and J. Michel, "Fusion approaches for land cover map production using high resolution image time series without reference data of the corresponding period," *Rem. Sens.*, vol. 9, no. 11, 2017. [Online]. Available: https://www.mdpi.com/2072-4292/9/11/1151

[10] ——, "Assessment of optimal transport for operational land-cover mapping using high-resolution satellite images time series without reference data of the mapping period," *Remote. Sens.*, vol. 11, no. 9, p. 1047, 2019.

[11] J. Nyborg, C. Pelletier, S. Lefèvre, and I. Assent, "Timematch: Unsupervised cross-region adaptation by temporal shift estimation," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 188, pp. 301–313, 2022.

[12] G. Wilson and D. J. Cook, "A survey of unsupervised deep domain adaptation," *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 5, pp. 51:1–51:46, 2020.

[13] W. M. Kouw and M. Loog, "A review of domain adaptation without target labels," *IEEE Trans. on PAMI*, vol. 43, no. 3, pp. 766–785, 2021.

[14] 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.

[15] 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.

[16] 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.

[17] Y. J. E. Gbodjo, O. Montet, D. Ienco, R. Gaetano, and S. Dupuy, "Multisensor land cover classification with sparsely annotated data based on convolutional neural networks and self-distillation," *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, vol. 14, pp. 11 485–11 499, 2021.

[18] A. Stoian, V. Poulain, J. Inglada, V. Poughon, and D. Derksen, "Land cover maps production with high resolution satellite image time series and convolutional neural networks: Adaptations and limits for operational systems," *Remote. Sens.*, vol. 11, no. 17, p. 1986, 2019.

[19] V. S. F. Garnot and L. Landrieu, "Panoptic segmentation of satellite image time series with convolutional temporal attention networks," in *ICCV*. IEEE, 2021, pp. 4852–4861.

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

[21] D. Ienco, R. Interdonato, R. Gaetano, and D. Ho Tong 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. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271619302278

[22] 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.

[23] M. Rußwurm and M. Körner, "Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images," in *CVPR Workshop*, 2017, pp. 1496–1504.

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

[25] L. Zhong, L. Hu, and H. Zhou, "Deep learning based multi-temporal crop classification," *Remote Sensing of Environment*, vol. 221, pp. 430–443, 2019.

[26] V. S. F. Garnot, L. Landrieu, S. Giordano, and N. Chehata, "Satellite image time series classification with pixel-set encoders and temporal self-attention," in *CVPR*. IEEE, 2020, pp. 12 322–12 331.

[27] M. Amini, V. Feofanov, L. Pauletto, E. Devijver, and Y. Maximov, "Self-training: A survey," *CoRR*, vol. abs/2202.12040, 2022.

[28] J. E. van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Mach. Learn.*, vol. 109, no. 2, pp. 373–440, 2020.

[29] A.Blum and T. M. Mitchell, "Combining labeled and unlabeled data with co-training," in *COLT*. ACM, 1998, pp. 92–100.

[30] Z. Zhou and M. Li, "Tri-training: Exploiting unlabeled data using three classifiers," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 11, pp. 1529–1541, 2005.

[31] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *NIPS*, 2017, pp. 1195–1204.

[32] B. Banerjee, F. Bovolo, A. Bhattacharya, L. Bruzzone, S. Chaudhuri, and B. K. Mohan, "A new self-training-based unsupervised satellite image classification technique using cluster ensemble strategy," *IEEE Geosci. Remote. Sens. Lett.*, vol. 12, no. 4, pp. 741–745, 2015.

[33] Y. Wu, G. Mu, C. Qin, Q. Miao, W. Ma, and X. Zhang, "Semi-supervised hyperspectral image classification via spatial-regulated self-training," *Remote. Sens.*, vol. 12, no. 1, p. 159, 2020.

[34] C. Paris, L. Orlandi, and L. Bruzzone, "An interactive strategy for the training set definition based on active self-paced learning implemented on a cloud-computing platform," *IEEE Geosci. Remote. Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[35] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. Raffel, E. D. Cubuk, A. Kurakin, and C. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," in *NeurIPS*, 2020.

[36] D. Ienco, D. P. dos Santos, and A. C. P. L. F. de Carvalho, "Evaluate pseudo labeling and CNN for multi-variate time series classification in low-data regimes," in *ICANN*, vol. 12895, 2021, pp. 126–137.

[37] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, 2021.

[38] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *2012 IEEE CVPR*, 2012, pp. 2066–2073.

[39] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *2017 IEEE CVPR*, 2017, pp. 2962–2971.

[40] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[41] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," in *NeurIPS*, 2018, pp. 1647–1657.

[42] Y. Zhang, T. Liu, M. Long, and M. I. Jordan, "Bridging theory and algorithm for domain adaptation," in *ICML*, vol. 97, 2019, pp. 7404–7413.

[43] M. Chen, S. Zhao, H. Liu, and D. Cai, "Adversarial-learned loss for domain adaptation," in *AAAI*, 2020, pp. 3521–3528.

[44] D. Tuia, C. Persello, and L. Bruzzone, "Recent advances in domain adaptation for the classification of remote sensing data," *Geoscience and Remote Sensing Magazine*, vol. 4, no. 2, pp. 41–57, 2016.

[45] J. Chen, J. Zhu, Y. Guo, G. Sun, Y. Zhang, and M. Deng, "Unsupervised domain adaptation for semantic segmentation of high-resolution remote sensing imagery driven by category-certainty attention," *IEEE Trans. Geosci. Remote. Sens.*, vol. 60, pp. 1–15, 2022.

[46] B. Tardy, J. Inglada, and J. Michel, "Assessment of optimal transport for operational land-cover mapping using high-resolution satellite images time series without reference data of the mapping period," *Remote Sensing*, vol. 11, no. 9, 2019. [Online]. Available: https://www.mdpi.com/2072-4292/11/9/1047

[47] J. Wang, A. Ma, Y. Zhong, Z. Zheng, and L. Zhang, "Cross-sensor domain adaptation for high spatial resolution urban land-cover mapping: From airborne to spaceborne imagery," *Remote Sensing of Environment*, vol. 277, p. 113058, 2022.

[48] Z. Wang, H. Zhang, W. He, and L. Zhang, "Phenology alignment network: A novel framework for cross-regional time series crop classification," in *CVPR Workshop*, 2021, pp. 2940–2949.

[49] M. Martini, V. Mazzia, A. Khaliq, and M. Chiaberge, "Domain-adversarial training of self-attention-based networks for land cover classification using multi-temporal sentinel-2 satellite imagery," *Remote. Sens.*, vol. 13, no. 13, p. 2564, 2021.

[50] J. N. Kundu, N. Venkat, A. Revanur, R. M. V., and R. V. Babu, "Towards inheritable models for open-set domain adaptation," in *CVPR*, 2020, pp. 12 373–12 382.

[51] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, 2013, p. 896.

[52] Y. Li, D. Huang, D. Qin, L. Wang, and B. Gong, "Improving object detection with selective self-supervised self-training," in *ECCV*, 2020, pp. 589–607.

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

[54] J. Inglada, A. Vincent, M. Arias, and B. Tardy, "iota2-a25386," Jul. 2016. [Online]. Available: https://doi.org/10.5281/zenodo.58150

[55] A. Jolivot, V. Lebourgeois, M. Ameline, V. Andriamanga, B. Bellon, M. Castets, A. Crespin-Boucaud, P. Defourny, S. Diaz, M. Dieye, S. Dupuy, R. Ferraz, R. Gaetano, M. Gely, C. Jahel, B. Kabore, C. Lelong, G. Le Maire, L. Leroux, D. Lo Seen, M. Muthoni, B. Ndao, T. Newby, C. L. M. De Oliveira Santos, E. Rasoamalala, M. Simoes, I. Thiaw, A. Timmermans, A. Tran, and A. Begue, "Harmonized in situ JECAM datasets for agricultural land use mapping and monitoring in tropical countries," 2021. [Online]. Available: https://doi.org/10.18167/DVN1/P7OLAP

[56] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24–31, 2016.

[57] C. Persello, J. D. Wegner, R. Hänsch, D. Tuia, P. Ghamisi, M. Koeva, and G. Camps-Valls, "Deep learning and earth observation to support the sustainable development goals: Current approaches, open challenges, and future opportunities," *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, no. 2, pp. 172–200, 2022.

[58] N. Karasiak, J.-F. Dejoux, C. Monteil, and D. Sheeren, "Spatial dependence between training and test sets: another pitfall of classification accuracy assessment in remote sensing," *Machine Learning*, 2021.

[59] Y. Li, N. Wang, J. Shi, J. Liu, and X. Hou, "Revisiting batch normalization for practical domain adaptation," in *ICLR*, 2017.

[60] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, ser. JMLR Workshop and Conference Proceedings, vol. 37, 2015, pp. 448–456.

[61] A. de Mathelin, F. Deheeger, G. Richard, M. Mougeot, and N. Vayatis, "Adapt: Awesome domain adaptation python toolbox," *arXiv preprint arXiv:2107.03049*, 2021.

[62] W. Shi, Y. Gong, C. Ding, Z. Ma, X. Tao, and N. Zheng, "Transductive semi-supervised deep learning using min-max features," in *ECCV*, vol. 11209, 2018, pp. 311–327.

[63] L. van der Maaten and G. E. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[64] J. Choi, Y. Choi, J. Kim, J. Chang, I. Kwon, Y. Gwon, and S. Min, "Visual domain adaptation by consensus-based transfer to intermediate domain," in *AAAI*, 2020, pp. 10 655–10 662.

[65] D. Zhang, M. Ye, Y. Liu, L. Xiong, and L. Zhou, "Multi-source unsupervised domain adaptation for object detection," *Inf. Fusion*, vol. 78, pp. 138–148, 2022. [Online]. Available: https://doi.org/10.1016/j.inffus.2021.09.011

Emmanuel Capliez received the engineer's degree in computer science with specialty in computer vision and remote sensing from the ENSTA Paris IP engineering school, Paris, France, in 2004. He is currently working toward his Ph.D. in computer science at the UMR TETIS laboratory, INRAE, Montpellier working on temporal domain adaptation approaches devoted to manage remote sensing data for land cover mapping.

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. From 2009 to 2015, he conducted postdoctoral research on fundamental image processing mainly applied to remote sensing for different institutions (INRIA, MTA SZTAKI, Telecom Paristech, University of Naples). 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 oriented to the design of operational methods for information extraction from multi-sensor and multi-temporal imagery (e.g. land cover and land use mapping).

Nicolas Baghdadi Nicolas Baghdadi received his Ph.D. degree from the University of Toulon, France in 1994. From 1995 to 1997, he was a postdoctoral researcher at INRS Ete – Water Earth Environment Research Centre, Quebec University, Canada. From 1998 to 2008, he was with the French geological Survey (BRGM), Orleans, France. Since 2008, he is a Research Director at the French Research Institute of Science and Technology for Environment and Agriculture (IRSTEA, now INRAE). He is the editor of two series of books: Land Surface Remote Sensing set and QGIS in remote sensing set http://www.iste.co.uk/subject.php?id=NJNK His main field of interest is the analysis of remote sensing data (mainly radar and lidar) and the retrieval of environmental parameters (e.g. soil moisture content, soil roughness, canopy height, forest biomass, etc.). From 2013 to 2022, Nicolas Baghdadi has been the Scientific Director of the French Land Data Center.

Adrien Hadj Salah received the engineer's degree from the ENAC engineering school, Toulouse, France, and also the Master's degree in air and space operations technology from the Technical University of Berlin, in 2012. He is currently the Computer Vision and Advanced Studies team leader for Airbus Defence and Space, Toulouse, France.

APPENDIX

A. Ablation analysis for the (2018 → 2021) and the (2020 → 2021) transfer tasks

The results obtained pour (2018 → 2021) and (2020 → 2021) transfer tasks, and detailed in Figure VIII and IX, confirm the findings presented for (2018 → 2020) transfer task in V-C3

TABLE VIII
F1-SCORE, ACCURACY AND KAPPA SCORE OF THE *SpADANN* ABLATIONS FOR THE TRANSFER TASK (2018 → 2021).

Method	F1-score	Accuracy	Kappa
DANN	70.2	70.7	63.9
$SpADANN_{noST}$	70.8	71.3	64.6
$SpADANN_{Th}$	70.5	71.3	64.7
$SpADANN_{onlyC1}$	75.6	78.0	72.7
SpADANN	**78.9**	**80.5**	**75.9**

TABLE IX
F1-SCORE, ACCURACY AND KAPPA SCORE OF THE *SpADANN* ABLATIONS FOR THE TRANSFER TASK (2020 → 2021).

Method	F1-score	Accuracy	Kappa
$DANN$	72.8	72.4	65.9
$SpADANN_{noST}$	72.3	71.7	65.2
$SpADANN_{Th}$	73.6	73.1	66.9
$SpADANN_{onlyC1}$	80.9	80.4	76.0
SpADANN	**81.6**	**81.1**	**76.8**

B. Land cover maps for the (2018 → 2021) and the (2020 → 2021) transfer tasks

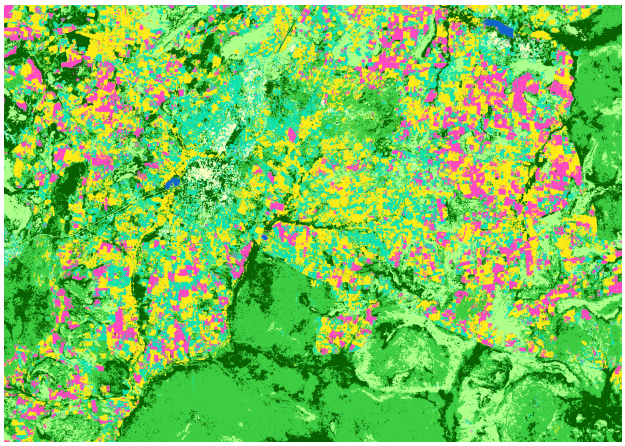
As with (2018 → 2020), the challenging transfer task (2018 → 2021) is affected by a longer term changes in seasonal vegetation dynamics that appear after a 3 year delay, and a redistribution of the proportions among the different crop classes. That is why we can formulate after visual analysis of Figure 18, 19 and 20 the same findings as for transfer task (2018 → 2021).

In the case of task (2020 → 2021), Using the direct transfer strategy (ONLY $\mathcal{D}_s$) with RF, the provided map mainly shows a significant reduction of the surface covered by the $cotton$ and $cereals$ class with respect to the baseline to the profit of the $oleaginous/legouminous$ class. An increased noise in the natural vegetation areas is also present. When a UDA approach is used things improve neatly: using DANN, the extent of the $cotton$ class is mainly restored, but some disproportion is present between the $cereals$ and $oleaginous/leguminous$ class, whose discrimination is more challenging. Finally, *SpADANN* seems to visually provide the best results, with an almost completely restored balance among crop classes, as well as improved details over natural vegetation areas. To better appreciate the spatial precision of the *SpADANN* maps w.r.t. its direct competitor, we also report some zoomed-in areas in Figure 22 and23. In both cases, spatial details better emerge in the maps provided by *SpADANN*, both over agricultural fields, with more structured and less noisy plots (in terms of salt and pepper error) especially over the $cotton$ and $cereals$ classes, and over natural spaces.

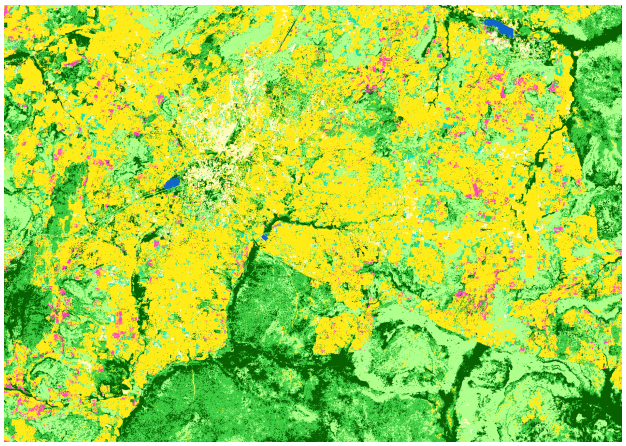
C. Visualisation of internal feature representations for the (2018 → 2021) and the (2020 → 2021) transfer tasks

Regarding statements mentionned in V-D2, they also apply to t-SNE visualisation for transfer task (2018 → 2021) and (2020 → 2021), which are respectively detailed in Figure 24 and 25.

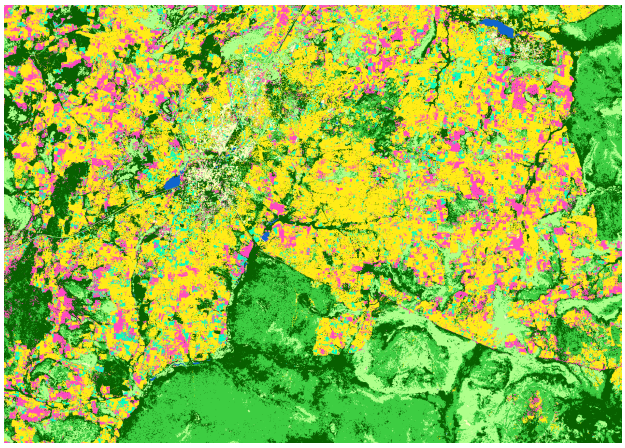
(a) Sentinel-2 image acquired on September 29, 2021.



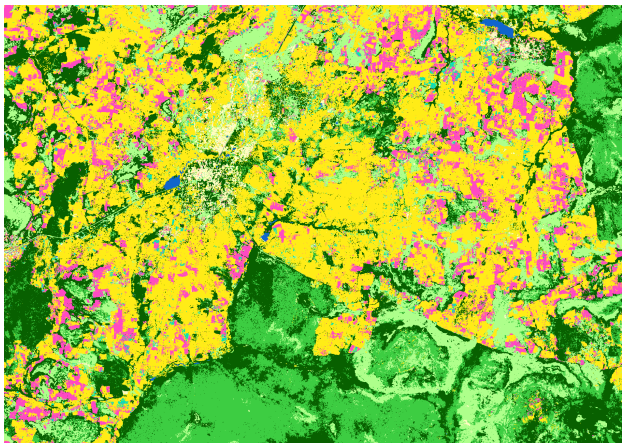
(b) (ONLY $\mathcal{D}_t$) RF.



(c) (ONLY $\mathcal{D}_s$) RF.



(d) DANN.

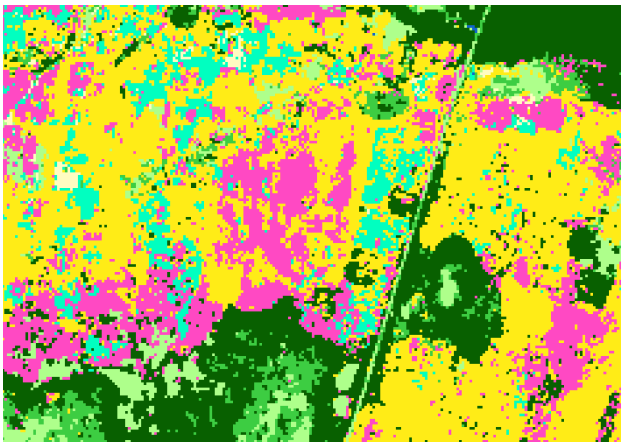


(e) *SpADANN*.

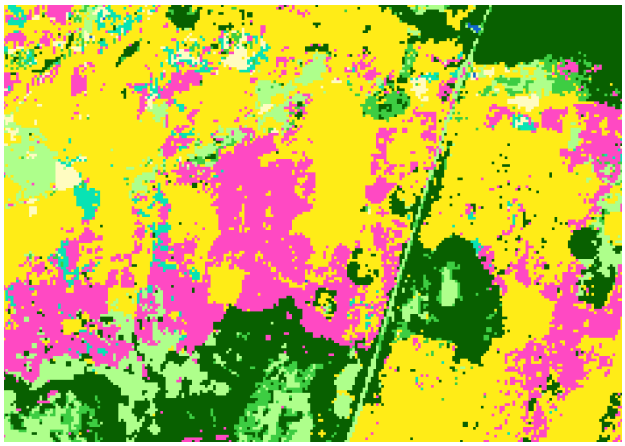
(f)

Fig. 18. Qualitative investigation of land cover maps produced by the RF methods - (ONLY $\mathcal{D}_t$) and (ONLY $\mathcal{D}_s$), DANN and *SpADANN* for the transfer task (2018 $\rightarrow$ 2021): zoom on the Koumbia city.

(a) Sentinel-2 image acquired on September 29, 2021.



(b) DANN.

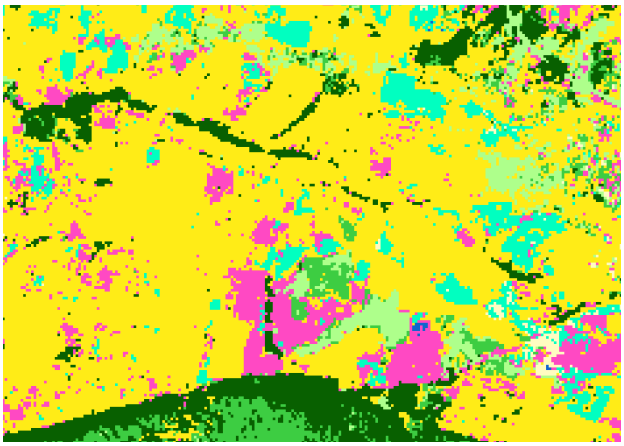


(c) *SpADANN*.

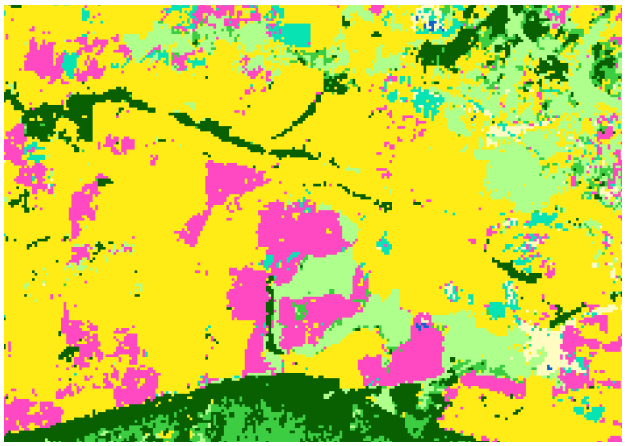
(d)

Fig. 19. Qualitative investigation of land cover maps produced by DANN and *SpADANN* for the transfer task (2018 → 2021): zoom on Area n°1.

(a) Sentinel-2 image acquired on September 29, 2021.



(b) DANN.

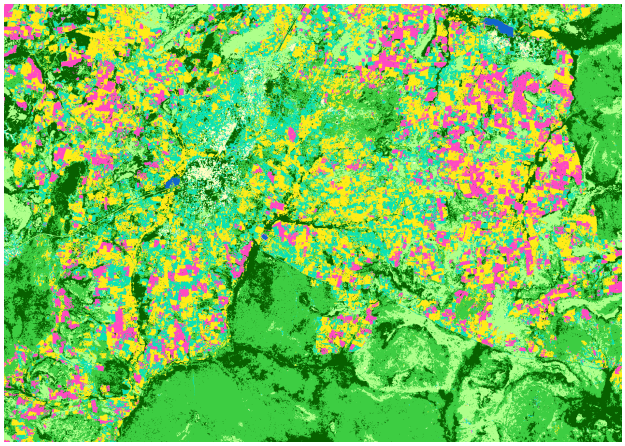


(c) *SpADANN*.

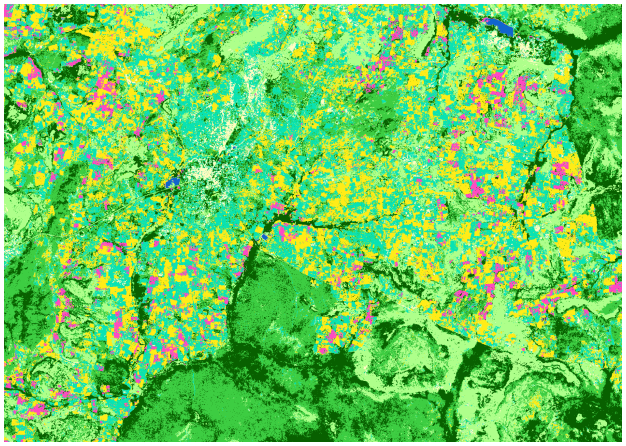
(d)

Fig. 20. Qualitative investigation of land cover maps produced by DANN and *SpADANN* for the transfer task (2018 → 2021): zoom on Area n°2.

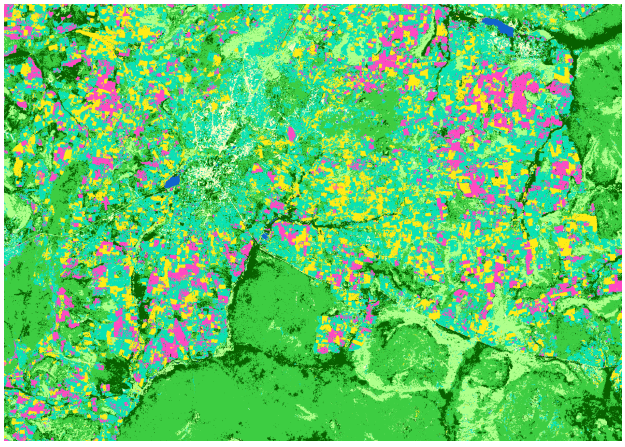
(a) Sentinel-2 image acquired on September 29, 2021.



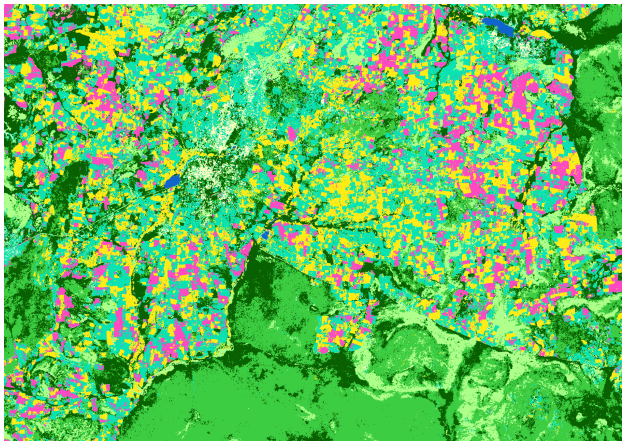
(b) (ONLY $\mathcal{D}_t$) RF.



(c) (ONLY $\mathcal{D}_s$) RF.



(d) DANN.

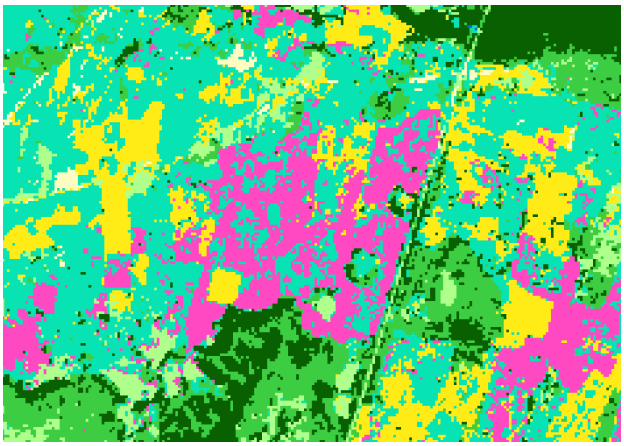


(e) *SpADANN*.

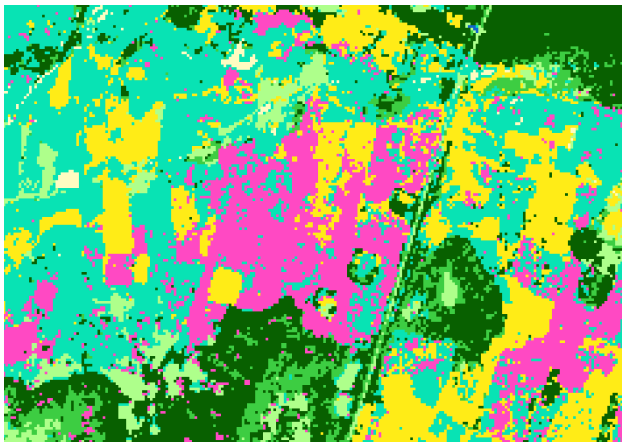
(f)

Fig. 21. Qualitative investigation of land cover maps produced by the RF methods - (ONLY $\mathcal{D}_t$) and (ONLY $\mathcal{D}_s$), DANN and *SpADANN* for the transfer task (2020 $\rightarrow$ 2021): zoom on the Koumbia city.

(a) Sentinel-2 image acquired on September 29, 2021.



(b) DANN.

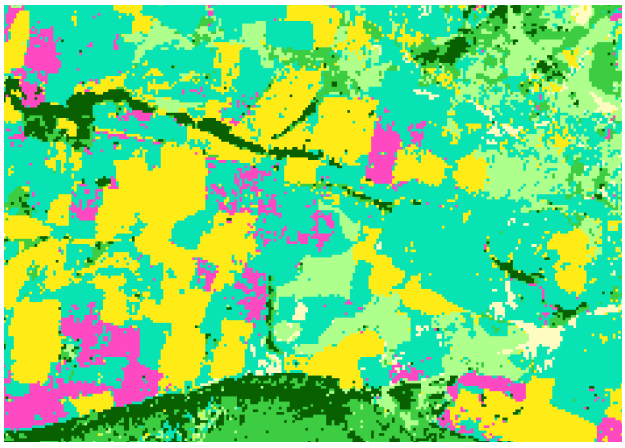


(c) *SpADANN*.

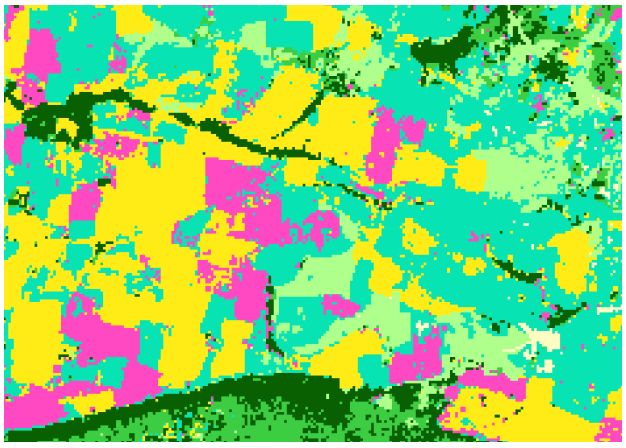
(d)

Fig. 22. Qualitative investigation of land cover maps produced by DANN and *SpADANN* for the transfer task (2020 → 2021): zoom on Area n°1.

(a) Sentinel-2 image acquired on September 29, 2021.



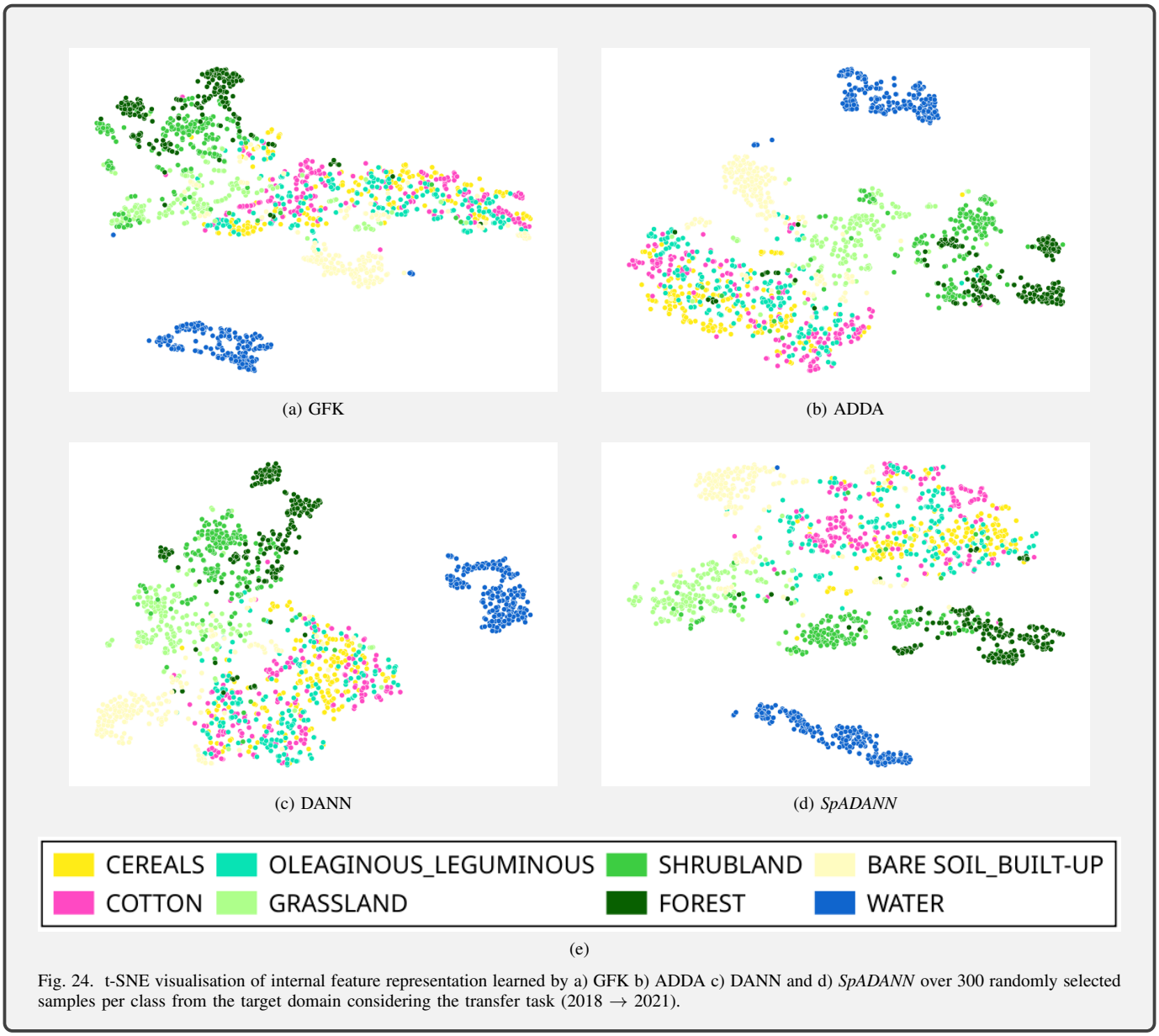
(b) DANN.



(c) *SpADANN*.

(d)

Fig. 23. Qualitative investigation of land cover maps produced by DANN and *SpADANN* for the transfer task (2020 → 2021): zoom on Area n°2.

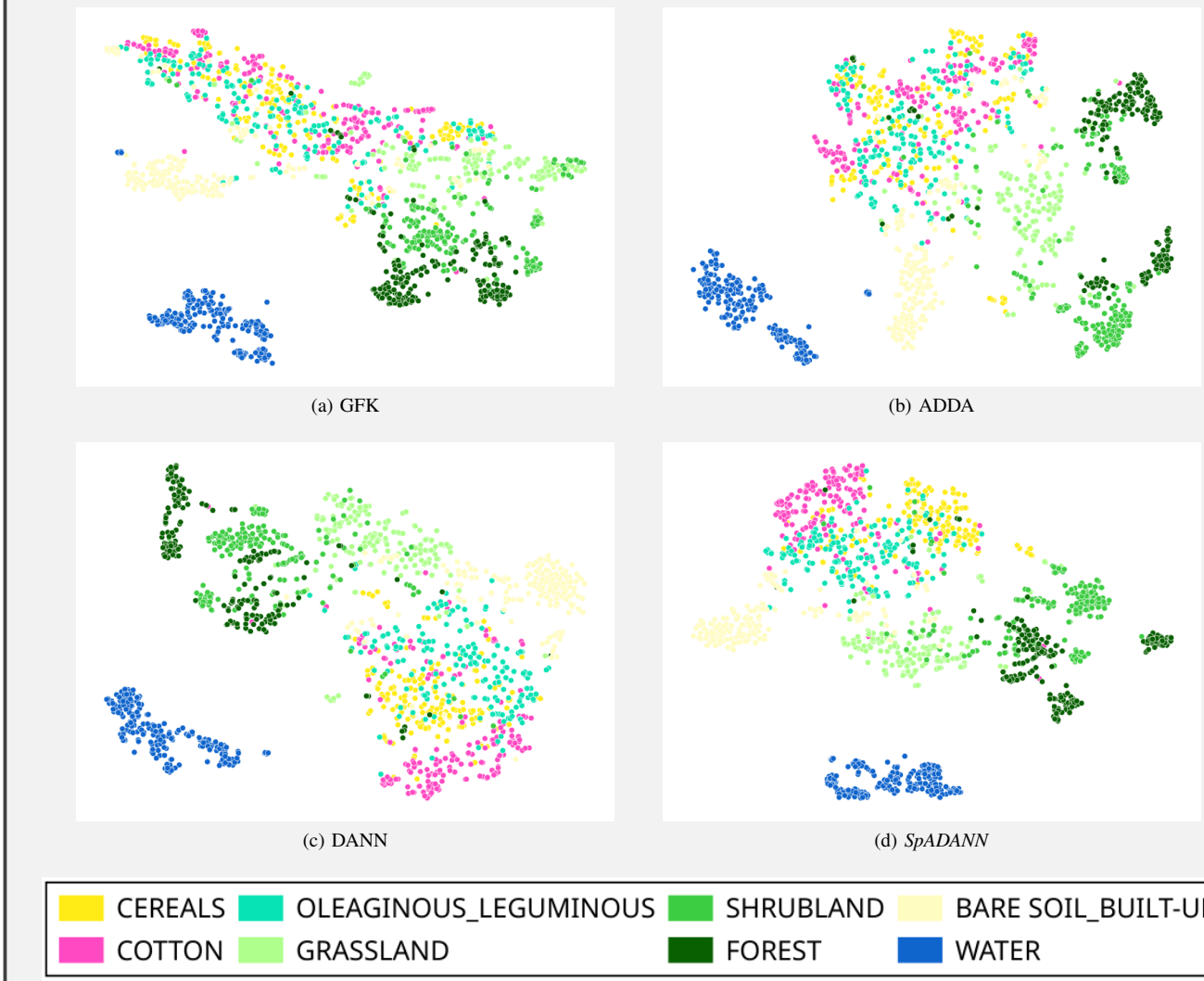


(a) GFK (b) ADDA

(c) DANN (d) *SpADANN*

(e)

Fig. 24. t-SNE visualisation of internal feature representation learned by a) GFK b) ADDA c) DANN and d) *SpADANN* over 300 randomly selected samples per class from the target domain considering the transfer task (2018 $\rightarrow$ 2021).



(a) GFK (b) ADDA

(c) DANN (d) *SpADANN*

(e)

Fig. 25. t-SNE visualisation of internal feature representation learned by a) GFK b) ADDA c) DANN and d) *SpADANN* over 300 randomly selected samples per class from the target domain considering the transfer task (2020 $\rightarrow$ 2021).