

TimeMatch: Unsupervised Cross-Region Adaptation by Temporal Shift Estimation

Joachim Nyborg^{a,c}, Charlotte Pelletier^b, Sébastien Lefèvre^b, Ira Assent^a

^aDepartment of Computer Science, Aarhus University, Aarhus, Denmark

^bIRISA UMR 6074, Université Bretagne Sud, Vannes, France

^cFieldSense A/S, Aarhus, Denmark

Abstract

The recent developments of deep learning models that capture complex temporal patterns of crop phenology, have greatly advanced crop classification from Satellite Image Time Series (SITS). However, when applied to target regions spatially different from the training region, these models perform poorly without any target labels due to the temporal shift of crop phenology between regions. Although various unsupervised domain adaptation techniques have been proposed in recent years, no method explicitly learns the temporal shift of SITS and thus provides only limited benefits for crop classification. To address this, we propose TimeMatch, which explicitly accounts for the temporal shift for improved SITS-based domain adaptation. In TimeMatch, we first estimate the temporal shift from the target to the source region using the predictions of a source-trained model. Then, we re-train the model for the target region by an iterative algorithm where the estimated shift is used to generate accurate target pseudo-labels. Additionally, we introduce an open-access dataset for cross-region adaptation from SITS in four different regions in Europe. On our dataset, we demonstrate that TimeMatch outperforms all competing methods by 11% in average F1-score across five different adaptation scenarios, setting a new state-of-the-art in cross-region adaptation. Our source code and dataset are available at <https://github.com/jnyborg/timematch>.

Keywords: Satellite Image Time Series, Temporal Shift, Crop Classification, Domain Adaptation, Deep Learning

1. Introduction

Today, the availability of satellite image time series (SITS) data is rapidly increasing. For instance, the twin Sentinel-2 satellites provide imagery of the entire Earth every two to five days [1]. A frequent acquisition of images is crucial for vegetation-related remote sensing applications such as crop type classification [2, 3]. Multi-temporal data enables capturing the phenological development of crops (*i.e.*, the progressions of crop growth), a key dimension to discriminate each crop type [4]. Recently, the increasing availability of SITS along with advances in deep learning has led to crop classifiers with temporal neural architectures using convolutions [5, 6], recurrent units [7–10], self-attention [11, 12], or combinations thereof [13, 14].

These crop classification models achieve impressive performance by capturing the temporal structure of the problem but rely on the existence of a large amount of labeled training data. While unlabeled SITS are plenty, access to labels in the region of interest (the *target* domain) is often either costly or otherwise unavailable. A possible solution is to train a model in a region with labels available (the *source* domain) and apply it to the unlabeled target region. However, when the two regions are geographically different, the dissimilarity between the source and target data distributions can cause a source-trained model to perform poorly when applied to the target region [15–17].

Addressing the distributional shift problem to adapt a source-trained model to an unlabeled target domain is in

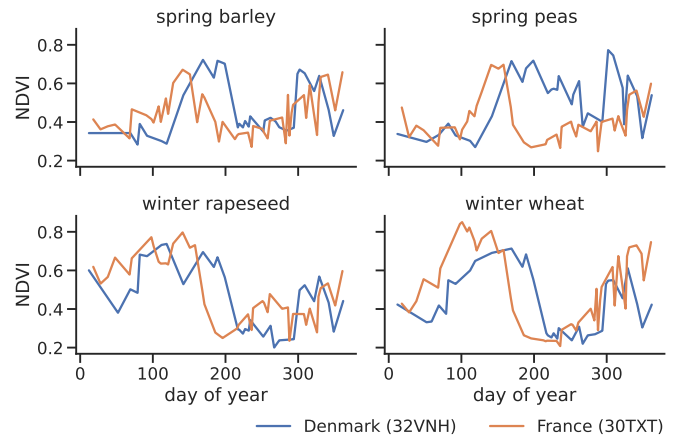


Figure 1: Normalized difference vegetation index (NDVI) time series for crops from two different Sentinel-2 tiles in Europe, indicating the growth of four crop types. Crops develop similarly in different regions, but the patterns are temporally shifted, *e.g.* if crops ripen at different times of the year.

machine learning known as unsupervised domain adaptation (UDA) [15, 18, 19]. Here, we consider the cross-region UDA problem for SITS [20], where we are provided with labeled data from a source region and unlabeled data from a target region. In this setting, the source and target data distributions differ due to changes in local conditions, such as the soil, climate, and farmer practices, which cause spectral and temporal shifts [15].

Addressing the temporal shift is of particular importance when adapting crop classifiers to new regions, as we illustrate in Figure 1. While crops in different regions have similar growth patterns, the timing of key growth stages, such as the peak of greenness, is shifted along the temporal axis. As crops are classified primarily by their unique growth patterns, the temporal shift may cause misclassifications when a source-trained model is applied to a target region. For example, the shift in time could cause the phenology of spring barley to appear similar to that of winter barley in the target. Thus, accounting for the temporal shift is a key factor in cross-region adaptation.

A possible approach could be to train models that are invariant to temporal shifts, such as by applying random temporal shifts to the training data. However, as the temporal shift could be the main feature that separates two crop types, shift-invariant models have reduced classification ability compared to shift-variant models.

Another approach is to apply existing UDA methods. Typically, these methods address domain adaptation by constraining the classifier to operate on domain-invariant features [21]. This is achieved by training the classifier to perform well on the source domain while minimizing a divergence measure between features extracted from the source and the target domains [20, 22, 23]. While these methods have been successfully applied in various applications [19, 24], they do not directly account for the temporal shift in SITS and have thus been reported to provide limited benefits in cross-region UDA [25]. More recently, self-training methods have emerged as a promising alternative to domain-invariant methods [26–30]. Self-training iteratively generates pseudo-labels [31] for the target domain and then uses them to retrain the model with target data. To account for noisy pseudo-labels caused by the domain shift, these methods typically incorporate a refinement step where the noise is reduced in various ways, such as with generative models [28] or learned confusion matrices [29]. Still, no method considers the particular case of SITS where the pseudo-label noise is caused by a temporal shift.

In this paper, we propose *TimeMatch*, a self-training method for cross-region UDA where we directly account for the temporal shift. *TimeMatch* consists of two components: (i) the temporal shift estimation and (ii) the *TimeMatch* learning algorithm.

Estimating the temporal shift directly from the target data is difficult, as the lack of labels hinders *e.g.* the comparison of class-wise vegetation indices as in Figure 1. To address this, we propose an unsupervised method where we estimate the temporal shift from target to source with a source-trained model. First, we obtain the softmax predictions of the model when input target data with different temporal shifts. Then, we choose the temporal shift with high prediction confidences across a diverse set of classes. We show that this approach corresponds well to the actual climatic differences between the two regions. Moreover, as correctly classified examples tend to have higher prediction confidence [32], the estimated shift enables us to

generate more accurate pseudo-labels in the target domain for self-training.

In *TimeMatch* learning, we therefore use self-training to adapt a model to the target domain. We propose an iterative algorithm where we alternate between temporal shift estimation and re-training the model for the target domain by learning from both labeled source data and pseudo-labeled target data. By doing so, the model learns discriminative target features for accurate crop classification in the target region.

Lastly, we present the *TimeMatch* dataset, a challenging new open-access dataset for training and evaluating cross-region models on SITS with over 300,000 annotated parcels from four different regions in Europe. Evaluated on this dataset, our approach outperforms all competing methods by 11% F1-score on average across five different cross-region UDA experiments.

In summary, our contributions are as follows:

- We propose a method for estimating the temporal shift between a labeled source region and an unlabeled target region to reduce their temporal discrepancy.
- We propose *TimeMatch*, a novel UDA method designed for the cross-region problem of SITS, where crop classification models are adapted to an unlabeled target region by self-training on temporally shifted data for improved performance compared to existing methods. Our source code is available at <https://github.com/jnyborg/timematch>.
- We release the *TimeMatch* dataset [33], a new dataset for training and evaluating cross-region UDA models on SITS from four different European regions.

This paper is organized as follows. Section 2 describes the existing literature related to our work. Section 3 describes the proposed method for temporal shift estimation and the *TimeMatch* learning algorithm. Section 4 presents our dataset and the experimental setup, and Section 5 the experimental results. Lastly, Section 6 concludes this work.

2. Related Work

TimeMatch is related to the existing work in unsupervised domain adaptation of learning domain-invariant features, time-series adaptation, cross-region adaptation, and self-training.

2.1. Domain-Invariant Methods

The predominant approach in UDA is to train the classifier to rely only on domain-invariant features [21, 24]. To this end, several works consider adversarial training [22, 34, 35]. In domain adversarial neural networks (DANN) [22, 34], the feature extractor is adversarially trained to produce domain-invariant features that are indistinguishable by a domain discriminator. Conditional domain adversarial networks (CDAN) [35] improves upon

DANN by conditioning the domain discriminator on classifier predictions in addition to features to enable the alignment of multimodal data distributions.

Another approach is to align the feature distributions directly by minimizing a divergence measure. Choices for divergence measure include maximum mean discrepancy (MMD) [23], correlation alignment [36], or optimal transport [37, 38]. Recently, JUMBOT [38] achieves state-of-the-art UDA results by using mini-batch unbalanced optimal transport to minimize the domain discrepancy of joint deep feature and label distributions.

While domain-invariant methods achieve strong results on computer vision datasets, they do not explicitly handle the temporal dimensions of SITS data and time series in general.

2.2. Time-Series Unsupervised Domain Adaptation

Few methods tackle the challenge of time series UDA. Current methods for time series are typically also based on learning domain-invariant features [39–41]. Recurrent domain adversarial neural network (R-DANN) and variational recurrent adversarial deep domain adaptation (VRADA) explore long short-term memory and variational recurrent neural networks as feature extractors, respectively, and learn domain-invariant features using the DANN method [39]. Likewise, the convolutional deep domain adaptation model for time series data (CoDATS) learns domain-invariant features with a temporal convolutional network with the DANN method [40]. However, while these methods are effective at learning domain-invariant features for time series, they are not designed to learn the temporal shift present in SITS.

2.3. Cross-Region Crop Classification

Lucas *et al.* [25] reports that existing UDA methods, including existing domain-invariant methods [42, 43], perform poorly when applied to cross-region UDA of SITS due to the temporal shift problem and the change in class distribution between the two regions. Recently, Wang *et al.* [20] proposed the phenology alignment network (PAN) as the first method for cross-region UDA of SITS. PAN learns domain-invariant features with MMD [23] and a feature extractor consisting of gated recurrent units and self-attention. Still, as PAN learns domain-invariant features, the temporal shift problem is not directly addressed.

2.4. Self-Training Methods

Semi-supervised learning (SSL) is a similar task to domain adaptation, but where the labeled and unlabeled data are sampled from the same data distribution [44]. Many SSL methods are based on pseudo-labeling [31] (also called self-training [45]), where the model’s own high-confidence predictions are used as labels for the unlabeled samples. In Mean Teacher [46], the model assumes a dual role as *teacher* and *student*. The student is updated by gradient descent with pseudo-labels generated by the teacher,

whereas the teacher is updated by an exponential moving average (EMA) of student parameters to reduce pseudo-label noise. FixMatch [45] generates pseudo-labels for weakly-augmented inputs, and uses confident pseudo-labels to self-train the model on strongly-augmented inputs, regularizing the model to output consistent pseudo-labels for random augmentations of the input.

Recently, self-training has emerged for UDA as an alternative to domain-invariant methods [26–30]. 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. However, since the domain shift often results in increased pseudo-label noise compared to SSL, existing methods introduce various refinement methods to reduce the noise, such as co-training [47], tri-training [26], conditional generative models [28], or confidence regularization [30]. Recently, Adversarial-Learned Loss for Domain Adaptation (ALDA) [29] proposes to refine the pseudo-labels with a noise-correcting domain discriminator.

Similar to this line of work, our approach is based on self-training. By directly accounting for the temporal shift, we can temporally align the target SITS with that of the source, which enables the generation of more accurate pseudo-labels compared to existing self-training methods that do not.

3. TimeMatch

In this section, we describe our proposed method TimeMatch for cross-region UDA. We begin by formally defining the problem setting, followed by an overview of how TimeMatch addresses it. We then give the details of the two TimeMatch components: temporal shift estimation and TimeMatch learning.

3.1. Problem Setting

In crop classification, the input is a sequence of satellite images $\mathbf{x}_i = (\mathbf{x}_i^{(1)}, \dots, \mathbf{x}_i^{(T_i)})$ of length T_i to be classified into one of the K crop classes. In object-based classification, which we focus on in this work, each $\mathbf{x}_i \in \mathbb{R}^{T_i \times N_i \times C}$ contains a sequence of N_i pixels of C spectral bands within a homogeneous, agricultural plot of land which we refer to as a *parcel*.

Each \mathbf{x}_i is accompanied by a sequence $\boldsymbol{\tau}_i = (\tau_i^{(1)}, \dots, \tau_i^{(T_i)})$ indicating the time $\tau_i^{(j)}$ at which each observation $\mathbf{x}_i^{(j)}$ is sampled. In practice, $\tau_i^{(j)}$ is typically represented by the day-of-year [8, 12], and makes it possible to account for the irregular temporal sampling of most satellites. The goal of the crop classification task is to learn a model which predicts class probabilities $p(y | (\mathbf{x}_i, \boldsymbol{\tau}_i)) \in \mathbb{R}^K$, typically learned with supervision from labels $y \in \{1, \dots, K\}$.

In this work, we consider the problem of cross-region UDA. We are given a source domain $\mathcal{D}^s = \{(\mathbf{x}_i^s, \boldsymbol{\tau}_i^s, y_i^s)\}_{i=1}^{n^s}$ of n^s labeled SITS and a target domain $\mathcal{D}^t = \{\mathbf{x}_i^t, \boldsymbol{\tau}_i^t\}_{i=1}^{n^t}$

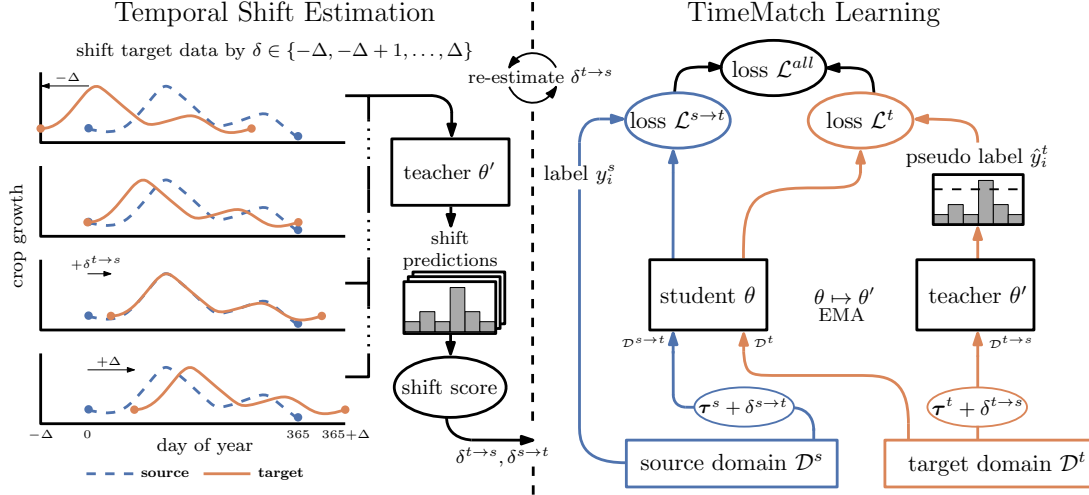


Figure 2: Overview of TimeMatch. Both the student and teacher are pre-trained on the source domain. *Temporal Shift Estimation*: We input shifted target data to the teacher model and obtain its predictions for each shift. We then score each shift by the confidence and diversity of the teacher predictions, and the shift with the best score is output as the temporal shift estimate $\delta^{t \rightarrow s}$ and $\delta^{s \rightarrow t} = -\delta^{t \rightarrow s}$. *TimeMatch Learning*: The teacher generates pseudo-labels for unlabeled target data shifted by $\delta^{t \rightarrow s}$. Then, the student is updated for (non-shifted) target data using the pseudo-labels, and for source data shifted by $\delta^{s \rightarrow t}$ using the available source labels. As a result, the student is adapted to the target domain with both generated target labels and actual source labels. After the student parameters have been updated with gradient descent, the teacher parameters are updated as an exponential moving average (EMA) of the student parameters. As both models adapt to the temporal shift of the target domain, the best shift for pseudo-labeling with the teacher changes and must be re-estimated. The EMA ensures the teacher adapts slowly which enables $\delta^{t \rightarrow s}$ to be re-estimated each epoch only for improved training efficiency and pseudo-label accuracy.

of n^t unlabeled SITS. We assume both the source and target domains consist of SITS acquired over a single year (January 1 to December 31) and in geographically different locations. The two domains can be associated with different data distributions, as changes in local conditions, *e.g.* soil, weather, climate, or farmer practices, cause domain discrepancies [15]. In this work, we focus on the domain discrepancies caused by temporal shifts (Section 1). Although not explicitly addressed in this work, there are other sources of discrepancies that might occur. For example, the local topography or soil conditions could impact not only the temporal development of crop growth, but also the spectral values, which could change the spectral signature of the same crop type in different regions.

Because of these data discrepancies, models which are trained with the labeled source domain can fail when applied to the unlabeled target domain [16], thus hindering the large-scale application of crop classifiers. To this end, our goal is to adapt a model trained on \mathcal{D}^s to make accurate predictions on \mathcal{D}^t . We do so by explicitly estimating the temporal shift between the two regions to generate accurate pseudo-labels for \mathcal{D}^t . Then, we re-train the model with target data using the pseudo-labels, thereby adapting the model to the spectral and temporal properties of the target region. We note that the classes in the source may not be exactly the same as the classes in the target. This complicates UDA, which typically assumes a closed-set setting [48], where the set of classes in the source and target domains are equal. For simplicity, we focus on a closed-set setting by adapting a classifier trained for the main $K - 1$ crop types in the source region, plus an “unknown” class

containing all remaining source data. This ensures that all target examples can be classified to either one of the $K - 1$ crop classes or “unknown”.

3.2. Approach Overview

Here we give an overview of how TimeMatch addresses the cross-region UDA problem before describing the full details. A visual presentation of TimeMatch is given in Figure 2. TimeMatch consists of two components (i) temporal shift estimation and (ii) TimeMatch learning.

We aim to estimate the temporal shift between the source and target regions to reduce their domain discrepancy (see Figure 1). We represent the temporal shift by a scalar $\delta^{t \rightarrow s} \in \mathbb{Z}$ (as the number of days), here in the direction from target to source. Note that the shift in the opposite direction is obtained by $\delta^{s \rightarrow t} = -\delta^{t \rightarrow s}$, so we only have to estimate one shift. To shift the target domain by $\delta^{t \rightarrow s}$, we write $\tau^t + \delta^{t \rightarrow s}$, meaning $\delta^{t \rightarrow s}$ is added element-wise to each target day-of-year. With our proposed method for temporal shift estimation (Section 3.3), we obtain estimates for $\delta^{t \rightarrow s}$ and $\delta^{s \rightarrow t}$.

In TimeMatch learning (Section 3.4), we use $\delta^{s \rightarrow t}$ to construct a target-shifted source domain $\mathcal{D}^{s \rightarrow t} = \{(\mathbf{x}_i^s, \tau_i^s + \delta^{s \rightarrow t}), y_i^s\}_{i=1}^{n^s}$, which has reduced domain discrepancy to the unlabeled target domain \mathcal{D}^t due to the temporal alignment. We therefore use self-training to learn from the labeled $\mathcal{D}^{s \rightarrow t}$ and unlabeled \mathcal{D}^t . To do so, TimeMatch learning unifies temporal shift estimation with the loss function of FixMatch [45] and the exponential moving average (EMA) training of Mean Teacher [46], as we explain next.

We first obtain source-trained parameters by training a crop classifier with \mathcal{D}^s . We then duplicate the trained classifier into two models: the *teacher* and the *student*. Our TimeMatch learning algorithm aims to adapt both the teacher and the student to the new target region with self-training. The teacher generates pseudo-labels for the target domain to train the student, and the knowledge learned by the student is then updated back to the teacher, thus the pseudo-labels used to train the student itself are improved. We generate pseudo-labels by using $\delta^{t \rightarrow s}$ to create an adapted target domain $\mathcal{D}^{t \rightarrow s} = \{\mathbf{x}_i^t, \boldsymbol{\tau}_i^t + \delta^{t \rightarrow s}\}_{i=1}^{n^t}$. As $\mathcal{D}^{t \rightarrow s}$ is temporally aligned with \mathcal{D}^s , the source-initialized teacher generates more accurate pseudo-labels for $\mathcal{D}^{t \rightarrow s}$ than \mathcal{D}^t . The student is then trained with labeled $\mathcal{D}^{s \rightarrow t}$ and pseudo-labeled \mathcal{D}^t via the FixMatch loss [45], thereby leveraging both the available source labels and the target pseudo-labels to adapt the student to the target domain.

After updating the student, the teacher is updated via an EMA of the student parameters. As the two models adjust to the temporal shift of the target domain, the best shift $\delta^{t \rightarrow s}$ for pseudo-labeling with the teacher gradually moves to zero during TimeMatch learning. To adjust to the changing shift and ensure the pseudo-labels are consistently accurate, it is necessary to re-estimate the temporal shift of the teacher as it learns. However, repeating temporal shift estimation is computationally expensive, and drastically increases training time if done each training iteration. Therefore, in Section 3.4.3, we discuss how EMA training alleviates this issue by enabling the re-estimation to be done only once per epoch.

Next, we first describe our method for estimating the temporal shift before describing the loss function and learning algorithm of TimeMatch learning.

3.3. Temporal Shift Estimation

Estimating the temporal shift directly from the data is difficult, as labels are not available in the target domain. Without labels, we cannot separate the target data into each crop type, which prevents the computation of *e.g.* vegetation indices to compare the source and target phenology of each crop type directly.

Instead, we propose to estimate the temporal shift by calculating statistics on the predictions of a source-trained model when input temporally shifted target data. By doing so, we estimate the shift that aligns the target data with the source crop phenology learned by a model, leveraging the classification ability of the trained model to estimate the shift from unlabeled data. Another benefit of this approach is that it enables re-estimation of the best temporal shift for pseudo-labeling as the learned phenology of the model changes from source to target in TimeMatch learning.

One possible value to measure is the confidence of the model predictions. Intuitively, when a source-trained model is applied to correctly shifted target data, it should output more confident predictions than for incorrectly shifted target data. As correctly classified examples tend to have

more confident predictions than wrongly classified or out-of-distribution examples [32], we argue that a confident temporal shift indicates a better alignment of the target domain with the source which results in accurate pseudo-labels and reduced domain discrepancy.

The confidence of a model for a particular shift $\delta^{t \rightarrow s}$ can be measured by the expected entropy:

$$\mathbb{E}_{(\mathbf{x}^t, \boldsymbol{\tau}^t) \sim \mathcal{D}^t} [H(p_\theta(y | (\mathbf{x}^t, \boldsymbol{\tau}^t + \delta^{t \rightarrow s})))] , \quad (1)$$

where H denotes the entropy, here computed over the predictions of the model θ when input temporally shifted target data sampled from \mathcal{D}^t .

To estimate a temporal shift with entropy, Equation 1 should be computed for each possible shift $\delta^{t \rightarrow s} \in \{-\Delta, -\Delta + 1, \dots, \Delta\}$, and the estimated shift is then the one with lowest entropy. Here, Δ defines the maximum possible shift (in days) to estimate between the source and target regions.

However, due to the class imbalance of SITS, relying on expected entropy alone could result in choosing a shift where the model outputs confident predictions for only the most frequent classes while ignoring the less frequent classes. This would hinder the adaptation of the model for the less frequent target classes. To address this problem, the diversity of the predicted marginal distribution should also be considered in the estimation. The marginal is given by:

$$p_\theta(y) = \mathbb{E}_{(\mathbf{x}^t, \boldsymbol{\tau}^t) \sim \mathcal{D}^t} [p_\theta(y | (\mathbf{x}^t, \boldsymbol{\tau}^t + \delta^{t \rightarrow s}))] , \quad (2)$$

that is, the expected predictions of the model (parameterized by θ) when input shifted target data.

Ideally, the marginal distribution should match the class distribution of the target domain, as this indicates a shift where the model predicts a diverse set of classes according to their actual frequency. But since target labels are unavailable, so is the target class distribution. Instead, inspired by metrics for evaluating image generative models, we consider two options to address this: the Inception score [49] (IS), and the activation maximization score [50] (AM). Both metrics consider the entropy and marginal of a pre-trained model, but IS scores the marginal distribution by its similarity to a uniform distribution, whereas AM uses the actual class distribution.

As these metrics were originally proposed to evaluate the quality of generated images, we describe next how we repurpose them for temporal shift estimation. Finally, we describe an algorithm where IS is used to initialize the temporal shift for estimating the target class distribution with pseudo-labels and enable a better temporal shift estimate with AM.

3.3.1. Inception Score

IS is computed for a temporal shift δ by:

$$\begin{aligned} \text{IS}(\delta^{t \rightarrow s}, \theta) &= \mathbb{E}_{(\mathbf{x}^t, \boldsymbol{\tau}^t)} [D_{\text{KL}}(p_\theta(y | (\mathbf{x}^t, \boldsymbol{\tau}^t + \delta^{t \rightarrow s})) \parallel p_\theta(y))] \quad (3) \\ &= H(p_\theta(y)) - \mathbb{E}_{(\mathbf{x}^t, \boldsymbol{\tau}^t)} [H(p_\theta(y | (\mathbf{x}^t, \boldsymbol{\tau}^t + \delta^{t \rightarrow s})))] \quad (4) \end{aligned}$$

where $D_{\text{KL}}(\cdot \parallel \cdot)$ is the KL-divergence between two distributions, here the conditional distribution $p_\theta(y|(\mathbf{x}^t, \boldsymbol{\tau}^t + \delta))$ and marginal distribution $p_\theta(y)$ predicted with model parameters θ . Higher values of IS indicate a better δ , as when the conditional and marginal distributions are different, this corresponds to a temporal shift where the former has low entropy (*i.e.*, the model is confident), and the latter has high entropy (*i.e.*, the model predicts a diverse set of classes). Hence, the temporal shift $\delta^{t \rightarrow s}$ is estimated by:

$$\delta_{IS}^{t \rightarrow s}(\theta^s) = \underset{\delta^{t \rightarrow s} \in \{-\Delta, \dots, \Delta\}}{\operatorname{argmax}} \quad \text{IS}(\delta^{t \rightarrow s}, \theta^s), \quad (5)$$

where the estimated temporal shift maximizes IS for a source-trained model parameterized by θ^s when applied to target data.

3.3.2. AM Score

A shortcoming of IS is that the highest score is achieved when $p_\theta(y)$ is uniform [51], which corresponds to an even distribution of classes in the target domain. For SITS, where the class distribution is often highly imbalanced, this may cause IS to estimate a suboptimal shift. AM [50] addresses this issue by taking the actual target class distribution C^t into account:

$$\begin{aligned} \text{AM}(\delta^{t \rightarrow s}, \theta, C^t) &= \mathbb{E}_{(\mathbf{x}^t, \boldsymbol{\tau}^t)} [H(p_\theta(y|(\mathbf{x}^t, \boldsymbol{\tau}^t + \delta^{t \rightarrow s})))] \\ &\quad + D_{\text{KL}}(C^t \parallel p_\theta(y)). \end{aligned} \quad (6)$$

AM consists of two terms: the first term is an entropy term on the conditional distribution, and the second is the KL-divergence between the underlying class distribution C^t and the marginal distribution. Lower values of AM indicate a better δ , as the model is confident in its predictions, and the actual class distribution of the data matches the predicted distribution of classes. The temporal shift $\delta^{t \rightarrow s}$ is estimated by:

$$\delta_{AM}^{t \rightarrow s}(\theta^s, C^t) = \underset{\delta^{t \rightarrow s} \in \{-\Delta, \dots, \Delta\}}{\operatorname{argmin}} \quad \text{AM}(\delta^{t \rightarrow s}, \theta^s, C^t). \quad (7)$$

where the estimated temporal shift minimizes AM.

3.3.3. Algorithm for Estimating Temporal Shift

While AM is more accurate at estimating the temporal shift, it requires knowledge of the target class distribution C^t , which is not available. To address this, we propose to approximate the target class distribution for AM by pseudo-labels obtained with IS. We show our approach in Algorithm 1. First, we use IS (Equation 5) to estimate an initial shift $\delta^{t \rightarrow s}$ (line 3). This initial estimate allows us to shift the target domain so that more accurate pseudo-labels can be generated with a source-trained model. We then use the pseudo-labels to estimate the target class distribution \hat{C}^t (lines 4-5). Finally, we re-estimate the temporal shift more accurately with AM and \hat{C}^t (line 6).

Algorithm 1: ESTIMATETEMPORALSHIFT

- 1 **Input:** Source-trained parameters θ^s , target domain \mathcal{D}^t , target class distribution estimate \hat{C}^t
 - 2 **if** $\hat{C}^t = \mathbf{0}$ **then**
 - 3 Estimate temporal shift $\delta^{t \rightarrow s} \leftarrow \delta_{IS}^{t \rightarrow s}(\theta^s)$ (Eq. 5)
 - 4 Compute pseudo labels for each $(\mathbf{x}_i^t, \boldsymbol{\tau}_i^t) \in \mathcal{D}^t$:
 $\hat{y}_i^t \leftarrow \operatorname{argmax}_y (p_{\theta^s}(y|\mathbf{x}_i^t, \boldsymbol{\tau}_i^t + \delta^{t \rightarrow s}))$
 - 5 Estimate class distribution $\hat{C}_y^t \leftarrow \frac{1}{n^t} \sum_{i=1}^{n^t} \mathbf{1}_{\hat{y}_i^t=y}$
for $y \in \{1, \dots, K\}$
 - 6 Estimate temporal shift $\delta^{t \rightarrow s} \leftarrow \delta_{AM}^{t \rightarrow s}(\theta^s, \hat{C}^t)$ (Eq. 7)
 - 7 **Output:** Temporal shift $\delta^{t \rightarrow s}$
-

3.4. TimeMatch Learning

With our method for estimating the temporal shift, we can reduce the domain discrepancy between the source and target domains. The TimeMatch learning algorithm uses the temporal shift to train the student model for the target domain from teacher-generated pseudo-labels via the FixMatch loss [45] and EMA training [46]. We present the complete TimeMatch algorithm in Algorithm 2, and describe the details of each step in the following.

3.4.1. Pre-training on the Source Domain

As we rely on the teacher to generate pseudo-labels to train the student, it is important to obtain a good initialization for both models. Additionally, temporal shift estimation requires a source-trained model. Thus, we first use the labeled source domain to obtain source-trained model parameters θ^s . Given a batch of labeled source data from \mathcal{D}^s , we optimize the following loss function:

$$\mathcal{L}^s = \frac{1}{B} \sum_{i=1}^B L(p_{\theta^s}(y|(\mathbf{x}_i^s, \boldsymbol{\tau}_i^s)), y_i^s), \quad (8)$$

where $L(\cdot, \cdot)$ is a classification loss (*e.g.* cross-entropy or focal loss [52]) and B the batch size. After pre-training, we initialize the parameters of the student θ and teacher θ' from θ^s (line 2).

3.4.2. TimeMatch Loss

The TimeMatch loss consists of two terms: a supervised loss $\mathcal{L}^{s \rightarrow t}$ applied to the adapted source domain $\mathcal{D}^{s \rightarrow t}$ and an unsupervised loss \mathcal{L}^t applied to the unlabeled target domain \mathcal{D}^t . Our loss is based on the FixMatch loss [45]. To regularize the model to predict consistent pseudo-labels on randomly augmented versions of the same inputs, FixMatch applies two types of augmentation functions: *weakly*-augmented $a(\cdot)$ and *strongly*-augmented $A(\cdot)$, corresponding to simple and extensive augmentations of the input. We describe the form of augmentations we use for $a(\cdot)$ and $A(\cdot)$ in Section 4.4.

Let $\delta^{s \rightarrow t}$ and $\delta^{t \rightarrow s}$ be temporal shifts estimated given by Algorithm 1 using the teacher (line 5-7). To compute the

Algorithm 2: TIMEMATCH

```

1 Input: Labeled source domain  $\mathcal{D}^s$ , unlabeled target domain  $\mathcal{D}^t$ , source-trained parameters  $\theta^s$ , total epochs  $n$  and
   iterations  $m$ , pseudo label threshold  $\epsilon$ , trade-off value  $\lambda$ , EMA decay rate  $\alpha$ , learning rate  $\eta$ 
2 Initialize student parameters  $\theta \leftarrow \theta^s$  and teacher parameters  $\theta' \leftarrow \theta^s$ 
3 Initialize estimated target class distribution  $\hat{C}^t = \mathbf{0}$ 
4 for epoch = 1 to  $n$  do
5   Estimate temporal shift with teacher:  $\delta^{t \rightarrow s} \leftarrow \text{ESTIMATE\_TEMPORAL\_SHIFT}(\theta', \mathcal{D}^t, \hat{C}^t)$ 
6   if epoch = 1 then
7     Initialize  $\delta^{s \rightarrow t} \leftarrow -\delta^{t \rightarrow s}$ 
8   for iteration = 1 to  $m$  do
9     Sample mini-batches of size  $B$  from source  $\mathcal{S} = \{(\mathbf{x}_i^s, \boldsymbol{\tau}_i^s, y_i^s)\}_{i=1}^B$  and target  $\mathcal{T} = \{(\mathbf{x}_i^t, \boldsymbol{\tau}_i^t)\}_{i=1}^B$ 
10    With  $\mathcal{S}$  shifted by  $\delta^{s \rightarrow t}$ , compute source loss  $\mathcal{L}^{s \rightarrow t}$  (Eq. 9)
11    For each example in  $\mathcal{T}$  shifted by  $\delta^{t \rightarrow s}$ , generate teacher prediction  $\mathbf{q}_i^t$  and pseudo labels  $\hat{y}_i^t$  (Eq. 10 and 11)
12    With  $\mathcal{T}$  and confident pseudo labels  $\hat{y}_i^t$  with  $\max(\mathbf{q}_i^t) > \epsilon$ , compute target loss  $\mathcal{L}^t$  (Eq. 12)
13    Update student by gradient:  $\theta \leftarrow \theta - \gamma \nabla_{\theta}(\mathcal{L}^{s \rightarrow t} + \lambda \mathcal{L}^t)$ 
14    Update teacher by EMA:  $\theta' \leftarrow (1 - \alpha)\theta + \alpha\theta'$ 
15  Re-estimate class distribution:  $\hat{C}_y^t \leftarrow \frac{1}{mB} \sum_i \mathbf{1}_{\hat{y}_i^t=y}$  for  $y \in \{1, \dots, K\}$  (using all pseudo labels from epoch)
16 Output: Student parameters  $\theta$ 

```

supervised loss on the source domain, we use $\delta^{s \rightarrow t}$ to align the source domain with the target domain and optimize:

$$\mathcal{L}^{s \rightarrow t} = \frac{1}{B} \sum_{i=1}^B L(p_{\theta}(y|A(\mathbf{x}_i^s, \boldsymbol{\tau}_i^s + \delta^{s \rightarrow t})), y_i^s), \quad (9)$$

using source labels y_i^s to update the student θ on strongly augmented source data shifted by $\delta^{s \rightarrow t}$. This loss makes it possible for the student to learn the target phenology from shifted source data (line 10).

To generate pseudo-labels for the target domain, we obtain the predicted class distribution from the teacher when input source-shifted target data:

$$\mathbf{q}_i^t = p_{\theta'}(y|A(\mathbf{x}_i^t, \boldsymbol{\tau}_i^t + \delta^{t \rightarrow s})), \quad (10)$$

where the teacher θ' is input a weakly-augmented target sample, shifted by $\delta^{t \rightarrow s}$. Then, we use

$$\hat{y}_i^t = \text{argmax}(\mathbf{q}_i^t) \quad (11)$$

as pseudo-label (line 11). The student θ is then updated on strongly-augmented target data for confident pseudo-labels (line 10):

$$\mathcal{L}^t = \frac{1}{B} \sum_{i=1}^B \mathbf{1}_{\max(\mathbf{q}_i^t) > \epsilon} L(p_{\theta}(y|A(\mathbf{x}_i^t, \boldsymbol{\tau}_i^t)), \hat{y}_i^t), \quad (12)$$

where $\mathbf{1}$ is the indicator function, and ϵ is the confidence threshold for using a pseudo-label. With this loss, the student is trained with target data using pseudo-labels. The total loss minimized by the student in TimeMatch is:

$$\mathcal{L}^{all} = \mathcal{L}^{s \rightarrow t} + \lambda \mathcal{L}^t, \quad (13)$$

where λ is a scalar hyperparameter to control the trade-off between the supervised and the unsupervised loss (line 13).

3.4.3. EMA training and re-estimating temporal shift

By optimizing \mathcal{L}^{all} , the student and teacher are trained only for the target phenology, as $\mathcal{L}^{s \rightarrow t}$ shifts the time of the source to the target, while \mathcal{L}^t keeps the target in its original time. This loss enables a source-trained model to adapt to the crop phenology of the target domain.

However, by doing so, the source domain is gradually “forgotten”, and as a result, it becomes unnecessary to apply the temporal shift $\delta^{t \rightarrow s}$ for pseudo-labeling the target domain with the teacher. This causes $\delta^{t \rightarrow s}$ to gradually move to zero during TimeMatch learning. Thus, if $\delta^{t \rightarrow s}$ is fixed to the same shift, the target samples will be wrongly shifted, which results in incorrect pseudo-labels. To address this, we re-estimate the temporal shift for the teacher during TimeMatch learning. As Algorithm 1 chooses the shift based on the confidence and diversity of model predictions, re-estimating the temporal shift with the teacher ensures the generated pseudo-labels remain accurate during training.

However, if the teacher is a direct copy of the student, the model will rapidly adapt to the target domain, which requires the temporal shift to be re-estimated every few iterations. But doing so drastically increases training time, as Equation 7 requires forwarding a large sample of target data for each possible temporal shift. We address this by introducing EMA training, where the teacher is slowly updated via an EMA of the student parameters (line 14):

$$\theta' \leftarrow (1 - \alpha)\theta + \alpha\theta', \quad (14)$$

where α is a decay rate. By choosing α close to 1, we reduce the rate at which the teacher adapts to the target domain, enabling the re-estimation of $\delta^{t \rightarrow s}$ to be done only once each epoch (line 5). Moreover, by averaging model weights via EMA, we also obtain less noisy pseudo-labels [46].

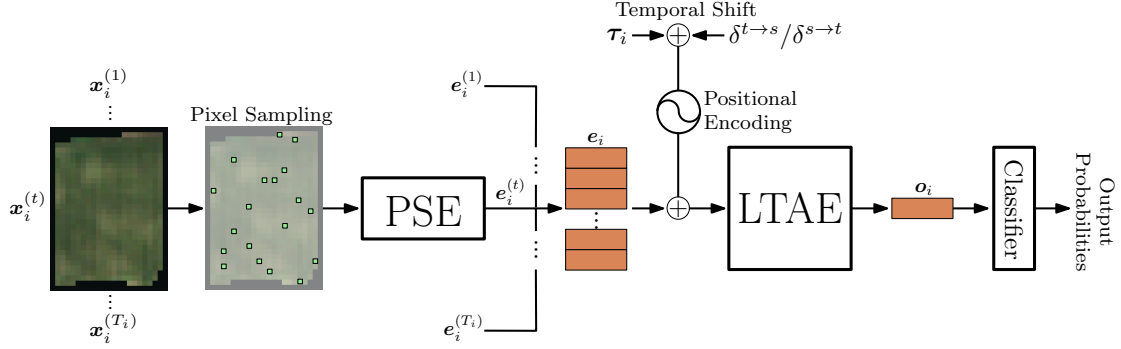


Figure 3: Overview of the PSE+LTAE model [12, 53]. Given SITS of an agricultural parcel, the PSE module process each time step independently by embedding a random sample of pixels. The results are then concatenated into a sequence of embeddings \mathbf{e}_i . The observation dates τ_i , which we add temporal shifts to, are input to the model by adding their positional encoding to \mathbf{e}_i . The result is temporally processed by LTAE to a single embedding \mathbf{o}_i which is then passed to the classifier.

By re-estimating the temporal shift, the teacher and the student can both evolve jointly during training, resulting in better pseudo-labels for improved cross-region adaptation. Note that $\delta^{s \rightarrow t}$ is not re-estimated (line 7). The first shift estimate represents the shift of the data, whereas the re-estimated shift represents the shift of the teacher. By fixing $\delta^{s \rightarrow t}$ to the initial estimate, the source domain is kept aligned with the target domains during training, which enables semi-supervised learning.

4. Dataset and Materials

This section presents the TimeMatch dataset [33] and the materials for our experiments. We first introduce the crop classification model we use, followed by a description of the dataset and its pre-processing. Then, we describe the competitors and our implementation. Our source code is publicly available, and contains the implementation of TimeMatch and the competitors, a link to download our dataset, and the full experimental results: <https://github.com/jnyborg/timematch>.

4.1. Network Architecture

As model, we use the existing object-based crop classifier PSE+LTAE introduced by Sainte Fare Garnot *et al.* [12, 53]. The network consists of two modules: the pixel-set encoder (PSE) and the lightweight temporal attention encoder (LTAE). See Figure 3 for an overview.

The PSE module handles the spatial and spectral context of SITS. Given SITS of an agricultural parcel, PSE samples a random pixel-set of size S among the N_i available pixels within the parcel. The PSE is efficient compared to *e.g.* convolutions, which are time and memory-consuming when applied to irregularly sized parcels. As spatial information is lost by doing so, the PSE supports an optional extra input with various geometrical properties of the given parcel, such as its area. We do not input this extra feature to avoid biasing the model towards the shapes of parcels in the source region, which typically change depending on the local farmer practices. Thus, we only input the sequence

$\mathbf{x}_i \in \mathbb{R}^{T_i \times N_i \times C}$, which is then embedded by the PSE for each time step independently.

The LTAE module [53] handles the temporal context by applying self-attention [54] with modifications to output a single embedding. It improves the accuracy and computational efficiency compared to the original TAE [12] by a channel grouping strategy and a learnable master query. The additional input τ_i is input to LTAE by encoding the days with sinusoidal positional encoding [54] and adding the result to the output of PSE. As the positional encoding does not support negative inputs, we input negative temporal shifts by offsetting each τ_i by the maximum temporal shift Δ . Given the sequence of PSE-embeddings and the encoded τ_i , LTAE outputs a single embedding \mathbf{o}_i , which is then classified by a multi-layer perceptron to produce class probabilities $p(y | (\mathbf{x}_i, \tau_i)) \in \mathbb{R}^K$.

4.2. The TimeMatch Dataset

The TimeMatch dataset [33] contains SITS from Sentinel-2 Level-1C products in top-of-atmosphere reflectance. Four Sentinel-2 tiles are chosen in various climates: 33UVP (Austria), 32VNH (Denmark), 30TXT (mid-west France), and 31TCJ (southern France), abbreviated as AT1, DK1, FR1, and FR2, respectively. A map of the tiles is shown in Figure 4. We use all available observations with cloud coverage $\leq 80\%$ and coverage $\geq 50\%$ between January 2017 and December 2017. Figure 5 shows the resulting acquisition dates for the four tiles. We leave out the atmospheric bands (1, 9, and 10), keeping $C = 10$ spectral bands. The 20m bands are bilinearly interpolated to 10m.

For ground truth data, we retrieve geo-referenced parcel shapes and their crop type labels from the openly available Land Parcel Identification System (LPIS) records in Denmark¹, France², and Austria³. We select 15 major crop classes in Europe and label any remaining parcels as unknown. Figure 6 shows the selected classes and their frequency in each tile.

¹<https://kortdata.fvm.dk/download> (“Marker”)

²<http://professionnels.ign.fr/rpg> (“RPG”)

³<https://www.data.gv.at> (“INVEKOS Schlge”)

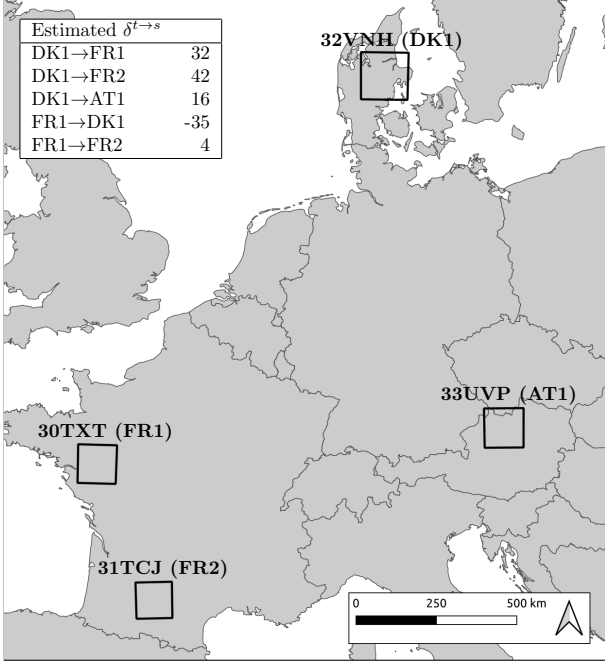


Figure 4: Locations of the four European Sentinel-2 tiles in the TimeMatch dataset. In the upper left corner, we show the temporal shifts $\delta^{t \rightarrow s}$ estimated by Algorithm 1 with a source-trained model.

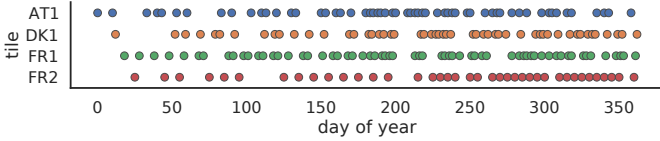


Figure 5: Acquisition dates for each Sentinel-2 tile in our dataset. The inputs are irregularly sampled with variable temporal length.

We pre-process the parcels by applying 20m erosion and removing all parcels with an area of less than 1 hectare. This reduces label noise by removing pixels near the border of parcels, which are often less representative of the given crop class compared to the pixels in the middle, and also by removing small or thin polygons, which are typically miscellaneous classes such as field borders. The SITS are pre-processed for object-based classification by cropping the pixels within each parcel to input sequences $\mathbf{x}_i \in \mathbb{R}^{T_i \times N_i \times 10}$. Each input is then randomly assigned to the train/validation/test sets of each Sentinel-2 tile by a 70%/10%/20% ratio. Note that this process assumes knowledge of parcel shapes in the target region. If this is not available, TimeMatch may instead be applied for pixel-based classification by inputting single pixels ($S = 1$) to PSE+LTAE. We choose five different cross-region tasks (written as “source \rightarrow target”): DK1 \rightarrow FR1, DK1 \rightarrow FR2, DK1 \rightarrow AT1, FR1 \rightarrow DK1, and FR1 \rightarrow FR2. When a Sentinel-2 tile is chosen as source, all labels of the train and validation sets are available for training. When a tile is the target region, no labels are available, except for the final evaluation on the test set. In contrast, many existing UDA methods assume that a labeled validation set is available for

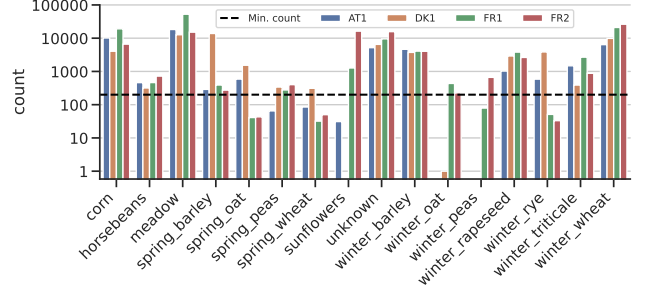


Figure 6: Class frequencies (log scale) for each Sentinel-2 tile in the TimeMatch dataset. The dashed line indicates the threshold for the source region when selecting a class as part of the K classes.

the target domain, and use it during training *e.g.* to select the best model [23, 29, 34, 35]. However, this assumption is unrealistic, as if labels were available in real-world scenarios, they would be better used for training the model. Instead, we report all cross-region UDA test results with the model output at the end of training. Still, hyperparameters must be chosen with a labeled validation set. Thus, we tune hyperparameters with the target validation set for only one task, DK1 \rightarrow FR1, and apply the found hyperparameters to all remaining tasks (as done in [38]).

The class distributions between regions differ significantly, and there may not be enough examples of a crop type in the source region for a model to learn their classification. Thus, when pre-training models on source data, we only use a subset of the available crop types with at least 200 examples in the source region (as indicated by the dashed line in Figure 6). The remaining classes are set as “unknown”. When evaluating on the target data, we report results on the same selection of source classes no matter their frequency in the target.

4.3. Comparisons

Baselines. We consider the following baseline methods:

- *Source-Trained* is PSE+LTAE trained on the source domain and applied to the target domain without domain adaptation. This result represents the lower bound cross-region performance of the model.
- *Target-Trained* is PSE+LTAE trained with labeled target data using the same classes as the source-trained. We note that by training with the source classes, which is required for comparison, infrequent classes may not be learned properly which increases the variance of the results. This result can be seen as the upper bound cross-region performance if labels were available in the target region.

Competing UDA Methods. We compare TimeMatch to five existing UDA methods. We reproduce these methods for SITS by replacing the original feature extractor with PSE+LTAE. For domain-invariant methods, we align the LTAE feature vector input to the final classifier (*i.e.*, \mathbf{o}_i in Figure 3), similar to the original approach in these methods.

We compare to the following methods:

- *FixMatch* [45] is TimeMatch without the temporal shift estimation. As this method is semi-supervised learning, it shows whether UDA or SSL is more beneficial for cross-region adaptation.
- *MMD* [23] learns domain-invariant features by minimizing the maximum mean discrepancy metric.
- *DANN* [22] uses a domain classifier to learn domain-invariant features with adversarial training.
- *CDAN+E* [35] improves upon DANN by conditioning the domain classifier on the classification output and minimizing an entropy loss on target data.
- *ALDA* [29] is a self-training method where pseudo-labels are refined by a noise-correcting domain discriminator. This method is in essence the most similar to TimeMatch.
- *JUMBOT* [38] learns domain-invariant features by a discrepancy measure based on optimal transport.

We note that time-series domain adaptation methods R-DANN, VRADA [39] and CoDATS [40] also employ DANN to align the features extracted by temporal network architectures. Thus, the only difference between VRADA, CoDATS, and the DANN approach mentioned here is the backbone architecture, which in our case is the temporal model PSE+LTAE.

PAN [20], a UDA method for SITS, learns domain-invariant features by minimizing the MMD loss for a temporal crop classification network. Unfortunately, we were unable to gain access to the source code of PAN for comparison. As an alternative, we include the MMD comparison, which is similar to PAN, except the crop classifier is changed to PSE+LTAE.

ShiftAug. To verify the benefits of estimating the temporal shift compared to training models that are invariant to temporal shifts, we implement a simple data augmentation technique to train shift-invariant models that we name *ShiftAug*. During training, ShiftAug uniformly samples $\delta \sim \mathcal{U}(-\Delta, \Delta)$ for each training example and shifts the example by $(\mathbf{x}_i, \tau_i + \delta)$. By extending the training data to contain all valid temporal shifts with uniform probability, ShiftAug enables training models with invariance towards shifts. Note that ShiftAug is incompatible with the temporal shift estimation presented in Algorithm 1, which requires a shift-variant model.

We implement all competing methods with and without ShiftAug. This reveals the degree at which existing methods can implicitly learn shift-invariance.

4.4. Implementation Details

All experiments are implemented in PyTorch [55] and trains on a single NVIDIA 1080 Ti GPU. Our implementation is based on the source code of PSE+LTAE [53].

Source-training. To initialize models on the labeled source domain, we follow the original training approach of PSE+LTAE [12]. We train for 100 epochs with the Adam [56] optimizer with an initial learning rate of 0.001 and we decay the learning rate using a cosine annealing schedule [57]. We use weight decay of 0.0001, batch size 128, and focal loss $\gamma = 1$. Inputs are normalized to $[0, 1]$ by dividing by the max 16-bit pixel value $2^{16} - 1$. The best source-trained model is selected using the source validation set. We augment the inputs by randomly sub-sampling 30 time steps. The pixel-set size of PSE is set to $S = 64$ during training. The same setup is used for the target-trained model. For the final evaluation, we do not sample time steps or pixels, and instead input all available time steps ($T = T_i$) and pixels ($S = S_i$) for each example to the model. This ensures deterministic test results, and we also observe slightly improved results by doing so.

ShiftAug. When training with ShiftAug, all training data (both source and target) are randomly shifted during training as described in Section 4.3. ShiftAug is disabled during evaluation.

TimeMatch. We use the same training setup as the source-trained model but instead train for 20 epochs with a lower initial learning rate of 0.0001. We define an epoch as 500 iterations to fix the frequency in which the temporal shift is re-estimated. We use maximum temporal shift $\Delta = 60$ days, as we did not observe shifts greater than 2 months for our dataset in Europe. We set the trade-off hyperparameter $\lambda = 2.0$ in Equation 13, EMA keep-rate $\alpha = 0.9999$, and pseudo-label threshold $\tau = 0.9$. A sensitivity analysis of these hyperparameters is provided in Section 5.4. For the FixMatch [45] augmentations, we use the identity function for the weak $a(\cdot)$ in Equation 10 and randomly sub-sample time steps for the strong $A(\cdot)$ in Equations 9 and 12. These are used for simplicity, and we leave the use of more advanced augmentations for SITS to future work. At each iteration, we sample two mini-batches of size 128, one from the source and one from the target, in order to calculate the TimeMatch objective in Equation 13. We use a class-balanced mini-batch sampler for the source domain to ensure each source mini-batch contains roughly the same number of samples for each class. This reduces the class imbalance problem for the source domain for improved performance. Additionally, we apply domain-specific batch normalization [58–60] by forwarding the source and target mini-batches separately instead of concatenated. This ensures the batch normalization [61] statistics are calculated separately for each domain, for improved adaptation.

Competing Methods. We re-implement the competitors MMD, DANN and CDAN+E following the domain adaptation library in [62], and use the original source codes for ALDA [29] and JUMBOT [38]. FixMatch [45] follows our re-implementation for TimeMatch with an EMA teacher

Method	ShiftAug	DK1→FR1	DK1→FR2	DK1→AT1	FR1→DK1	FR1→FR2	Avg.
Source-trained	✗	28.3±1.9	29.0±5.2	43.4±4.0	24.9±2.0	70.3±1.9	39.2±3.0
	✓	40.9±0.8	37.4±2.3	48.9±2.8	47.3±1.9	70.5±1.1	49.0±1.8
FixMatch [45]	✗	24.2±4.0	28.2±6.9	37.4±5.6	26.2±1.8	70.4±0.9	37.3±3.8
	✓	48.2±1.3	44.2±3.2	57.4±2.2	51.3±1.6	67.7±0.2	53.7±1.7
MMD [23]	✗	36.6±0.7	35.5±0.6	49.7±2.0	32.5±2.0	61.6±2.6	43.2±1.6
	✓	42.2±0.4	39.5±0.8	48.9±2.4	42.8±2.3	59.0±2.7	46.5±1.7
DANN [22]	✗	38.7±0.7	37.3±0.6	52.0±1.4	34.0±1.8	71.0±0.2	46.6±0.9
	✓	45.3±2.2	44.1±1.4	52.4±1.4	42.9±2.5	68.7±0.5	50.7±1.6
CDAN+E [35]	✗	39.3±0.6	37.9±0.3	51.5±2.9	36.5±1.3	71.7±0.6	47.4±1.1
	✓	46.5±2.3	45.2±1.3	55.0±1.3	46.9±0.5	70.7±1.3	52.9±1.3
ALDA [29]	✗	36.9±0.2	33.1±1.9	47.2±3.9	35.0±1.0	55.3±3.1	41.5±2.0
	✓	42.8±2.1	36.2±0.6	51.5±2.2	40.7±1.3	53.8±3.9	45.0±2.0
JUMBOT [38]	✗	36.8±0.2	33.6±1.3	50.5±0.6	35.6±3.0	63.7±3.0	44.0±1.6
	✓	42.7±0.1	38.3±1.2	49.7±4.2	41.5±0.5	62.2±1.2	46.9±1.4
TimeMatch	✗	57.4±1.5	47.0±0.9	61.7±4.9	52.1±1.4	73.0±0.5	58.2±1.8
Target-trained	✗	74.6±0.6	72.4±1.4	86.9±2.7	90.6±4.3	85.7±0.7	82.0±1.9

Table 1: Macro F1-score (%) results on our dataset for unsupervised cross-region adaptation. We consider five adaptation tasks across four Sentinel-2 tiles: DK1=32VNH (Denmark), FR1=30TXT (mid-west France), FR2=31TCJ (southern France), and AT1=33UVP (Austria).

and the student as the final model. All methods are initialized from a source-trained model. ShiftAug versions are initialized from the corresponding ShiftAug source-trained model, and we continue to use ShiftAug during training. As in TimeMatch, we train for 20 epochs and tune the hyper-parameters of these methods on the task DK1→FR1. The full details can be found in our source code.

5. Experimental Results

5.1. Main Results

Table 1 shows the performance obtained with our approach and the re-implemented baselines and competitors. We report the mean and standard deviation of macro F1 scores, calculated from the results of three runs with different dataset splits.

We observe that source-trained models transfer very poorly to new target regions, with an average F1-score of 39% on target data. In comparison, target-trained models on the same classes achieve 82% on average. We observe that training shift-invariant models with ShiftAug improves domain generalization, leading to an increased average score of 49%. This greatly motivates addressing the temporal shift in UDA.

Existing UDA methods, however, only slightly increase the performance of source-trained models, with the best result obtained by CDAN+E [35] with 47%. By incorporating our ShiftAug, we observe a performance boost across all evaluated methods, indicating that existing methods are unable to implicitly handle the temporal shift.

Our approach TimeMatch, where we explicitly estimate the temporal shift, outperforms all competing methods by 11% on average and 5% for their ShiftAug variants. This shows that accounting for the temporal shift is a key component for the cross-region adaptation problem of SITS. Moreover, the shift-variant approach of TimeMatch outperforms the shift-invariance strategy. We hypothesize that training for shift-invariance may complicate crop classification, as the classification of certain crop types is shift-variant. For example, spring barley and winter barley develop similarly over time but shifted, as also discussed in Section 1.

Comparing TimeMatch to the results of the target-trained model, we observe that our approach—without any target labels—recovers a significant part of the highest achievable performance if target labels were available, but we also find that there is room for improvement. From our results, we see that methods which explicitly account for the temporal shift, such as TimeMatch and the ShiftAug variants of competing methods, generally outperform methods which do not. We therefore believe that further improvements can be gained by considering stronger forms of temporal alignment than shifts, such as class-wise alignments or time warping. We leave this interesting direction to future work.

Lastly, we highlight the results of the semi-supervised learning method FixMatch [45]. This method is similar to TimeMatch, but without temporal shift estimation. We observe that without ShiftAug, FixMatch obtains results worse than the source-trained model. This indicates semi-

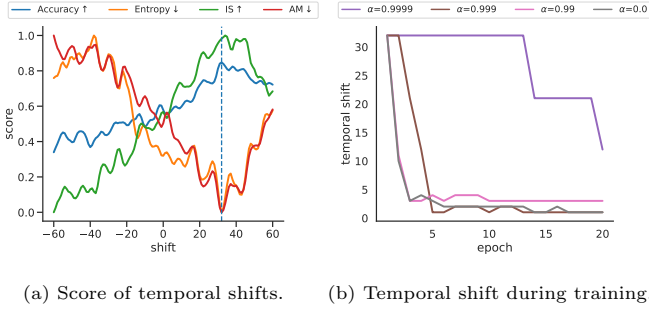


Figure 7: (a) Overall accuracy, entropy, IS, and AM scores of a source-trained model when applied to the target domain with different shifts. The dashed line indicate the most accurate shift. (b) The re-estimated temporal shifts of the teacher model during TimeMatch learning with different EMA decay rates.

supervised learning cannot address the cross-region task alone. With ShiftAug, however, the results are greatly improved on average. Interestingly, the performance is worse without ShiftAug for all tasks *except* FR1→FR2. Here, the source and target regions are the most geographically close, and as result, the temporal shift is also closer to zero (see the top-left table in Figure 4). This indicates that ShiftAug (controlled by Δ) is a trade-off between better long-range classification results and worse short-range results. In contrast, by estimating the temporal shift directly, TimeMatch does not have this issue and outperforms shift-invariance at both short and long distances.

5.2. Analysis of Temporal Shift Estimation

In Figure 7a, we show the change in the overall accuracy of a source-trained model when applied to target data with different temporal shifts for DK1→FR1. We also show the change in entropy, IS, and AM scores of the model. We observe a significant increase in accuracy by temporally shifting the target data. Calculating the statistics of entropy, IS, and AM from the predictions of the model works well as an unlabeled proxy to accuracy. We aim to estimate the shift with the highest accuracy (dashed blue line) for the highest quality pseudo-labels. For the shown example, the minimum of both entropy and AM correspond to the best shift. However, we find AM to be the most consistent across different adaptation tasks.

In Figure 7b, we show the rate at which the estimated temporal shift for the teacher goes to zero in TimeMatch learning when training with different EMA decay rates. When the shift changes, the previous estimate becomes sub-optimal for generating accurate pseudo-labels. We address this by re-estimating the temporal shift during training. We observe that low decay rates (*e.g.* 0.99) require the shift to be re-estimated after a few iterations, which is inefficient. In comparison, a decay rate of 0.9999 allows us to only re-estimate the shift only once every epoch.

The table in the upper left corner of Figure 4 shows the initial temporal shifts estimated by our method. We find the estimated shifts are connected to the climatic differences

Ablation	DK1→FR1
No EMA ($\alpha = 0.0$)	49.9±3.7
No source temporal shift ($\delta^{s \rightarrow t} = 0$)	51.9±1.9
No balanced batch sampler for source	53.3±3.6
IS instead of AM	56.3±2.6
Entropy instead of AM	56.9±1.8
No domain-specific batch norm.	56.9±4.1
TimeMatch	57.4±1.5

Table 2: Ablation study of TimeMatch components, sorted by increasing F1-score (%).

between regions. For example, the temporal shift ($\delta^{t \rightarrow s}$) from the warmer FR1 (mid-west France) to the colder DK1 (Denmark) is estimated as 32 days. Due to the warmer climate, crops in FR1 mature earlier than in DK1, and a positive shift is required to align the former with the latter. In the other direction, the opposite is true, and indeed, we estimate a negative temporal shift of −35 days. Note that these are off by 3 days due to estimation variance. Here, the two source-trained models used to estimate the temporal shift in each direction are trained with two completely separate source region, yet their estimated shifts are still roughly inverses. This indicates that the temporal shift learned by these models is connected to the phenological properties of their respective source regions.

5.3. Ablation Study

To better understand how TimeMatch is able to obtain state-of-the-art results, we perform an ablation study on the different components for the task DK1→FR1. We report the results in Table 2. We first study the impact of the EMA training. Instead of EMA, we set the teacher as a direct copy of the student (No EMA). We observe that training without EMA introduces a significant drop in F1-score. This shows that EMA is important to ensure high pseudo-label accuracy. Setting $\delta^{s \rightarrow t} = 0$ disables the temporal shift of the source domain, and the student is trained with datasets with different temporal shifts. We observe a significant decrease in F1-score as a result. Disabling the balanced mini-batch sampler for the source domain also leads to a degradation of the performance. If the model is trained with class imbalanced source data, the teacher will make biased pseudo-labels for the samples from the target domain [64]. This hinders the TimeMatch learning process, as pseudo-labels for infrequent classes in the source domain are less likely to be generated for the target. By applying a balanced mini-batch sampler for the source, we address this problem by ensuring each source batch contains roughly the same number of samples for each category. Estimating the temporal shift with IS or entropy instead of AM results in a slight performance drop. Domain-specific batch normalization is simple to implement, as it just requires forwarding source and target batches separately instead of concatenated. Disabling this

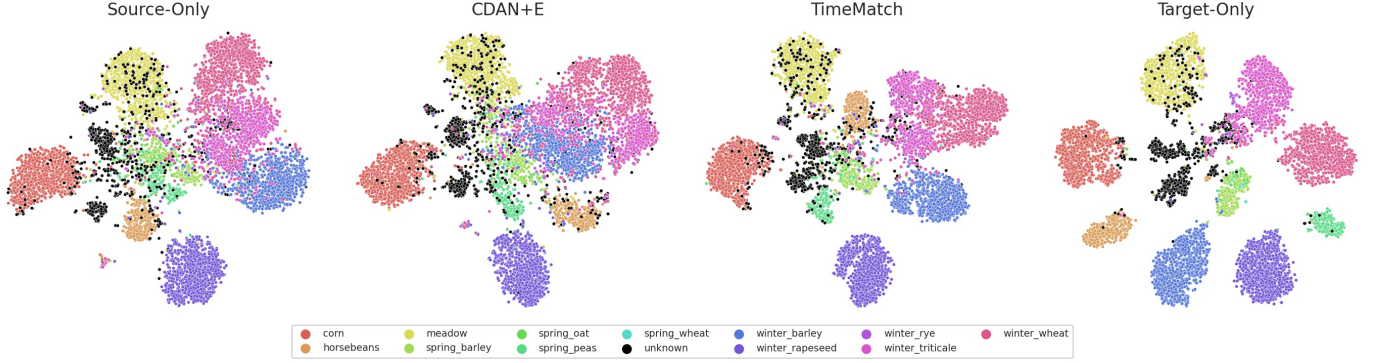


Figure 8: Visualization with t-SNE [63] of target features for the DK1→FR1 task. TimeMatch shows improved clustering of target features compared to existing approaches.

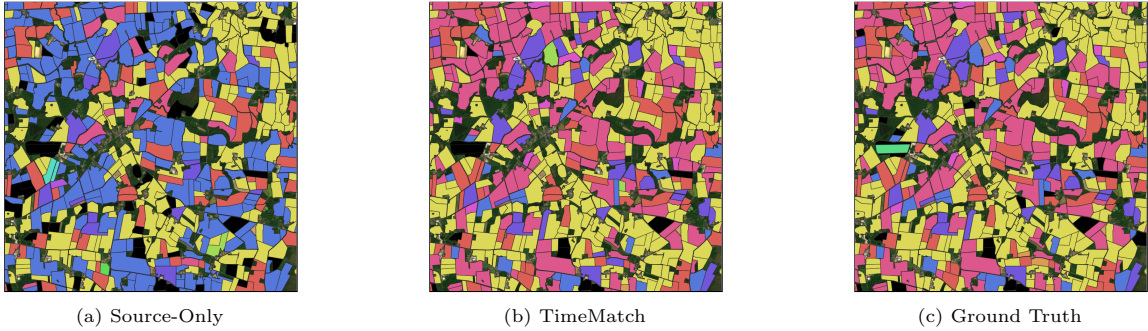


Figure 9: Parcel predictions for an example target area (6 km²) from the DK1→FR1 task, comparing (a) Source-Only, (b) TimeMatch, and (c) the corresponding ground truth. The figure shows the combination of multiple individual parcel predictions in the target region. The colors map to the classes in Figure 8.

component results in a small average performance loss with notably higher variance.

5.4. Sensitivity Analysis

Here we study the sensitivity of the TimeMatch hyperparameters. The results are shown in Figure 10. Higher values of α lead to better results, with a decay rate of 0.9999 being the best. However, increasing it to 1.0, so the teacher is not updated, results in a drop in F1-score, as the teacher cannot benefit from the knowledge learned by the student. The confidence threshold ϵ controls the trade-off between the quality and quantity of pseudo-labels. A threshold of 0.9 gives the best F1-score and further increasing the threshold to 0.95 drops performance as a result of too few pseudo-labels, which particularly decreases performance for the less frequent classes. Finally, the trade-off parameter λ controls the importance of the source domain loss $\mathcal{L}^{s \rightarrow t}$ with respect to the target domain loss \mathcal{L}^t . We observe that this hyperparameter is less important than the other two, but setting $\lambda = 2.0$ gives the best results.

5.5. Visual Analysis

Finally, we visualize the ability of TimeMatch in learning discriminative features for the target domain. In Figure 8, we visualize t-SNE [63] embeddings of target domain features from source-trained, CDAN+E (the best competitor

on average), TimeMatch, and target-trained models on the task DK1→FR1. The colors of the points represent their class (black is the unknown class). With TimeMatch, the

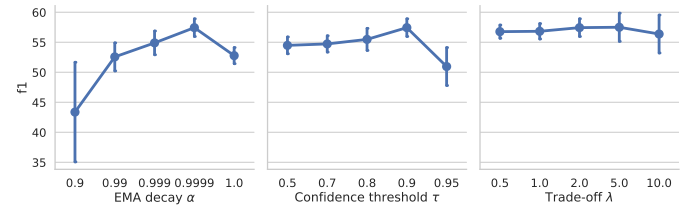


Figure 10: Sensitivity analysis of TimeMatch for the EMA decay rate, pseudo-label confidence threshold, and the trade-off in Eq. 13. The error bars show standard deviation.

target features are better clustered into their respective classes compared CDAN+E, which does not result in much better feature separation than the source-trained model. The target-trained plot shows the best possible learned features when training with all available target labels. Even with labels, the classes are not perfectly separated, *e.g.* for unknown/meadow or winter triticale/winter wheat.

Figure 9 shows example parcel predictions in a small area for the source-trained and TimeMatch models compared to the ground truth. The colors represent the same classes as before. We observe a large class confusion for the source-trained model, in particular between winter barley

(blue) and winter wheat (dark pink), which are also not separated well in Figure 8. Without using any target labels, TimeMatch resolves this issue, resulting in clusters that better resemble the ground truth.

6. Conclusion

This paper presented TimeMatch, a novel cross-region adaptation method for SITS. Unlike previous methods that solely match the feature distributions across domains, TimeMatch explicitly captures the underlying temporal discrepancy of the data by estimating the temporal shift between two regions. Through TimeMatch learning, we adapt a crop classifier trained in a labeled source region to an unlabeled target region. This is achieved by a learning algorithm that combines temporal shift estimation with self-training, where target pseudo-labels are generated using the estimated temporal shift from target to source. Lastly, we presented the TimeMatch dataset, a new large-scale cross-region UDA dataset with SITS from four different regions in Europe. Evaluated on this dataset, TimeMatch outperforms all existing approaches by 11% F1-score on average across five different adaptation tasks, setting a new state of the art in unsupervised cross-region adaptation. While this demonstrates that TimeMatch reaches strong results, there is still a gap with the performance obtained by fully supervised approaches. To overcome this limitation, we hypothesize that stronger temporal alignments, *e.g.* class-wise alignments or time warping, could further improve the performance. Another possibility is to perform domain adaptation across both time and space, which in addition to the temporal aspect also brings new considerations, such as the change in parcel shapes over time and crop rotations. We hope our proposed method and released dataset will encourage the remote sensing community to consider the challenging cross-region adaptation problem and its temporal aspect.

7. Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments and constructive feedback. The work of Joachim Nyborg was funded by the *Innovation Fund Denmark* under reference *8053-00240*.

References

- [1] M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, et al., Sentinel-2: ESA’s optical high-resolution mission for GMES operational services, *Remote Sensing of Environment* 120 (2012) 25–36.
- [2] B. C. Reed, J. F. Brown, D. VanderZee, T. R. Loveland, J. W. Merchant, D. O. Ohlen, Measuring phenological variability from satellite imagery, *Journal of Vegetation Science* 5 (5) (1994) 703–714.
- [3] F. Vuolo, M. Neuwirth, M. Immitzer, C. Atzberger, W.-T. Ng, How much does multi-temporal Sentinel-2 data improve crop type classification?, *International Journal of Applied Earth Observation and Geoinformation* 72 (2018) 122–130.
- [4] J. B. Odenweller, K. I. Johnson, Crop identification using landsat temporal-spectral profiles, *Remote Sensing of Environment* 14 (1) (1984) 39–54. doi:10.1016/0034-4257(84)90006-3.
- [5] C. Pelletier, G. I. Webb, F. Petitjean, Temporal convolutional neural network for the classification of satellite image time series, *Remote Sensing* 11 (5) (2019) 523. doi:10.3390/rs11050523.
- [6] L. Zhong, L. Hu, H. Zhou, Deep learning based multi-temporal crop classification, *Remote Sensing of Environment* 221 (2019) 430–443.
- [7] E. Ndikumana, D. Ho Tong Minh, N. Baghdadi, D. Courault, L. Hossard, Deep recurrent neural network for agricultural classification using multitemporal SAR Sentinel-1 for Camargue, France, *Remote Sensing* 10 (8) (2018) 1217.
- [8] M. Rußwurm, M. Körner, Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 11–19.
- [9] D. Ienco, R. Gaetano, C. Dupaquier, P. Maurel, Land cover classification via multitemporal spatial data by deep recurrent neural networks, *IEEE Geoscience and Remote Sensing Letters* 14 (10) (2017) 1685–1689.
- [10] D. H. T. Minh, D. Ienco, R. Gaetano, N. Lalande, E. Ndikumana, F. Osman, P. Maurel, Deep recurrent neural networks for winter vegetation quality mapping via multitemporal SAR Sentinel-1, *IEEE Geoscience and Remote Sensing Letters* 15 (3) (2018) 464–468.
- [11] M. Rußwurm, M. Körner, Self-attention for raw optical satellite time series classification, *ISPRS Journal of Photogrammetry and Remote Sensing* 169 (2020) 421–435. doi:10.1016/j.isprsjprs.2020.06.006.
- [12] V. Sainte Fare Garnot, L. Landrieu, S. Giordano, N. Chehata, Satellite image time series classification with pixel-set encoders and temporal self-attention, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12325–12334.
- [13] M. Rußwurm, M. Körner, Multi-temporal land cover classification with sequential recurrent encoders, *ISPRS International Journal of Geo-Information* 7 (4) (2018) 129. doi:10.3390/ijgi7040129.
- [14] R. Interdonato, D. Ienco, R. Gaetano, K. Ose, DuPLO: A Dual view Point deep Learning architecture for time series classification, *ISPRS Journal of Photogrammetry and Remote Sensing* 149 (2019) 91–104. doi:https://doi.org/10.1016/j.isprsjprs.2019.01.011.
- [15] D. Tuia, C. Persello, L. Bruzzone, Domain adaptation for the classification of remote sensing data: An overview of recent advances, *IEEE Geoscience and Remote Sensing Magazine* 4 (2) (2016) 41–57.
- [16] B. Lucas, C. Pelletier, D. Schmidt, G. I. Webb, F. Petitjean, A bayesian-inspired, deep learning-based, semi-supervised domain adaptation technique for land cover mapping, *Machine Learning* (2021) 1–33.
- [17] L. Kondmann, A. Toker, M. Rußwurm, A. Camero, D. Peressuti, G. Milcinski, P.-P. Mathieu, N. Longépé, T. Davis, G. Marchisio, L. Leal-Taixé, X. X. Zhu, DENETHOR: The DynamicEarthNET dataset for harmonized, inter-operable, analysis-ready, daily crop monitoring from space, in: *Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.
- [18] S. J. Pan, Q. Yang, A survey on transfer learning, *IEEE Transactions on Knowledge and Data Engineering* 22 (10) (2009) 1345–1359. doi:10.1109/TKDE.2009.191.
- [19] B. Kellenberger, O. Tasar, B. Bhushan Damodaran, N. Courty, D. Tuia, *Deep Domain Adaptation in Earth Observation*, John Wiley & Sons, Ltd, 2021, Ch. 7, pp. 90–104.
- [20] Z. Wang, H. Zhang, W. He, L. Zhang, Phenology alignment network: A novel framework for cross-regional time series crop

- classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2940–2949.
- [21] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, J. W. Vaughan, A theory of learning from different domains, *Machine Learning* 79 (1) (2010) 151–175. doi:10.1007/s10994-009-5152-4.
 - [22] Y. Ganin, V. Lempitsky, Unsupervised domain adaptation by backpropagation, in: International Conference on Machine Learning, PMLR, 2015, pp. 1180–1189.
 - [23] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, T. Darrell, Deep domain confusion: Maximizing for domain invariance, *CoRR abs/1412.3474* (2014).
 - [24] G. Wilson, D. J. Cook, A survey of unsupervised deep domain adaptation, *ACM Transactions on Intelligent Systems and Technology* 11 (5) (2020) 1–46. doi:10.1145/3400066.
 - [25] B. Lucas, C. Pelletier, D. Schmidt, G. I. Webb, F. Petitjean, Unsupervised domain adaptation techniques for classification of satellite image time series, in: International Geoscience and Remote Sensing Symposium (IGARSS), IEEE, 2020, pp. 1074–1077. doi:10.1109/IGARSS39084.2020.9324339.
 - [26] K. Saito, Y. Ushiku, T. Harada, Asymmetric tri-training for unsupervised domain adaptation, in: International Conference on Machine Learning, PMLR, 2017, pp. 2988–2997.
 - [27] R. Shu, H. H. Bui, H. Narui, S. Ermon, A dirt-t approach to unsupervised domain adaptation, in: International Conference on Learning Representations, 2018.
 - [28] P. Morerio, R. Volpi, R. Ragonesi, V. Murino, Generative pseudo-label refinement for unsupervised domain adaptation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 3130–3139.
 - [29] M. Chen, S. Zhao, H. Liu, D. Cai, Adversarial-learned loss for domain adaptation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 3521–3528.
 - [30] Y. Zou, Z. Yu, X. Liu, B. Kumar, J. Wang, Confidence regularized self-training, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5982–5991.
 - [31] D.-H. Lee, et al., Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks, in: Workshop on Challenges in Representation Learning, ICML, Vol. 3, 2013, p. 896.
 - [32] D. Hendrycks, K. Gimpel, A baseline for detecting misclassified and out-of-distribution examples in neural networks, in: International Conference on Learning Representations, 2017.
 - [33] J. Nyborg, C. Pelletier, S. Lefèvre, I. Assent, The TimeMatch Dataset (2021). doi:10.5281/zenodo.5636422.
 - [34] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, Domain-adversarial training of neural networks, *The Journal of Machine Learning Research* 17 (1) (2016) 2096–2030.
 - [35] M. Long, Z. Cao, J. Wang, M. I. Jordan, Conditional adversarial domain adaptation, in: Advances in Neural Information Processing Systems, 2018, pp. 1647–1657.
 - [36] B. Sun, K. Saenko, Deep coral: Correlation alignment for deep domain adaptation, in: European Conference on Computer Vision, Springer, 2016, pp. 443–450.
 - [37] B. B. Damodaran, B. Kellenberger, R. Flamary, D. Tuia, N. Courty, DeepJDOT: Deep joint distribution optimal transport for unsupervised domain adaptation, in: European Conference on Computer Vision, 2018, pp. 447–463.
 - [38] K. Fatras, T. Séjourné, N. Courty, R. Flamary, Unbalanced minibatch optimal transport; applications to domain adaptation, in: International Conference on Machine Learning, 2021.
 - [39] S. Purushotham, W. Carvalho, T. Nilanon, Y. Liu, Variational recurrent adversarial deep domain adaptation, in: International Conference on Learning Representations, 2017.
 - [40] G. Wilson, J. R. Doppa, D. J. Cook, Multi-source deep domain adaptation with weak supervision for time-series sensor data, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 1768–1778.
 - [41] A. Bailly, L. Chapel, R. Tavenard, G. Camps-Valls, Nonlinear time-series adaptation for land cover classification, *IEEE Geoscience and Remote Sensing Letters* 14 (6) (2017) 896–900.
 - [42] B. Gong, Y. Shi, F. Sha, K. Grauman, Geodesic flow kernel for unsupervised domain adaptation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 2066–2073.
 - [43] B. Fernando, A. Habrard, M. Sebban, T. Tuytelaars, Unsupervised visual domain adaptation using subspace alignment, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2013, pp. 2960–2967.
 - [44] O. Chapelle, B. Scholkopf, A. Zien, Semi-supervised learning, *IEEE Transactions on Neural Networks* 20 (3) (2009) 542–542.
 - [45] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, C.-L. Li, FixMatch: Simplifying semi-supervised learning with consistency and confidence, *Advances in Neural Information Processing Systems* 33 (2020).
 - [46] A. Tarvainen, H. Valpola, Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, in: Advances in Neural Information Processing Systems, 2017, pp. 1195–1204.
 - [47] M. Chen, K. Q. Weinberger, J. Blitzer, Co-training for domain adaptation., in: Nips, Vol. 24, Citeseer, 2011, pp. 2456–2464.
 - [48] P. Panareda Busto, J. Gall, Open set domain adaptation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 754–763.
 - [49] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training GANs, *Advances in Neural Information Processing Systems* 29 (2016) 2234–2242.
 - [50] Z. Zhou, H. Cai, S. Rong, Y. Song, K. Ren, W. Zhang, J. Wang, Y. Yu, Activation maximization generative adversarial nets, in: International Conference on Learning Representations, 2018.
 - [51] S. Barratt, R. Sharma, A note on the inception score, in: Workshop on Theoretical Foundations and Applications of Deep Generative Models, ICML, 2018.
 - [52] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2017, pp. 2980–2988.
 - [53] V. Sainte Fare Garnot, L. Landrieu, Lightweight temporal self-attention for classifying satellite images time series, in: International Workshop on Advanced Analytics and Learning on Temporal Data, Springer, 2020, pp. 171–181.
 - [54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.
 - [55] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., PyTorch: An imperative style, high-performance deep learning library, *Advances in Neural Information Processing Systems* 32 (2019) 8026–8037.
 - [56] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: International Conference on Learning Representations, 2015.
 - [57] I. Loshchilov, F. Hutter, SGDR: Stochastic gradient descent with warm restarts, in: International Conference on Learning Representations, 2017.
 - [58] Y. Li, N. Wang, J. Shi, J. Liu, X. Hou, Revisiting batch normalization for practical domain adaptation, *Pattern Recognition* 80 (03 2016). doi:10.1016/j.patcog.2018.03.005.
 - [59] W.-G. Chang, T. You, S. Seo, S. Kwak, B. Han, Domain-specific batch normalization for unsupervised domain adaptation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7354–7362.
 - [60] K. Saito, D. Kim, S. Sclaroff, T. Darrell, K. Saenko, Semi-supervised domain adaptation via minimax entropy, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8050–8058.
 - [61] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: International Conference on Machine Learning, PMLR, 2015, pp. 448–456.

- [62] J. Jiang, B. Chen, B. Fu, M. Long, Transfer learning library, <https://github.com/thuml/Transfer-Learning-Library> (2020).
- [63] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE., *Journal of Machine Learning Research* 9 (11) (2008).
- [64] H. He, E. A. Garcia, Learning from imbalanced data, *IEEE Transactions on Knowledge and Data Engineering* 21 (9) (2009) 1263–1284.