Multi-Modal Temporal Domain Adaptation for Land-Cover Prediction

Master Thesis

For obtaining the degree

Master of Science (MSc)

Department of Geoinformatics, Faculty of Digital and Analytical Sciences
Paris-Lodron-University Salzburg
&
Faculté Science & Science de l'Ingénieur
Université Bretagne Sud

Submitted by:

Edgar Joao Manrique Valverde
12120831

Supervisors:
Dino Ienco, PhD
INRAE, UMR TETIS, Montpellier, France
Assoc. Prof. Charlotte Pelletier
Université Bretagne Sud, Vannes, France
Raffaele Gaetano, PhD
CIRAD, UMR TETIS, Montpellier, France
Assoc. Prof. Dirk Tiede
Paris-Lodron-University Salzburg, Salzburg, Austria

June, 2023

Contents

List of Figures

List of Tables

List of Acronyms

ALDA	Adversarial-Learned Loss for Domain Adaptation
CNN	Convolutional Neural Network
DANN	Domain Adversarial Neural Networks
DNN	Deep Neural Networks
DL	Deep Learning
EO	Earth Observation
ESA	European Space Agency
FAO	Food and Agriculture Organization
GRD	Ground Range Detected
GRL	Gradient Reversal Layer
GRU	Gated Recurrent Unit
LCM	Land Cover Maps
LSTM	Long Short Term Memory
MPC	Microsoft Planetary Computer
MS	Multispectral
MSI	Multi-spectral Instrument
PSE-TAE	Pixel Set Encoder Temporal Attention Encoder
RNN	Recurrent Neural Network
RF	Random Forest
RTC	Radiometrically Terrain Corrected
RS	remote sensing
SCL	Scene Classification Layer
SAR	Synthetic Aperture Radar
SITS	Satellite Image Time Series

SpADANN	Spatially-Aligned Domain-Adversarial Neural Network
STAC	SpatioTemporal Asset Catalog
SVM	Support Vector Machines
S1	Sentinel-1
S2	Sentinel-2
UDA	Unsupervised Domain Adaptation
SWIR	Shortwave Infrared
VNIR	Visible/Near Infrared

Abstract

Satellite remote sensing have, and continues to, produce huge amounts of data that are useful to characterize the earth's surface, being the land cover maps one of the most useful products to monitor a variety of applications and services. Traditionally, single-sensor imagery was used to produce such maps, however, more recently the benefits of combining multi-sensor data like optical and radar satellite images have demonstrated to produce more reliable land cover maps. There is an opportunity to use a diverse source of information in order to exploit their complementarity and produce more accurate land cover maps, which includes using Satellite Image Time Series (SITS), multi-sensor information, and the ability to re-use reference data which is expensive to acquire. For the latter point, Unsupervised Domain Adaptation (UDA) have been recently explored in the context of multi-temporal remote sensing analysis to produce land cover maps without reference data from the year in which that map needs to be produced (known as target domain), thus, re-using reference data from a previous year (known as source domain).

In this study, we expanded the Spatially-Aligned Domain-Adversarial Neural Network (SpADANN) framework (Capliez et al., 2023) that combines adversarial learning and self-training to transfer a classification model from a time period to a successive one on a specific study area, to be able to use optical and radar data from Sentinel-2 and Sentinel-1 SITS specifically, thus, dealing with the multimodal temporal transfer scenario in the context of land cover mapping. The results from the experimental assessment on a study area in Burkina Faso, show an increase in the performance of the SpADANN framework by up to 1 point of F1 score compared to using only single-sensor SITS. Additionally, the increase in performance of the multimodal framework if more evident when analyzed with the macro F1 score, suggesting that the multimodality brings added value for the low represented classes. The results contribute to demonstrate the importance of using multimodal information in the framework of UDA, where it is still under-explored.

Keywords: Satellite image time series, Unsupervised domain adaptation, Multimodal, Deep learning, Remote sensing

Acknowledgements

I would like to thank my supervisors for this project, Dino Ienco, PhD; Assoc. Prof. Charlotte Pelletier; Raffaele Gaetano, PhD; and Assoc. Prof. Dirk Tiede.

Dino and Charlotte were critical in leading me patiently in the implementation of the methods and rigorously reviewing the work done in a timely manner. Their advice and feedback throughout the meetings (and outside of meetings) made the completion of the project possible. Raffaele's and Dirk's input were crucial in the interpretation of results and to keep the remote sensing perspective in mind.

Beyond my supervisors, I would also like to thank my family, Maria Valverde Valverde, Miguel Manrique Flores, Paulo Manrique Valverde, and Manuela Herrera Varela, for their continuous encouragement during the MSc programme. To my friends of the CDE programme (students, teachers, coordinators), thank you for all the experiences and your support.

This work was supported by the French National Research Agency under France 2030 program, referred as ANR-16-CONV-0004; and supported by the Erasmus+ Programme of the European Union, through the Erasmus Mundus Joint Master Degree Copernicus Master in Digital Earth.

1 Introduction

Food security is defined as the accessibility to safe and nutritious food. The Food and Agriculture Organization (FAO) report of 2022, shows that the prevalence of severe food insecurity in the world has increased in 4% during the period of 2014 (7.7%) to 2021 (11.7%), being Africa and Latin America and the Caribbean the regions with the most increase (16.7% to 23.4% and 7.5% to 12.8% respectively) (FAO et al., 2022). Rainfall and temperature play a decisive role in Africa's food security, and the context of climate change, food security is threatened by extreme weather events that are forecast to occur more frequently, posing a particular challenge for countries that rely on rain-fed agriculture for their supply (Pickson & Boateng, 2022).

The possibility to obtain timely and up-to-date Land Cover Maps (LCM) and more specifically crop maps are of paramount importance to develop economically and ecologically sustainable agriculture (Wardlow & Egbert, 2008; You et al., 2014). These maps should include crop type information across multiple years and regions, enabling the mapping of crop sequences as an indicator of agricultural land use intensity. The duration and diversity of crop sequences directly impact landscape complexity (Tscharntke et al., 2021).

The Sentinel-1 (S1) and Sentinel-2 (S2) missions are of particular interest, since they provide publicly available radar and optical/Multispectral (MS) satellite imagery with a high revisiting time and spatial resolution (∼10m pixel size). Multitemporal images can be organized as Satellite Image Time Series (SITS), which comprises images of the same area acquired at different dates, and are of particular interest in crop studies and land cover mapping. They enable opportunities for studying the seasonality or evolution of objects through time and aid to their discrimination. The single use of optical or radar has been successful to produce field scale crop maps at national and continental scales (Defourny et al., 2019; d'Andrimont et al., 2021). For example, Defourny et al. (2019) developed Sen2-Agri, an operational system for generating crop type maps and vegetation status from Landsat 8 and Sentinel-2 time series. In their study, they generated crop type maps for five major crop types in three countries (Ukraine, Mali, and South Africa) at a spatial resolution of 10m, and found that accuracies improved with clear-sky observations during the growing season. These types of large scale systems rely of traditional machine learning algorithms (e.g. random forest, support vector machines) due to the effective performance they provide (Pelletier et al., 2016; Son et al., 2018). More recently, the use of deep learning approaches that are able to explicitly exploit the temporal features of SITS are gaining more attention (Rußwurm & Körner, 2018; Ndikumana et al., 2018; Pelletier et al., 2019).

Including Synthetic Aperture Radar (SAR) data in classification models may improve crop mapping accuracies because of the increased cloud cover independent data availability and the physical and structural properties of the SAR signal over plant canopies which complements optical information from MS sensors (Blickensdörfer et al., 2022). In recent years, Sentinel-1 (S1) and Sentinel-2 (S2) have been successfully exploited together for LCM, performing better than the use of a single sensor and thus, demonstrating how the availability and usage of both sensors is beneficial (Ienco et al., 2019; Blickensdörfer et al., 2022; J. Li et al., 2022). Recently, the European Space Agency (ESA) has moved the WorldCereal[1] project to the operational phase. The project aims to demonstrate the feasibility of global crop mapping at field scale using open Earth Observation (EO) datasets, such as Sentinel-1, Sentinel-2, Sentinel-3, and Landsat 8. The project will exploit the complementarity between optical and radar time series, with radar data providing structural information and optical data sensitive to biophysical parameters.

Multi-modal image fusion in remote sensing, and specifically for SITS, have shown that most fusion schemes outperform single-sensor models. Although, among the fusion schemes, they have advantages and drawbacks at specific settings (Ofori-Ampofo et al., 2021; Sainte Fare Garnot et al., 2022). For example, Ofori-Ampofo et al. (2021) explored three types of fusion for Sentinel-2 (S2) and Sentinel-1 (S1) SITS classification, finding that layer-level fusion overall performed the best, but decision-level fusion performed better in dominant classes while layer-level fusion gained the advantage of performance in minority classes. Additionally, Sainte Fare Garnot et al. (2022) explored four types of fusion for Sentinel-2 (S2) and Sentinel-1 (S1) SITS on three classification tasks, finding that for all tasks using both modalities improved the overall performance, however, the late fusion scheme outperformed the others on parcel-based classification. These findings highlight the difficulty in choosing a fusion approach, as the specific settings related to the reference data (e.g. class imbalance) or classification task (e.g. parcel-based classification, semantic segmentation, etc) will require a specific fusion scheme to achieve the best performance.

Although most research and development of methods to improve LCM have increased significantly due to the large availability of multi-sensor SITS datasets, the majority of these methods still belong to the supervised classification setting (Hong et al., 2021; J. Li et al., 2022). The availability and quality of reference data needed for creating these maps are rarely sufficient for large areas and/or over long timespans (Capliez et al., 2023), and the need for reference data is required and exacerbated for deep learning. In standard supervised classification, the reference data are only valid for the period (reference year, for instance) and geographical area corresponding to their acquisition.

[1] `https://esa-worldcereal.org/en`

Training a classifier with imagery from a different period than the reference data can produce poor LCM due to differences in weather conditions, cloud cover, and other factors between the two time periods, leading to inaccurate classifications and limiting their usage for future LCM in the following years or in other areas (Tardy et al., 2017).

These differences between time periods and regions due to environmental conditions result in a shift in the probability distribution of the source domain and target domain (e.g. different time or region) (Capliez et al., 2023; Nyborg et al., 2022). Here, we will focus on the scenario where the shift is produced by a change in the time period (i.e year) the information comes from, on the same region, and same sensor information (multi-sensor SITS). One approach to addressing this issue is to reuse models trained on past years when reference data was available. This is particularly important because obtaining updated reference data can be costly. By leveraging previous efforts, the need for fresh reference data, which may be difficult to collect, can be reduced. In addition, transfer learning strategies become crucial to achieving this goal.

In the field of computer vision, the Unsupervised Domain Adaptation (UDA) approach provides methods and strategies to cope with distribution shifts, for a model trained in a labelled source domain and transferred to an unlabelled target domain (Wilson & Cook, 2020). If this approach is successful, we can train a classification model that can provide a reliable LCM using a SITS on a given year for which no specific reference data is provided (target domain), as well as both sparsely annotated reference data and SITS from a previous year (source domain). However, still many approaches deal with UDA in the context of optical/MS image analysis, but only a few deal with SITS and multi-sensor SITS. In the context of temporal UDA, a new framework that combines UDA and self-training approaches has been proposed by Capliez et al. (2023), where pseudo-labels are selected from the target domain (on which three deterministic conditions have to be satisfied) to improve the domain-invariant representation of the features learned by the model.

Unsupervised Domain Adaptation (UDA) is a challenging task but has shown promising results in remote sensing to produce land cover maps in an unlabelled target domain. However, situations like reference data scarcity and class imbalance in the source domain and, changes in the class distribution and in general land cover change towards the target domain, make the task even more challenging. Here we hypothesize that having more than one modality of SITS could help to provide a more robust feature representation, thus, the study is set up in the multi-sensor (Sentinel-2 and Sentinel-1) temporal UDA problem, where both source and target domain are multi-sensor. The goal is to train a multi-sensor land cover classifier with labelled samples from the source domain and make inference on the unlabelled target domain. Building on previous research, the objective of this study is to evaluate the potential of combining dense S2 and S1 time series for improving crop type mapping in supervised and UDA settings. More specifically, this study addresses the following research questions:

1. How does the use of single-sensor SITS and multi-sensor SITS (the fusion of Sentinel-2 (S2) (optical) and Sentinel-1 (S1) (radar)) affect the supervised classification model performance?

2. Does the performance based on optical-radar layer fusion differs from decision fusion?

3. Hoes does the use of single-sensor SITS and multi-sensor SITS (the fusion of Sentinel-2 (S2) (optical) and Sentinel-1 (S1) (radar)) helps to improve the classification model performance in a Unsupervised Domain Adaptation (UDA) setting?

4. How does the multi-sensor SITS approach can be implemented into the framework of Unsupervised Domain Adaptation (UDA)?

Experimental evaluation is carried out on a rural study site located in Burkina Faso, referred as Koumbia site and characterized by a mostly agricultural land cover types (crop types as well as natural and built-up classes). We consider multi-sensor (Sentinel-2 (S2) and Sentinel-1 (S1)) data coming from three different years (2018, 2020 and 2021) and perform a transfer assessment to each pair of years.

The rest of the manuscript is organized as follows: Section 2 presents the literature on multi-sensor land cover classification and unsupervised domain adaptation. Section 3 describes the study site and associated reference and multi-sensor SITS data. Section 4 describes the architecture of the models used for the experiments as well as the experimental settings and Section 5 presents the results of the experiments. Finally, Section 6 will provide the conclusions.

2 Theoretical background

Satellite Image Time Series (SITS) comprises images of the same area acquired at different dates by the same sensor (or constellation of sensors). The latest satellite constellations are capable of acquiring SITS with high spectral, spatial and temporal resolutions (Drusch et al., 2012). For instance, the two Sentinel-2 satellites provide, since March 2017, worldwide images every five days (at the equator), freely distributed, within 13 spectral bands at spatial resolutions varying from 10 to 60 m. The Landsat programme has been taking images since July 1972, currently operating with Landsat 7, 8, and 9 satellites. And the two Sentinel-1 satellites, since April 2015, provide data from a dual-polarized C-band SAR enabling imaging through clouds every 6 days, although currently Sentinel-1B is no longer operational, increasing the revisiting time to 12 days[2].

In this section, we will present advances in processing SITS for land cover mapping using deep learning methods in three settings: Supervised classification, multimodal data fusion, and Unsupervised Domain Adaptation.

2.1 Supervised classification of satellite image time series for land cover mapping

A supervised classification task of SITS is defined by set of training samples (X, Y), such as $(X, Y) = \{(\boldsymbol{x_1}, y_1), ..., (\boldsymbol{x_n}, y_n)\}$ where n is the number of instances or samples. The pair $(\boldsymbol{x_i}, y_i)$ represents the training sample i where $\boldsymbol{x_i}$ is a D-variate time series pixel of length T associated with the label $y_i \in Y = \{1, ..., K\}$ for K classes. More in detail, $\boldsymbol{x_i}$ can be expressed by $\boldsymbol{x_i} = (\boldsymbol{x_i(1)}, ..., \boldsymbol{x_i(T)})$, where $x_i(t) = (x_i^1(t), ..., x_i^D(t))$ for a timestamp t. The goal is to build a classifier h parameterized by Θ such that $h_\Theta : X \rightarrow Y$. In supervised classification, the model is meant to be applied on test data that is drawn from the same distribution of the training samples it was trained on.

The current state-of-the-art algorithms used for producing maps are classical machine learning methods (i.e. Support Vector Machines (SVM) and Random Forest (RF)) (Khatami et al., 2016; Phiri et al., 2020). These algorithms are typically applied at the pixel-level on the stack of multi-spectral images found in the SITS. However, these algorithms are not aware of the temporal dimension that structures SITS. This means that the temporal order of the images has no impact on the results (Pelletier et al., 2019). Surprisingly, even if the images were rearranged in the series, the model and accuracy would remain the same. As a result, there is a loss of temporal behavior for classes

[2] https://sentinels.copernicus.eu/web/sentinel/-/end-of-mission-of-the-copernicus-sentinel-1b-satellite/

that evolve over time, such as various forms of vegetation that are subject to seasonal changes.

To overcome this issue, researchers have explored the use of Deep Neural Networks (DNN) that are able to learn feature representations that capture time dependencies. Deep learning has been applied using Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN) for handling the temporal dimension (Ndikumana et al., 2018; Pelletier et al., 2019; Zhong et al., 2019). Ndikumana et al. (2018) assessed the usefulness of Sentinel-1 (S1) SITS for crop mapping comparing the performance of two RNN approaches using Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU), against classical machine learning approaches. Their results show that S1 SITS provided good performances in all models, although both versions of RNN outperforms k-nearest neighbor, random forest and support vector machines. Pelletier et al. (2019) propose the use of one dimensional (temporal) CNN for SITS based LCM, 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. Similarly, Zhong et al. (2019) evaluated two models, one based on LSTM and the other on 1D-CNN. Both Pelletier et al. (2019) and Zhong et al. (2019) show that temporal convolutions outperform RNN for SITS classification. Garnot, Landrieu, Giordano, and Chehata (2019) propose a modified self-attention-based mechanism architecture named Pixel Set Encoder Temporal Attention Encoder (PSE-TAE) developed for crop mapping. Their results show that PSE-TAE is able to extract more expressive features exploiting both the spatial and temporal dimensions than CNN and GRU, resulting in better performance. Other approaches have also explored the use of transformers (Yuan et al., 2022), and 3D-CNN (R. Li et al., 2022). Rußwurm, Pelletier, Zollner, Lefèvre, and Körner (2020) presented a dataset for SITS classification and made a comparison of different approaches, finding that the attention-based transformer model outperformed the recurrent models (i.e LSTM and RNN based models) and the CNN based models (TempCNN and InceptionTime).

2.2 Multimodal SITS data fusion

In a broad sense, the term multimodal data fusion in remote sensing (RS) as defined by J. Li et al. (2022), includes: multisource data fusion, which comprises the technical specifications of the sensor and determines the internal characteristics of the product (e.g. imaging mechanism and the resolutions); and multitemporal data fusion, which is defined by the conditions of the acquisition determined by external properties (e.g. acquisition time or observation angle). In this study, the term multimodal comprises multisource (a.k.a multisensor) data fusion of optical (Sentinel-2) and SAR (Sentinel-1) SITS for land cover classification.

Although a large amount of multimodal RS data has become readily available, each modality can only capture few specific properties and hence cannot fully describe the observed scenes, which poses a great constraint on subsequent applications (J. Li et al., 2022). Naturally, multimodal RS data fusion is a feasible way to break out of the dilemma induce by unimodal data. By integrating the complementary information extracted from multimodal data, a more robust and reliable decision can be made in many tasks, including, as is our interest, land cover classification. More specifically, including both optical and SAR data in classification models can enrich the information for discriminating targets and improve performances due to contributing information on reflectance (optical) and the physical and structural properties (SAR), additionally, the cloud cover independent characteristic of SAR helps to have a better representation of the changes over time compared to using only optical data (Ofori-Ampofo et al., 2021).

Multi-modal image fusion can be grouped into input (a.k.a pixel), layer (a.k.a feature), and decision levels (Zhou et al., 2019). In remote sensing, the input fusion combines image bands from multi-sensor data, usually through resampling and concatenation, or image to image co-registration. Layer fusion requires the extraction and concatenation of high-level features allowing the model to learn a feature representation for each modality. Lastly, in decision fusion each modality is independently processed by a model to generate class confidence scores (in the classification task), which are combined statistically (e.g. averaged) to yield a final fused decision. Figure 1 depicts the three data fusion scenarios in a general Deep Learning (DL) framework.

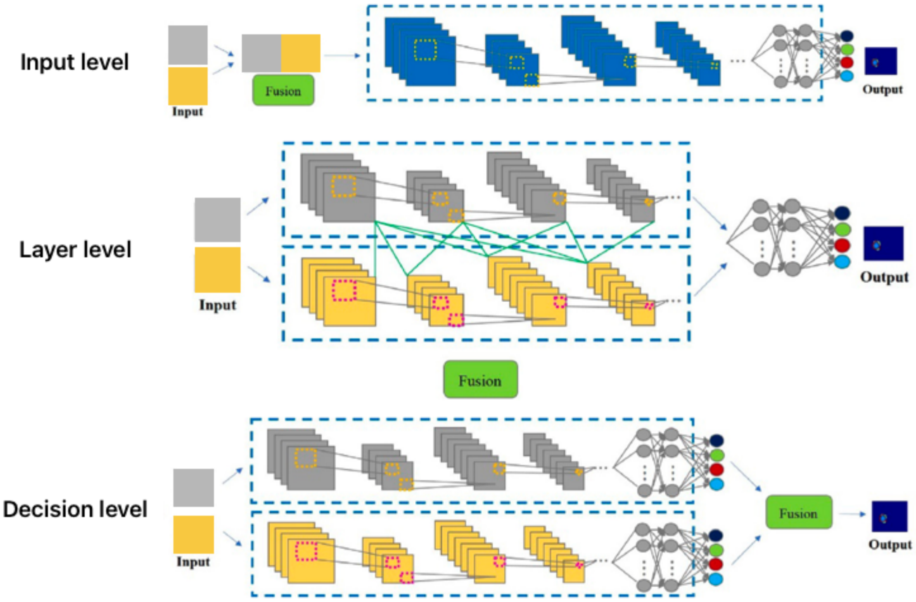


Figure 1: Levels of multimodal data fusion approaches in deep learning (Zhou et al., 2019)

Although the use of Sentinel-1 (S1) and Sentinel-2 (S2) have increased the performance of classification models, most of the studies use traditional machine learning approaches (e.g. random forest) and very few studies use deep learning approaches to leverage the use of these datasets. Ienco et al. (2019) combined Sentinel-1 (S1) and Sentinel-2 (S2) SITS processing the information on independent encoders, where each encoder has also two branches that processes the SITS with a CNN and attentive GRU to extract spatial and temporal features separately. The feature representation that the network was able to learn allowed to have better performances compared to other competing methods and input level fusion using random forest. Ofori-Ampofo et al. (2021) evaluated the use of Sentinel-2 and Sentinel-1 SITS from crop type mapping using PSE-TAE, and assessed different model configurations to perform data fusion at the three levels (input, layer, and decision). All approaches were able to perform better that a single modality, however, input and layer fusion performed better than decision fusion for underrepresented classes. Similarly, Sainte Fare Garnot et al. (2022) explored four types of fusion for Sentinel-2 and Sentinel-1 SITS, but on three classification tasks (parcel classification, pixel-based segmentation, and panoptic parcel segmentation), finding that the multimodal approach based on attention-based models can outperform single modality models. A different approach is done by Gbodjo, Montet, Ienco, Gaetano, and Dupuy (2021), where they combine Sentinel-2 and Sentinel-1 SITS and SPOT VHR image, for land cover mapping through a three-branch patch-based CNN model, using different per-source encodersto deal with the specificity of the input signals.

2.3 Unsupervised domain adaptation

A bottleneck of supervised classification in remote sensing is the need of training the model on reference data that are specific to every image acquisition (Tuia et al., 2016), posing serious challenges to their use in situations characterized by a reduced amount of (or unavailable) reference data (Capliez et al., 2023). Applying a model trained on an image with it's corresponding reference data to a new image, often provides poor results. This is because of differences in the spectra observed in the new image, even though representing the same types of objects. The differences can be related to a series of deformations, or shifts, related to a variety of effects such as a biased sampling in the spatial domain (typically if the ground sampling has been focused on a region non-representative of the new scene), changes in the acquisition conditions like weather (Tardy et al., 2017).

In the general field of computer vision, a domain consists of a feature space and a marginal probability distribution (i.e. the features of the data and the probability distribution of those features in the dataset). By this definition, a change in domain may result from either a change in feature space or a change in the marginal probability distribution (Wilson & Cook, 2020). In remote sensing, a change in the feature space might be due to a change in climate, weather, or environmental change, and a change in the marginal probability distribution might be due to a change in the geographical region. Figure 2 presents some scenarios where there is a change in domain in remote sensing imagery and there is a need to adapt a model to new images, on the left there are changes in the spatial domain, on the middle changes in image acquisitions and possible sensor, and on the right includes changes in the spatial and temporal domains.

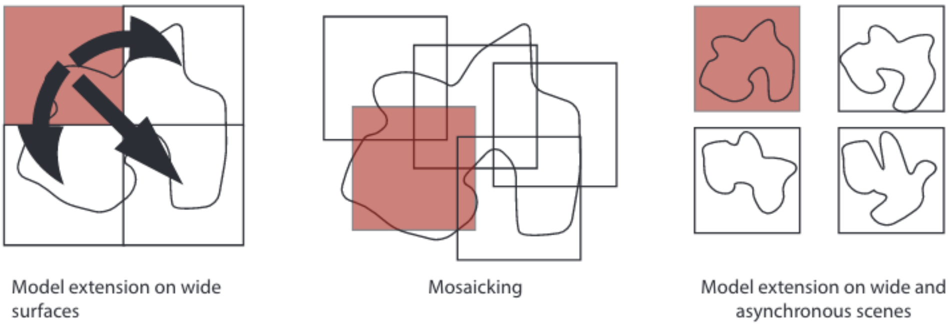

Model extension on wide surfaces
Mosaicking
Model extension on wide and asynchronous scenes

Figure 2: Domain Adaptation scenarios in remote sensing
(Tuia et al., 2016)

Learning a discriminative classifier or other predictor in the presence of a shift between training and test distributions is known as Domain Adaptation (DA) (Liu et al., 2022). Unsupervised domain adaption (UDA) methods have the main objective to transfer a model trained on a labelled source domain to an unlabelled target domain. We consider classification tasks where X is the input space and $Y = \{1,2,...,K\}$ is the set of K possible labels. Moreover, we have two different distributions over $X \times Y$, called the source domain $\boldsymbol{D}^s$ and the target domain $\boldsymbol{D}^t$. An unsupervised domain adaptation learning algorithm is then provided with a labeled source sample X^s drawn i.i.d. from $\boldsymbol{D}^s$, and an unlabeled target sample X^t drawn i.i.d. from $\boldsymbol{D}^t$, providing only labeled samples on the source domain Y^s, thus, $\boldsymbol{D}^s = \{X^s, Y^s\}$ and $\boldsymbol{D}^t = \{X^t\}$.

$$X^s = \{(\boldsymbol{x}_i^s, y_i^s)\}_{i=1}^n \sim (D^s)^n; \quad X^t = \{\boldsymbol{x}_i^t\}_{i=n+1}^N \sim (D^t)^{n'},$$

with $N = n + n'$ being the total number of samples. The goal of the learning algorithm is to build a classifier $h_\Theta : X \to Y$ with a low target risk while having no information about the labels of $\boldsymbol{D}^t$. Adapting a model trained on an image (or set of images, e.g. time series) to another can be performed in different ways. In remote sensing, two approaches have shown promising results in the last years, domain invariant (e.g. adversarial learning) (Ganin et al., 2016) and self-training (semi-supervised learning) (Chapelle et al., 2009) methods.

The goal of these methods is to generate similar feature distributions for the source and target data, known as domain-invariant features. In this regard, the adversarial learning setting uses a domain discriminator to minimize the divergence between the features of the source and target domains, in other words, it aims to generate the desired domain-invariant features. A feature extractor $G_f(\cdot) \in \mathbb{R}$ is applied onto x_i to extract a feature representation $G_f(\boldsymbol{x}_i) \in \mathbb{R}^k$, minimizing $d[G_f(\boldsymbol{x}_i^s), G_f(\boldsymbol{x}_i^t)]$, where d is the divergence. In addition to training a label predictor $G_y(\cdot)$ on source data, $G_f(\cdot)$ is also optimized to generate similar feature distributions for the source and target data, following the supervision signal of a domain discriminator $G_d(\cdot) : \mathbb{R}^k \rightarrow (0, 1)$. The remote sensing field has intensively investigated one of these techniques called Domain Adversarial Neural Networks (DANN) (Ganin et al., 2016), where the domain classifier is associated with a Gradient Reversal Layer (GRL) that enforces the features extracted by the encoder to be invariant to the domains. In the context of SITS, Wang, Zhang, He, and Zhang (2021) proposed the Phenology Alingment Network, a cross-region UDA method for SITS, that learns domain invariant features by minimizing the Maximum Mean Discrepancy loss. For the temporal UDA scenario, Tardy, Inglada, and Michel (2019) uses Optimal Transport for the case of temporal unsupervised domain adaptation from multiple source domain (multiple annual SITS) to a specific target domain (annual SITS). Recently, Capliez et al. (2023) leveraged the specificity of the temporal UDA scenario conceiving a process based on the spatial consistency between the two SITS pixels allowing them to use both adversarial learning and self-training. 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.

Inspired by the semi-supervised learning strategies, in the self-training setting (Sohn et al., 2020), 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. By learning from both labeled source data and pseudo-labeled target data, self-training methods implicitly encourage feature alignment for each class without restricting the model to operate on domain-invariant features (Zou et al., 2020; Morerio et al., 2020). However, the domain shift often results in increased pseudo-label noise and there has been approaches deal with this issue e.g. co-training (Chen et al., 2011), tri-training (Saito et al., 2017), conditional generative models (Morerio et al., 2020), confidence regularization (Zou et al., 2020), or Adversarial-Learned Loss for Domain Adaptation (ALDA) which uses a noise correction domain discriminator (Chen et al., 2020). Recently, Nyborg et al. (2022) proposed a method called TimeMatch where crop classification models are adapted to an unlabeled target region by self-training on temporally shifted SITS in a cross-region UDA scenario. The model explicitly captures the underlying temporal discrepancy of the data by estimating the temporal shift between two regions by generating pseudo-labels using the estimated temporal shift from target to source.

3 Data

In this section, we will describe the study area and the associated reference data, as well as the procedure to query, download, and preprocess the Sentinel-2 and Sentinel-1 SITS.

3.1 Study area

The study area is situated in the sub-humid sudanian zone, around the town of Koumbia in the southwest of Burkina Faso. It covers an area of about 2338 km^2, with forests and natural Savannah covering most of the surface, interspersed with about 35 % of land used for rain-fed agricultural production, mainly smallholder farming. The main crops cultivated in this area are cotton and cereals such as maize, sorghum, and millet, with leguminous and oleaginous crops also being grown. Figure 3 shows the study site with the locations of the reference data.

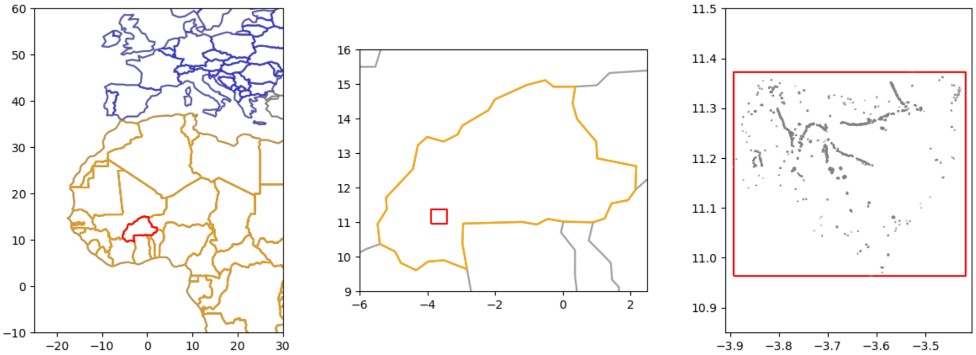


Figure 3: Koumbia Study area

3.2 Data availability

3.2.1 Reference data

Reference data for 2018, 2020, and 2021 was obtained from a large agricultural land cover dataset available online (Jolivot et al., 2021). Field surveys were conducted yearly around the growing peak of the cropping season from 2013 to 2021, although during 2019 no fieldwork was conducted. GPS way-points were gathered following an opportunistic sampling approach along the roads or tracks according to their accessibility, while ensuring representation of the existing cropping practices in place.

As previously described in (Capliez et al., 2023), records were provided on different types of non-crop classes (e.g. natural vegetation, settlement areas, water bodies) to differentiate crop and non-crop classes. Additional non-crop reference polygons were obtained by photo-interpretation of very high-resolution optical satellite images. The final reference data was assembled in a GIS vector file, containing a collection of polygons, each attributed with a land cover category. To ensure consistency, the same surface was kept for the three reference years by performing a year by year intersection of the polygons of the original database. The polygons were rasterized with a pixel sized of 10 m and aligned with the satellite imagery described in section 3.2.2. A summary of the reference data can be found in Table 1.

The changes occurring in the reference data from one year to another, which mainly occur between crop classes (Figure 4), were measured. For the year 2018, the surface of cotton was about two times that of oleaginous/leguminous when this ratio is balanced for years 2020 and 2021, however, as can be seen in Figure 5, this does not mean that there were not many changes in those two classes in the years 2020 and 2021, in fact, only a small fraction of the cotton crops from 2020 remained in 2021 (22%), while most of the cotton from 2020 was turned into cereals in 2021 (65%) and 38% of cereals from 2020 were turned into cotton in 2021. Figure 5 quantifies the changes in terms of land cover classes between each couple of reference years. Bare soil/built-up and water classes remained mostly unchanged, and few changes occurred on non-crop classes, mainly due to occasional shifts in the density of natural vegetation or conversion to active cropland.

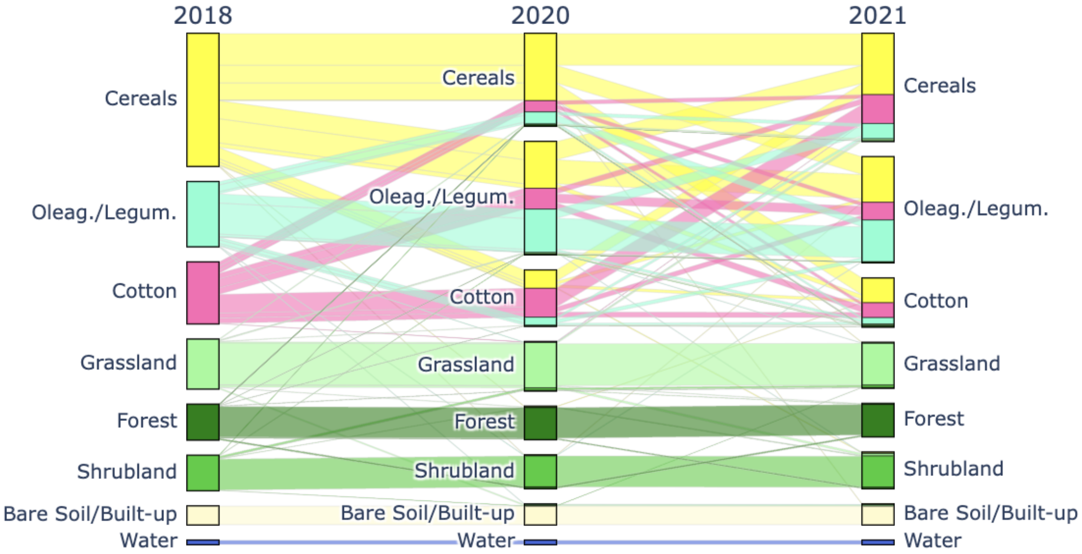


Figure 4: Land Cover Change at the polygon level

3.2.2 Satellite imagery

Satellite Image Time Series (SITS) of Sentinel-1 RTC and Sentinel-2 L2A images were accessed from Microsoft Planetary Computer (MPC)[3]. Briefly, MPC is a platform for environmental sustainability that provides access to petabyte-scale geospatial data and

[3] `https://planetarycomputer.microsoft.com`

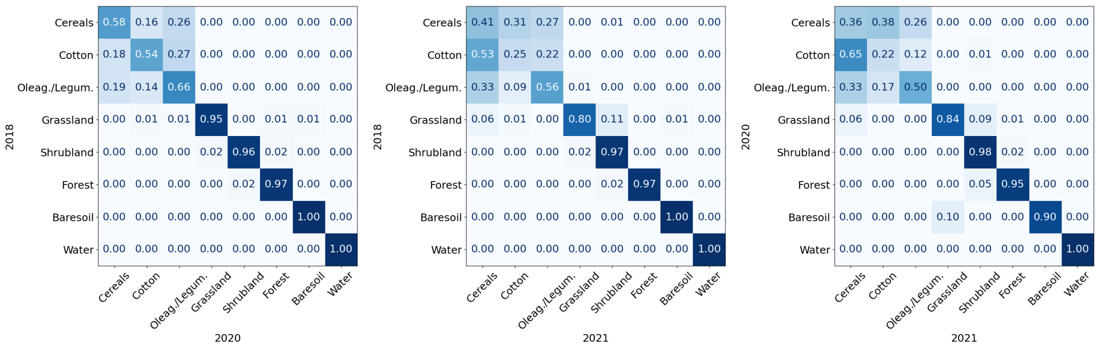


Figure 5: Land Cover Change at the pixel level

Table 1: Reference data summary over the years 2018, 2020, and 2021

Class name	Class ID	2018	2020	2021
# Polygons (% Polygons)				
Cereals 🟨	1	330 (33.07)	230 (23.05)	268 (26.85)
Cotton 🟪	2	153 (15.33)	139 (13.93)	121 (12.12)
Oleag./Legum. 🟦	3	161 (16.13)	281 (28.16)	263 (26.35)
Grassland 🟩	4	123 (12.32)	122 (12.22)	113 (11.32)
Shrubland 🟩	5	87 (8.72)	83 (8.32)	90 (9.02)
Forest 🟩	6	88 (8.82)	82 (8.22)	82 (8.22)
Baresoil ⬜	7	46 (4.61)	51 (5.11)	51 (5.11)
Water 🟦	8	10 (1.00)	10 (1.00)	10 (1.00)
Total			998	
# Pixels (% Pixels)				
Cereals 🟨	1	13056 (16.33)	9731 (12.17)	11435 (14.30)
Cotton 🟪	2	7672 (9.59)	6971 (8.72)	6575 (8.22)
Oleag./Legum. 🟦	3	3595 (4.50)	7950 (9.94)	7316 (9.15)
Grassland 🟩	4	13108 (16.39)	12998 (16.26)	11100 (13.88)
Shrubland 🟩	5	23122 (28.92)	22547 (28.20)	24325 (30.42)
Forest 🟩	6	17369 (21.72)	17435 (21.80)	16984 (21.24)
Baresoil ⬜	7	835 (1.04)	1125 (1.41)	1022 (1.28)
Water 🟦	8	1205 (1.51)	1205 (1.51)	1205 (1.51)
Total			79962	

machine learning tools. It is a cloud-based service that enables researchers, policy-makers, and other stakeholders to analyze and model environmental data at a global scale. MPC uses the SpatioTemporal Asset Catalog (STAC)[4] API to query the satellite imagery, making it easier to filter and search imagery based on spatial and temporal criteria.

Sentinel-2 (S2) is a satellite mission launched by the European Space Agency (ESA) as part of the Copernicus program. It consists of two twin satellites in the same orbit, but phased 180º, with a repetition frequency of 5 days or less. Each satellite carries a Multi-spectral Instrument (MSI) with 13 spectral channels (bands) in the Visible/Near Infrared (VNIR) and Shortwave Infrared (SWIR) part of the electromagnetic spectrum. The MSI acquires images at 3 spatial resolutions: 10 m, 20 m, and 60 m. The radiometric and spectral resolutions of each band are listed in Table 2.

Table 2: Sentinel-2 bands and their spatial and radiometric resolutions

Band	**Resolution (m)**	**Central wavelength (nm)**	**Bandwidth (nm)**	**Description**
B1*	60	443	21	Ultra blue (Coastal and aerosol)
B2	10	490	66	Blue
B3	10	560	36	Green
B4	10	665	31	Red
B5	20	705	15	Visible and near infrared (VNIR)
B6	20	740	15	Visible and near infrared (VNIR)
B7	20	783	20	Visible and near infrared (VNIR)
B8	10	842	106	Visible and near infrared (VNIR)
B8A	20	865	21	Visible and near infrared (VNIR)
B9*	60	940	20	Short wave infrared (SWIR)
B10*	60	1375	31	Short wave infrared (SWIR)
B11	20	1610	91	Short wave infrared (SWIR)
B12	20	2190	175	Short wave infrared (SWIR)

* 60 m bands were excluded from the study

Sentinel-2 (S2) L2A imagery has a 5 day revisit time in the study area, making a total of 73 time-steps per year (Figure 6). Imagery was queried for the bounding box of the study area and the three years corresponding to the study period (2018, 2020, 2021). In addition, the 60 m resolution bands were discarded and 20 m bands were re-sampled to 10 m using nearest neighbor interpolation. No cloud cover filter was used at this stage, collecting all the imagery available, however, the Scene Classification Layer (SCL), provided by ESA in the S2 L2A product, was used to mask pixels corresponding to the following categories: no data pixels, saturated, cloud shadows, unclassified pixels,

[4] https://stacspec.org/en

medium probability clouds, high probability clouds, and cirrus clouds. Linear interpolation was used to impute masked pixels and missing acquisitions. Thus, harmonizing the SITS to have the 73 time-steps on each of the three years.

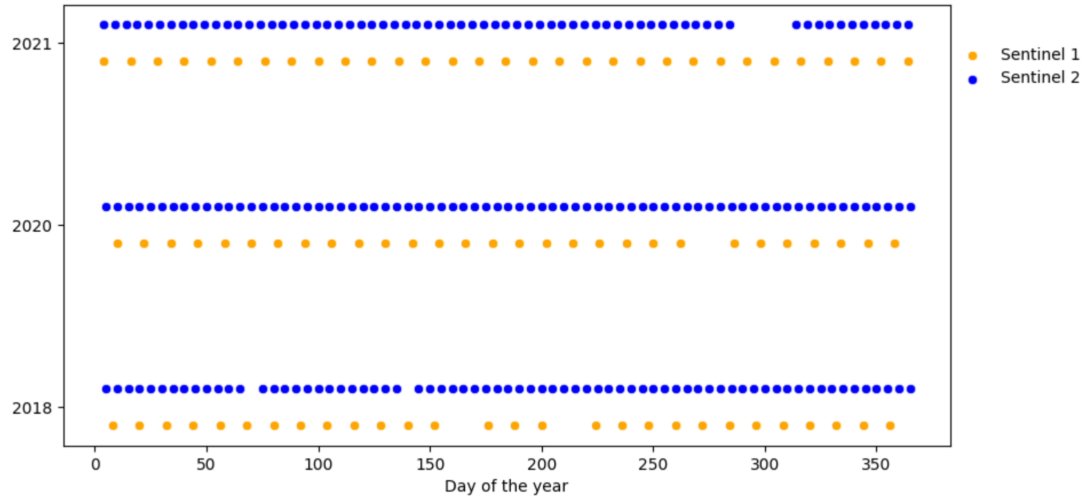


Figure 6: Sentinel-1 & 2 time series acquisitions and time gaps. Time gaps in both Sentinel-2 and Sentinel-1 relate to no acquisition being available in the L2A and RTC products respectively.

The Sentinel-1 mission is a constellation of two polar-orbiting satellites, operating day and night performing C-band synthetic aperture radar imaging. MPC provides Sentinel-1 Radiometrically Terrain Corrected (RTC) data derived from the Ground Range Detected (GRD) Level-1 products produced by ESA and provides the images at 10m pixel size. Radiometric Terrain Correction RTC accounts for terrain variations that affect both the position of a given point on the Earth's surface and the brightness of the radar return, as expressed in radar geometry. Without treatment, the hill-slope modulations of the radiometry threaten to overwhelm weaker thematic land cover-induced backscatter differences.

Sentinel-1 RTC imagery has a 6 day revisit time, however, due to the lost of Sentinel 1-B on 2021, we only included data for Sentinel 1-A, having a 12 day revisit time and making a total of 31 time-steps per year (Figure 6) for the ascending pass. The dataset contains VH and VV polarizations. Similar to the S2 SITS, linear interpolation was used to impute missing acquisitions. Thus, harmonizing the SITS to have the 31 time-steps on the three years.

As suggested by Pelletier et al. (2019), to preserve the overall shape of the time series, feature normalization was performed with a mix-max normalization per feature across all images, using the 2% (98 %) percentile instead of the minimum (maximum) value. This helps to overcome the sensitivity of this method to extreme values. Finally, pixel extraction of the SITS were done using the pixels of the reference data, for S2 the dimensions of the final pixel time series are: 79962 pixels, 10 spectral bands, and 73 time-steps. While for S1, we considered a 9x9 neighborhood of the reference pixels to diminish the speckle effect and stacked the VH and VV polarization bands into a single one, having the time series the following dimensions: 79962 patches, 9x9 spatial neighborhood, 2 polarizations, 31 timestamps.

4 Methods

This section presents the model architectures and details of their implementations used for the experiments. First, we briefly describe the base model architectures used to process Sentinel-2 and Sentinel-1 SITS. Then we proceed to describe unsupervised domain adaptation setting, and the framework we use to deal with the transfer tasks. We continue to describe our multimodal data fusion implementations for the experiments in the supervised and UDA settings. And finally, we provide the details of the experimental settings which include the splitting procedure of the data, and the implementation settings of the experiments.

4.1 Baseline model architectures

We use random forest as the state-of-the-art algorithm for SITS classification, and TempCNN and InceptionTime as the leading deep learning algorithms SITS classification, the three algorithms were used with Sentinel-2 SITS. Additionally, a 2D-CNN was used for Sentinel-1 SITS. All algorithms are briefly introduced.

4.1.1 Random Forest

Random Forest is an ensemble of decision trees (Breiman, 2001) that has been widely used for SITS classification. The algorithm works by randomly selecting a subset of features and a subset of the training data. Then, it creates a decision tree based on the selected features and data. This process is repeated multiple times, and the outputs of all the decision trees are combined to make the final prediction.

4.1.2 TempCNN

The model proposed by Pelletier et al. (2019) has been developed by applying 1D convolutional layers that are capable of extracting relevant features from time series data. The extracted features are then passed through a fully connected layers for classification. The use of batch normalization is another important feature of the TempCNN model. This technique helps to stabilize the training process and improve generalization performance. By normalizing the input to each layer, batch normalization enables the model to learn more effectively, leading to better results.

4.1.3 InceptionTime

InceptionTime is a deep neural network architecture that was proposed by Fawaz et al. (2020) for time series classification inspired by the Inception network (Szegedy et al., 2015). The core building block of InceptionTime is the InceptionModule, which efficiently extracts features from time series data using a combination of convolutional filters of different sizes. The InceptionModule is designed to capture both local and global patterns in time series data, allowing the model to build a hierarchical structure

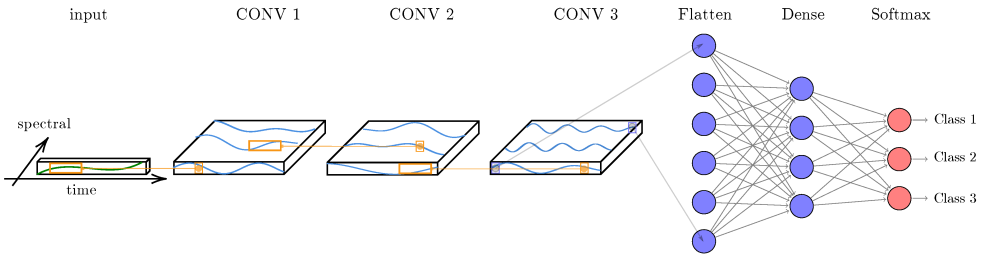


Figure 7: TempCNN architecture
(Pelletier et al., 2019)

to model short-term and long-term dependencies in time series data. InceptionTime consists of multiple InceptionBlocks, which are stacked together to form the network. Each InceptionBlock contains one or more InceptionModules, and is designed to capture increasingly complex patterns in the time series data as the network deepens. The model uses residual connections to be able to implement a deeper architecture, avoiding the vanishing/exploding gradient problem.

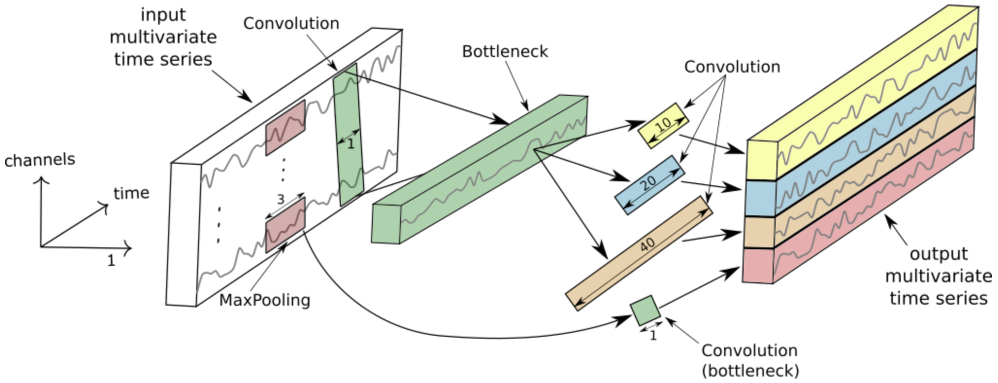


Figure 8: Inception module
(Fawaz et al., 2020)

4.1.4 Sentinel-1 patch-based CNN

Sentinel-1 (S1) data will mainly be used to complement Sentinel-2 (S2) in a multimodal approach, however, a S1-only model will also be trained and evaluated for reference.

We consider a two-dimensional convolutional neural network (2D-CNN), as proposed by (Gbodjo et al., 2021). As described in section 3.2.2 the data is organized as a stacked image with as many bands as the number of time-steps times two, since S1 data have backscatter values with two polarizations: VV and VH. Patches of size 9x9 pixels are extracted from the stacked image are then concatenated and constitute the input information for the model.

4.2 Unsupervised domain adaptation

In this section we focus on the Unsupervised Domain Adaptation (UDA) setting. We begin describing the Domain Adversarial Neural Networks (DANN) model which aims to generate domain-invariant features, and then continue with the Spatially-Aligned Domain-Adversarial Neural Network (SpADANN) framework, which expands DANN by including a self-training strategy to improve the transfer taks to the target domain.

4.2.1 Domain Adversarial Neural Network (DANN)

Domain Adversarial Neural Networks (DANN) is a neural network architecture that is trained on labeled data from the source domain and unlabeled data from the target domain (Ganin et al., 2016). The network uses a Gradient Reversal Layer (GRL) which has no parameter associated with it. During the forward pass the GRL acts as an identity transformation, while during the backpropagation takes the gradient from the subsequent level and changes its sign, i.e., multiplies it by -1, before passing it to the preceding layer. This GRL is inserted between the feature extractor $G_f(\cdot)$ and the domain classifier $G_d(\cdot)$, resulting in the architecture shown in Figure 9. Equation 1 shows the loss function.

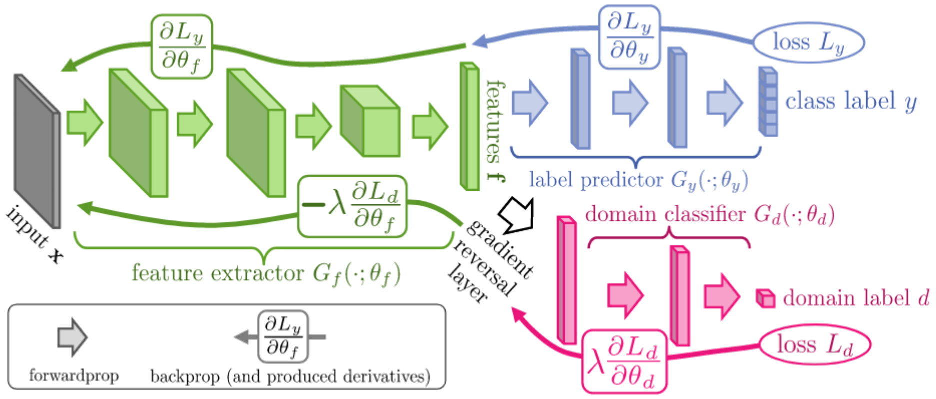

$\frac{\partial L_y}{\partial \theta_f}$
$\frac{\partial L_y}{\partial \theta_y}$
loss L_y
class label y
features $\mathbf{f}$
label predictor $G_y(\cdot;\theta_y)$
$-\lambda \frac{\partial L_d}{\partial \theta_f}$
input x
feature extractor $G_f(\cdot;\theta_f)$
gradient reversal layer
domain classifier $G_d(\cdot;\theta_d)$
domain label d
$\lambda \frac{\partial L_d}{\partial \theta_d}$
loss L_d
forwardprop
backprop (and produced derivatives)

Figure 9: Domain Adversarial Neural Network architecture
(Ganin et al., 2016)

$$L_{DANN}(X^s,Y^s,X^t|\Theta_f,\Theta_y,\Theta_d) = L_y(X^s,Y^s|\Theta_f,\Theta_y) - \lambda L_d(X^s,X^t|\Theta_f,\Theta_d) \quad (1)$$

where $L_y(X^s,Y^s|\Theta_f,\Theta_y)$ is the loss associated to the label classifier, while $\lambda L_d(X^s,X^t|\Theta_f,\Theta_d)$ 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.

For the scenario of SITS classification, each sample $x_i \in \mathbb{R}^{T\times B}$ is a SITS pixel defined over T timestamps and characterized by B spectral bands.

4.2.2 Spatially-Aligned DANN (SpADANN)

The Spatially-Aligned Domain-Adversarial Neural Network (SpADANN) framework has been recently proposed for domain adaptation in SITS classification (Capliez et al., 2023). In order to improve the classification of pixels from the target domain, the SpADANN framework leverages a self-training strategy to associate pseudo-labels with a subset of data from the target domain. This is done to inject pseudo-supervision into the training process and improve the performance of the land-cover classifier subnetwork. Instead of using a threshold to select high-confidence samples, target pixels are selected based on two criteria: spatial consistency and correct land-cover prediction on the corresponding source pixel. The resulting pseudo-labeled target samples act as anchor points between the source and target domains, exploiting model output stability and reducing the distribution gap between domains. This process is specific to the land-cover mapping temporal UDA problem and does not require a hyperparameter threshold. Figure 10 shows the self-training procedure used by SpADANN.

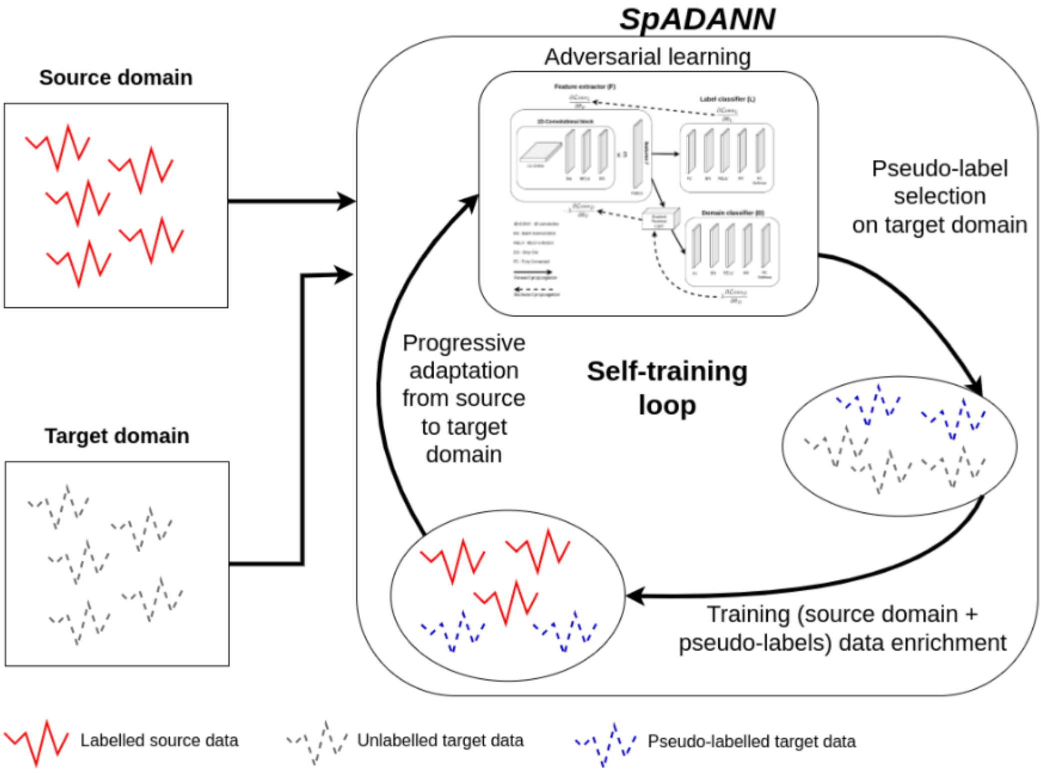


Figure 10: Self-training loop for the SpADANN framework
(Capliez et al., 2023)

Capliez et al. (2023) leveraged the specificity of the land cover mapping temporal UDA problem conceiving a process based on the spatial consistency between the two SITS pixels $\boldsymbol{x}_i^s$ and $\boldsymbol{x}_i^t$ sharing the same spatial location ($location(\boldsymbol{x}_i^t) = location(\boldsymbol{x}_i^s)$). On that way, they focused the pseudo-labelling selection process based on two criteria. The first criteria is based on spatial consistency:

$$G_y(\boldsymbol{x}_i^s|\theta_f,\theta_y) = G_y(\boldsymbol{x}_i^t|\theta_f,\theta_y)$$

given that ($location(\boldsymbol{x}_i^t) = location(\boldsymbol{x}_i^s)$), and the second criteria requires that the land cover classifier G_y supplies the correct prediction for the source sample:

$$G_y(\boldsymbol{x}_i^s|\theta_f,\theta_y) = y_i^s$$

Thus, this selection process chooses target samples that remain stable in terms of model output prediction, allowing to select pseudo-labelled samples to act as anchor points between the source and the target domains (Capliez et al., 2023). The loss associated with the pseudo-labels is defined as:

$$L_p(X^s,X^t,Y^s,\hat{Y}^t|\Theta_f,\Theta_y) = \sum_{\boldsymbol{x}_i^t \in Xt} 1_{\{G_y(\boldsymbol{x}_i^s)=G_y(x_i^t) and G_y(x_i^s)=y_i^s\}} H(\hat{y}_i^t, G_{y_{prob}}(\boldsymbol{x}_i^t)) \quad (2)$$

where 1_{cond} is an indicator function that returns 1 if the condition is verified and 0 otherwise, $G_{y_{prob}}(\cdot)$ provides the model output distribution over the possible land cover set, $H(\cdot)$ 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 $G_{y_{prob}}(\boldsymbol{x}_i^t)$ for the pixel $\boldsymbol{x}_i^t$ coming from the target domain.

Combining L_p from equation 2 into the L_{DANN} defined in equation 1, we get the final loss implemented for SpADANN.

$$L_{SpADANN} = (1-\alpha)\times L_{DANN}(X^s,X^t,Y^s|\Theta_f,\Theta_y,\Theta_d)+\alpha\times L_p(X^s,X^t,Y^s,\hat{Y}^t|\Theta_f,\Theta_y) \quad (3)$$

where the hyper-parameter α associated to the progressive transfer strategy gets updated as epochs progresses with the aim to vary their importance during the learning procedure, defined as:

$$\alpha = \beta \times \frac{\text{epoch}}{\text{total number of epochs}}$$

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 the model.

4.3 Multimodal data fusion

Here we describe the multimodal strategies for the supervised and the unsupervised domain adaptation settings.

4.3.1 Supervised setting

We evaluate two data fusion approaches discussed in Section 2, **layer and decision fusion**. To accomplish this, we used a TempCNN encoder for S2 data and the CNN for S1 data. These networks create a feature representation of both datasets independently. The **layer fusion** approach concatenates these features and then pass them through a fully connected layer to generate the logits and ultimately make a decision. In contrast, the **decision fusion** approach keeps the features separated and generates logits from each of them, with independent fully connected layers, then class probabilities are computed and those probabilities are averaged to have a combined decision. More details of the implementation can be seen in Figure 11, auxiliary classifiers are added to the network to generate a combined loss between each of the features and the fused. For the auxiliary classifier we apply a linear fully connected layer with the one neuron for each land cover class.

$$L = L_{S2+S1} + \lambda_1 \times L_{S2} + \lambda_2 \times L_{S1} \tag{4}$$

where λ_1 and λ_2 are used to weight the contribution of the auxiliary classifiers.

4.3.2 Unsupervised domain adaptation setting

Here we tested different ways to adapt the loss L_d from the domain discriminator in a multimodal setting. As each encoder will have it's own feature representation, we explored what will be the best location to place the domain classifier, to use the fused features, each individual features, or a combination of the three. Figure 12 provides more details of the implementation. Similar to how the loss is combined in the supervised classification, here we use the same loss for the label classifier, and perform a similar approach for the domain loss:

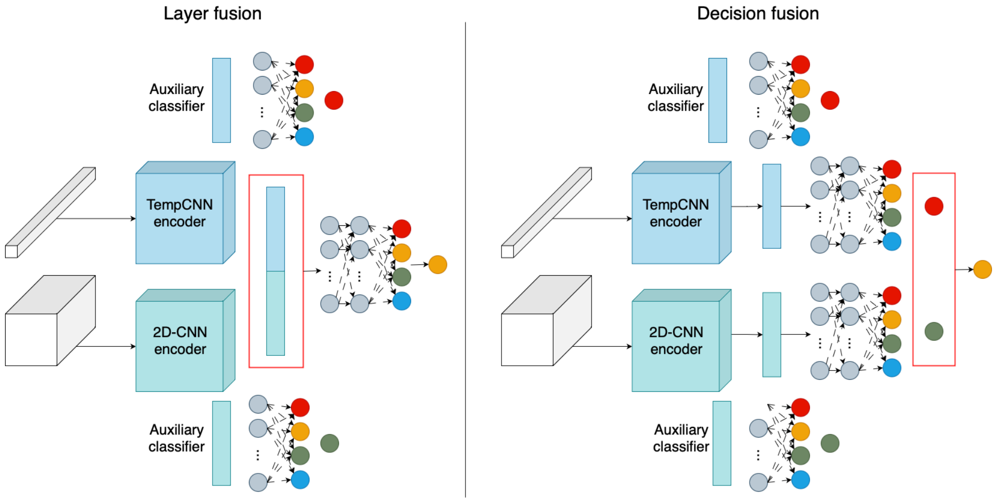


Figure 11: Multimodal strategy for the layer and decision fusion approaches to combine S2 and S1 data.

$$L_d^{v1} = L_{d_{S1+S2}} \tag{5}$$

$$L_d^{v2} = L_{d_{S1}} + L_{d_{S2}} \tag{6}$$

$$L_d^{v3} = L_{d_{S1+S2}} + L_{d_{S1}} + L_{d_{S2}} \tag{7}$$

4.4 Experimental settings

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 were implemented in Python with the Pytorch library and the code is available in a GitHub repository[5]. In the following subsections we will provide the details for the splitting procedure and the implementation of the experiments.

4.4.1 Splitting procedure

Supervised SITS classification models were be trained using either only S2 or S1 data. The data for each year was split into training, validation and test sets following a proportion of 50%, 30% and 20% respectively. Additionally, with the aim to avoid possible spatial bias in the evaluation (Karasiak et al., 2022), we ensured that all pixels belonging to the same object to be exclusively associated with one of the sets (training, validation or test). The splitting procedure was repeated five times. To also take into account the variations from weights' initialization, we trained the models five times on each

[5] `https://github.com/EdgarJoao30/mtda`

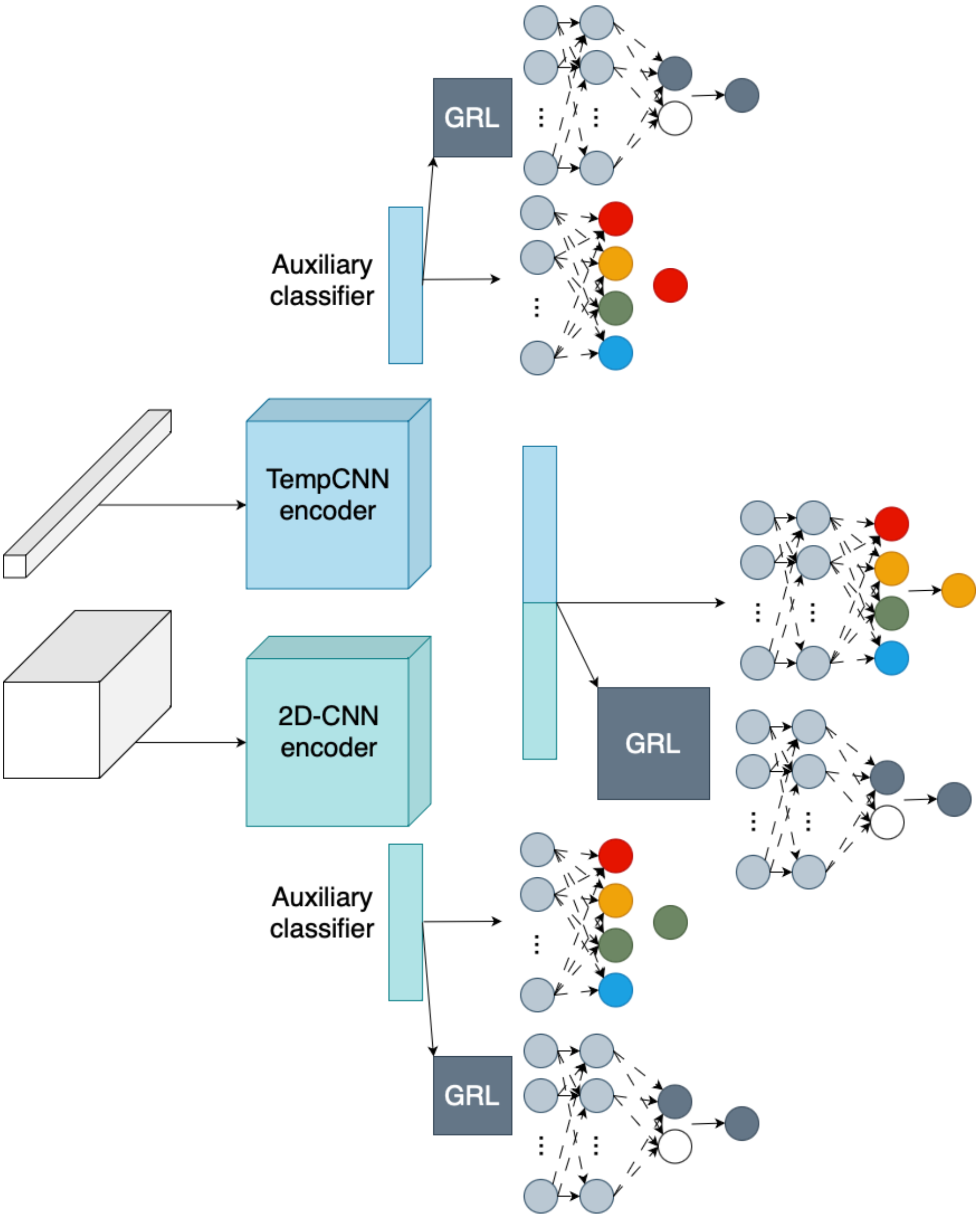


Figure 12: Multimodal UDA strategy for the domain discriminator location approaches to combine S2 and S1 data.

training/validation/test split.

For the UDA setting, models are trained exploiting the whole set of source and target samples while having access only to label information from the source domain and tested on all the target samples. To also take into account the variations from weights' initialization, we trained the models five times.

4.4.2 Implementation details

The models evaluated in this part were described in Section 4.1, here their parameter choices are listed. The general hyperparameter settings for all models used are shown in Table 3.

Table 3: Hyperparameter settings of the evaluated approaches

Hyperparameter	**Setting or value**
Epochs	400
Learning rate	0.0005
Optimizer	Adam
Dropout rate	0.5**
Batchsize	256

** InceptionTime uses a dropout rate of 0.2

In total there are six classification tasks that we evaluated, these are direct classification and the transfer tasks. The direct classification tasks correspond to the supervised classification models applied to the SITS for each year (i.e. 2018, 2020, 2021) using their corresponding train, validation and test sets. In the results section the direct classification tasks will be refered using the specific year (e.g. 2018 → 2018), or as $D_{t \to t}$ when placed together with the transfer tasks or when aggregating results. The transfers tasks are: 2018 → 2020, 2018 → 2021, and 2020 → 2021, which will be refered in the results in Section 5 using the specific years involved (e.g. 2018 → 2021), or as $D_{s \to t}$. It is important to mention that the direct classification tasks will be done exclusively in the supervised setting, while the UDA setting will be used exclusively in the transfer tasks. However, the supervised models will also be evaluated in the transfer tasks for reference.

Supervised setting: The implementation of **Random Forest** (Breiman, 2001) from Scikit-Learn was used with standard parameter settings (Pelletier et al., 2016): 500 trees at maximum depth, and a number of randomly selected variables per node equals to the square root of the total number of features. For **TempCNN**, the proposed default values from Pelletier et al. (2019) were kept, consisting of three convolutional layers

(64 units), one dense layer (256 units), and the filter size of convolutions is set to 5. Similarly, for **InceptionTime**, most of the proposed values from Fawaz et al. (2020) were kept, but the Inception module had to be adapted to the number of timesteps of our data, the module has 3 sets of filters each with 32 filters of lengths 5, 9, and 13 (producing a maximum receptive field of 73), and the bottleneck size value was set to 32. The models describe above were used with the S2 SITS data, additionally, a **CNN** was used to process S1 SITS data, following the findings of Ienco et al. (2019) and Gbodjo et al. (2021), using a 2D-CNN for S1 data helps to alleviate possible issues induced by spatial speckle phenomena that usually affects SAR signal and improves the added value for combining S2 and S1. As presented in Section 4.3.1 we evaluated **layer** and **decision** fusion approaches. For this, we used a TempCNN encoder for S2 data and the CNN for S1 data. When presenting the results in Section 5, these models will be named as *Layer fusion* and *Decision fusion*.

UDA setting: We evaluated three transfer scenarios ($D_s \rightarrow D_t$): 2018 → 2020, 2018 → 2021, and 2020 → 2021. The UDA baselines were trained only with S2 or S1 data at a time. Here, we used the **SpADANN** framework. The model was trained having a TempCNN encoder for the only S2 experiments, and the CNN for the only S1 experiments. We set the β parameter to 1, to allow a full transfer to the target domain at the end of the training process. And also set the λ parameter, from the GRL to 1, to prevent an accumulated effect with α, which already has the role to gradually balance the transfer task with the use of L_p over the training process.

Ablation analysis: DANN is the backbone of SpADANN, thus, it will be used as a competing method during the experiments but will also be seen as an ablation since it doesn't use the self-training approach. Additionally, an ablation where the pseudo labels are selected with a confidence threshold will be tested, where the largest probability for the class prediction will be compared with a threshold of 0.9. Finally, an ablation where only the condition regarding both source and target having the same prediction is preserved, thus, redefining L_p as:

$$L_p(X^s, X^t, Y^s, \hat{Y}^t | \Theta_f, \Theta_y) = \sum_{\boldsymbol{x}_i^t \in X^t} 1_{\{G_y(\boldsymbol{x}_i^s) = G_y(\boldsymbol{x}_i^t)\}} H(\hat{y}_i^t, G_{y_{prob}}(\boldsymbol{x}_i^t)) \tag{8}$$

4.4.3 Performance metrics

The assessment of the model performances was done considering the test set and the following metrics: accuracy (global precision), F1 score (harmonic mean of precision and recall). For the temporal domain adaptation approach, the target domain was considered as the test set. The feature representation of the models is qualitatively assessed via a t-SNE visualization (Maaten, 2014). Finally, land cover maps are visually investigated and compared with each other.

5 Results and discussion

In this section we will present the results of the experiments for the supervised classification scenario and the unsupervised domain adaptation scenario. For this, we evaluated the models on a set of transfer tasks defined as follow: $D_{s \to t}$ models trained on source domain and tested on target domain (i.e. 2018 → 2020, 2018 → 2021, and 2020 → 2021), and $D_{t \to t}$ models trained on target domain and tested on target domain, thus, no transfer was performed (i.e. 2018 → 2018, 2020 → 2020, and 2021 → 2021). We start with the results of the supervised setting, it is important to note that here we also evaluated the models on $D_{s \to t}$ for reference, even though the supervised setting does not assess the transfer tasks in any way, thus, the performance is expected to decrease significantly. We continue with the results on the unsupervised temporal domain adaptation setting, all models are evaluated in $D_{s \to t}$, and the results of the supervised setting on $D_{s \to t}$ and $D_{t \to t}$ are added here for reference. We extend the results by analyzing the internal feature representation that the models are able to achieve, the confusion on the classification tasks, and finally the land cover maps are investigated.

5.1 Supervised satellite image time series classification

Here we evaluate the performance of two types of multimodal fusion, layer and decision fusion for S2 and S1 SITS. To arrive a this point, we started evaluating some baseline models to test whether our data and our models were performing as expected. Table 4 shows the results of such experiments, where we compare the performance of random forest, TempCNN and InceptionTime. On average, TempCNN performed better on the $D_{t \to t}$ scenario and InceptionTime performed better on the D_s scenario. As expected $D_{s \to t}$ performances are significantly lower than $D_{t \to t}$ performances. For TempCNN, the year that had the lowest performance in $D_{t \to t}$ was 2020 with an F1-score of 90.08 ± 2.85 , while the transfer task $D_{s \to t}$ with the lowest performance was 2018 → 2021. These results are consistent to what was found by Rußwurm et al. (2020), where they also compared the performances of random forest, TempCNN, and InceptionTime, among other models, in a dataset they elaborated. When tested on Sentinel-2 data, TempCNN performed consistently better than InceptionTime. Here TempCNN also performs better than InceptionTime on average in all $D_{t \to t}$, although with no significant differences. In addition, here we evaluated the transfer tasks $D_{s \to t}$, where InceptionTime performed better, on average, than random forest and TempCNN.

Table 5 shows the results for the next set of experiments where we introduced multimodality to the supervised classification setting. The model CNN$_{S1}$ was trained with S1 data to be used as baseline together with the TempCNN$_{S2}$ trained with S2 from the previous experiments. Layer fusion and decision fusion models correspond to models

Table 4: Supervised SITS classification

Transfer task	Random Forest		TempCNN		InceptionTime	
	F1-score	**Accuracy**	**F1-score**	**Accuracy**	**F1-score**	**Accuracy**
2018 → 2018	88.77 ± 1.40	88.77 ± 1.50	91.48 ± 1.98	89.99 ± 2.03	90.97 ± 1.77	89.42 ± 1.90
2018 → 2020	65.10 ± 4.89	66.88 ± 4.22	69.14 ± 7.37	65.14 ± 6.79	68.36 ± 4.54	64.31 ± 4.05
2018 → 2021	61.80 ± 6.15	63.03 ± 5.78	63.10 ± 5.50	58.03 ± 5.19	70.14 ± 7.09	66.03 ± 7.92
2020 → 2020	88.53 ± 2.46	88.67 ± 2.35	90.08 ± 2.85	89.12 ± 2.52	89.42 ± 2.39	88.42 ± 2.30
2020 → 2021	60.12 ± 8.14	62.04 ± 6.42	64.91 ± 7.08	61.32 ± 7.55	72.23 ± 4.81	67.58 ± 4.55
2021 → 2021	89.88 ± 2.40	90.00 ± 2.15	92.95 ± 1.61	91.39 ± 1.51	90.67 ± 2.85	88.55 ± 3.36
Avg. $D_{s\to t}$	62.34 ± 6.53	63.98 ± 5.55	65.72 ± 6.70	61.50 ± 6.58	**70.24 ± 5.60**	**65.97 ± 5.77**
Avg. $D_{t\to t}$	89.06 ± 2.14	89.15 ± 2.03	**91.50 ± 2.21**	**90.17 ± 2.06**	90.35 ± 2.38	88.80 ± 2.59

using a TempCNN encoder for S2 data and a CNN for S1 data, as depicted in Figure 11. Overall, CNN$_{S1}$ was able to perform almost as good as TempCNN$_{S2}$ and even having a better performance than S2 on the transfer task 2018 → 2021 which was the lowest performance one for TempCNN$_{S2}$. There was not an improvement on the $D_{t\to t}$ scenario, although also not a significant decrease compared to TempCNN$_{S2}$. However, where we see an increase is in the $D_{s\to t}$ transfer tasks, where layer fusion performs on average almost 6 points higher than TempCNN$_{S2}$ and 2 points higher than decision fusion. These results are similar to those found by Ofori-Ampofo et al. (2021), where they tested S2 and S1 data fusion at input, layer and decision level, finding no significant differences in the overall F1 score. However, when classes of interest are underrepresented they found that it is better to use input or layer fusion. Here we also present the results on the $D_{s\to t}$ transfer tasks, where layer fusion is performing better. Due to this findings, layer fusion will be used as a reference on the next section, where we deal with the unsupervised domain adaptation setting.

Table 5: Multimodal SITS classification

Transfer task	TempCNN$_{S2}$		CNN$_{S1}$		Layer fusion		Decision fusion	
	F1-score	**Accuracy**	**F1-score**	**Accuracy**	**F1-score**	**Accuracy**	**F1-score**	**Accuracy**
2018 → 2018	91.48 ± 1.98	89.99 ± 2.03	88.91 ± 2.95	86.90 ± 3.29	91.89 ± 1.55	90.32 ± 1.47	92.08 ± 1.53	90.40 ± 1.35
2018 → 2020	69.14 ± 7.37	65.14 ± 6.79	66.22 ± 7.88	61.23 ± 8.02	74.41 ± 1.15	70.05 ± 1.48	72.45 ± 2.70	67.99 ± 3.20
2018 → 2021	63.10 ± 5.50	58.03 ± 5.19	67.55 ± 4.75	62.79 ± 4.47	64.65 ± 10.81	59.25 ± 10.36	61.76 ± 10.92	56.57 ± 9.20
2020 → 2020	90.08 ± 2.85	89.12 ± 2.52	89.73 ± 1.64	88.10 ± 1.48	88.68 ± 2.12	87.57 ± 1.30	89.83 ± 3.07	88.75 ± 2.66
2020 → 2021	64.91 ± 7.08	61.32 ± 7.55	60.33 ± 8.23	53.24 ± 8.39	75.38 ± 6.89	71.85 ± 7.13	73.35 ± 7.62	71.53 ± 8.37
2021 → 2021	92.95 ± 1.61	91.39 ± 1.51	89.73 ± 1.60	87.10 ± 2.28	91.58 ± 1.57	89.98 ± 1.06	90.26 ± 0.64	88.39 ± 0.75
Avg. $D_{s\to t}$	65.72 ± 6.70	61.50 ± 6.58	64.70 ± 7.13	59.09 ± 7.18	**71.48 ± 7.43**	**67.05 ± 7.31**	69.19 ± 7.84	65.36 ± 7.41
Avg. $D_{t\to t}$	**91.50 ± 2.21**	**90.17 ± 2.06**	89.46 ± 2.16	87.37 ± 2.46	90.72 ± 1.77	89.29 ± 1.29	90.72 ± 2.00	89.18 ± 1.77

5.2 Unsupervised temporal domain adaptation for satellite image time series classification

For the temporal domain adaptation setting, we compared the different versions of SpADANN w.r.t the data modalities, i.e. using S2 (SpADANN$_{S2}$), S1 (SpADANN$_{S1}$) or the multimodal scenarios, where we assessed the number and location of domain discriminators, referred here as mmSpADANN$_{v1}$, mmSpADANN$_{v2}$, and mmSpADANN$_{v3}$, for one, two or three domains discriminators on the fused data, on each feature representation or on all fused and separated features respectively. We also added the previous results from the supervised setting, more specifically, TempCNN$_{S2}$, CNN$_{S1}$, and layer fusion, to be used as reference. We used $D_{s \to t}$ transfer tasks on the supervised setting as a direct transfer, or a worst case scenario where there is no information about the target domain and a supervised model trained in a different year is applied directly to the target year. Whereas $D_{t \to t}$ is used as a best case scenario, where no transfer is needed as the supervised model is trained on the same year to where it's applied.

Table 6 shows the results for these set of experiments, both SpADANN$_{S2}$ and SpADANN$_{S1}$ are able to achieve good performances on average, reaching the performance of $D_{t \to t}$ supervised models, SpADANN$_{S1}$ is on average 0.43 points higher than SpADANN$_{S2}$, and 0.88 points lower than TempCNN$_{S2}$ on $D_{t \to t}$. All of the mmSpADANN approaches are able to improve SpADANN, however, mmSpADANN$_{v2}$ and mmSpADANN$_{v3}$ are on average higher than mmSpADANN$_{v1}$ with an average F1 score of 91.32, but still lower than TempCNN$_{S2}$ on $D_{t \to t}$ which has an average F1 score of 91.99. Additionally, we also trained a multimodal DANN model with the same settings as mmSpADANN$_{v3}$ as a competing method, and we can observe that even though mmDANN$_{v3}$ is able to achieve a better performance than the supervised models on $D_{s \to t}$, mmSpADANN$_{v3}$ still outperforms mmDANN$_{v3}$ by 8.49 points on F1 score, highlighting the importance of the self-training component in SpADANN.

We also evaluated per-class performances for each of the transfer tasks. Table 7 shows the results for the transfer task of 2018 → 2021, the results for the other transfer tasks can be seen in the supplementary material. At a first glance on the supervised models, both on $D_{s \to t}$ and $D_{t \to t}$, there is no obvious benefit on combining S2 and S1 data, the models perform very similar and the standard deviation of the per-class performance is large enough to no be able to determine significant differences, except on two classes in $D_{s \to t}$, oleaginous and grassland, here the layer fusion approach is able to outperform the single modality models. However, the contribution in the temporal domain adaptation setting is clearer, mmSpADANN$_{v3}$ is able to increase the performance of all classes and even achieving a higher performance than $D_{t \to t}$ in cereals, grassland, forest and baresoil. In general, all UDA approaches either with single modality or multimodal are

Table 6: Unsupervised temporal domain adaptation

Strategy	Method	2018 → 2020		2018 → 2021		2020 → 2021		Avg.	
		F1 score	**Accuracy**	**F1 score**	**Accuracy**	**F1 score**	**Accuracy**	**F1 score**	**Accuracy**
$D_{s\to t}$	CNN$_{S1}$	66.22 ± 7.88	61.23 ± 8.02	67.55 ± 4.75	62.79 ± 4.47	60.33 ± 8.23	53.24 ± 8.39	64.70 ± 7.13	59.09 ± 7.18
	TempCNN$_{S2}$	69.14 ± 7.37	65.14 ± 6.79	63.10 ± 5.50	58.03 ± 5.19	64.91 ± 7.08	61.32 ± 7.55	65.72 ± 6.70	61.50 ± 6.58
	Layer fusion	74.41 ± 1.15	70.05 ± 1.48	64.65 ± 10.81	59.25 ± 10.36	75.38 ± 6.89	71.85 ± 7.13	**71.48 ± 7.43**	**67.05 ± 7.31**
UDA	mmDANN$_{\upsilon3}$	84.74 ± 1.50	78.76 ± 2.22	80.29 ± 0.41	73.12 ± 0.49	83.47 ± 0.32	75.90 ± 0.39	82.83 ± 0.92	75.93 ± 1.33
	SpADANN$_{S2}$	88.79 ± 0.15	87.49 ± 0.24	90.56 ± 0.27	89.02 ± 0.32	92.71 ± 0.18	91.07 ± 0.20	90.68 ± 0.21	89.19 ± 0.26
	SpADANN$_{S1}$	89.38 ± 0.18	89.13 ± 0.10	91.03 ± 0.73	90.47 ± 1.37	92.93 ± 0.23	92.70 ± 0.51	91.11 ± 0.45	90.77 ± 0.84
	mmSpADANN$_{\upsilon1}$	89.41 ± 0.06	89.22 ± 0.06	91.42 ± 0.28	91.13 ± 0.49	93.05 ± 0.11	92.86 ± 0.19	91.29 ± 0.18	91.07 ± 0.30
	mmSpADANN$_{\upsilon2}$	89.35 ± 0.17	89.12 ± 0.30	91.58 ± 0.01	91.42 ± 0.03	93.04 ± 0.11	92.91 ± 0.17	**91.32 ± 0.12**	**91.15 ± 0.20**
	mmSpADANN$_{\upsilon3}$	89.30 ± 0.07	89.05 ± 0.11	91.61 ± 0.04	91.38 ± 0.08	93.06 ± 0.07	92.96 ± 0.13	**91.32 ± 0.06**	91.13 ± 0.11
$D_{t\to t}$	CNN$_{S1}$	89.73 ± 1.64	88.10 ± 1.48	89.73 ± 1.60	87.10 ± 2.28	89.73 ± 1.60	87.10 ± 2.28	89.73 ± 1.61	87.43 ± 2.05
	Layer fusion	88.68 ± 2.12	87.57 ± 1.30	91.58 ± 1.57	89.98 ± 1.06	91.58 ± 1.57	89.98 ± 1.06	90.61 ± 1.77	89.18 ± 1.33
	TempCNN$_{S2}$	90.08 ± 2.85	89.12 ± 2.52	92.95 ± 1.61	91.39 ± 1.51	92.95 ± 1.61	91.39 ± 1.51	**91.99 ± 2.11**	**90.63 ± 1.91**

able to significantly outperform the supervised models in $D_{s\to t}$, and as mentioned, even outperform $D_{t\to t}$ in well represented classes. On average mmSpADANN$_{\upsilon3}$ is able to outperform the supervised models in $D_{s\to t}$ by at least 24.06 points of F1 score.

Table 7: Class-wise F1-score for the transfer task 2018→2021

Class name	$D_{s\to t}$			UDA			$D_{t\to t}$		
	TempCNN$_{S2}$	**CNN$_{S1}$**	**Layer fusion**	**SpADANN$_{S2}$**	**SpADANN$_{S1}$**	**mmSpADANN$_{\upsilon3}$**	**TempCNN$_{S2}$**	**CNN$_{S1}$**	**Layer fusion**
🟨 Cereals	**55.69 ± 4.39**	45.96 ± 15.06	53.37 ± 4.05	86.81 ± 0.36	91.79 ± 0.89	**92.38 ± 0.41**	**90.70 ± 1.90**	85.99 ± 3.54	88.17 ± 4.21
🟪 Cotton	58.90 ± 4.93	**60.76 ± 4.41**	57.91 ± 3.41	64.76 ± 0.72	68.35 ± 0.89	**69.08 ± 0.30**	87.91 ± 5.56	87.44 ± 4.62	**90.08 ± 7.21**
🟦 Oleag./Legum.	26.27 ± 6.13	23.71 ± 8.65	**47.49 ± 5.65**	57.60 ± 2.29	65.22 ± 0.22	**65.74 ± 0.45**	**85.73 ± 4.26**	80.66 ± 2.90	79.64 ± 1.62
🟩 Grassland	64.12 ± 11.82	61.58 ± 8.74	**72.29 ± 5.11**	91.41 ± 0.32	91.22 ± 0.72	**91.48 ± 0.34**	88.04 ± 5.69	89.16 ± 2.29	**90.27 ± 2.10**
🟩 Shrubland	51.89 ± 14.42	**70.67 ± 10.77**	45.77 ± 27.68	96.00 ± 0.16	96.06 ± 2.15	**97.27 ± 0.16**	**92.92 ± 2.63**	88.03 ± 3.63	92.04 ± 2.87
🟩 Forest	58.27 ± 19.95	54.16 ± 21.10	**58.75 ± 16.40**	98.40 ± 0.23	97.53 ± 2.71	**99.03 ± 0.19**	**91.24 ± 8.02**	81.70 ± 17.61	88.21 ± 10.48
⬜ Baresoil	**74.76 ± 12.38**	33.82 ± 20.09	60.61 ± 27.25	88.20 ± 0.34	88.83 ± 2.11	**90.38 ± 1.30**	**88.44 ± 10.29**	80.48 ± 24.46	83.21 ± 23.25
🟦 Water	**99.92 ± 0.10**	97.91 ± 1.97	99.58 ± 0.84	99.99 ± 0.02	**100.00 ± 0.00**	**100.00 ± 0.00**	98.82 ± 2.04	96.70 ± 5.12	**99.09 ± 0.81**
Total	61.10 ± 5.50	67.55 ± 4.75	64.65 ± 10.81	90.56 ± 0.27	91.03 ± 0.73	91.61 ± 0.04	92.95 ± 1.61	89.73 ± 1.60	91.58 ± 1.57

These results can be further explained by seeing Figure 13. Here, the improvement of mmSpADANN$_{\upsilon3}$ compare to layer fusion in $D_{s\to t}$ is evident, and the confusion between the crop classes are further described: 25% of cotton is misclassified as cereals, and 37% of oleaginous is misclassified as cotton. As previously seen in Table 1, cotton and oleaginous classes are the least represented of the crop classes, and overall with the exception of bare soil and water. On Figure 14 we can see how the crop classes on mmSpADANN$_{\upsilon3}$ are still not as well defined as in layer fusion in $D_{t\to t}$, but present a better cluster configuration than layer fusion in $D_{s\to t}$.

It is important to consider the aggregation method of the metrics on interpreting the results. Table 6 shows a small improvement in the multimodal framework compared to the single modalities versions, however, when we see the results on the specific classes in Table 7, they show a systematic improvement in all classes compared to the single modalities models. This is because the metrics are aggregated using a weighted aver-

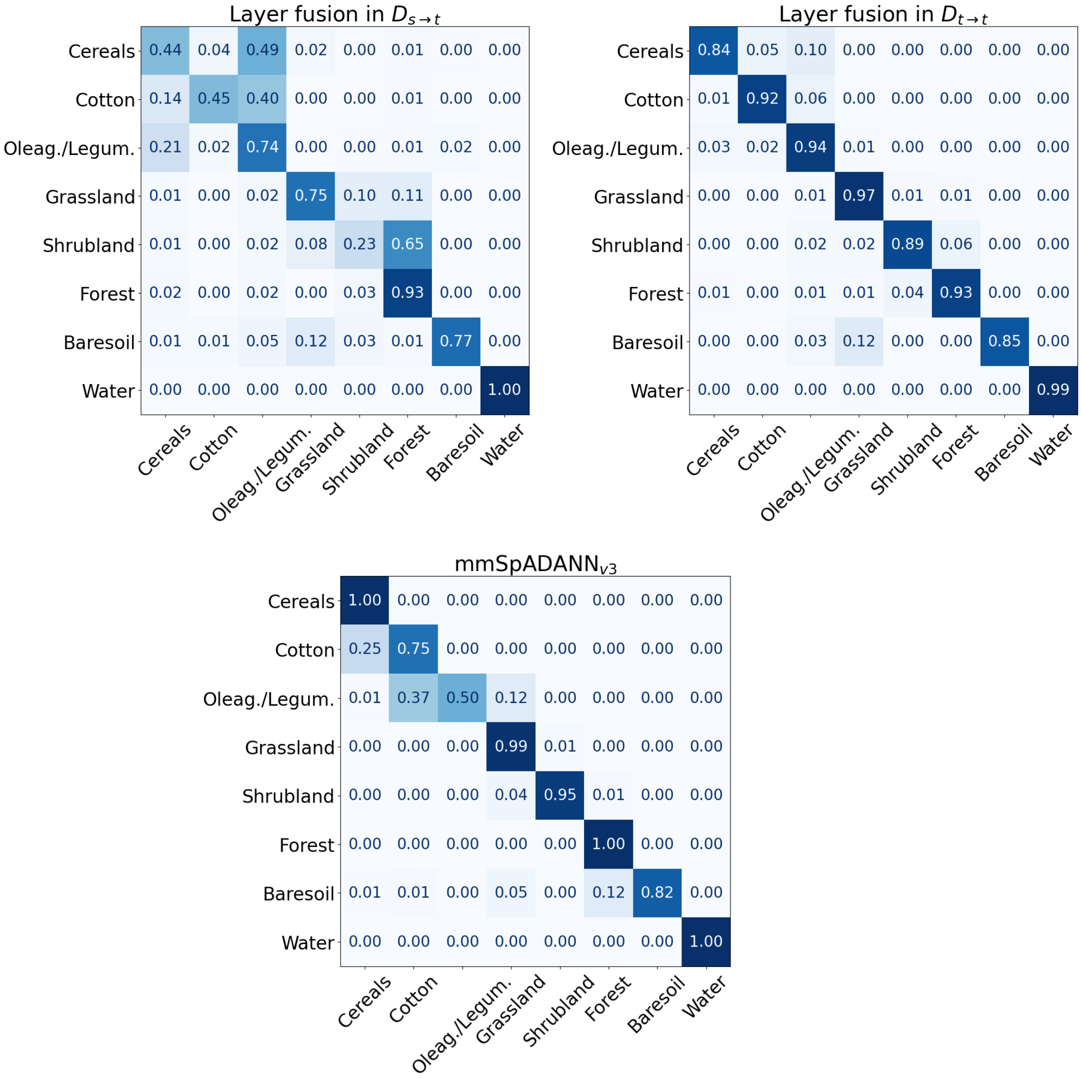


Figure 13: Confusion matrix 2018 → 2021

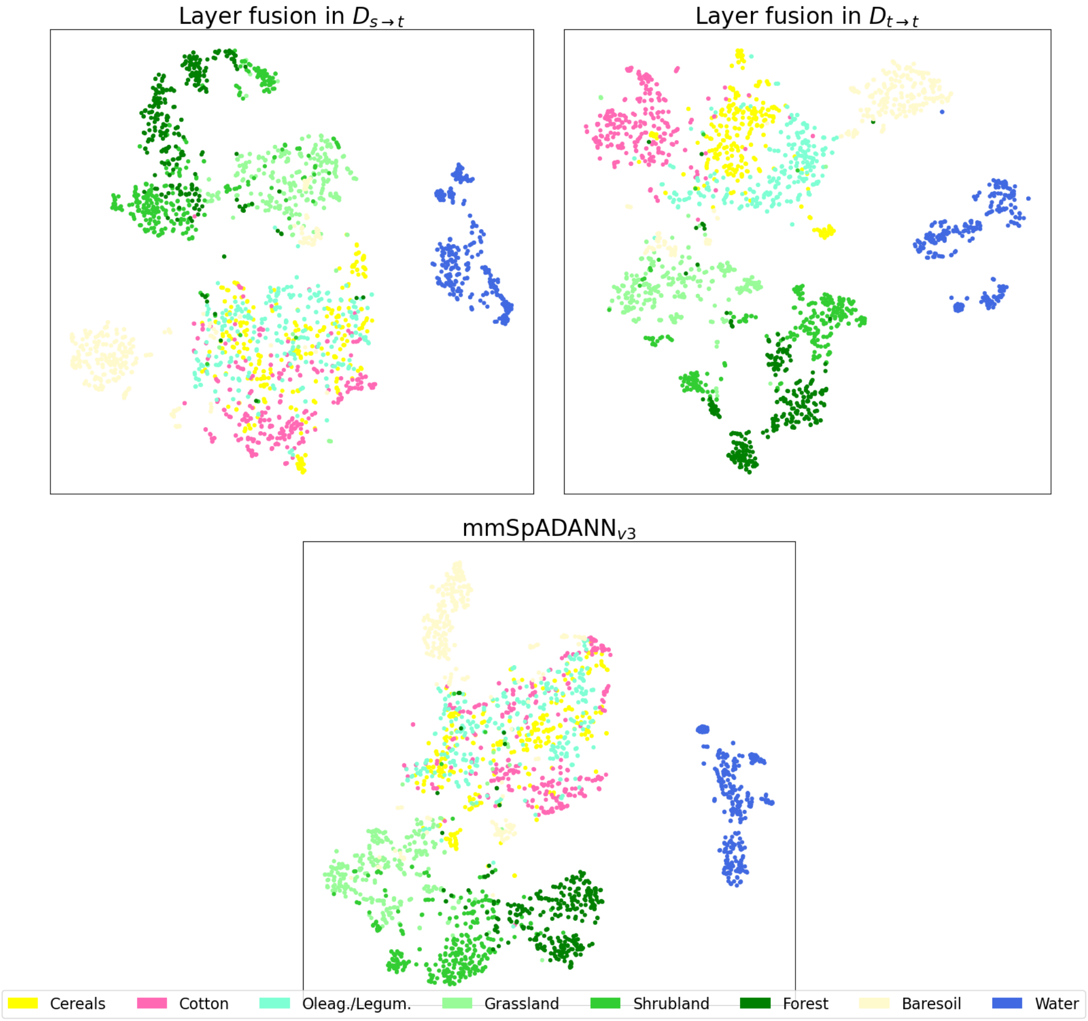


Figure 14: t-SNE 2018 → 2021

age, which accounts for the number of instances of each class. This is done to account for class imbalance, but at the same time hides the improvement that we aimed to see for those underrepresented classes. Table 8 shows the results of the SpADANN framework with the different aggregation procedures to compute the average F1 score. Where *Weighted* takes into account class imbalance, *Micro* calculates the metric globally, and *Macro* doesn't perform any weighting. There, we can see that in the Macro F1 score the improvement of the model is more evident.

Table 8: F1-score 2018 → 2021

Method	**Weighted**	**Micro**	**Macro**
SpADANN$_{S2}$	90.56 ± 0.27	89.01 ± 0.28	56.78 ± 0.90
SpADANN$_{S1}$	91.03 ± 0.73	90.47 ± 1.23	72.40 ± 7.02
mmSpADANN$_{v3}$	91.61 ± 0.04	91.38 ± 0.07	78.34 ± 3.04

The ablation analysis consisted on comparing the performance of the multimodal approach on the different components of SpADANN. Table 9 summarizes the results, where all models use the mmSpADANN$_{v3}$ multimodal approach with three domain classifiers. mmDANN$_{v3}$ takes out the self-training component, mmSpADANN$_{v3}^{TH}$ uses a traditional thresholding approach where predictions with more than 90% of confidence in the target domain are used as pseudo-labels, and mmSpADANN$_{v3}^{C1}$ uses only the condition that both source and target domain have to give the same prediction to be used as pseudo-labels. We can note that effectively mmSpADANN$_{v3}$ provides the best performance compared to its ablations followed very closely by mmSpADANN$_{v3}^{C1}$, and mmDANN$_{v3}$ and mmSpADANN$_{v3}^{TH}$ provides the lowest performance. These results highlight the importance of guiding the transfer process with the use of pseudo-labels, as the domain-invariant features are not sufficient to provide a good performance on its own. However, the selection criteria for the pseudo-labels also plays an important role, solely relying in a threshold value will add too much noise from wrongly classified pseudo-labels and ultimately degrade the performance of the model, making it even worse than not using them in the first place. Having only the condition for the same prediction on the source and target domain achieves almost the same performance than adding the condition of having the correct prediction on the source domain, this is probably due to the high performance that the model is able to achieve at the beginning of the training process in the source domain. These results are consistent with the ablation study performed in the original publication of the SpADANN framework by Capliez et al. (2023), confirming that the extension of the framework to deal with multimodal remote sensing data does not change the behaviour of the SpADANN components.

Table 9: mmSpADANN ablations 2018 → 2021

Method	**F1-score**	**Accuracy**
mmDANN$_{v3}$	80.29 ± 0.41	73.12 ± 0.49
mmSpADANN$_{v3}^{TH}$	79.16 ± 0.74	71.48 ± 1.22
mmSpADANN$_{v3}^{C1}$	91.56 ± 0.08	91.30 ± 0.11
mmSpADANN$_{v3}$	91.61 ± 0.04	91.38 ± 0.08

Similarly to what we did in the previous analysis, to qualitatively assess the Land Cover Maps (LCM) we compared mmSpADANN$_{v3}$ with layer fusion in $D_{s\to t}$ and $D_{t\to t}$. Additionally, we also compared it with the single modality versions SpADANN$_{S2}$ and SpADANN$_{S1}$. Figure 15 provides the LCM for an area of 500 ha, the main observation to notice is the underestimation of the oleaginous class (compared to $D_{t\to t}$), which is consistent with the results described in Figure 13, and the overestimation of the cereals class. However, it is important to notice that even though we using $D_{t\to t}$ as a sort of upper bound, it cannot be fully considered as a true reference map.

The lower performance of the oleaginous class and its resulting underestimation depicted in the LCM can be due to the imbalance of the crop classes in the reference data, where oleaginous reference samples represent only the 4.5% of the total amount of reference data during the year 2018, and 9.15% during the year 2021. Additionally, as was seen on Figure 5 in Section 3.2 only 56% of the reference data from the oleaginous class in 2018 remains on the same class in 2021, due to the self-training procedure of SpADANN, this limits even more the amount of reference data used for the oleaginous class, as the pseudo-labels are selected only from pixels that belong to the same class in both the source and target domain. Even though the framework aims to generate domain invariant features, it requires objects or classes associated to those features to be the same. Meaning that the criteria to define what is considered to be e.g. forest in the source domain is the same criteria in the target domain. This is strongly dependent on how field campaigns collect reference data work over long time-spans. In the case of general crop classes that group many specific crops types, as is the case in the oleaginous/leguminous class, represents an additional difficulty to generate features that are able to correctly classify the class across different years. One indication that this could be the case for the oleaginous class is the increase in the number of polygons from 2018 to 2020 and 2021. There is an increase of 63.35% in the number of polygons collected as reference data in the oleaginous class, while increasing the average surface of each individual polygon by 25.53%. This change from small sparse crops (e.g. ground nuts, cow peas) during 2018 to larger crops (e.g. soybeans) in 2021, could explain the difficulty in that we encountered in having good performances for this class compared to the others, as those different crops grouped into the same category have different phenological cycles. When comparing to the single modality versions, mmSpADANN$_{v3}$ is able to produce less noisy maps, demonstrating the contribution of integrating S2 and S1 data to provide more consistent LCM.

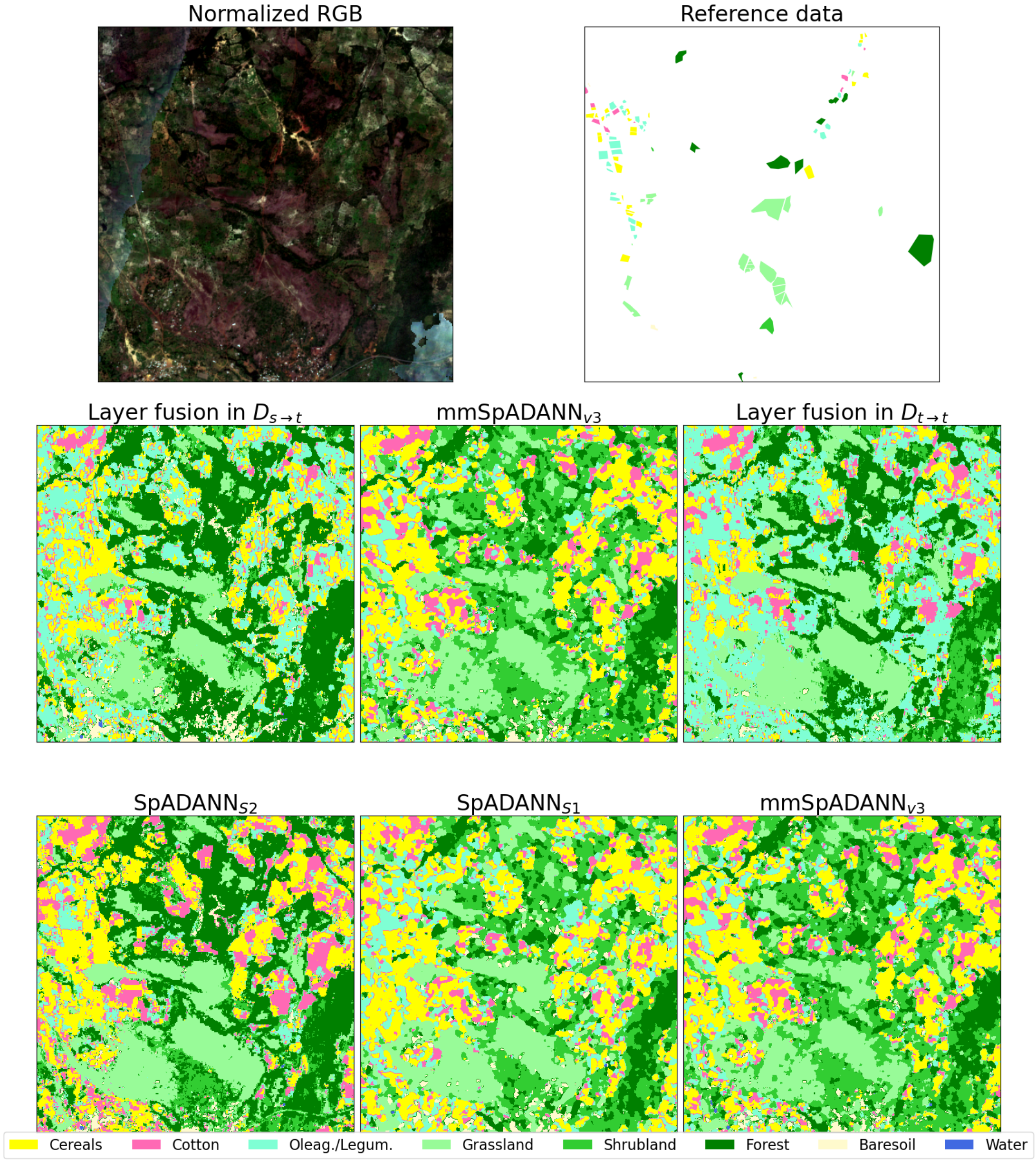


Figure 15: Land cover map 2018 → 2021

6 Conclusion

In this work we have extended the SpADANN framework that deals with temporal unsupervised domain adaptation of satellite image time series, to be able to use multimodal (Sentinel-2 and Sentinel-1) data and take advantage of both optical and radar satellite images to improve land cover mapping performances. More precisely, we used a layer fusion approach to combine the domain-invariant feature representations of Sentinel-2 and Sentinel-1 time series, where for Sentinel-2 we focused on extracting temporal patterns using a TempCNN encoder, while for Sentinel-1 we used a 2D-CNN to cope with the speckle effect associated with radar images.

The results obtained on the Koumbia study site have shown that on both supervised and unsupervised domain adaption settings, there is an improvement on using multimodal data. On a supervised setting, layer and decision fusion performed similarly, however, the layer fusion approach is able to achieve a better performance in a direct transfer task. In the unsupervised temporal domain adaptation setting our framework that expands SpADANN with a layer fusion approach of Sentinel-2 and Sentinel-1 SITS, takes advantage of the spatio-temporal features related to the underlying multi-sensor remote sensing data, and is able to improve the performance compared to single-sensor data. The UDA setting is of great importance due to their ability to transfer a model to an unlabeled domain, resulting in saving resources and having products in a timely manner. To be able to use as much information as possible in an integrated manner helps to improve the performances of land cover maps. In this regard, there is already huge advances in having analysis ready data in remote sensing, but the quality of reference data is a factor that cannot be overlooked. As we have seen in our results, underrepresented classes that are a group of other sub-classes can introduce confusion in the classification, specially if the class is not well defined, or if the class is so broad that introduces sub-classes with different spectral and temporal signatures.

Further research could explore the possibility to include other modalities of remote sensing data (e.g. very high resolution images, elevation) or in general earth observation data (e.g. weather re-analysis) that can provide additional context to the domains, or even non-remote sensing data in the form of metadata. Also, the scenario of multi-source unsupervised domain adaptation remains as a possible extension of the SpADANN framework, which could be able to deal with the issue of classes than change definition, or group different sub-classes, in different years. Finally, another plausible scenario is to have limited reference data on some important classes in the target domain, there, a semi-supervised domain adaptation framework could be able to use data from both the source and target domain jointly.

References

Blickensdörfer, L., Schwieder, M., Pflugmacher, D., Nendel, C., Erasmi, S., & Hostert, P. (2022, February). Mapping of crop types and crop sequences with combined time series of Sentinel-1, Sentinel-2 and Landsat 8 data for Germany. *Remote Sensing of Environment*, *269*, 112831. Retrieved 2023-04-26, from https://www.sciencedirect.com/science/article/pii/S0034425721005514 doi: 10.1016/j.rse.2021.112831

Breiman, L. (2001, October). Random Forests. *Machine Learning*, *45*(1), 5–32. Retrieved 2023-05-09, from https://doi.org/10.1023/A:1010933404324 doi: 10.1023/A:1010933404324

Capliez, E., Ienco, D., Gaetano, R., Baghdadi, N., & Salah, A. H. (2023). Temporal Domain Adaptation for Satellite Image Time Series Land Cover Mapping With Adversarial Learning and Spatially-Aware Self-Training. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 1–36. (Conference Name: IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing) doi: 10.1109/JSTARS.2023.3263755

Chapelle, O., Scholkopf, B., & Zien, A., Eds. (2009, March). Semi-Supervised Learning (Chapelle, O. et al., Eds.; 2006) [Book reviews]. *IEEE Transactions on Neural Networks*, *20*(3), 542–542. (Conference Name: IEEE Transactions on Neural Networks) doi: 10.1109/TNN.2009.2015974

Chen, M., Weinberger, K. Q., & Blitzer, J. (2011). Co-Training for Domain Adaptation. In *Advances in Neural Information Processing Systems* (Vol. 24). Curran Associates, Inc. Retrieved 2023-06-13, from https://papers.nips.cc/paper_files/paper/2011/hash/93fb9d4b16aa750c7475b6d601c35c2c-Abstract.html

Chen, M., Zhao, S., Liu, H., & Cai, D. (2020, April). Adversarial-Learned Loss for Domain Adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(04), 3521–3528. Retrieved 2023-06-13, from https://ojs.aaai.org/index.php/AAAI/article/view/5757 (Number: 04) doi: 10.1609/aaai.v34i04.5757

Defourny, P., Bontemps, S., Bellemans, N., Cara, C., Dedieu, G., Guzzonato, E., … Koetz, B. (2019, February). Near real-time agriculture monitoring at national scale at parcel resolution: Performance assessment of the Sen2-Agri automated system in various cropping systems around the world. *Remote Sensing of Environment*, *221*, 551–568. Retrieved 2023-05-18, from https://www.sciencedirect.com/science/article/pii/S0034425718305145 doi: 10.1016/j.rse.2018.11.007

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., … Bargellini, P. (2012, May). Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sensing of Environment*, *120*, 25–36. Retrieved 2023-06-10, from https://www.sciencedirect.com/science/article/pii/S0034425712000636 doi: 10.1016/j.rse.2011.11.026

d'Andrimont, R., Verhegghen, A., Lemoine, G., Kempeneers, P., Meroni, M., & van der Velde, M. (2021, December). From parcel to continental scale – A first European crop type map based on Sentinel-1 and LUCAS Copernicus in-situ observations. *Remote Sensing of Environment*, *266*, 112708. Retrieved 2023-05-19, from https://www.sciencedirect.com/science/article/pii/S0034425721004284 doi: 10.1016/j.rse.2021.112708

FAO, IFAD, UNICEF, WFP, & WHO. (2022). *The State of Food Security and Nutrition in the World 2022: Repurposing food and agricultural policies to make healthy diets more affordable* (No. 2022). Rome, Italy: FAO, IFAD, UNICEF, WFP, WHO. Retrieved 2023-06-10, from https://www.fao.org/documents/card/en/c/cc0639en doi: 10.4060/cc0639en

Fawaz, H. I., Lucas, B., Forestier, G., Pelletier, C., Schmidt, D. F., Weber, J., … Petitjean, F. (2020, November). InceptionTime: Finding AlexNet for Time Series Classification. *Data Mining and Knowledge Discovery*, *34*(6), 1936–1962. Retrieved 2023-03-03, from http://arxiv.org/abs/1909.04939 (arXiv:1909.04939 [cs, stat]) doi: 10.1007/s10618-020-00710-y

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., … Lempitsky, V. (2016, May). *Domain-Adversarial Training of Neural Networks.* arXiv. Retrieved 2023-01-10, from http://arxiv.org/abs/1505.07818 (arXiv:1505.07818 [cs, stat])

Garnot, V. S. F., Landrieu, L., Giordano, S., & Chehata, N. (2019, November). *Satellite Image Time Series Classification with Pixel-Set Encoders and Temporal Self-Attention.* arXiv. Retrieved 2023-05-28, from http://arxiv.org/abs/1911.07757 (arXiv:1911.07757 [cs]) doi: 10.48550/arXiv.1911.07757

Gbodjo, Y. J. E., Montet, O., Ienco, D., Gaetano, R., & Dupuy, S. (2021). Multisensor Land Cover Classification With Sparsely Annotated Data Based on Convolutional Neural Networks and Self-Distillation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *14*, 11485–11499. (Conference Name: IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing) doi: 10.1109/JSTARS.2021.3119191

Hong, D., Gao, L., Yokoya, N., Yao, J., Chanussot, J., Du, Q., & Zhang, B. (2021, May). More Diverse Means Better: Multimodal Deep Learning Meets Remote-Sensing Imagery Classification. *IEEE Transactions on Geoscience and Remote Sensing*, *59*(5), 4340–4354. (Conference Name: IEEE Transactions on Geoscience and Remote Sensing) doi: 10.1109/TGRS.2020.3016820

Ienco, D., Interdonato, R., Gaetano, R., & Ho Tong Minh, D. (2019, December). Combining Sentinel-1 and Sentinel-2 Satellite Image Time Series for land cover mapping via a multi-source deep learning architecture. *ISPRS Journal of Photogrammetry and Remote Sensing*, *158*, 11–22. Retrieved 2023-04-17, from `https://www.sciencedirect.com/science/article/pii/S0924271619302278` doi: 10.1016/j.isprsjprs.2019.09.016

Jolivot, A., Lebourgeois, V., Leroux, L., Ameline, M., Andriamanga, V., Bellón, B., … Bégué, A. (2021, December). Harmonized in situ datasets for agricultural land use mapping and monitoring in tropical countries. *Earth System Science Data*, *13*(12), 5951–5967. Retrieved 2023-05-29, from `https://essd.copernicus.org/articles/13/5951/2021/` (Publisher: Copernicus GmbH) doi: 10.5194/essd-13-5951-2021

Karasiak, N., Dejoux, J.-F., Monteil, C., & Sheeren, D. (2022, July). Spatial dependence between training and test sets: another pitfall of classification accuracy assessment in remote sensing. *Machine Learning*, *111*(7), 2715–2740. Retrieved 2023-05-11, from `https://doi.org/10.1007/s10994-021-05972-1` doi: 10.1007/s10994-021-05972-1

Khatami, R., Mountrakis, G., & Stehman, S. V. (2016, May). A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research. *Remote Sensing of Environment*, *177*, 89–100. Retrieved 2023-05-28, from `https://www.sciencedirect.com/science/article/pii/S0034425716300578` doi: 10.1016/j.rse.2016.02.028

Li, J., Hong, D., Gao, L., Yao, J., Zheng, K., Zhang, B., & Chanussot, J. (2022, August). Deep learning in multimodal remote sensing data fusion: A comprehensive review. *International Journal of Applied Earth Observation and Geoinformation*, *112*, 102926. Retrieved 2023-04-26, from `https://www.sciencedirect.com/science/article/pii/S1569843222001248` doi: 10.1016/j.jag.2022.102926

Li, R., Zheng, S., Duan, C., Wang, L., & Zhang, C. (2022, April). Land cover classification from remote sensing images based on multi-scale fully convolutional network. *Geo-spatial Information Science*, *25*(2), 278–294. Retrieved 2023-06-13, from https://doi.org/10.1080/10095020.2021.2017237 (Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10095020.2021.2017237) doi: 10.1080/10095020.2021.2017237

Liu, X., Yoo, C., Xing, F., Oh, H., Fakhri, G. E., Kang, J.-W., & Woo, J. (2022, August). Deep Unsupervised Domain Adaptation: A Review of Recent Advances and Perspectives. *APSIPA Transactions on Signal and Information Processing*, *11*(1). Retrieved 2023-05-16, from https://www.nowpublishers.com/article/Details/SIP-2022-0019 (Publisher: Now Publishers, Inc.) doi: 10.1561/116.00000192

Maaten, L. v. d. (2014). Accelerating t-SNE using Tree-Based Algorithms. *Journal of Machine Learning Research*, *15*(93), 3221–3245. Retrieved 2023-06-11, from http://jmlr.org/papers/v15/vandermaaten14a.html

Morerio, P., Volpi, R., Ragonesi, R., & Murino, V. (2020, January). *Generative Pseudo-label Refinement for Unsupervised Domain Adaptation.* arXiv. Retrieved 2023-06-13, from http://arxiv.org/abs/2001.02950 (arXiv:2001.02950 [cs]) doi: 10.48550/arXiv.2001.02950

Ndikumana, E., Ho Tong Minh, D., Baghdadi, N., Courault, D., & Hossard, L. (2018, August). Deep Recurrent Neural Network for Agricultural Classification using multitemporal SAR Sentinel-1 for Camargue, France. *Remote Sensing*, *10*(8), 1217. Retrieved 2023-05-28, from https://www.mdpi.com/2072-4292/10/8/1217 (Number: 8 Publisher: Multidisciplinary Digital Publishing Institute) doi: 10.3390/rs10081217

Nyborg, J., Pelletier, C., Lefèvre, S., & Assent, I. (2022, June). TimeMatch: Unsupervised cross-region adaptation by temporal shift estimation. *ISPRS Journal of Photogrammetry and Remote Sensing*, *188*, 301–313. Retrieved 2023-04-20, from https://www.sciencedirect.com/science/article/pii/S0924271622001216 doi: 10.1016/j.isprsjprs.2022.04.018

Ofori-Ampofo, S., Pelletier, C., & Lang, S. (2021, January). Crop Type Mapping from Optical and Radar Time Series Using Attention-Based Deep Learning. *Remote Sensing*, *13*(22), 4668. Retrieved 2023-04-17, from https://www.mdpi.com/2072-4292/13/22/4668 (Number: 22 Publisher: Multidisciplinary Digital Publishing Institute) doi: 10.3390/rs13224668

Pelletier, C., Valero, S., Inglada, J., Champion, N., & Dedieu, G. (2016, December). Assessing the robustness of Random Forests to map land cover with high resolution satellite image time series over large areas. *Remote Sensing of Environment*, *187*, 156–168. Retrieved 2023-06-09, from `https://www.sciencedirect.com/science/article/pii/S0034425716303820` doi: 10.1016/j.rse.2016.10.010

Pelletier, C., Webb, G. I., & Petitjean, F. (2019, January). Temporal Convolutional Neural Network for the Classification of Satellite Image Time Series. *Remote Sensing*, *11*(5), 523. Retrieved 2023-03-03, from `https://www.mdpi.com/2072-4292/11/5/523` (Number: 5 Publisher: Multidisciplinary Digital Publishing Institute) doi: 10.3390/rs11050523

Phiri, D., Simwanda, M., Salekin, S., Nyirenda, V. R., Murayama, Y., & Ranagalage, M. (2020, January). Sentinel-2 Data for Land Cover/Use Mapping: A Review. *Remote Sensing*, *12*(14), 2291. Retrieved 2023-06-10, from `https://www.mdpi.com/2072-4292/12/14/2291` (Number: 14 Publisher: Multidisciplinary Digital Publishing Institute) doi: 10.3390/rs12142291

Pickson, R. B., & Boateng, E. (2022, March). Climate change: a friend or foe to food security in Africa? *Environment, Development and Sustainability*, *24*(3), 4387–4412. Retrieved 2023-06-10, from `https://doi.org/10.1007/s10668-021-01621-8` doi: 10.1007/s10668-021-01621-8

Rußwurm, M., & Körner, M. (2018, April). Multi-Temporal Land Cover Classification with Sequential Recurrent Encoders. *ISPRS International Journal of Geo-Information*, *7*(4), 129. Retrieved 2023-03-28, from `https://www.mdpi.com/2220-9964/7/4/129` (Number: 4 Publisher: Multidisciplinary Digital Publishing Institute) doi: 10.3390/ijgi7040129

Rußwurm, M., Pelletier, C., Zollner, M., Lefèvre, S., & Körner, M. (2020, May). *BreizhCrops: A Time Series Dataset for Crop Type Mapping*. arXiv. Retrieved 2023-03-20, from `http://arxiv.org/abs/1905.11893` (arXiv:1905.11893 [cs, stat]) doi: 10.48550/arXiv.1905.11893

Sainte Fare Garnot, V., Landrieu, L., & Chehata, N. (2022, May). Multi-modal temporal attention models for crop mapping from satellite time series. *ISPRS Journal of Photogrammetry and Remote Sensing*, *187*, 294–305. Retrieved 2023-06-09, from `https://www.sciencedirect.com/science/article/pii/S0924271622000855` doi: 10.1016/j.isprsjprs.2022.03.012

Saito, K., Ushiku, Y., & Harada, T. (2017, July). Asymmetric Tri-training for Unsupervised Domain Adaptation. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 2988–2997). PMLR. Retrieved 2023-06-13, from `https://proceedings.mlr.press/v70/saito17a.html` (ISSN: 2640-3498)

Sohn, K., Berthelot, D., Li, C.-L., Zhang, Z., Carlini, N., Cubuk, E. D., … Raffel, C. (2020, November). *FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence.* arXiv. Retrieved 2023-06-13, from http://arxiv.org/abs/2001.07685 (arXiv:2001.07685 [cs, stat]) doi: 10.48550/arXiv.2001.07685

Son, N.-T., Chen, C.-F., Chen, C.-R., & Minh, V.-Q. (2018, June). Assessment of Sentinel-1A data for rice crop classification using random forests and support vector machines. *Geocarto International*, *33*(6), 587–601. Retrieved 2023-06-09, from https://doi.org/10.1080/10106049.2017.1289555 (Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10106049.2017.1289555) doi: 10.1080/10106049.2017.1289555

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015, June). Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1–9). (ISSN: 1063-6919) doi: 10.1109/CVPR.2015.7298594

Tardy, B., Inglada, J., & Michel, J. (2017, November). Fusion Approaches for Land Cover Map Production Using High Resolution Image Time Series without Reference Data of the Corresponding Period. *Remote Sensing*, *9*(11), 1151. Retrieved 2023-05-18, from https://www.mdpi.com/2072-4292/9/11/1151 (Number: 11 Publisher: Multidisciplinary Digital Publishing Institute) doi: 10.3390/rs9111151

Tardy, B., Inglada, J., & Michel, J. (2019, January). 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*, *11*(9), 1047. Retrieved 2023-05-22, from https://www.mdpi.com/2072-4292/11/9/1047 (Number: 9 Publisher: Multidisciplinary Digital Publishing Institute) doi: 10.3390/rs11091047

Tscharntke, T., Grass, I., Wanger, T. C., Westphal, C., & Batáry, P. (2021, October). Beyond organic farming – harnessing biodiversity-friendly landscapes. *Trends in Ecology & Evolution*, *36*(10), 919–930. Retrieved 2023-06-10, from https://www.sciencedirect.com/science/article/pii/S016953472100183X doi: 10.1016/j.tree.2021.06.010

Tuia, D., Persello, C., & Bruzzone, L. (2016, June). Domain Adaptation for the Classification of Remote Sensing Data: An Overview of Recent Advances. *IEEE Geoscience and Remote Sensing Magazine*, *4*(2), 41–57. (Conference Name: IEEE Geoscience and Remote Sensing Magazine) doi: 10.1109/MGRS.2016.2548504

Wang, Z., Zhang, H., He, W., & Zhang, L. (2021, June). Phenology Alignment Network: A Novel Framework for Cross-Regional Time Series Crop Classification. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 2934–2943). Nashville, TN, USA: IEEE. Retrieved 2023-05-22, from `https://ieeexplore.ieee.org/document/9522702/` doi: 10.1109/CVPRW53098.2021.00329

Wardlow, B. D., & Egbert, S. L. (2008, March). Large-area crop mapping using time-series MODIS 250 m NDVI data: An assessment for the U.S. Central Great Plains. *Remote Sensing of Environment*, *112*(3), 1096–1116. Retrieved 2023-06-10, from `https://www.sciencedirect.com/science/article/pii/S0034425707003458` doi: 10.1016/j.rse.2007.07.019

Wilson, G., & Cook, D. J. (2020, October). A Survey of Unsupervised Deep Domain Adaptation. *ACM Transactions on Intelligent Systems and Technology*, *11*(5), 1–46. Retrieved 2023-05-16, from `https://dl.acm.org/doi/10.1145/3400066` doi: 10.1145/3400066

You, L., Wood, S., Wood-Sichra, U., & Wu, W. (2014, May). Generating global crop distribution maps: From census to grid. *Agricultural Systems*, *127*, 53–60. Retrieved 2023-06-10, from `https://www.sciencedirect.com/science/article/pii/S0308521X14000110` doi: 10.1016/j.agsy.2014.01.002

Yuan, Y., Lin, L., Liu, Q., Hang, R., & Zhou, Z.-G. (2022, February). SITS-Former: A pre-trained spatio-spectral-temporal representation model for Sentinel-2 time series classification. *International Journal of Applied Earth Observation and Geoinformation*, *106*, 102651. Retrieved 2023-06-13, from `https://www.sciencedirect.com/science/article/pii/S0303243421003585` doi: 10.1016/j.jag.2021.102651

Zhong, L., Hu, L., & Zhou, H. (2019, February). Deep learning based multi-temporal crop classification. *Remote Sensing of Environment*, *221*, 430–443. Retrieved 2023-03-14, from `https://www.sciencedirect.com/science/article/pii/S0034425718305418` doi: 10.1016/j.rse.2018.11.032

Zhou, T., Ruan, S., & Canu, S. (2019, September). A review: Deep learning for medical image segmentation using multi-modality fusion. *Array*, *3-4*, 100004. Retrieved 2023-05-01, from `https://www.sciencedirect.com/science/article/pii/S2590005619300049` doi: 10.1016/j.array.2019.100004

Zou, Y., Yu, Z., Liu, X., Kumar, B. V. K. V., & Wang, J. (2020, July). *Confidence Regularized Self-Training*. arXiv. Retrieved 2023-06-13, from `http://arxiv.org/abs/1908.09822` (arXiv:1908.09822 [cs]) doi: 10.48550/arXiv.1908.09822

Appendix

A Supplementary material

Table 10: Class-wise F1-score for the transfer task 2018→2020

Class		D_s			**UDA**			D_t	
name	**TempCNN$_{S2}$**	**CNN$_{S1}$**	**lf(S2+S1)**	**SpADANN$_{S2}$**	**SpADANN$_{S1}$**	**mmSpADANN$_{\upsilon3}$**	**TempCNN$_{S2}$**	**CNN$_{S1}$**	**lf(S2+S1)**
■ Cereals	**60.57 ± 6.26**	48.14 ± 6.46	49.31 ± 15.17	80.67 ± 0.37	**85.06 ± 0.19**	84.73 ± 0.32	**89.82 ± 5.84**	84.43 ± 3.60	84.75 ± 5.49
■ Cotton	56.40 ± 6.48	58.45 ± 5.80	**62.04 ± 8.95**	49.53 ± 0.40	**49.98 ± 0.76**	49.44 ± 0.25	**92.21 ± 2.70**	90.62 ± 3.43	90.71 ± 6.09
■ Oleag./Legum.	37.27 ± 6.77	25.06 ± 8.27	**50.94 ± 9.47**	56.12 ± 1.52	61.80 ± 0.28	**62.03 ± 0.36**	**85.54 ± 4.99**	78.52 ± 3.94	83.03 ± 6.73
■ Grassland	79.31 ± 6.17	66.07 ± 7.33	**83.38 ± 7.09**	96.62 ± 0.31	97.46 ± 0.06	**97.65 ± 0.13**	89.37 ± 7.29	90.83 ± 1.72	**92.21 ± 2.36**
■ Shrubland	64.07 ± 27.07	**74.57 ± 3.28**	72.20 ± 15.20	97.47 ± 0.25	**98.52 ± 0.08**	98.43 ± 0.22	88.34 ± 10.05	**91.16 ± 3.05**	89.54 ± 5.27
■ Forest	62.08 ± 7.21	52.87 ± 23.38	**71.43 ± 15.35**	97.74 ± 0.52	**98.12 ± 0.11**	98.02 ± 0.28	**86.07 ± 10.96**	84.86 ± 12.08	82.26 ± 10.50
■ Baresoil	53.58 ± 13.37	33.38 ± 15.93	**54.49 ± 16.54**	83.07 ± 0.24	83.77 ± 1.47	**84.79 ± 1.85**	**81.07 ± 23.88**	78.98 ± 12.89	77.11 ± 16.35
■ Water	**99.79 ± 0.46**	99.27 ± 1.07	99.62 ± 0.76	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
Total	69.14 ± 7.37	66.22 ± 7.88	74.41 ± 1.15	88.79 ± 0.15	89.38 ± 0.18	89.30 ± 0.07	90.08 ± 2.85	89.73 ± 1.64	88.68 ± 2.12

Table 11: Class-wise F1-score for the transfer task 2020→2021

Class		D_s			**UDA**			D_t	
name	**TempCNN$_{S2}$**	**CNN$_{S1}$**	**lf(S2+S1)**	**SpADANN$_{S2}$**	**SpADANN$_{S1}$**	**mmSpADANN$_{\upsilon3}$**	**TempCNN$_{S2}$**	**CNN$_{S1}$**	**lf(S2+S1)**
■ Cereals	60.80 ± 5.47	52.62 ± 7.39	**62.61 ± 6.56**	86.37 ± 0.50	91.08 ± 0.50	**91.20 ± 0.34**	**90.70 ± 1.90**	85.99 ± 3.54	88.17 ± 4.21
■ Cotton	**59.28 ± 14.44**	56.42 ± 12.83	55.69 ± 12.34	74.40 ± 1.02	77.26 ± 0.69	**77.51 ± 0.29**	87.91 ± 5.56	87.44 ± 4.62	**90.08 ± 7.21**
■ Oleag./Legum.	30.76 ± 7.88	27.55 ± 7.66	**47.28 ± 7.48**	82.93 ± 0.37	86.36 ± 0.50	**86.63 ± 0.13**	**85.73 ± 4.26**	80.66 ± 2.90	79.64 ± 1.62
■ Grassland	62.29 ± 13.61	52.88 ± 9.61	**78.28 ± 7.53**	**92.73 ± 0.43**	92.33 ± 0.35	92.30 ± 0.22	88.04 ± 5.69	89.16 ± 2.29	**90.27 ± 2.10**
■ Shrubland	57.51 ± 9.79	51.00 ± 6.73	**68.27 ± 17.91**	94.70 ± 0.05	95.85 ± 0.55	**96.11 ± 0.10**	**92.92 ± 2.63**	88.03 ± 3.63	92.04 ± 2.87
■ Forest	61.40 ± 17.65	57.18 ± 16.46	**68.29 ± 19.78**	97.32 ± 0.12	97.56 ± 1.21	**98.16 ± 0.15**	**91.24 ± 8.02**	81.70 ± 17.61	88.21 ± 10.48
■ Baresoil	**72.17 ± 20.06**	23.60 ± 20.32	69.01 ± 10.14	93.46 ± 0.58	96.94 ± 1.96	**97.23 ± 1.77**	**88.44 ± 10.29**	80.48 ± 24.46	83.21 ± 23.25
■ Water	99.78 ± 0.43	98.45 ± 1.49	**99.95 ± 0.10**	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	98.82 ± 2.04	96.70 ± 5.12	**99.09 ± 0.81**
Total	64.91 ± 7.08	60.33 ± 8.23	75.38 ± 6.89	92.71 ± 0.18	92.93 ± 0.23	93.06 ± 0.07	92.95 ± 1.61	89.73 ± 1.60	91.58 ± 1.57

Table 12: F1-score 2018 → 2020

Method	**Weighted**	**Micro**	**Macro**
SpADANN$_{S2}$	88.79 ± 0.15	87.49 ± 0.21	57.34 ± 1.12
SpADANN$_{S1}$	89.38 ± 0.18	89.13 ± 0.09	76.90 ± 2.38
mmSpADANN$_{\upsilon3}$	89.30 ± 0.07	89.05 ± 0.10	75.09 ± 1.69

Table 13: F1-score 2020 → 2021

Method	**Weighted**	**Micro**	**Macro**
SpADANN$_{S2}$	92.71 ± 0.18	91.07 ± 0.18	55.52 ± 1.80
SpADANN$_{S1}$	92.93 ± 0.23	92.70 ± 0.45	78.15 ± 5.95
mmSpADANN$_{v3}$	93.06 ± 0.07	92.96 ± 0.12	81.44 ± 3.08

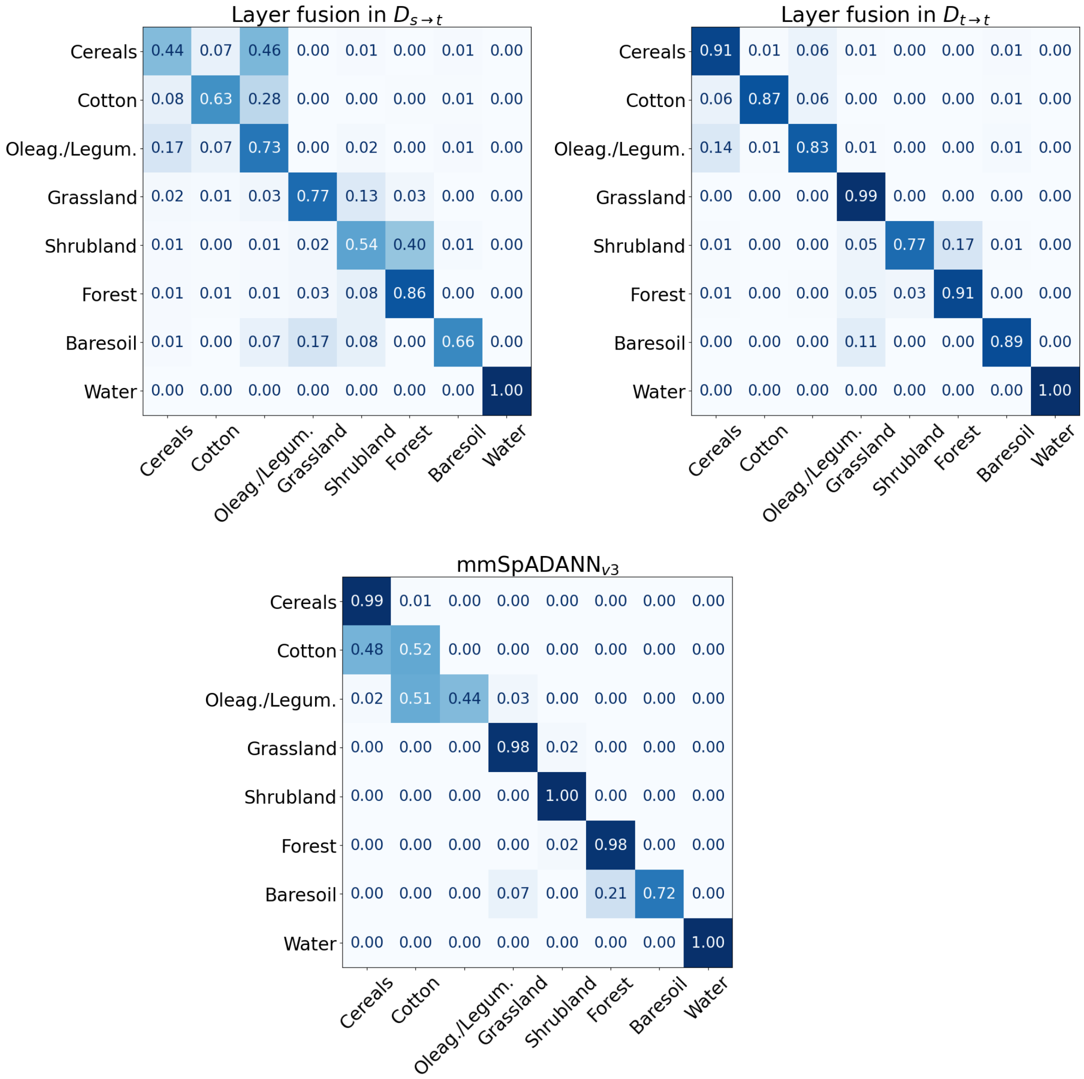


Figure 16: Confusion matrix 2018 → 2020

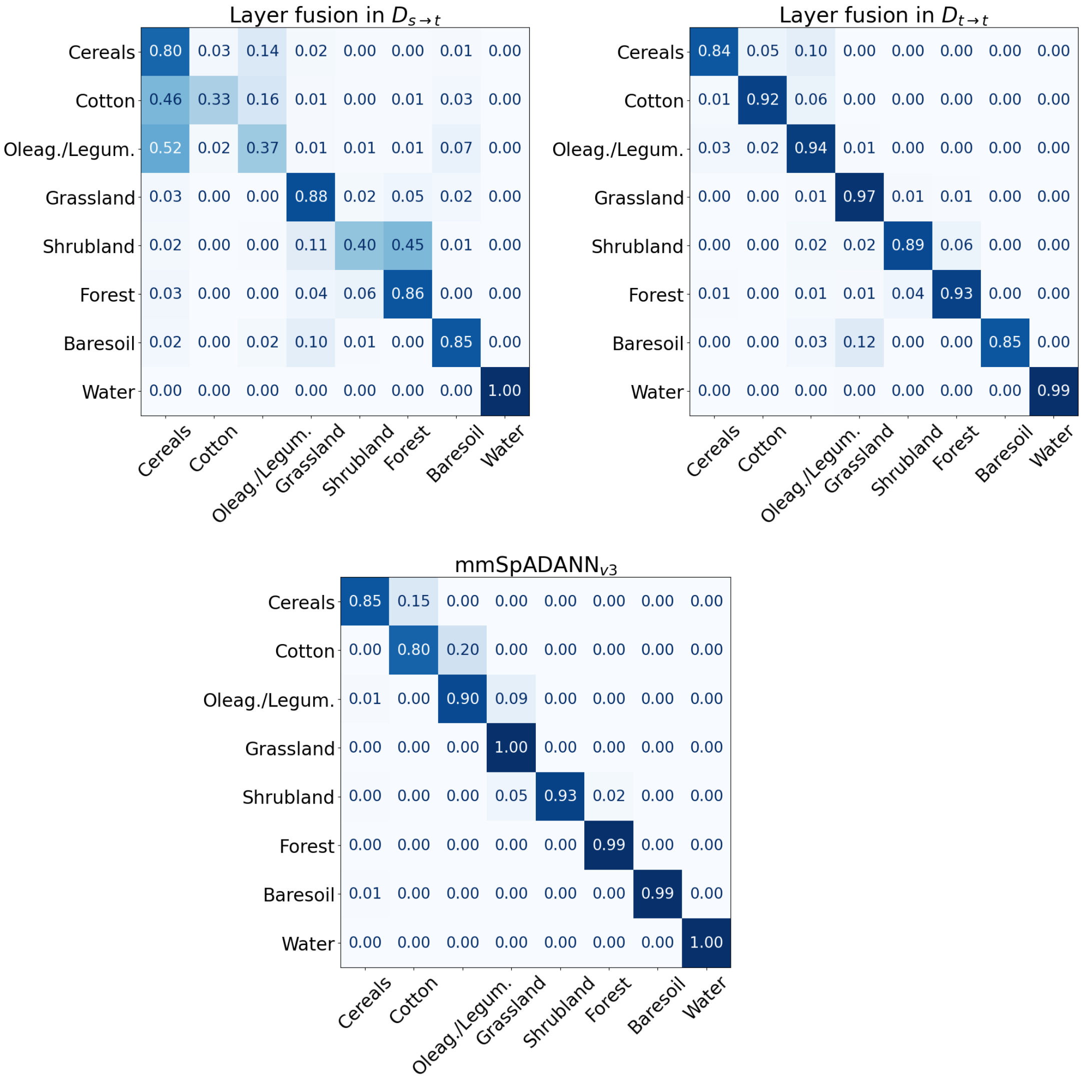
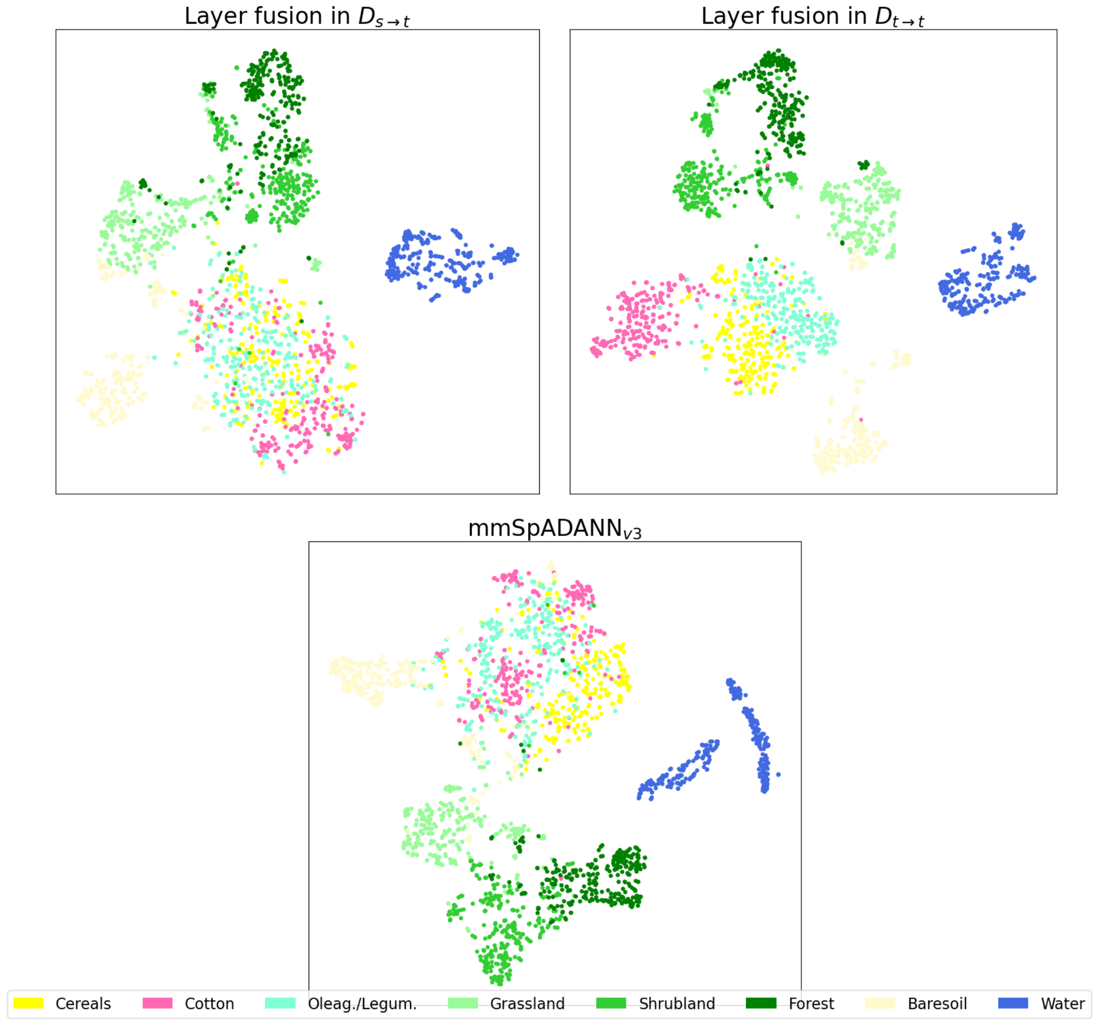


Figure 17: Confusion matrix 2020 → 2021



Layer fusion in $D_{s\to t}$

Layer fusion in $D_{t\to t}$

mmSpADANN$_{v3}$

Figure 18: t-SNE 2018 → 2020

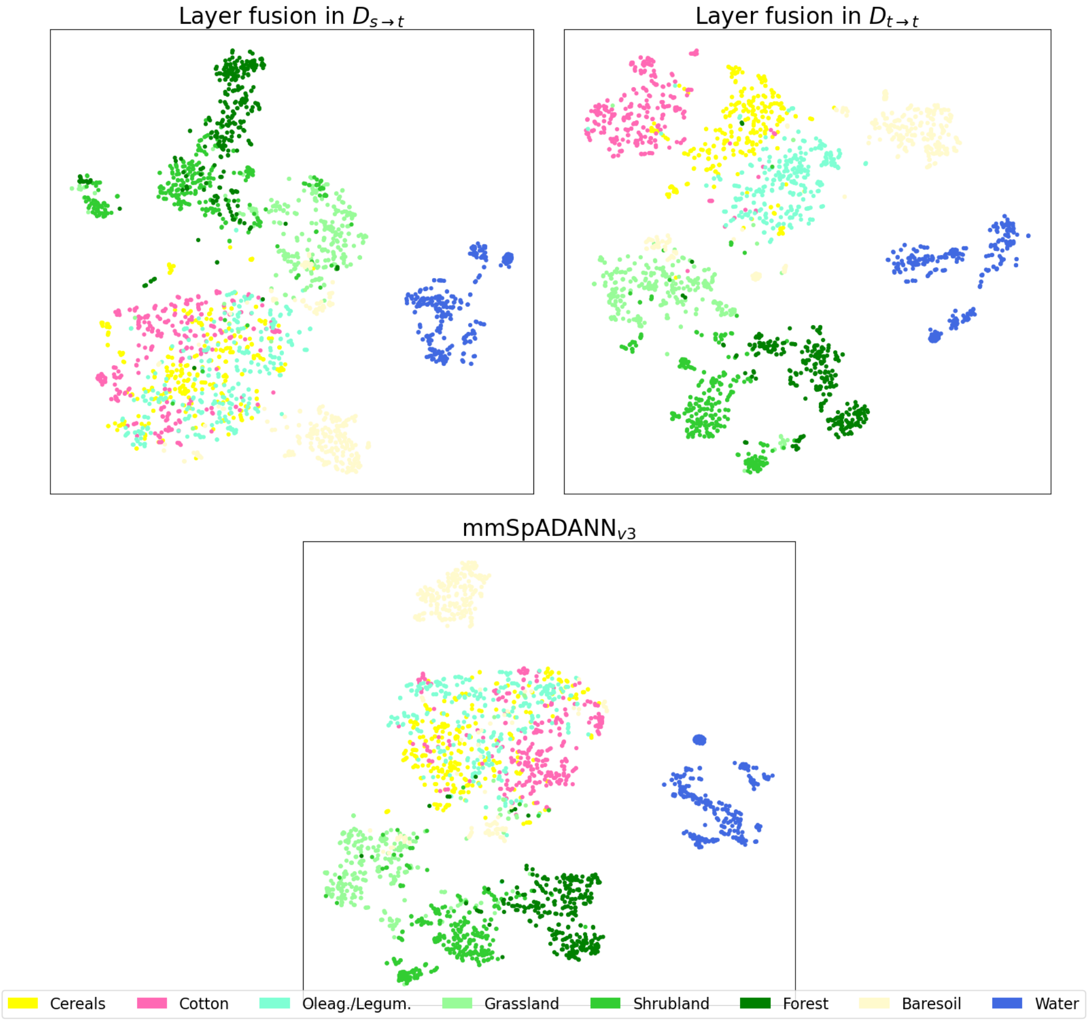


Figure 19: t-SNE 2020 → 2021