Domain-Adversarial Training of Self-Attention-Based Networks for Land Cover Classification Using Multi-Temporal Sentinel-2 Satellite Imagery

remote sensing

MDPI

Article

Domain-Adversarial Training of Self-Attention-Based Networks for Land Cover Classification Using Multi-Temporal Sentinel-2 Satellite Imagery

Mauro Martini [1,2], Vittorio Mazzia [1,2,3], Aleem Khaliq [2,4,*] and Marcello Chiaberge [1,2]

1 Department of Electronics and Telecommunications, Politecnico di Torino, 10124 Turin, Italy; mauro.martini@polito.it (M.M.); vittorio.mazzia@polito.it (V.M.); marcello.chiaberge@polito.it (M.C.)
2 PIC4SeR, Interdepartmental Centre for Service Robotics, Politecnico di Torino, 10129 Turin, Italy
3 SmartData@PoliTo, Big Data and Data Science Laboratory, 10129 Turin, Italy
4 Department of Electrical Engineering, International Islamic University, Islamabad 44000, Pakistan
* Correspondence: aleem.khaliq@polito.it; Tel.: +39-389-041-9074

Abstract: The increasing availability of large-scale remote sensing labeled data has prompted researchers to develop increasingly precise and accurate data-driven models for land cover and crop classification (LC&CC). Moreover, with the introduction of self-attention and introspection mechanisms, deep learning approaches have shown promising results in processing long temporal sequences in the multi-spectral domain with a contained computational request. Nevertheless, most practical applications cannot rely on labeled data, and in the field, surveys are a time-consuming solution that pose strict limitations to the number of collected samples. Moreover, atmospheric conditions and specific geographical region characteristics constitute a relevant domain gap that does not allow direct applicability of a trained model on the available dataset to the area of interest. In this paper, we investigate adversarial training of deep neural networks to bridge the domain discrepancy between distinct geographical zones. In particular, we perform a thorough analysis of domain adaptation applied to challenging multi-spectral, multi-temporal data, accurately highlighting the advantages of adapting state-of-the-art self-attention-based models for LC&CC to different target zones where labeled data are not available. Extensive experimentation demonstrated significant performance and generalization gain in applying domain-adversarial training to source and target regions with marked dissimilarities between the distribution of extracted features.

Keywords: domain adaptation; Transformers; deep learning; land cover classification

Citation: Martini, M.; Mazzia, V.; Khaliq, A.; Chiaberge, M. Domain-Adversarial Training of Self-Attention-Based Networks for Land Cover Classification Using Multi-Temporal Sentinel-2 Satellite Imagery. *Remote Sens.* **2021**, *13*, 2564. https://doi.org/10.3390/rs13132564

Academic Editor: Dino Ienco

Received: 31 March 2021
Accepted: 28 June 2021
Published: 30 June 2021

1. Introduction

In the past few decades, the launch of many satellite missions with short revisit time and comparatively high-resolution sensors has offered an extensive repository of remote sensing images. Availability of the open-source data by many Earth-observation satellites has made remote sensing very easy and obtainable [1]. Open-source data sets are available free of cost from several satellite missions such as the Sentinel-2 and Landsat [2]. These satellites are equipped with multi-spectral sensors with short revisit time, and good spatial and spectral resolution, allowing researchers to test modern image analysis techniques to extract more detailed information of the target object. It is quite possible to monitor the dynamic processes on Earth [3,4]. Additionally, it has become easier to estimate and classify biophysical parameters using several data sources [5–7]. Overall, the new scenario has led to the opportunity for the land cover monitoring, change detection, image mosaicking, and large-scale processing using multi-temporal and multi-source images [1,8–10].

The most essential and critical remote sensing application is land cover and crops classification (LC&CC). It facilitates labeling the cover such as forest, ocean, and agricultural land. Moreover, mapping can also be done manually using satellite images, but the

process is quite tedious, costly, and time-consuming. Finally, an exquisite global cover map is not available as yet, but there is a land cover map with the name Corine Land Cover (CLC) [11] which provides land cover information with 100 m per pixel resolution. However, the problem with this map is that it only covers the European area and is updated once in six years. There are several ways to perform land classification automatically. In general, the classification involves the creation of a training dataset that consists of annotated samples of the corresponding class labels, training a model using the training dataset, and evaluating the resulting predictions. The number and quality of training samples play a pivotal role in defining the performance of the trained model. From a remote sensing prospective, training sample collection requires a ground survey or visual photo-interpretation by an expert [12]. Ground surveying involves GIS expert knowledge, human resource that is not typically economical, while visual interpretation is not appropriate to be used for some applications, such as finding chlorophyll concentration [13] and classification of tree species [14]. Most of the machine learning (ML) algorithms such as random forest, support vector machines, logistic regression performs well in the context of classification of remote sensing images. However, performance of these ML algorithms are not satisfied when learning features from different sources such as active and passive sensors [15]. It was shown in [16,17] that Convolutional Neural Networks (CNN) are better than traditional land cover classification techniques. In the land segmentation section of the deep globe challenge [18], the Deep Neural networks (DNN) completely dominate the leaderboards. The best examples of land cover classification using Deep Neural Networks are ResNet and DenseNet [19,20].

Since there is a difference in the land covers of different locations, the model trained in one area cannot be deployed for the other areas. Additionally, the satellite imagery of different satellites is not the same. That phenomenon is due to the difference in their resolution, capture time, and other radiometric parameters. Due to these multiple changing variables, the dataset taken from a satellite covering one region and another satellite dataset covering the same or other regions leads to a domain shift between the datasets. One way to achieve a reliable outcome is possibly to train a model with a huge amount of training samples to generalize its behavior for all classes of all the regions. However, that needs an enormous labeled dataset that is time and labor-intensive.

Another method to deal with the shift between the datasets is termed Domain Adaptation (DA), in which a model is trained on one dataset (source data) and predictions are made on the other dataset (target domain). The distribution shift between the target and source dataset is mainly due to temporal differences in the acquisition, differences in the acquisition sensors, and geographical differences such as variations of objects at the Earth's surface. The domain shift affects the performance of a model trained on a source dataset and applied on the target dataset. Domain adaptation methods often rely on learning domain-invariant models that keep comparable performances on the two datasets. Existing domain adaptation techniques may be classified as supervised, unsupervised, and semisupervised. In supervised DA methods, it is presumed that labeled data are available for both source and target domains [21]. In a semisupervised domain, the labeled data for the target domain is assumed to be small while an unsupervised method contains labeled data for the source domain only. For example in [22], a semisupervised visual domain adaptation was proposed to address classification of very high-resolution remote sensing images. To deal with the variation in features distribution between the source and target domains, multiple kernel learning domain adaptation method was employed. Another example [23], in which domain adaptation based on semisupervised transfer component analysis was employed to extract features for knowledge transfer from source image to target image for land cover classification of remotely sensed images.

Tuia et al. divides the domain adaptation methodologies into four different categories: domain-invariant feature selection, adapting data distribution, adapting classifiers, and adaptive classifiers using active learning methods [12]. Many studies discuss the unsupervised domain adaptation in the context of classification and segmentation of the

remotely sensed satellite and aerial imagery. For example, In [24], an unsupervised adversarial domain adaptation method was proposed based on boosted domain confusion network (ADA-BDC) which focuses on feature extraction to enhance the transferability of classifier which is trained by source domain images and tested on target domain images. In [25], an unsupervised domain adaptation was used using generative adversarial networks (GANs)for semantic segmentation of aerial images. A multi-source domain adaptation (MDA) for scene classification was proposed to transfer knowledge from the multiple-source domains to the target domain [26]. Most of the studies presented in the literature related to DA-based classification have used single date images of source and target domain. However, in [27], first approach was proposed in the context of DA for classification of multi-temporal satellite images in which Bayesian classifier-based DA was employed with only two images of Landsat-5 satellite.

This work investigates adversarial training of deep neural networks to bridge the domain discrepancy between distinct geographical zones. In particular, we perform a thorough analysis of domain adaptation applied to challenging multi-spectral, multi-temporal data, highlighting the advantages of adapting state-of-the-art self-attention-based models for LC(&)CC to different target zones where labeled data are not available. We choose to experiment our methodology on the BreizhCrops dataset, a large-scale time series benchmark dataset introduced in 2020 by Rußwurm et al., [28], for supervised classification of field crops from satellite data. Figure 1 shows the visual representation of the crop prediction performed on a sub-region of Brittany, highlighting the benefit provided by the proposed methodology.

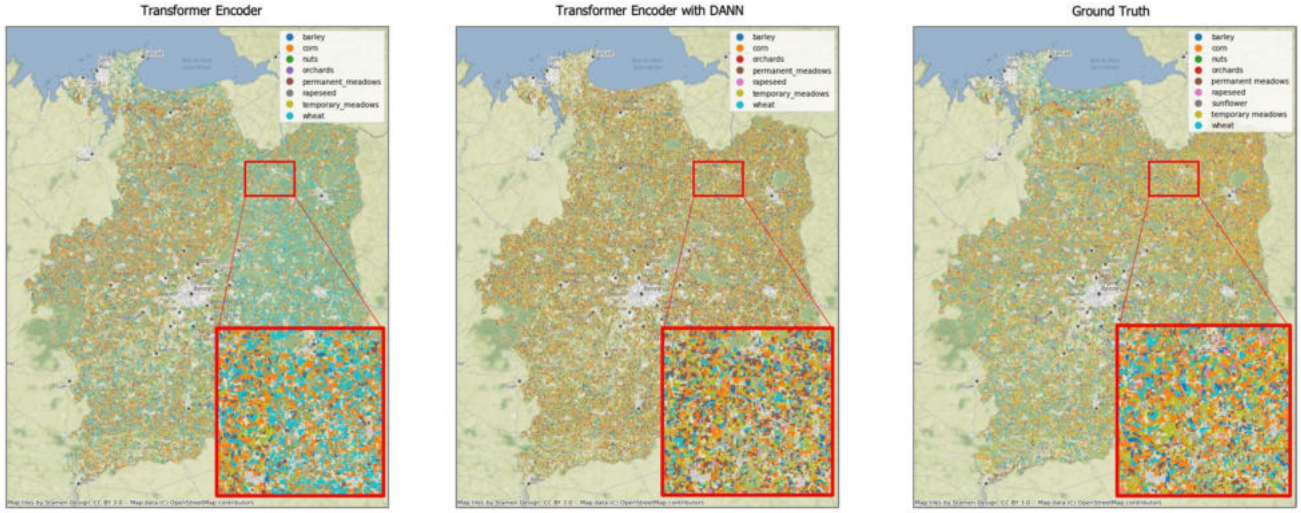

Figure 1. Visual representation of land crops classification on zone 3 (Ille-et-Vilain) of the BreizhCrops dataset. For each sub-image we show the complete region and a sub-area to facilitate the visualization of the advantage obtained by the proposed methodology. In particular, on the left the crops predictions without our domain adaptation mechanism are shown, while in the center the same predictions performed adopting DANN are proposed. On the right, ground truth labeled crops can be visualized. The improvement in the classification with DANN is evident, especially in the reduction of misclassification of wheat and meadows.

This article is organized as follows. Section 2 covers the related work on domain adaptation and its developments in techniques for LC&CC. Section 3 describes the dataset. A detailed description of the proposed method is presented in Section 4. The experimental setup, the results and related discussion are reported in Section 5. Finally, Section 6 draws some conclusions and future directions.

2. Related Work

2.1. Land Cover and Crop Classification

LC&CC has been the subject of many studies in the past. A widely used classification method makes use of time series of vegetation indices (VI) derived from remotely sensed imagery to extract temporal features and phenological metrics. There are also some thresholds and simple statistical techniques that help calculate the time of peak VI, Maximum VI, and other vegetation related metrics [29,30]. Moranduzzo et al. and Hao et al. [31,32] Illustrate the older image classification methods using handcrafted features for image representation and training classifiers such as support vector machine and random forest. Machine learning methods self-learn how to extract the features from the data with massive datasets available and improved computing devices. Random Forest (RF)-based classifiers is another common approach for remote sensing applications [32], though it should be noted that multiple features need to be derived and fed to the RF classifier for more effective output.

One of the newest and most powerful concepts integrated into mapping is a branch of machine learning known as Deep Learning (DL). DL is a type of machine learning based on artificial neural networks in which multiple layers of processing are used to extract progressively higher-level features from data. DL can be used to solve a wide range of problems such as signal processing, computer vision, image processing, and natural language processing [33]. DL has shown significant contribution in remote sensing image classification due to its ability to represent features and its competence of mechanization for end-to-end learning. Autoencoders are type of artificial neural networks and are often used to represent features of data [34,35]. In the remote sensing field, object detection and image segmentation have been performed extensively using two-dimensional CNNs [36,37] to perform spatial feature extraction from high-resolution images. 2D CNN proved better than 1D CNN in crop classification [38]. In remote sensing, two-dimensional CNN can be used effectively for image classification where the correlation between the morphological details and the target classes exists. For example in [39], a 2D-CNN is used to obtain the spatial features of the hyperspectral imagery (HSI), analyzing the continuity of land covers in the spatial domain. Often relation among spectral bands of HSI is not linear, in that case, 2D-CNNs are normally used together with 1D-CNNs to incorporate the spectral and spatial domain of features [40].

Indeed, the classification task becomes quite challenging when dealing with high-dimensional hyperspectral data with few labeled samples. Recently, generative adversarial networks (GANs) have been exploited for sample generation, though it is not easy to acquire high-quality samples with authenticity. In this context, the generative adversarial networks (GANs) aim to generate more labeled samples by mimicking labeled data and provide high-quality realistic data to increase the number of training samples [41]. Generally, GANs are comprised of two adversarial modules: a generator that obtains the original data distribution and a discriminator that differentiates between the generated labeled data and the original ones [42]. For this purpose, an unsupervised 1D GAN was aimed to capture the spectral distribution while increasing the training samples for HSI classification [43]. It was trained on unlabeled samples, which were then transformed as a classifier in a semisupervised setting. Hence, it is difficult to learn class features during the training process. Modified versions of GANs have considered the label information, such as conditional GAN (CGAN) [44], InfoGAN [45], deep convolutional GAN (DCGAN) [46], and categorical GAN (CatGAN) [47].

The aforementioned versions of GANs are susceptible to noises and disregard the relationships between spectral bands. Additionally, the generated samples are usually very different in the spectral domain from the original ones which fail in increasing classification results. This problem has been addressed in [48], authors developed a self-attention generative adversarial adaptation network (SaGAAN) to produce high-quality labeled samples in the spectral domain for hyperspectral image classification.

2.2. Domain Adaptation

The method of domain adaptation aims to reduce the domain shift between source and target datasets. Domain adaptation has three possible approaches according to [49–51]. The primary approach consists of reducing the difference in the feature space among the target and source data. For this purpose, maximum mean discrepancy (MMD) is often used as a cost function to minimize the distance or to check a consistent feature extraction in both source and target domains [51]. Other investigations focus on feature extraction; however, Nielsen et al. [52] performed change detection by aligning both domains using canonical correlation analysis (CCA). The work is extended with a semisupervised approach, where change detection is performed on multi-scale data obtained from different sensors [53]. In [54], the domain alignment is achieved through an eigenproblem aiming at preserving the mismatch of labels and the geometric structure. The second approach uses Generative Adversarial Networks (GANs) [42] for an adversarial domain adaptation. The purpose of the GANs is to make both the source and target datasets spectral characteristics similar. Tzeng et al. [55] shows an example where the target dataset is translated to the source dataset using GANs. The translation contains a discriminator that recognizes the two datasets. Most of the studies employ a feature extraction network to generate feature sets for source and target domain [56–58]. The feature extraction network acts as a generator to reduce the classification loss for the source domain and concurrently maximize the loss of the discriminator. Based on these approaches, Adversarial Discriminative Domain Adaptation (ADDA) was employed to learn feature extraction networks for the source as well as for target domains [55]. In [57], an adversarial feature augmentation method was proposed to achieve DA in which the encoder is trained for the source and target domains. Inspired by the concept used in ADDA, Mesay et al. [59] implemented GAN-based DA for object classification in the remote sensing data. The last approach of domain adaptation creates a shared representation of both domains. In this method, one domain can be translated to another, and both domains can be translated into a common space. The method also provides a transfer function that facilitates the translation of one domain to another and translating back to the original state. CycleGAN provides the third approach and involves two discriminators that are used to translate one domain to another and converse [60].

The general methods of domain adaptations are not well interpreted for semantic segmentation [61]. Thus, adversarial and reconstruction procedures are chosen. Adversarial and constraint-based adaptations are performed at pixel level using architectures that exploit adversarial domain adaptation using GANs to transform source-like images [62]. Then, the images are segmented using a network that has been trained on the source dataset. In [63], Domain-Invariant Structure Extraction (DISE) structure was adopted to transform images into the domain-invariant structure and domain-specific texture representations. The bidirectional method prevents the translational model to reach a point where the discriminator fails to identify the image from the same distribution setup and fails to align correctly [64].

3. Study Area and Data

To promote reproducibility of our experimentation, we rely on BreizhCrops, a large-scale time series benchmark dataset introduced in 2020 by Rußwurm et al., [28], for supervised classification of field crops from satellite data. The dataset comprises multivariate time series examples in the Region of Brittany, France, of the season 2017, from January 1 to December 31. In particular, the authors of the dataset exploited all available Sentinel 2 images from Google Earth Engine, [65], and farmer surveys collected by France National Institute of Forest and Geography Information (IGN) to collect more than 600 k samples divided into 9 classes with 45 temporal steps and 13 spectral bands. Most importantly, as shown in Figure 2, acquired data are equally split into distinct regional areas. Indeed, as regulated by the Nomenclature des unites territoriales statistiques (NUTS), the overall dataset is divided into the four NUTS-3 regions Côes-d'Armor, Finistère, Ille-et-Vilaine,

and Morbihan. That, in conjunction with the challenging nature of the dataset, makes BreizhCrops an ideal benchmark to test domain adaptation for multi-spectral and multi-temporal data for LC&CC.

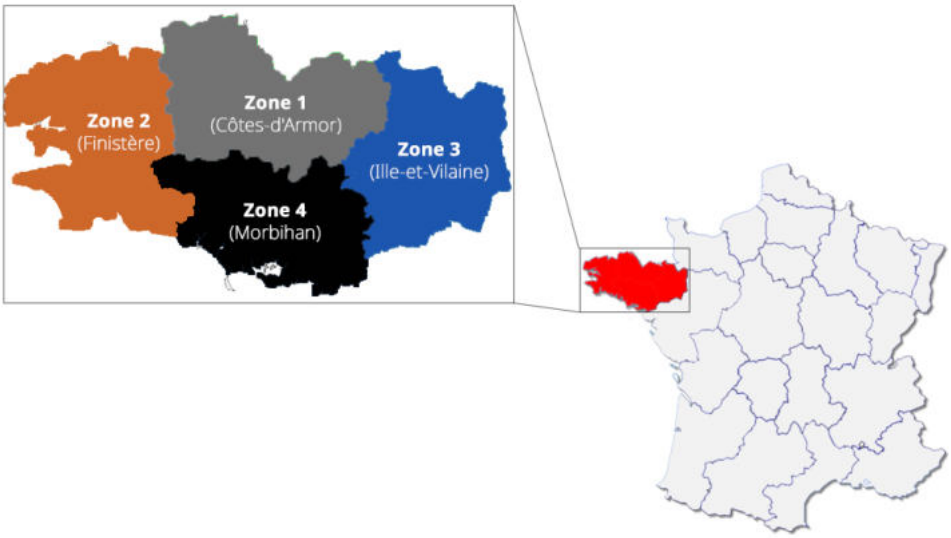

Figure 2. Magnified view of the four NUTS-3 regions of Brittany, located in the northwest of France and covering 27,200 km². The strict division of the supervised BreizhCrops dataset in the four regions allows the performance of a formal and controlled analysis on domain adaptation for LC&CC with multi-spectral and multi-temporal data.

As summarized in Figure 3, even if the authors of the dataset avoided broad categories, due to the nature of agricultural production, which focuses on a few dominant crop types, a class imbalance can be observed in the collected parcels. That constitutes a challenge for every classifier type, but it reflects the strong imbalance in real-world crop-type-mapping datasets. On the other hand, sample classes in the different regions are balanced, making BreizhCrops a perfect bench for testing domain adaptation strategies. Finally, to disentangle the performed domain adaptation analysis from the influence of the random variation of the atmospheric conditions, we exclusively make use of L2A bottom-of-atmosphere imagery where data acquired over time and space share the same reflectance scale. Adjacent and slope effects are corrected by the MAJA processing chain [66] that employs 60-meter spectral bands to apply atmospheric rectification and detect clouds. Therefore, only ten spectral features are available for each parcel. Table 1 is presented as a summary of the number of samples collected for the domain adaptation experimentation divided into classes and regions. In conclusion, multi-spectral, multi-temporal pixels are individually extracted for each parcel and are constituted by 10 spectral bands and 45 temporal steps each. The class imbalanced highlighted by the number of parcels of Figure 3 is reflected in the number of samples of Table 1 used for all experimentation.

Table 1. Summary of the number of samples per class divide in the four NUTS-3 regions of Brittany. Instances are derived by L2A bottom-of-atmosphere parcels to disentangle our analysis with variation of the atmospheric conditions.

	Barley	Wheat	Rapeseed	Corn	Sunflower	Orchards	Nuts	Permanent Meadows	Temporary Meadows
Zone 1	13,051	30,380	5596	44,003	1	937	10	32,641	52,013
Zone 2	10,736	15,026	2349	36,620	6	348	18	36,536	39,143
Zone 3	7154	27,202	3557	42,011	10	1217	10	32,524	52,682
Zone 4	5981	17,009	3244	31,361	2	552	11	26,134	38,141

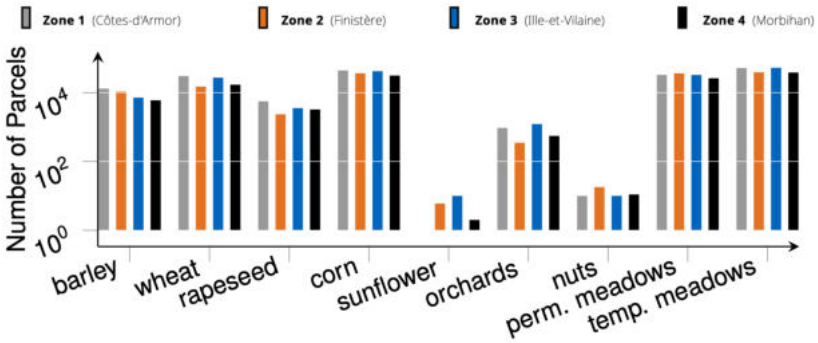

Figure 3. Class frequencies divided in the four NUTS-3 regions of Brittany. The respective number of parcels highlights the strong class imbalance, reflecting the substantial imbalance in real-world crop-type-mapping datasets. However, samples per class in the four regions are equally divided.

4. Methodology

In this work, unsupervised domain adaptation is considered in the field of land cover classification from satellite images. The study aims to tackle the problem of low generalization capability of classifiers only trained on a peculiar geographical region dataset. Moreover, the lack of rich available datasets of labeled satellite images increases the interest towards this challenge. In particular, the proposed methodology is intended to investigate the application of representation learning (RL) techniques for domain adaptation when dealing with multi-temporal data. For this purpose, a Transformer Encoder-based classifier is adapted to a Domain-Adversarial Neural Networks (DANN) architecture and trained accordingly.

In this section, a thorough description of the methodology is provided. First, we frame domain adaptation with the DANN method. Then, we briefly explain the Transformer Encoder structure with self-attention adopted for the multi-temporal crops classification. Finally, we describe the resulting architecture of the attention-based DANN, which is used to train a classifier with improved domain generalization.

4.1. Domain-Adversarial Neural Networks

Classifiers obtained with Deep Neural Networks often suffer from a lack of generalization related to possible variations in the appearance of the same objects. This problem is usually identified as a domain gap. In the land cover classification task, this situation is very recurrent and can be associated with the spectral shift affecting the data collected in different regions at different times. The shift is often related to photogrammetric distortion or visual differences in the appearance of lands. Furthermore, when dealing with satellite images, a dataset usually needs to be created by labeling images for a specific region to train a classification model. Despite this time-expensive procedure, standard training does not guarantee satisfying performance on images of different regions.

Domain-Adversarial Neural Networks (DANN) is a representation learning technique that allows a classifier to generalize better from a *source domain* to a *target domain*. This specific domain adaptation method consists of adding a branch to the original feed-forward architecture of the classifier and carry out an adversarial training. From a generic perspective, it is possible to identify three main components of the DANN: a *feature extractor* with parameters θ_f, a *label predictor* with parameters θ_y, and a *domain classifier* with parameters θ_d. The feature extractor is the first block of the DANN model. It is responsible for learning the function $G_f : X \rightarrow \mathbb{R}^d$, which maps the input samples X to a d-dimensional vector containing the extracted features. The label predictor function, $G_y(G_f(X))$, compute the label associated with the predicted class of the sample. The domain discriminator function $G_d(G_f(X))$ distinguishes between source and target domains given the extracted features. The combination of feature extractor and label predictor gives us the complete classifier model. The domain classifier is composed of a secondary branch, similar to the label predictor, which receives the extracted feature vector by the first block of the network.

Given these three main elements, the expression of the total loss used to train DANN is obtained by the following expression, according to the authors [67]:

$$\mathcal{L}\left(\theta_f, \theta_y, \theta_d\right) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_y^i\left(\theta_f, \theta_y\right) - \lambda\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_d^i\left(\theta_f, \theta_d\right) + \frac{1}{n'}\sum_{i=n+1}^{N}\mathcal{L}_d^i\left(\theta_f, \theta_d\right)\right) \quad (1)$$

The first term $\mathcal{L}_y$ is the label predictor loss, while the second one involves the domain discriminator loss $\mathcal{L}_d$. The hyper-parameter λ can be tuned to weigh the contribution of the two learning terms. A more detailed analysis of the choice of λ is proposed in the experiments section. n and n' are respectively the numbers of samples from the source and the target domains. Totally, we have $N = n + n'$ samples used in the training. The expression of the total loss function also describes the principal goals of DANN: first, we want to obtain a label predictor with low classification risk. Second, we are adding a regularization term for the domain adaptation. To this extent, we aim to find a set of parameters of the feature extractor θ_f that can map a generic input sample from either source or target domain to a new latent space of features, where the domain gap is reduced. On the other hand, the classification performance has not to be affected. For this reason, the extracted features should be discriminative as well as domain-invariant. According to this goal, the optimal choice of parameters θ_f and θ_y is represented by the one which minimizes the total loss function, keeping $\hat{\theta_d}$ unchanged. By contrast, the domain discriminator parameters θ_d are updated to maximize the loss while not changing the other ones.

$$\left(\hat{\theta}_f, \hat{\theta}_y\right) = \underset{\theta_f, \theta_y}{\operatorname{argmin}}\mathcal{L}\left(\theta_f, \theta_y, \hat{\theta}_d\right) \quad (2)$$

$$\hat{\theta}_d = \underset{\theta_d}{\operatorname{argmax}}\mathcal{L}\left(\hat{\theta}_f, \hat{\theta}_y, \theta_d\right). \quad (3)$$

In the original paper of DANN, the parameters of each piece of the neural network model are updated with a classical Stochastic Gradient Descent (SGD) optimizer. Here instead we use Adam (Adaptive momentum estimation), another popular optimization algorithm introduced by [68]. Parameters θ_f,θ_y and θ_d are updated according to its rules.

$$\theta_f \longleftarrow \theta_f - \eta\left(\frac{\hat{m}_{f,y}}{\sqrt{\hat{v}_{f,y}} + \epsilon} - \lambda\frac{\hat{m}_{f,d}}{\sqrt{\hat{v}_{f,d}} + \epsilon}\right) \quad (4)$$

$$\theta_y \longleftarrow \theta_y - \frac{\eta}{\sqrt{\hat{v}_y} + \epsilon}\hat{m}_y \quad (5)$$

$$\theta_d \longleftarrow \theta_d - \frac{\eta}{\sqrt{\hat{v}_d} + \epsilon}\hat{m}_d \quad (6)$$

As can be studied more in detail in the Adam original paper, the first (mean) and the second (uncentered variance) moments of Adam $\hat{m}$ and $\hat{v}$ are estimated as exponentially moving averages computed with the gradients obtained from each mini-batch. For the specific case of DANN, gradients used to estimate the Adam moments change for each element G_f, G_y, G_d of DANN structure. For example, the feature extractor gradients $(\partial\mathcal{L}_y^i/\partial\theta_f)$ and $(\partial\mathcal{L}_d^i/\partial\theta_f)$ are used to compute $\hat{m}_{f,y}$ and $\hat{m}_{f,d}$. Diversely, gradients obtained from label predictor $(\partial\mathcal{L}_y^i/\partial\theta_y)$ and domain discriminator $(\partial\mathcal{L}_d^i/\partial\theta_d)$ are only used to update their respective momentum $\hat{m}_y$ and $\hat{m}_d$.

The feature extractor and the domain discriminator play adversarial roles during the training process. A satisfying feature extractor can fool the domain discriminator by forwarding a vector of domain-invariant features. The role of the domain discriminator is to improve and evaluate this ability. A key intuition in the DANN method is to carry out the adversarial training with a standard backpropagation of the gradients, thanks to a custom Gradient Reversal Layer between the feature extractor and the domain discriminator. This

particular layer does not add other parameters to the model but changes the sign of the upstream gradients. The GRL operation can be formulated with $\mathcal{R}(\mathbf{x})$ in the following mathematical expressions for the forward and backpropagation step:

$$\mathcal{R}(\mathbf{x}) = \mathbf{x} \tag{7}$$

$$\frac{d\mathcal{R}}{d\mathbf{x}} = -\mathbf{I} \tag{8}$$

where $\mathbf{I}$ is the identity matrix. Hence, by performing optimization steps on the resulting DANN architecture, we can update parameters to reach saddle points of the total loss function reported in Equation (1).

4.2. Classification of Multi-Spectral Time Series Data with Self-Attention

Self-Attention, popularized by the Transformer model in 2017, [69], has provided a considerable boost in machine translation performance while being more parallelizable and requiring significantly less time to train. Nevertheless, the introspection capability behind the success of Transformers is not limited only to natural language processing, but can be adapted to any time series analysis to filter data and focus on more relevant repressions aspects.

A single sample pixel i-th of multi-spectral, multi-temporal acquisition can be represented as a matrix $\boldsymbol{X}^{(i)} \in \mathbb{R}^{t \times b}$ where t is the temporal dimension and b is given by the number of spectral bands. Therefore, it is a 1D sequence of tokens, $(\boldsymbol{x}_0, ..., \boldsymbol{x}_t)$, with $\boldsymbol{x}_t \in \mathbb{R}^b$, that can be easily linearly projected to feed a standard Transformer encoder. The encoder can map a temporal input sequence $\boldsymbol{X}_{t \times b}$ in a continuous representation $\boldsymbol{X}^L_{t \times d_{model}}$, where L is the output layer of the Transformer model and d_{model} is the constant latent dimension of the projection space.

Self-attention, through local multi-head dot-product self-attention blocks, can easily manipulate the temporal sequence finding correlations between different time-steps and completely avoiding the use of recurrent layers. The dot-product self-attention operation is composed on a trainable associative memory with key and value vector pairs of dimensions d. For a sequence of t query vectors, arranged in a matrix $Q \in \mathbb{R}^{t \times d}$, the self-attention operation is described by the following operation:

$$\text{Attention}(Q, K, V) = \text{Softmax}(QK^T/\sqrt{d})V \tag{9}$$

where the Softmax function is applied over each row of the input matrix and $K \in \mathbb{R}^{t \times d}$ and $V \in \mathbb{R}^{t \times d}$ are the key and value vector matrices, respectively. Query, key and values matrices are themselves computed from a sequence of t input vectors with dimension d_{model} using linear transformations: $Q = XW_Q, K = XW_K, V = XW_V$ where $X \in \mathbb{R}^{t \times d_{model}}$. Finally, multi-head dot-product self-attention is defined by considering applying h self-attention functions to the input X. Each head provides a sequence of size $t \times d$. These h sequences are rearranged into a $t \times dh$ sequence that is linearly projected into $t \times d_{model}$.

Subsequently, after the transformer encoder, the output representation, $\boldsymbol{X}^L_{t \times d_{model}}$,can be exploited to perform a classification of the input sequence. Indeed, that can be achieved by further processing the output encoder matrix and feeding a classification head trained to map the hidden representation to one of the k classes.

Several approaches have been proposed in the literature to obtain this result; in [70,71] they pre-append to the input sequence a learnable embedding, whose state at the output of the Transformer encoder serves as a hidden representation of the membership class. Indeed, only that output token is fed to the classification head to obtain the final prediction. On the other hand, the output sequence can be averaged or processed with a max operation on the temporal dimension [72]. Nevertheless, despite the type of processing applied to $\boldsymbol{X}^L$, the encoder will adapt to elaborate the sequence properly and embed the needed information for the classification task. In conclusion, a Transformer encoder can

be repurposed to process a multi-spectral input sequence and find valuable correlations between the different time-steps to perform LC&CC with a high level of accuracy.

4.3. DANN for Land Cover and Crop Classification

We employ DANN in conjunction with self-attention-based models to bridge the domain gap between different geographical regions. The overall architecture of the adopted methodology is shown in Figure 4. First, an input sequence $\boldsymbol{X}_{t\times b}$ is linearly projected to the constant latent dimension of the Transformer model d_{model}. Moreover, a Transformer encoder does not contain recurrence or convolution to make use of the order of the sequence. Therefore, some positional encoding is injected about the relative or absolute position of the tokens in the sequence. The positional encodings have dimension d_{model} as the projected sequence, so that the two can be summed. Guided by experimentation, as in [71], we adopt a learnable positional encoding instead of the sine and cosine functions with different frequencies of [69]. The resulting pre-processed input sequence $\boldsymbol{X}^{l_0}_{t\times d_{model}}$ feeds the Transformer encoder, parameterized by Θ_f, that provides as output a continuous representation $\boldsymbol{X}^{L}_{t\times d_{model}}$. Subsequently, we make use of the max function, over the temporal axis, to extract a token, $\boldsymbol{x}^{L}_{d_{model}}$, from the output sequence.

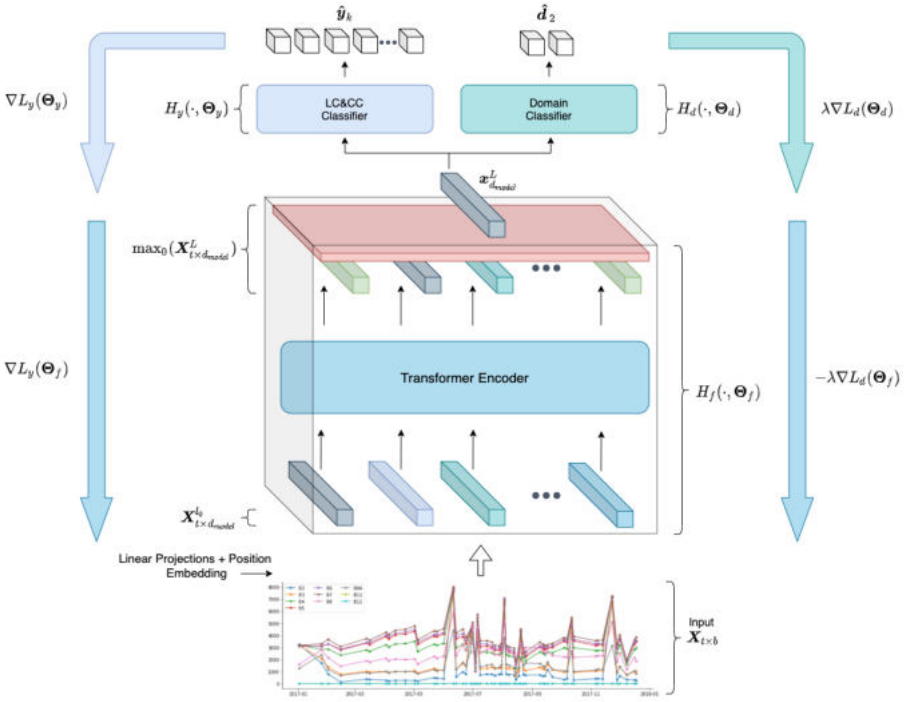

Figure 4. Overview of the overall framework to train a Transformer encoder with domain-adversarial training. The multi-spectral temporal sequence $\boldsymbol{X}_{t\times b}$ is first linearly projected and fused with a position encoding. Subsequently, the self-attention-based model manipulates the input series and, through a max operation applied to the last layer of the encoder, is possible to extract a token $\boldsymbol{x}^{L}_{d_{model}}$ from the output sequence. Finally, gradients derived by LC&CC and Domain classifiers train the network while keeping close the distribution of source and target domains.

The extracted representation constitutes the input for either the LC&CC and domain multi-layer perceptron classifiers. The first network provides a probability distribution over the k different classes, $\hat{\boldsymbol{y}}_k$. On the other hand, the domain classifier outputs the probability, $\hat{\boldsymbol{d}}_2$, that the extracted representation $\boldsymbol{x}^{L}_{d_{model}}$ belongs to the target or source domain. Using the cross-entropy loss function for both classifiers, it is possible to compute the respective gradients and update the weights, Θ_f of the feature extractor. Indeed, inverting the sign of the gradients, $\nabla L_d(\Theta_d)$, derived from the domain classifier, and multiplying them for

a scale factor λ, we can increasingly reduce the distance between the latent space of the two domains while training the encoder on the classification task. Overall, the proposed training framework provides an effective solution to transfer the acquired knowledge of a model to a diverse region, exploiting only the original nature of the data.

5. Experiments and Discussion

We experiment with the proposed methodology on the four regions of the multi-temporal satellite BreizhCrops dataset presented in Section 3. As explained in the same section, we indicate this dataset as an optimal choice to train and test new domain adaptation methods exploiting labeled multi-temporal data. The first main objective of the conducted experimentation is to investigate how the classification performance of a state-of-the-art model for LC&CC model is affected by a lack of generalization towards different geographical regions. Then, we clearly highlight how adversarial training can mitigate the domain gap and significantly boost performance for source and target regions with marked distribution distance. It is important to remark that the method relies on the availability of samples of both source and target domains, whereas only source labels are required, not allowing direct applicability of transfer learning techniques. Finally, in the last part of the section, obtained results are discussed and inspected through dimensionality reduction techniques, validating the proposed method for practical use.

5.1. Experimental Settings

We carried out a complete set of experiments to compare the Transformer encoder classifier performance with and without DANN. The standard classifier is trained separately on each of the single regions of the dataset, then tested on the other ones. By contrast, DANN models are trained on each source-target pair to gain the desired adaptation capability and tested an all the regions except for the source domain. No validation set is used for model selection. Tests are always performed with the model resulting from a fixed number of training epochs.

In the final architecture, the classifier model comprises a transformer encoder feature extractor and a final classification stage. In all experimentation, the transformer encoder receives as input a batch of 256 tensors with $t = 45$ temporal steps and $b = 10$ spectral bands in the image samples. Moreover, to linearly project the temporal sequence to the constant latent dimension of the encoder, the input is first passed to a dense layer with 64 units. Therefore, d_{model} is equal to 128. On the other hand, the multi-head attention Transformer encoder is defined with several layers and attention heads equal to $n_{layers} = 3$ and $n_{heads} = 2$. Finally, the dimension of internal fully connected layers $d_{inner} = 128$. Rectified linear units is the non-linear activation function used for all neurons of the encoder.

The LC&CC classification stage is a simple multi-layer perceptron head composed of a normalization layer, a fully connected layer with 128 units, ReLU as activation function, and a final layer with $k = 9$ neurons. On the other hand, for the DANN experimentation, the domain predictor is identical to the multi-layer perceptron head of the LC&CC classifier, with 128 units and a ReLU activation. However, the number of neurons in the final layer is set to $d = 2$, since we always perform a single target domain adaptation.

A cross-entropy loss function is chosen to train both the classifiers. The parameters of both models are updated using Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1 \times 10^{-7}$. A fixed number of epochs is always set to 250. The learning rate value is changed during training according to an exponential decay policy from a starting value of 0.001, with a decay scheduled for each epoch equal to 0.99^{epoch}. A key point in the experimental settings is related to the domain adaptation parameter λ. It acts as a regularization parameter, since it regulates the impact of the domain discriminator gradients on the feature extractor during training. Therefore, it can be considered to be the principal hyper-parameter to tune when using DANN. We always use a scheduling policy for λ, as suggested in the original publication of DANN:

$$\lambda_t = \lambda_{max} \left(\frac{2}{1 + e^{-\gamma t}} - 1 \right) \tag{10}$$

where λ_{max} is the plateau value reached. This is the actual value of λ used for the second half of the training, which affects the final performance of the model in terms of generalization. The parameter $\gamma = 10$ defines the slope of the curve and it is fixed to such value to let λ_{max} be reached in a suitable number of epochs. A scheduled value of λ allows the feature extractor to learn the basic features for the classification during the first epochs. It then adjusts the mapping function to let the source and target domain feature distributions to overlap at the end of the training process. As shown in Figure 5, different values of λ_{max} are tested to study the response of the model. To our knowledge, $\lambda_{max} = 0.2$ is the best value for a robust adaptation improvement of the classifier, at least among the set of tested λ_{max} values.

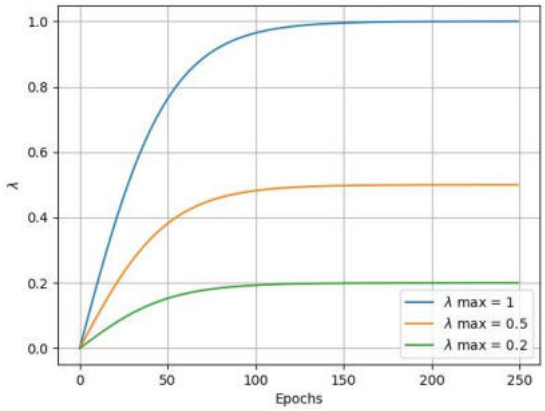

Figure 5. λ scheduling: the value of the domain adaptation parameter λ is changed during training according to an exponentially growing trend. This allows the feature extractor to learn basic features during the initial epochs. Different final λ_{max} values are tested to study the right level of adaptation required in the different cases: 1, 0.5 and 0.2. $\lambda_{max} = 0.2$ is the best choice for an overall adaptation improvement of the classifier in the different regions. The parameter γ influences the slope of the curve and it is kept constant to 10 to let λ reach the desired value in a suitable number of epochs.

As already explained at the beginning of the section, the classifiers are trained and tested on all the possible combinations of regions to quantify the existing domain gap.

The classification performance is evaluated using three different classification metrics, which are chosen among the ones proposed in the BreizhCrops dataset benchmarks: Accuracy, F1-score and K-score. This last metric is the Cohen's kappa [73], computed according $\kappa = (p_o - p_e)/(1 - p_e)$ where p_o and p_e are the empirical and expected probability of agreement on a label. In addition, we make use of Maximum Mean Discrepancy (MMD) metric, presented in Section 5.2, to quantitatively evaluate the distance between source and target distributions.

5.2. Maximum Mean Discrepancy

MMD is a statistical test originally proposed in [74] to determine a measure of the distance between two distributions. MMD is largely used in domain adaptation since it perfectly fits the need to understand whether the source and the target domain extracted features overlap. MMD can be directly exploited as a loss function for adversarial training of generative models or for domain adaptation purposes, as shown in [75,76]. However, in this works we limit its usage to show the results of the Transformer Encoder DANN in terms of reduction of feature distances.

Formally, MMD is a kernel-based difference between feature means. Given a set of m samples X with a probability measure P, the feature mean can be expressed as:

$$\mu_p\left(\phi(X)\right) = [E[\phi(X_1], \cdots, E[\phi(X_m]]^T \tag{11}$$

where $\phi(X)$ is the feature map that maps X to a new feature space $\mathcal{F}$. If it satisfies the necessary theoretical conditions, a kernel-based approach can be used to compute the inner product of two distributions of samples $X \sim P$ and $Y \sim Q$:

$$\langle \mu_P(\phi(X), \mu_Q(\phi(Y)\rangle_{\mathcal{F}} = E_{P,Q}[\langle \phi(X), \phi(Y)\rangle_{\mathcal{F}}] = E_{P,Q}[k(X,Y)] \tag{12}$$

At this point the MMD can be defined as the distance between the feature means of $X \sim P$ and $Y \sim Q$:

$$MMD^2(P,Q) = \|\mu_P - \mu_Q\|^2_{\mathcal{F}} \tag{13}$$

which can be expressed more in detail using Equation (12):

$$MMD^2(P,Q) = E_P[k(X,X)] - 2E_{P,Q}[k(X,Y)] + E_Q[k(Y,Y)] \tag{14}$$

However, an empirical estimate of MMD needs to be computed since in a real case only samples are available instead of the explicit formulation of the distributions. It is possible to obtain the MMD expression by considering the empirical estimates of the feature means based on their samples:

$$MMD^2(X,Y) = \frac{1}{m(m-1)}\sum_i\sum_{j\neq i} k(\mathbf{x_i}, \mathbf{x_j}) - 2\frac{1}{m.m}\sum_i\sum_j k(\mathbf{x_i}, \mathbf{y_j}) + \frac{1}{m(m-1)}\sum_i\sum_{j\neq i} k(\mathbf{y_i}, \mathbf{y_j}) \tag{15}$$

where $\mathbf{x_i}$ and $\mathbf{y_i}$ in this case are the image samples from source and target domains, m is the number of samples of the considered subsets. Finally, we specifically use a gaussian kernel with the following expression:

$$k(\mathbf{x_i}, \mathbf{x_j}) = \exp\left(\frac{-\|\mathbf{x_i} - \mathbf{x_j}\|^2}{2\sigma^2}\right) = \exp\left(\frac{-1}{\sigma^2}[\mathbf{x_i}^\intercal\mathbf{x_i} - 2\mathbf{x_i}^\intercal\mathbf{x_j} + \mathbf{x_j}^\intercal\mathbf{x_j}]\right) \tag{16}$$

5.3. Results Discussion and Applicability Study

In this section, we present the comparison results between the Transformer classifier with and without DANN, clearly highlighting the scenarios that present a definite advantage in applying adversarial training for training a classifier for LC&CC. From results in Tables 2 and 3, Figure 6, it is possible to notice that DANN adversarial training allows the classifier to improve knowledge transferability to other domains for most of the cases. Nonetheless, we investigate a potential criterion to decide if the transfer of learning from source to target can be effectively improved by DANN. More in detail, since DANN aims to overlap feature distributions, we look at the extracted features from a subset of 10,000 samples of each zone dataset that is considered representative of the total one.

We use the set of extracted features to compute a numerical evaluation of the distance decrease, and to give a graphical visualization of the effect of DANN. From a quantitative perspective, we propose Maximum Mean Discrepancy as the feature distance metrics to detect suitable conditions where DANN is an appropriate methodology. To compute MMD without considering the clustering of classes, we only need unlabeled image samples. We use PCA algorithm to compute the principal components of the extracted features and we exploit them to provide 2D and 3D visualization of relevant cases.

First, we can look at the MMD values obtained from both the Transformer encoder and DANN in Table 2. It is clear that DANN is always able to reduce the distance between feature distributions. However, this is not always associated with an increase in classification performance. We realize that key information is contained in the MMD value obtained from source and target features, extracted by the standard classifier. This simple test is crucial and can also be done without labels. The best improvement with DANN is reached considering zone 2 as the source domain and selecting zone 3 as the target domain. The percentage improvement shown in Table 3, with an increase of more than 30% of accuracy, correlates with an initial MMD value for this specific case is equal to 0.6700, reduced by DANN to 0.0104. What can be deduced by this observation is that high values

of the MMD indicate a lack of generalization of the classifier and a domain gap. It is also to consider that the geographical zones of interest are close to each other. Hence, it can be reasonable to find small domain gaps. A clear example is the case of zone 1, when chosen as source domain. This factor can be considered an additional difficulty of the study case. Therefore, it is possible that the same methodology applied to other regions on the planet, sharing the same categories of crops, can probably show greater results. Another peculiar case to be considered is: zones 4 (source) and 3 (target). The MMD value is low from the initial analysis of the case, without the intervention of DANN. However, a classification boost is always achieved.

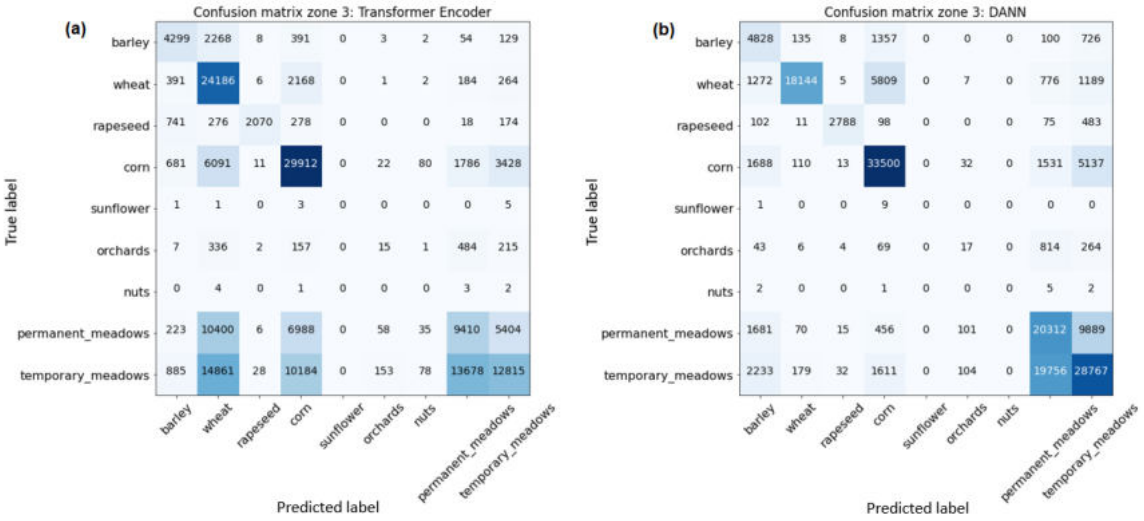

Figure 6. Class-wise comparison of classification results on zone 3 (target), selecting zone 2 as source domain. Confusion matrix obtained with Transformer encoder trained on zone 2 and tested on zone 3 is shown in (**a**) on the left. Figure (**b**) on the right shows classification results with DANN model tested on zone 3. The effect of DANN clearly mitigate the prediction error, with a particular focus on relevant classes such as Corn, Permanent and Temporary Meadows.

Table 2. Results of crops classification for the Transformer Encoder classifier trained with and without DANN using $\lambda_{max} = 0.2$. The two models are trained and tested on all the possible combinations of source/target domains available in BreizhCrops dataset. Accuracy, F1-Accuracy and K-score are the metrics used to compare the classification quality. Training accuracy is also reported for the Transformer encoder classifier. Maximum Mean Discrepancy computed on a subset of extracted features of source and target domain shows the successful reduction of features distance obtained with DANN.

Zone		Transformer Encoder					DANN			
Source Domain	Target Domain	Train Accuracy	Test Accuracy	F1-Accuracy	K-Score	MMD	Test Accuracy	F1-Accuracy	K-Score	MMD
1	2	0.8577	0.7877	0.5675	0.7229	0.1109	0.7628	0.5540	0.6950	0.0077
1	3	0.8577	0.7436	0.5266	0.6606	0.1620	0.7449	0.5080	0.6714	0.0183
1	4	0.8577	0.7941	0.5675	0.7294	0.0516	0.7960	0.5734	0.7343	0.0086
2	1	0.8951	0.7433	0.5309	0.6773	0.1577	0.7403	0.5161	0.6687	0.0208
2	3	0.8951	0.4967	0.3592	0.3642	0.6700	0.6505	0.4544	0.5483	0.0104
2	4	0.8951	0.6006	0.4395	0.4912	0.2536	0.7482	0.4832	0.6735	0.0416
3	1	0.8750	0.7767	0.5339	0.7122	0.1819	0.8045	0.5778	0.7488	0.0121
3	2	0.8750	0.6638	0.4594	0.5615	0.6254	0.7589	0.5334	0.6865	0.0277
3	4	0.8750	0.7348	0.5074	0.6504	0.1184	0.7968	0.5778	0.7338	0.0115
4	1	0.8870	0.7927	0.5551	0.7354	0.0339	0.8233	0.5822	0.7753	0.0039
4	2	0.8870	0.7600	0.5443	0.6870	0.0953	0.8003	0.5788	0.7399	0.0084
4	3	0.8870	0.7111	0.4961	0.6230	0.0960	0.7673	0.5443	0.6965	0.0062

Table 3. Comparison between Transformer Encoder Classifier with and without DANN, in terms of classification metrics reported in Table 2. This run of experiments is conducted with a scheduling of the adaptation parameter λ, with $\lambda_{max} = 0.2$.

Zone		Improvement [%]		
Source Domain	**Target Domain**	**Test Accuracy**	**F1-Accuracy**	**K-Score**
1	2	−3.1576	−2.3859	−3.8508
1	3	0.1762	−3.5378	1.6395
1	4	0.2296	1.0467	0.6773
2	1	−0.3996	−2.7935	−1.2698
2	3	30.9721	26.4916	50.5414
2	4	24.5690	9.9474	37.1046
3	1	3.5803	8.2152	5.1446
3	2	14.3204	16.1075	22.2539
3	4	8.4475	13.8791	12.8283
4	1	3.8705	4.8817	5.4228
4	2	5.3053	6.3384	7.6922
4	3	7.9018	9.7154	11.8067

We report a visual representation of the extracted features to add meaning to the previous considerations. In particular, Figures 7–9 show the 2D principal components obtained from the peculiar cases defined below:

- **case 1**: zone 2 (source), zone 3 (target). In this case DANN shows the greatest improvements with an initial high value of MMD. Features are visually reported in Figure 7: in (a,b) when extracted by standard Transformer encoder trained on the source domain, in (c,d) when extracted by DANN. The difference is visually clear. Features distributions are matched by DANN, with a resulting overlapping shape between source and target domain.
- **case 2**: zone 1 (source), zone 2 (target). In this case DANN shows the worst improvements with an initial low value of MMD. Features are visually reported in (a,b) of Figure 8 when extracted by standard Transformer encoder, in (c,d) of the same Figure 8 when extracted by DANN. They appear already similar also without DANN.
- **case 3**: zone 4 (source), zone 3 (target). In this case DANN shows noticeable improvements, regardless an initial low value of MMD. Features are visually reported in (a,b) of Figure 9 when extracted by standard Transformer encoder, in (c,d) of Figure 9 when extracted by DANN. As with case 1, the difference is visually clear, and the effect of DANN can be easily appreciated.

Finally, case 1 and case 2 defined above are also considered for a 3D representation. Figure 10 shows the obtained results. For each subplot in the figure, both source and target domain features are scattered. Thanks to this visual perspective, the effect of the DANN method is highlighted, considering both the worst and the best application scenario. In case 1, the difference between source and target features is shallow also without DANN, as shown in (a). By contrast, the situation from (c) to (d) is changed thanks to the adversarial training significantly.

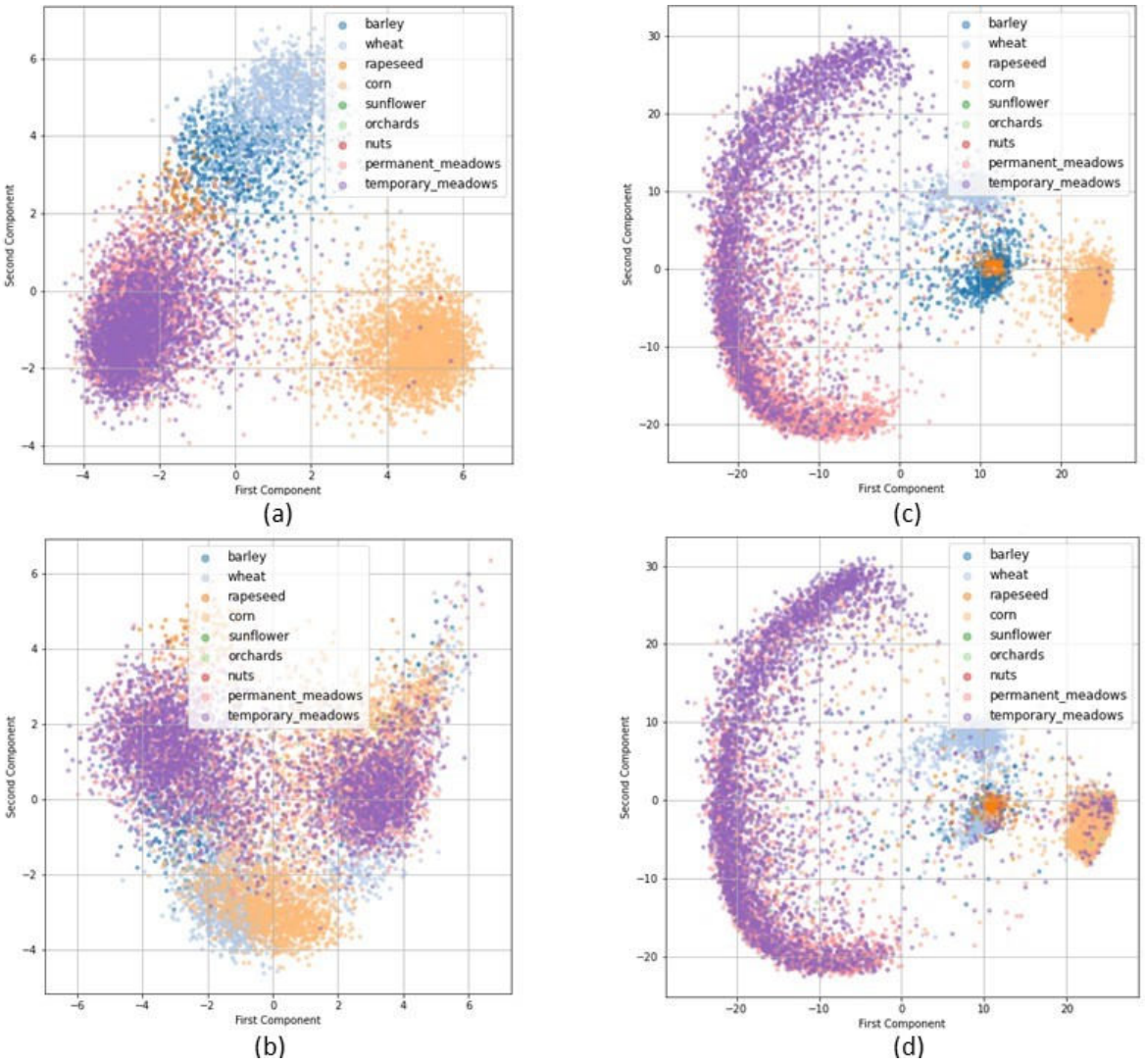

Figure 7. 2D feature visualization obtained with PCA, extracted with the Transformer Encoder trained on the source domain and with the Transformer DANN model trained on the specific source-target domains. A comparison between the 2D feature distributions is proposed for the case of zone 2 (source) and 3 (target). In (**a**,**b**) we have features extracted with the Transformer Encoder from source and target domains: (**a**) reports features of the source domain (zone 2) and (**b**) the ones extracted from the target domain (zone 3). In this case, features are mapped poorly in the target domain, with a consequent low accuracy in classification. In (**c**,**d**) the same features extracted with the Transformer DANN model are shown. The positive effect of DANN in terms of features overlapping is evident compared to (**a**,**b**).

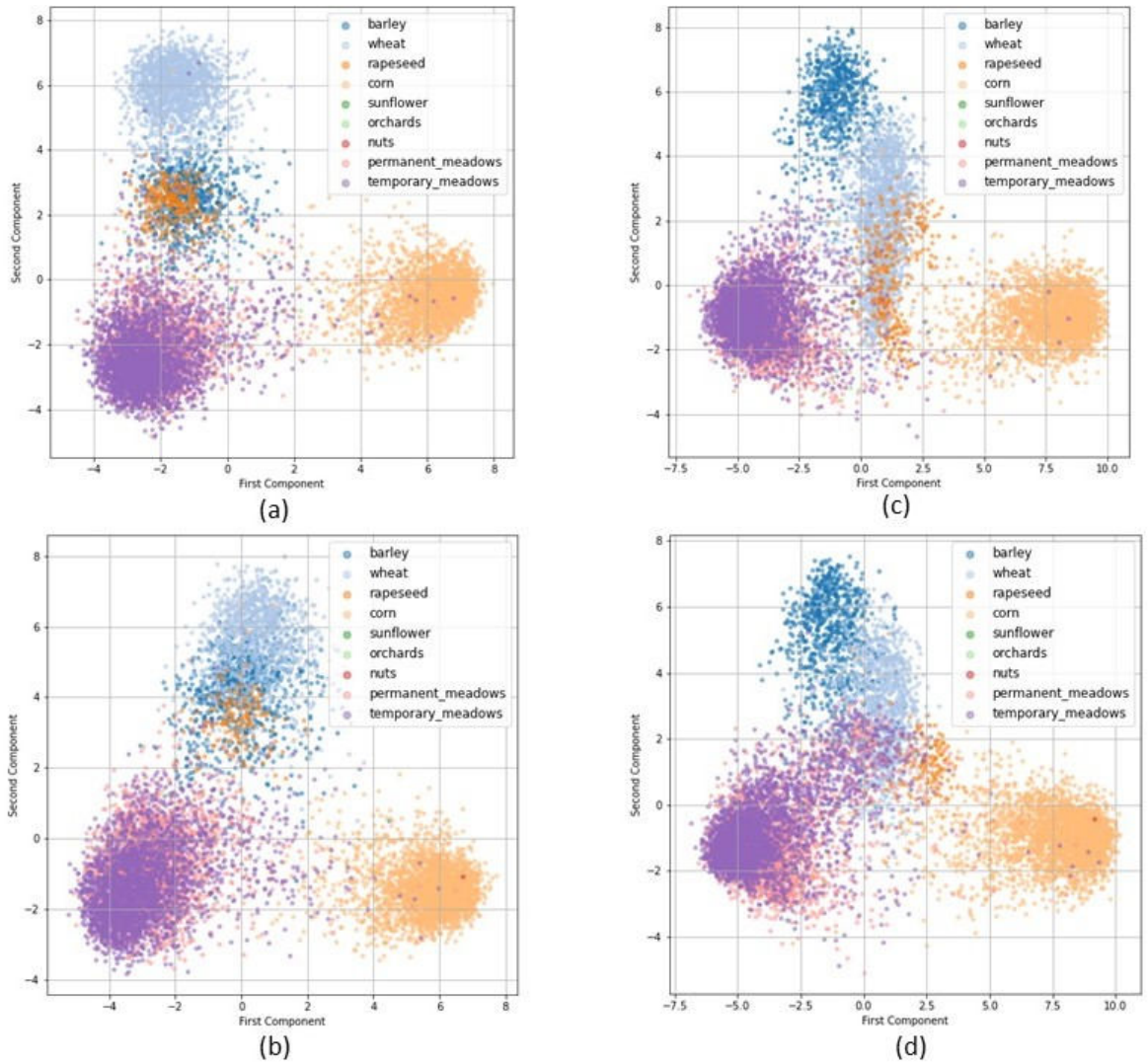

Figure 8. 2D feature visualization obtained with PCA, extracted with the Transformer Encoder trained on the source domain and with the Transformer DANN model trained on the specific source-target domains. A comparison between the 2D feature distributions is proposed for the case of zone 1 (source) and 2 (target). In (**a**,**b**) we have features extracted with the Transformer Encoder from source, (**a**), and target, (**b**), domain: a low MMD distance indicates no need for domain adaptation. In (**c**,**d**) the same features extracted with the Transformer DANN model are shown, with no substantial differences.

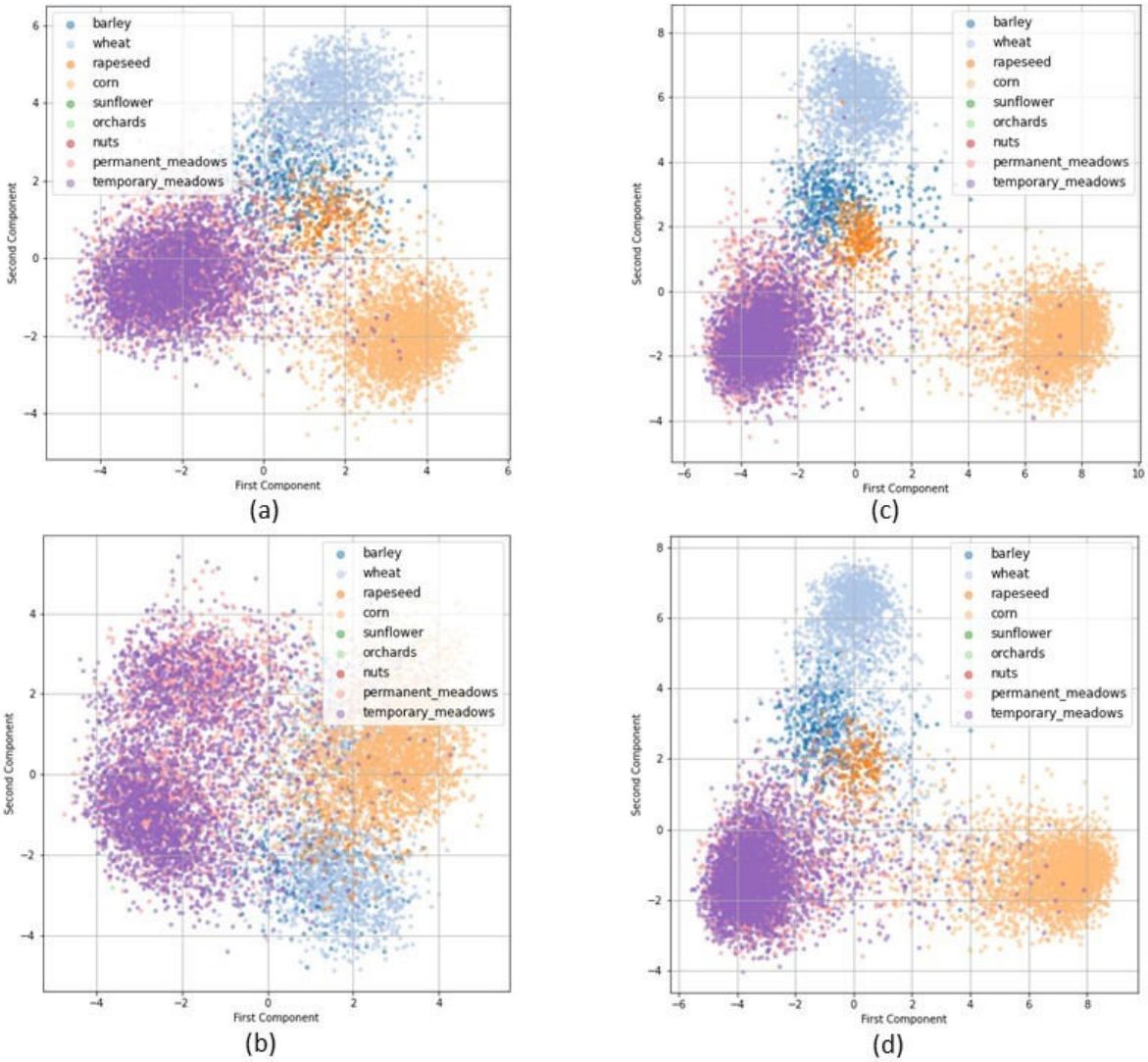

Figure 9. 2D feature visualization obtained with PCA, extracted with the Transformer Encoder trained on the source domain and with the Transformer DANN model trained on the specific source-target domains. A comparison between the 2D feature distributions is proposed for the case of zone 4 (source) and 3 (target). In (**a**,**b**) we have features extracted with the Transformer Encoder from source, (**a**), and target, (**b**), domains: regardless of an initial low MMD, the classifier accuracy can still be improved reducing the domain gap. In (**c**,**d**) the same features extracted with the Transformer DANN model are shown, with a clear improvement of the feature mapping, which result in very similar distributions from source to target domain.

The proposed discussion underlines some interesting insights on the correlation between reducing the domain gap and improving a classifier performance. The isolated cases considered provide a good reference example to decide if it is a reasonable and convenient choice to adopt the proposed DANN methodology for multi-spectral temporal sequences for Land Cover classification.

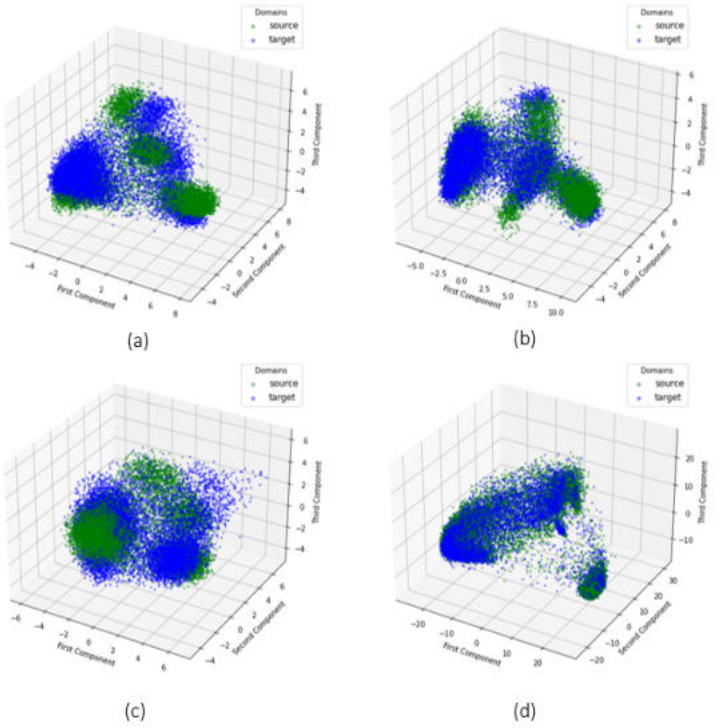

Figure 10. 3D feature visualization and comparison. (**a**,**b**) show the features extracted from zone 1 (source) and 2 (target). They are respectively obtained with transformer encoder and DANN. It is clear that the transformer encoder alone can correctly map features on both domains. By contrast, the improvement provided by DANN model is very evident in figures (**c**,**d**), representing the features extracted from zone 2 (source) and 3 (target), where the transformer encoder alone present both high values of MMD and low classification accuracy on target domain.

6. Conclusions

In this paper, we investigated adversarial training for domain adaptation with state-of-the-art self-attention-based models for LC&CC. Indeed, domain gaps between distinct geographical regions prevent the direct repurpose of the trained model on diverse areas of the training domain, and the practical difficulty of acquiring labeled data prevents the direct application of transfer learning techniques. Our extensive experimentation clearly highlights the advantages of applying the proposed methodology to transformer models trained on multi-spectral, multi-temporal data and the considerable gain in performance with considerable distribution distance between target and source regions. In particular, the best improvement obtained with DANN shows a percentage increase of more than 30% of classification accuracy, associated with an evident reduction of the features distance metrics MMD from 0.6700 to 0.0104. Moreover, our investigations conduct to a clear identification of the scenarios where it is advantageous to apply the DANN domain adaptation mechanism. More in detail we identified three different cases that highlight the strategy for a correct adoption of the methodology. A graphical visualization of the effect of DANN on the crop classification task has also been proposed and discussed exploiting the 2D class-wise and the 3D principal components of crops features distribution.

Future work may investigate the advantages and disadvantages of different domain adaptation techniques applied to LC&CC and extend our study to further geographical regions.

Author Contributions: Conceptualization, M.M., V.M. and A.K.; methodology, M.M. and V.M.; software, M.M. and V.M.; validation, M.M., V.M. and A.K.; formal analysis, M.M., V.M. and A.K.; investigation, M.M.; resources, M.M., V.M. and A.K.; data curation, M.M., V.M. and A.K.; writing—original draft preparation, M.M., V.M. and A.K.; writing—review and editing, M.M., V.M., A.K. and

M.C.; visualization, M.M.; supervision, M.M., V.M. and A.K.; project administration, M.C.; funding acquisition, M.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://breizhcrops.org/ (accessed on 18 May 2021). The code related to this study is openly available at https://github.com/maurom3197/Self-Attention-DANN-for-multi-temporal-Land-Cover (accessed on 18 May 2021).

Acknowledgments: This work has been developed with the contribution of the Politecnico di Torino Interdepartmental Centre for Service Robotics PIC4SeR (https://pic4ser.polito.it (accessed on 18 May 2021)) and SmartData@Polito (https://smartdata.polito.it (accessed on 18 May 2021)).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rudd, J.D.; Roberson, G.T.; Classen, J.J. Application of satellite, unmanned aircraft system, and ground-based sensor data for precision agriculture: A review. In Proceedings of the 2017 ASABE Annual International Meeting. American Society of Agricultural and Biological Engineers, Spokane, WA, USA, 16–19 July 2017.
2. Novelli, A.; Aguilar, M.A.; Nemmaoui, A.; Aguilar, F.J.; Tarantino, E. Performance evaluation of object based greenhouse detection from Sentinel-2 MSI and Landsat 8 OLI data: A case study from Almería (Spain). *Int. J. Appl. Earth Obs. Geoinf.* **2016**, *52*, 403–411. [CrossRef]
3. De Jong, R.; de Bruin, S.; de Wit, A.; Schaepman, M.E.; Dent, D.L. Analysis of monotonic greening and browning trends from global NDVI time-series. *Remote. Sens. Environ.* **2011**, *115*, 692–702. [CrossRef]
4. Pacifici, F.; Longbotham, N.; Emery, W.J. The importance of physical quantities for the analysis of multitemporal and multiangular optical very high spatial resolution images. *IEEE Trans. Geosci. Remote. Sens.* **2014**, *52*, 6241–6256. [CrossRef]
5. Amorós-López, J.; Gómez-Chova, L.; Alonso, L.; Guanter, L.; Zurita-Milla, R.; Moreno, J.; Camps-Valls, G. Multitemporal fusion of Landsat/TM and ENVISAT/MERIS for crop monitoring. *Int. J. Appl. Earth Obs. Geoinf.* **2013**, *23*, 132–141. [CrossRef]
6. Li, W.; Niu, Z.; Huang, N.; Wang, C.; Gao, S.; Wu, C. Airborne LiDAR technique for estimating biomass components of maize: A case study in Zhangye City, Northwest China. *Ecol. Indic.* **2015**, *57*, 486–496. [CrossRef]
7. Rembold, F.; Atzberger, C.; Savin, I.; Rojas, O. Using low resolution satellite imagery for yield prediction and yield anomaly detection. *Remote. Sens.* **2013**, *5*, 1704–1733. [CrossRef]
8. Khaliq, A.; Mazzia, V.; Chiaberge, M. Refining satellite imagery by using UAV imagery for vineyard environment: A CNN Based approach. In Proceedings of the 2019 IEEE International Workshop on Metrology for Agriculture and Forestry (MetroAgriFor), Naples, Italy, 24–26 October 2019; pp. 25–29.
9. Gomez, C.; White, J.C.; Wulder, M.A. Optical remotely sensed time series data for land cover classification: A review. *ISPRS J. Photogramm. Remote. Sens.* **2016**, *116*, 55–72. [CrossRef]
10. Khaliq, A.; Musci, M.A.; Chiaberge, M. Analyzing relationship between maize height and spectral indices derived from remotely sensed multispectral imagery. In Proceedings of the 2018 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), Washington, DC, USA, 9–11 October 2018; pp. 1–5.
11. Büttner, G.; Feranec, J.; Jaffrain, G.; Mari, L.; Maucha, G.; Soukup, T. The CORINE land cover 2000 project. *EARSeL eProceedings* **2004**, *3*, 331–346.
12. Tuia, D.; Persello, C.; Bruzzone, L. Domain adaptation for the classification of remote sensing data: An overview of recent advances. *IEEE Geosci. Remote. Sens. Mag.* **2016**, *4*, 41–57. [CrossRef]
13. Verrelst, J.; Alonso, L.; Caicedo, J.P.R.; Moreno, J.; Camps-Valls, G. Gaussian process retrieval of chlorophyll content from imaging spectroscopy data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2012**, *6*, 867–874. [CrossRef]
14. Ballanti, L.; Blesius, L.; Hines, E.; Kruse, B. Tree species classification using hyperspectral imagery: A comparison of two classifiers. *Remote. Sens.* **2016**, *8*, 445. [CrossRef]
15. Huang, X.; Ali, S.; Purushotham, S.; Wang, J.; Wang, C.; Zhang, Z. Deep multi-sensor domain adaptation on active and passive satellite remote sensing data. In *1st KDD Workshop on Deep Learning for Spatiotemporal Data, Applications, and Systems (DeepSpatial 2020)*; American Geophysical Union: Washington DC, USA, 2020.
16. Penatti, O.A.; Nogueira, K.; Dos Santos, J.A. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, 7–12 June 2015; pp. 44–51.
17. Mazzia, V.; Khaliq, A.; Chiaberge, M. Improvement in land cover and crop classification based on temporal features learning from Sentinel-2 data using recurrent-convolutional neural network (R-CNN). *Appl. Sci.* **2020**, *10*, 238. [CrossRef]

18. Demir, I.; Koperski, K.; Lindenbaum, D.; Pang, G.; Huang, J.; Basu, S.; Hughes, F.; Tuia, D.; Raskar, R. Deepglobe 2018: A challenge to parse the earth through satellite images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–23 June 2018; pp. 172–181.
19. Tian, C.; Li, C.; Shi, J. Dense fusion classmate network for land cover classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–23 June 2018; pp. 192–196.
20. Kuo, T.S.; Tseng, K.S.; Yan, J.W.; Liu, Y.C.; Frank Wang, Y.C. Deep aggregation net for land cover classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–23 June 2018; pp. 252–256.
21. Conjeti, S.; Katouzian, A.; Roy, A.G.; Peter, L.; Sheet, D.; Carlier, S.; Laine, A.; Navab, N. Supervised domain adaptation of decision forests: Transfer of models trained in vitro for in vivo intravascular ultrasound tissue characterization. *Med Image Anal.* **2016**, *32*, 1–17. [CrossRef] [PubMed]
22. Deng, Z.; Sun, H.; Zhou, S.; Ji, K. Semi-supervised cross-view scene model adaptation for remote sensing image classification. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 11–15 July 2016; pp. 2376–2379. [CrossRef]
23. Matasci, G.; Volpi, M.; Kanevski, M.; Bruzzone, L.; Tuia, D. Semisupervised Transfer Component Analysis for Domain Adaptation in Remote Sensing Image Classification. *IEEE Trans. Geosci. Remote. Sens.* **2015**, *53*, 3550–3564. [CrossRef]
24. Liu, W.; Su, F. A novel unsupervised adversarial domain adaptation network for remotely sensed scene classification. *Int. J. Remote. Sens.* **2020**, *41*, 6099–6116. [CrossRef]
25. Benjdira, B.; Bazi, Y.; Koubaa, A.; Ouni, K. Unsupervised domain adaptation using generative adversarial networks for semantic segmentation of aerial images. *Remote. Sens.* **2019**, *11*, 1369. [CrossRef]
26. Karimpour, M.; Saray, S.N.; Tahmoresnezhad, J.; Aghababa, M.P. Multi-source domain adaptation for image classification. *Mach. Vis. Appl.* **2020**, *31*, 1–19. [CrossRef]
27. Bahirat, K.; Bovolo, F.; Bruzzone, L.; Chaudhuri, S. A novel domain adaptation Bayesian classifier for updating land-cover maps with class differences in source and target domains. *IEEE Trans. Geosci. Remote. Sens.* **2011**, *50*, 2810–2826. [CrossRef]
28. Rußwurm, M.; Lefèvre, S.; Körner, M. Breizhcrops: A satellite time series dataset for crop type identification. In Proceedings of the International Conference on Machine Learning Time Series Workshop, Anchorage, AK, USA, 5 August 2019.
29. Walker, J.; De Beurs, K.; Wynne, R. Dryland vegetation phenology across an elevation gradient in Arizona, USA, investigated with fused MODIS and Landsat data. *Remote. Sens. Environ.* **2014**, *144*, 85–97. [CrossRef]
30. Arvor, D.; Jonathan, M.; Meirelles, M.S.P.; Dubreuil, V.; Durieux, L. Classification of MODIS EVI time series for crop mapping in the state of Mato Grosso, Brazil. *Int. J. Remote. Sens.* **2011**, *32*, 7847–7871. [CrossRef]
31. Moranduzzo, T.; Melgani, F. Automatic car counting method for unmanned aerial vehicle images. *IEEE Trans. Geosci. Remote. Sens.* **2013**, *52*, 1635–1647. [CrossRef]
32. Hao, P.; Zhan, Y.; Wang, L.; Niu, Z.; Shakir, M. Feature selection of time series MODIS data for early crop classification using random forest: A case study in Kansas, USA. *Remote. Sens.* **2015**, *7*, 5347–5369. [CrossRef]
33. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
34. Hubel, D.H.; Wiesel, T.N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* **1962**, *160*, 106–154. [CrossRef]
35. Szeliski, R. *Computer Vision: Algorithms and Applications*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2010.
36. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
37. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014; pp. 818–833.
38. Kussul, N.; Lavreniuk, M.; Skakun, S.; Shelestov, A. Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geosci. Remote. Sens. Lett.* **2017**, *14*, 778–782. [CrossRef]
39. Zhang, L.; Zhang, L.; Du, B. Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geosci. Remote. Sens. Mag.* **2016**, *4*, 22–40. [CrossRef]
40. Zhang, H.; Li, Y.; Zhang, Y.; Shen, Q. Spectral-spatial classification of hyperspectral imagery using a dual-channel convolutional neural network. *Remote. Sens. Lett.* **2017**, *8*, 438–447. [CrossRef]
41. He, Z.; Liu, H.; Wang, Y.; Hu, J. Generative adversarial networks-based semi-supervised learning for hyperspectral image classification. *Remote. Sens.* **2017**, *9*, 1042. [CrossRef]
42. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *arXiv* **2014**, arXiv:1406.2661
43. Zhan, Y.; Hu, D.; Wang, Y.; Yu, X. Semisupervised hyperspectral image classification based on generative adversarial networks. *IEEE Geosci. Remote. Sens. Lett.* **2017**, *15*, 212–216. [CrossRef]
44. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
45. Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; Abbeel, P. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *arXiv* **2016**, arXiv:1606.03657

46. Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv* **2015**, arXiv:1511.06434
47. Springenberg, J.T. Unsupervised and semi-supervised learning with categorical generative adversarial networks. *arXiv* **2015**, arXiv:1511.06390
48. Zhao, W.; Chen, X.; Chen, J.; Qu, Y. Sample generation with self-attention generative adversarial Adaptation Network (SaGAAN) for Hyperspectral Image Classification. *Remote. Sens.* **2020**, *12*, 843. [CrossRef]
49. Wang, M.; Deng, W. Deep visual domain adaptation: A survey. *Neurocomputing* **2018**, *312*, 135–153. [CrossRef]
50. Ma, W.; Pan, Z.; Yuan, F.; Lei, B. Super-resolution of remote sensing images via a dense residual generative adversarial network. *Remote. Sens.* **2019**, *11*, 2578. [CrossRef]
51. Bengana, N.M.; Heikkila, J. Improving land cover segmentation across satellites using domain adaptation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2020**, *14*, 1399–1410. [CrossRef]
52. Nielsen, A.A. The regularized iteratively reweighted MAD method for change detection in multi-and hyperspectral data. *IEEE Trans. Image Process.* **2007**, *16*, 463–478. [CrossRef]
53. Volpi, M.; Camps-Valls, G.; Tuia, D. Spectral alignment of multi-temporal cross-sensor images with automated kernel canonical correlation analysis. *ISPRS J. Photogramm. Remote. Sens.* **2015**, *107*, 50–63. [CrossRef]
54. Tuia, D.; Volpi, M.; Trolliet, M.; Camps-Valls, G. Semisupervised manifold alignment of multimodal remote sensing images. *IEEE Trans. Geosci. Remote. Sens.* **2014**, *52*, 7708–7720. [CrossRef]
55. Tzeng, E.; Hoffman, J.; Saenko, K.; Darrell, T. Adversarial discriminative domain adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7167–7176.
56. Hosseini-Asl, E.; Zhou, Y.; Xiong, C.; Socher, R. Augmented cyclic adversarial learning for low resource domain adaptation. *arXiv* **2018**, arXiv:1807.00374
57. Volpi, R.; Morerio, P.; Savarese, S.; Murino, V. Adversarial feature augmentation for unsupervised domain adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5495–5504.
58. Taigman, Y.; Polyak, A.; Wolf, L. Unsupervised cross-domain image generation. *arXiv* **2016**, arXiv:1611.02200
59. Bejiga, M.B.; Melgani, F. Gan-based domain adaptation for object classification. In Proceedings of the IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 26 September–2 October 2018; pp. 1264–1267.
60. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
61. Zhang, Y.; David, P.; Gong, B. Curriculum domain adaptation for semantic segmentation of urban scenes. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2020–2030.
62. Hoffman, J.; Wang, D.; Yu, F.; Darrell, T. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv* **2016**, arXiv:1612.02649.
63. Chang, W.L.; Wang, H.P.; Peng, W.H.; Chiu, W.C. All about structure: Adapting structural information across domains for boosting semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1900–1909.
64. Li, Y.; Yuan, L.; Vasconcelos, N. Bidirectional learning for domain adaptation of semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6936–6945.
65. Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote. Sens. Environ.* **2017**, *202*, 18–27. [CrossRef]
66. Hagolle, O.; Huc, M.; Villa Pascual, D.; Dedieu, G. A multi-temporal and multi-spectral method to estimate aerosol optical thickness over land, for the atmospheric correction of FormoSat-2, LandSat, VENμS and Sentinel-2 images. *Remote. Sens.* **2015**, *7*, 2668–2691. [CrossRef]
67. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; Lempitsky, V. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* **2016**, *17*, 2030–2096.
68. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
69. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
70. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
71. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 $\times$ 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
72. Rußwurm, M.; Körner, M. Self-attention for raw optical satellite time series classification. *ISPRS J. Photogramm. Remote. Sens.* **2020**, *169*, 421–435. [CrossRef]
73. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
74. Gretton, A.; Borgwardt, K.M.; Rasch, M.J.; Schölkopf, B.; Smola, A. A kernel two-sample test. *J. Mach. Learn. Res.* **2012**, *13*, 723–773.

75. Dziugaite, G.K.; Roy, D.M.; Ghahramani, Z. Training generative neural networks via maximum mean discrepancy optimization. *arXiv* **2015**, arXiv:1505.03906.
76. Long, M.; Zhu, H.; Wang, J.; Jordan, M.I. Deep transfer learning with joint adaptation networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 2208–2217.