# Reuse Out-of-Year Data to Enhance Land Cover Mapping via Feature Disentanglement and Contrastive Learning

Cássio F. Dantas, Raffaelle Gaetano, Claudia Paris, *Senior Member, IEEE*, and Dino Ienco, *Member, IEEE*

***Abstract*—Given the systematic acquisition of satellite data, it is possible to generate up-to-date land cover (LC) maps, essential for effective agricultural territory management, environmental monitoring, and informed decision-making. Typically, creating a LC map requires collecting high-quality labeled data, a process that is both costly and time-consuming. To mitigate the need to collect large volume of labeled data, we propose a deep learning framework called *REFeD* (data Reuse with Effective Feature Disentanglement for land cover mapping), which leverages already available out-of-year reference data to enhance the production of up-to-date LC maps. To this end, *REFeD* integrates remote sensing and reference data from different domains (e.g., historical and recent data) utilizing a disentanglement strategy based on contrastive learning. By separating domain-invariant and domain-specific features, *REFeD* isolates useful information associated to the downstream LC mapping task and mitigates distribution shifts between domains. Moreover, *REFeD* incorporates an effective supervision scheme to reinforce feature disentanglement through multiple levels of supervision at different granularities. Experimental evaluation on study areas characterized by diverse landscapes, including Koumbia (West Africa, Burkina Faso) and Centre-Val de Loire (central Europe, France), demonstrates the effectiveness of the proposed approach.**

***Index Terms*—Contrastive learning, data-centric artificial intelligence (data-centric AI), domain adaptation, land cover (LC) mapping, satellite image time series (SITS).**

## I. Introduction

THE unprecedented availability of Earth observation (EO) data regularly acquired through modern public and private EO Programmes and Missions (e.g., ESA Copernicus, NASA Landsat, and PlanetScope to cite a few) opens the opportunity to collect satellite image time series (SITS) over the same study area to characterize and study the underlying spatio-temporal dynamics and generate accurate land cover (LC) maps [1]. These maps have been demonstrated to be largely beneficial in a variety of different fields, such as ecology [2], agriculture [3], forestry [4], environmental monitoring [5], and facilitating well-informed and sustainable decision-making policies [6]. Typically, for the creation of LC maps over a region at a certain period of time, reference data are collected through expensive and time-demanding field campaigns or tedious manual annotation activity. These data are then utilized in conjunction with SITS information through advanced machine learning algorithms [7] to get the final LC map. While the access to high resolution EO data is no longer a major constraint, collecting up-to-date labeled reference data constitutes a consumable (neither enduring nor lasting) effort. Once served its purpose, reference data will be disregarded losing any further relevance. Furthermore, when the process is repeated (e.g., estimate agricultural production or potential biodiversity loss for a new year for the same or a related study site), new field campaigns or image photointerpretation activities must be afforded again with, in general, no way to profit from previous efforts.

To leverage existing reference data, it is common to apply the classification model trained on EO images with available reference data to new unlabeled EO acquisitions. However, when EO data from different acquisitions are combined under the same learning framework, challenges related to distribution shifts can impede the effective training of machine learning models [8]. To cope with this issue, the most widely studied setting is domain adaptation (DA) [9], where the main goal is to learn a model over a labeled source domain and transfer it to an unlabeled target one [10]. DA methods can be classified into three main categories based on the availability of labeled data: unsupervised, semisupervised and supervised [11]. Unsupervised domain adaptation (UDA) methods address scenarios where no target labels are available. This is achieved by minimizing the distribution gap between the source (where labeled data are available) and target domains (which lack labeled data) [12], [13].

In the field of remote sensing, significant effort has been devoted in developing UDA strategies specifically designed for EO data classification [14]. Recently, many deep-learning-based UDA methods have gained popularity [15]. While traditional UDA methods focus on aligning distributions at the instance [16], feature [17], or classifier levels [18],

Received 24 July 2024; revised 4 October 2024 and 31 October 2024; accepted 11 November 2024. Date of publication 21 November 2024; date of current version 13 December 2024. The work of Dino Ienco was supported by the National Research Agency (ANR) under the ANR-23-IAS1-0002 (GEO-ReSeT) project. This work was supported by the French Space Study Center (CNES, TOSCA 2024 Project) and the French National Research Agency under the Investments for the Future Program, referred as ANR-16-CONV-0004 (DigitAg). *(Corresponding author: Cássio F. Dantas.)*

Cássio F. Dantas and Dino Ienco are with the INRAE, Inria, UMR TETIS, University of Montpellier, 34090 Montpellier, France (e-mail: cassio.fraga-dantas@inrae.fr).

Raffaelle Gaetano is with the CIRAD, Inria, UMR TETIS, University of Montpellier, 34090 Montpellier, France.

Claudia Paris is with the Department of Natural Resources, ITC, University of Twente, 7522 NB Enschede, The Netherlands.

Digital Object Identifier 10.1109/JSTARS.2024.3503756

deep-learning-based UDA typically relies on the adversarial learning framework similar to the one used by generative adversarial networks (GANs) [13]. In this setting, the feature generator aims to create domain-invariant features from the input data, and the domain discriminator is designed to recognize whether the features come from the source or target domain [19], [20], [21]. Despite the efforts invested in designing and implementing UDA techniques, the success of UDA depends largely on the discrepancy between the source and target distributions, making these methods susceptible to potential pitfalls and limited generalizability. Moreover, only recently, strategies have begun to emerge to analyze SITS data for spatial [22], [23] and temporal transfer tasks [24]. Finally, these approaches typically assume the complete absence of reference data for the target domain. Although relevant in different operational settings, for the considered LC mapping task, it is often reasonable to assume access to a certain amount of labeled reference data because of the need to systematically generate up-to-date LC products.

SDA assumes that labeled samples are present in both the source and target domains. Typically a manifold alignment is performed to create a unified representation across domains by finding a projection to the common latent space, where the class separability is enhanced [25]. Several UDA strategies can be extended to the supervised setting, by increasing the reliability of the adaptation results obtained [17], [26]. However, SDA methods are particularly relevant when dealing with cross-sensor adaptation [27], [28] to ensure the possibility of effectively mitigating the severe domain shift. Moreover, the requirement for sufficient labeled samples in both domains can be a significant limitation [29]. For applications such as land-cover mapping, if a reasonable amount of target reference data are available, a supervised classifier can be effectively trained from scratch without relying on domain adaptation strategies [30]. To address this issue, Persello and Bruzzone [31] proposed a strategy to reduce the number of labeled samples needed from the target domain for effective supervised DA. Although promising, it requires active engagement with a supervisor responsible for accurately labeling the requested target samples.

In scenarios where only a limited amount of labeled data is available for the target domain, the paradigm shifts to semi-supervised domain adaptation (SSDA). Current SSDA approaches generally aim to align the target data with the labeled source data with feature space mapping and self-training assignments using pseudolabels [32]. In remote sensing, these techniques have been widely employed to expand training sets by leveraging the target domain's unlabeled data [11], [33], [34]. However, most SSDA approaches have been applied to scenarios where the labeled and unlabeled data belong to the same EO data, while little has been done to use them for different temporal EO acquisitions. Moreover, despite the SSDA setting holds potential for various real-world problems, it remains largely unexplored when dealing with SITS data [35] in the context of systematically producing up-to-date LC maps. Despite the long history of DA methods, recent efforts toward the systematic and effective exploitation of available high-quality labeled data have gained momentum under the framework of data-centric artificial intelligence (AI) [36]. Under this movement, the attention of researchers and practitioners is gradually shifting from advancing model design (model-centric AI) to enhancing the quality and quantity of the data (data-centric AI).

Considering geospatial and EO data, the data-centric AI perspective is even more important since it can steer the community toward developing methodologies to provide further improvements related to the generalization ability with impact on real-world relevant problems and applications [37]. Nevertheless, the two perspectives (model-centric and data-centric AI) play a complementary role in the larger machine learning deployment cycle since standard approaches still struggle to manage and exploit valuable data coming from different and heterogeneous distributions like, for instance, in the case of combining historical and up-to-date reference data for the downstream task of LC mapping [38], where distribution shifts can be related to the different environmental and/or climatic factors that determine the EO data acquisition conditions.

In addressing the significant challenge outlined above, more precisely take advantages of exploiting together both historical and recent EO data along with reference data to enhance LC mapping, we present a novel approach, namely *REFeD*, rooted on recent advances in the field of domain adaptation/generalization. *REFeD* adopts a model-centric AI perspective, aiming to fulfill a data-centric AI objective related to the effective exploitation of EO and reference data coming from two different domains (e.g., historical data and recent ones) with the aim to give value again to historical and/or overlooked reference data, and enhance the accuracy of the recent LC mapping result toward the systematic production of up-to-date LC products.

More precisely, *REFeD* built upon a pseudosiamese network with unshared parameters that, given a sample, extracts simultaneously domain-invariant and domain-specific features, using the former to make the final decision. The objective is to disentangle useful information for the downstream LC mapping task while isolating and discarding domain specific features that can hinder the learning process in the presence of data belonging to multiple domains with unaligned data distributions. The disentanglement process is achieved by shaping a representation manifold, via contrastive learning, that jointly structures both domain-invariant and domain-specific features. In addition, *REFeD* integrates an effective supervision strategy [39] that further enforces the disentanglement process via multiple levels of supervision at different granularities.

Experimental evaluations are carried out to assess the behavior of *REFeD* considering both baseline and domain adaptation/generalization approaches. To assess the behavior of our method, we perform both quantitative and qualitative evaluations considering two study sites covering extremely diverse and contrasted landscapes, namely *Koumbia* (located in the West-Africa region, in Burkina Faso) and *Centre-Val de Loire* (located in centre Europe, France). For the former study site, we consider a LC mapping task where data covering exactly the same geographical area are available for two different years (2020 and 2021) while, for the latter study site, we consider a crop type mapping task where data covering two closely related areas are available for the agricultural seasons 2018 and 2021. The obtained results highlight the potential of utilizing

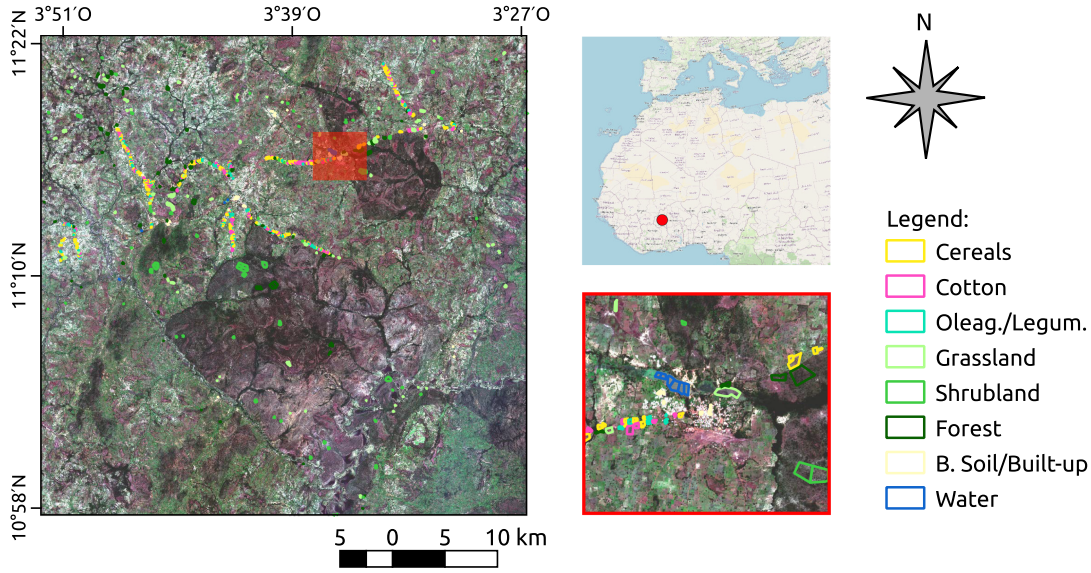


Fig. 1. View and location of Koumbia study site. The ground truth data coming from the 2020 year is superposed to a Sentinel-2 image covering the whole area. In the red box (bottom right), a more detailed view of the study site is depicted.

out-of-year information from the same or similar study sites to ameliorate the LC mapping process, underscoring the value of historical data in enhancing the mapping accuracy.

The rest of this article is organized as follows. Study sites and the associated information are described in Section II. Section III introduces the proposed framework based on feature disentanglement and contrastive learning to enhance LC mapping combining multiple reference data. The experimental evaluation and the related findings are reported and discussed in Section IV, while Section V draws the conclusions of this article.

## II. Dataset Description

To assess the effectiveness of the proposed method under diverse settings, we consider two different study areas, each with different LC nomenclature and different availability of historical data. We collected SITS of Sentinel-2 imagery via the Microsoft Planetary Computer platform[1] that allows to access level-2A Sentinel-2 products. We consider all bands at 10 and 20 m of spatial resolution for a total of 10 bands per image. We have conducted resampling of the SWIR 20 m bands to 10 m resolution, as well as image time series gap filling of cloudy pixels using multitemporal linear interpolation as explained in [40] and gap-filled images were generated at a regular 5-day frequency resulting in a sequence of 72 images for each study area and year.

### A. *Dataset 1:* Koumbia *Study Site*

The first study site covers an area around the town of *Koumbia*, in the Province of Tuy, *Hauts-Bassins* region, in the south-west of Burkina Faso. This area has a surface of about 2338 km$^2$, and is situated in the subhumid Sudanian zone. The surface is covered mainly by natural savannah (herbaceous and shrubby) and forests, interleaved with a large portion of land (around 35%) used for rainfed agricultural production (mostly smallholder farming). The main crops are cereals (maize, sorghum, and millet) and cotton, followed by oleaginous and leguminous. Several temporary watercourses constitute the hydrographic network around the city of Koumbia. Fig. 1 presents the study site with the 2020 reference data (ground truth) superposed on a Sentinel-2 image. A more detailed view corresponding to the red box in the overview is also depicted at the bottom right of the figure.

The ground truth data for 2020 and 2021 has been derived from a large agricultural LC dataset available online [41], mainly consisting of field data collected by local experts on several sites all over the tropics. For this study site, the field surveys were conducted around the growing peak of the cropping season. The ground truth data cover the exact same surface for the two reference years. Table I reports the statistics of the labeled reference data distribution for the years 2020 and 2021. This study site is characterized by eight LC classes, namely, "Cereals," "Cotton," "Oleaginous/Leguminous," "Grassland," "Shrubland," "Forest," "Bare soil/Built-up" and "Water."

Fig. 2 shows the NDVI profiles over 2020 and 2021, for the *Koumbia* study site, for some of the land cover classes: $Cereals$, $Oleaginous/Leguminous$ and $Forest$. We can observe that intrayear profiles are quite homogeneous, while interyear profiles, for the same land cover class, are shifted or delayed. This exploratory analysis on the interyear NDVI profiles provide insights on distribution shifts affecting remote sensing data coming from different time periods.

### B. *Dataset 2:* Centre Val De Loire *Study Site*

The second study site covers two spatially disjoint areas within the *Centre Val de Loire*, region located in the center of France. This region of France is characterized by intensive agricultural activity with agricultural surfaces representing around

[1]https://planetarycomputer.microsoft.com/

TABLE I
GROUND TRUTH STATISTICS FOR YEAR 2020 AND 2021 ON THE *KOUMBIA* STUDY SITE

| Class Name | Class ID. | 2020 | | 2021 | |
|---|---|---|---|---|---|
| | | # Polygons | # Pixels | # Polygons | # Pixels |
| Cereals | 1 | 230 | 9731 | 268 | 11435 |
| Cotton | 2 | 139 | 6971 | 121 | 6575 |
| Oleaginous / Leguminous | 3 | 281 | 7950 | 263 | 7316 |
| Grassland | 4 | 122 | 12998 | 113 | 11100 |
| Shrubland | 5 | 83 | 22546 | 90 | 24324 |
| Forest | 6 | 82 | 17435 | 82 | 16984 |
| Bare soil / Built-up | 7 | 51 | 1125 | 51 | 1022 |
| Water | 8 | 10 | 1205 | 10 | 1205 |
| | Total | 998 | 79961 | 998 | 79961 |

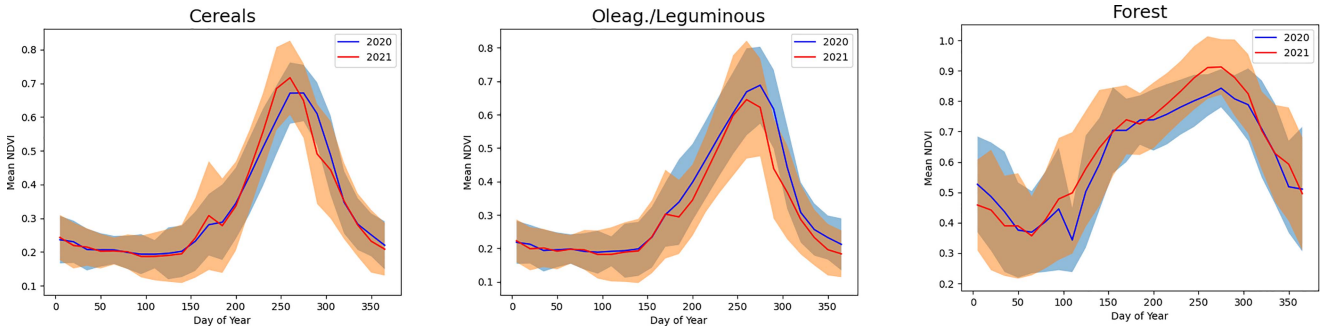


Fig. 2. Normalized Difference Vegetation Index (NDVI) profiles for some representative land cover classes: Cereal, Oleaginous/Leguminous and Forest over the years 2020 and 2021 on the *Koumbia* study site.

TABLE II
GROUND TRUTH STATISTICS FOR YEARS 2018 AND 2021 ON THE *CENTRE-VAL DE LOIRE* STUDY SITE

| Class Name | Class ID. | 2018 | | 2021 | |
|---|---|---|---|---|---|
| | | # Polygons | # Pixels | # Polygons | # Pixels |
| Soft wheat | 1 | 1048 | 9388 | 1341 | 9268 |
| Maize | 2 | 198 | 9442 | 264 | 9437 |
| Barley | 3 | 674 | 9345 | 527 | 9222 |
| Other cereals | 4 | 421 | 9288 | 360 | 9473 |
| Oleaginous / Proteaginous | 5 | 842 | 9222 | 775 | 9168 |
| Winter Fallows | 6 | 413 | 8390 | 611 | 7396 |
| Leguminous | 7 | 5 | 3418 | 12 | 4753 |
| Fodder | 8 | 255 | 9350 | 105 | 7422 |
| Meadow | 9 | 4132 | 8254 | 127 | 6030 |
| Other crops | 10 | 184 | 8183 | 355 | 9681 |
| | Total | 8172 | 84280 | 12649 | 81850 |

70% of the whole region with cereals and oleaginous as major crops. The two areas have a cumulative surface of about 840 $km^2$. Fig. 3 presents the two areas related to the *Centre-Val de Loire* study site depicting reference (ground truth) data for year 2018 and 2021 superposed on a Sentinel-2 image. On the right of the figure, a detail for each of the areas is proposed, in red for 2018 and in blue for 2021.

The ground truth data for the first area, related to 2018, was obtained through the EuroCrop dataset [42] while the ground truth data for the second area, related to 2021, are gathered from the *Registre Parcellaire Graphique* (RPG), the French land parcel identification system. The data covers only agricultural areas in order to set up a crop type mapping task. This second dataset covers a problem that implies both spatial and temporal transfer at the same time. Also note that a different temporal gap is considered in this dataset (three years) compared to the previous dataset (one-year gap). Table II reports the statistics of the labeled reference data distribution for the years 2018 and 2021. This study site is characterized by ten classes, namely, "Soft wheat," "Maize," "Barley," "Other cereals," "Oleaginous/Proteaginous,"

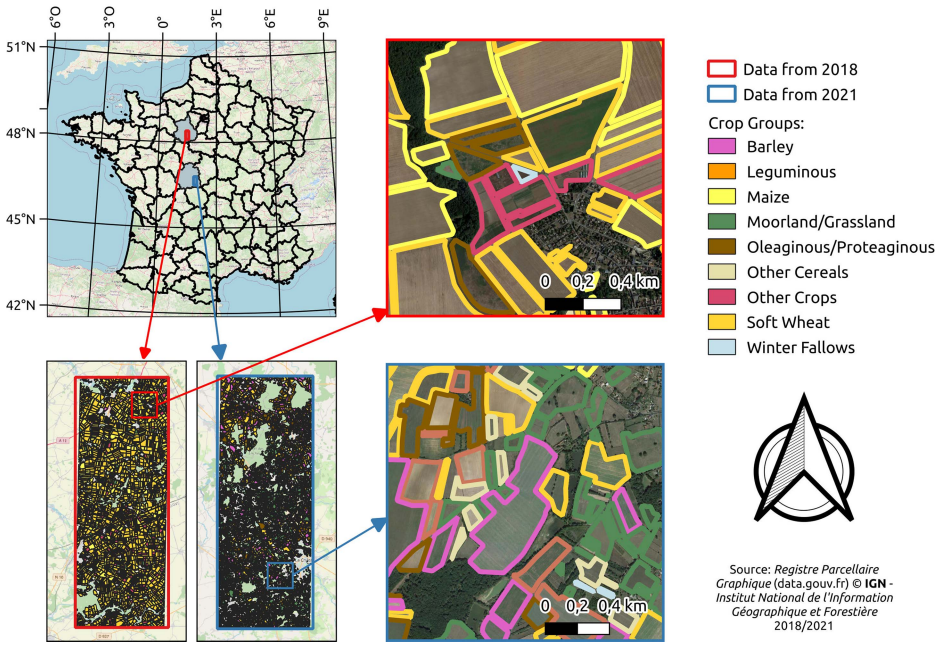


Fig. 3. View and location of the *Centre-Val de Loire* study site. Ground truth data coming from the 2018 and 2021 are superposed to a Sentinel-2 image. On the right, a detail for each of the areas is proposed, in red for 2018 and blue for 2021.

"Winter Fallows," "Leguminous," "Fodder," "Meadow," and "Other crops."

## III. Proposed Framework

### A. Problem Formulation and Notations

The proposed deep learning framework aims at improving the accuracy of LC mapping results obtained on recently acquired satellite data, i.e., target domain, by using pre-existing reference data coming from the same study site or a different yet correlated one, i.e., source domain. Typically, the source domain consists of some readily-available historical or out-of-year data, for instance. Differences in climate, weather, and other environmental conditions can lead to nonnegligible distribution shifts within SITS data from the different domains. These shifts may prevent the full exploitation of the source data as a naive direct enrichment of the target data in a standard supervised learning setup [43]. Moreover, in case the source data are more abundant, the learned classifier may likely be biased toward the source domain. Differently from the literature, the ultimate goal of *REFeD* is to maximize the classification performance in the target domain while taking full advantage of all the reference data available. In this work, we suppose we are given a set of $N_t$ labeled samples from a target domain $\mathcal{D}^t = \{(x_i^t, y_i^t)\}_{i=1}^{N_t}$, for which we want to train a classifier. Moreover, we dispose of additional labeled data (say $N_s$ samples) from a source domain $\mathcal{D}^s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ that we aim to exploit in order to improve the performance of our classifier on the target domain.

In our case, each sample $x_i \in \mathbb{R}^{T\times C}$ is the content of a pixel's $C$ spectral bands from a SITS defined over $T$ timestamps. The corresponding label $y_i \in \{1, \ldots, K\}$ is given by one of $K$ existing classes, shared between source and target domains—i.e., a closed-set scenario [44]. Depending on the considered classification task, the classes can be, for instance, different crops and/or LC types. Let us also define as $y_i' \in \{s, t\}$ the binary label associated with all the available labeled samples $\{(x_i, y_i')\}_{i=1}^{N_s+N_t}$, which specify from which domain each spectral samples $x_i$ belongs, i.e., $\mathcal{D}^t$ or $\mathcal{D}^s$.

### B. REFeD: Overview

Fig. 4 shows an overview of the proposed deep learning framework, by depicting the data needed during the training and inference stages. In the first stage, the supervised classifier is trained using the labeled data from both $\mathcal{D}^s$ and $\mathcal{D}^t$. To take full advantage of reference data coming from distinct domains, we propose to disentangle the information carried by the labeled input data into two parts: 1) domain-specific information, and 2) domain-invariant information (i.e., useful discriminative information for the subsequent classification task). The former is closely related to the domain to which the data belong, thus potentially hindering the learning model's ability to generalize. The second contains semantic information associated with the underlying classes, thus usable knowledge that can be exploited later for the classification process. Taking inspiration from current literature in domain adaptation/generalization fields [45], in the training stage we leverage two branches with separate encoders to generate the feature vectors, i.e., $g_{spe}$ and $g_{inv}$. Dedicated losses, $\mathcal{L}_{cl}$, $\mathcal{L}_{dom}$ and $\mathcal{L}_{con}$ (detailed in Section III-C) are employed to effectively disentangle domain-specific from domain-invariant information in each of the two obtained embeddings, by training a domain classifier $f'(\cdot)$ and task classifier $f(\cdot)$, respectively. This condition allows us to benefit from the

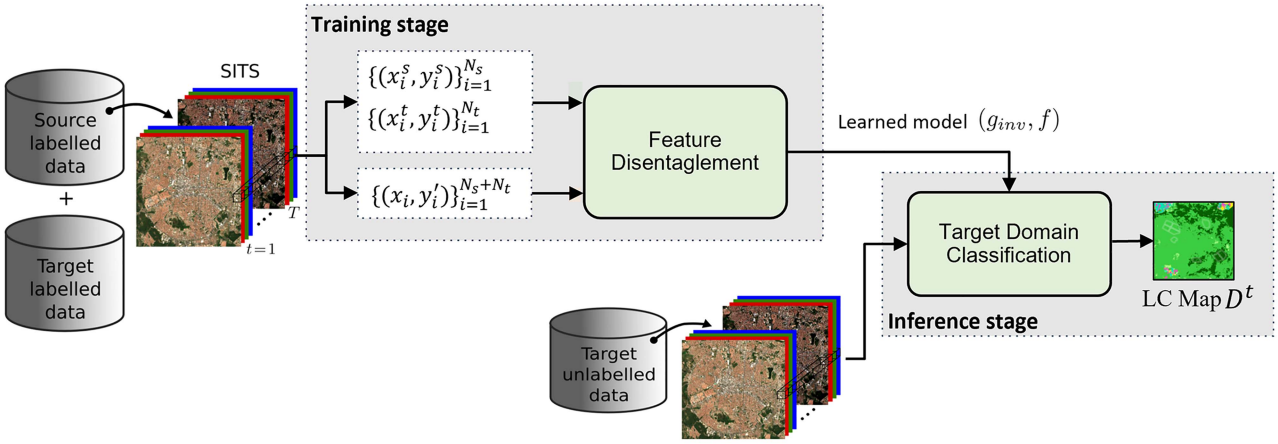


Fig. 4. Overview of the proposed framework. Training and inference stages are distinguished: while the former is performed on data coming from both domains, the latter is done exclusively on target data and uses only the domain-invariant branch $(g_{inf}, f)$ of the learned model.

labeled samples available in all domains, taking into account the domain to which each labeled sample belongs.

To maximize the classification results obtained in the target domain, in the inference stage, the domain-specific encoder $g_{\text{spe}}$ is discarded and only the domain-invariant encoder $g_{\text{inv}}$ is considered. In particular, we generate the LC map of the target domain using $g_{\text{inv}}$ for generating the feature representation to be classified along with the task classifier $f(\cdot)$ trained in the previous stage on the whole set of reference data. In the following, details are given.

### C. Feature Disentanglement

Fig. 5 depicts the dual-branch network, consisting of two encoders with identical architectures but unshared parameters, used for feature disentanglement. The two encoders, denoted $g_{\text{spe}}$ and $g_{\text{inv}} : \mathbb{R}^{T\times C} \rightarrow \mathbb{R}^D$ for domain-specific and domain-invariant, respectively, share the same architecture but are learned independently with unshared weights via different loss functions. In the domain-invariant branch, a task classifier $f(\cdot)$ is applied to the domain-invariant features extracted by $g_{\text{inv}}$. In the domain-specific branch, the domain-specific features obtained via $g_{\text{spe}}$ are fed to a domain classifier $f'(\cdot)$ which encourages the domain-discriminant information to be channeled to this branch.

*a) Domain Classifier:* The domain classifier aims to accurately predict domain labels $y_i' \in \{s, t\}$, i.e., determine if each sample $x_i$ belongs either to the source or the target domain, by minimizing a cross-entropy loss $\ell_{ce}$ as follows:

$$\mathcal{L}_{\text{dom}} = \frac{1}{N_s + N_t} \sum_{i=1}^{N_s+N_t} \ell_{ce}\left(f' \circ g_{\text{spe}}(x_i), y_i'\right). \quad (1)$$

*b) Task Classifier:* In its turn, the task classifier $f : \mathbb{R}^D \rightarrow \{1, \ldots, K\}$ maps the domain-invariant features onto one of the $K$ classes of interest guided by the following cross-entropy loss:

$$\mathcal{L}_{cl} = \frac{1}{N_s + N_t} \sum_{i=1}^{N_s+N_t} \ell_{ce}\left(f \circ g_{\text{inv}}(x_i), y_i\right). \quad (2)$$

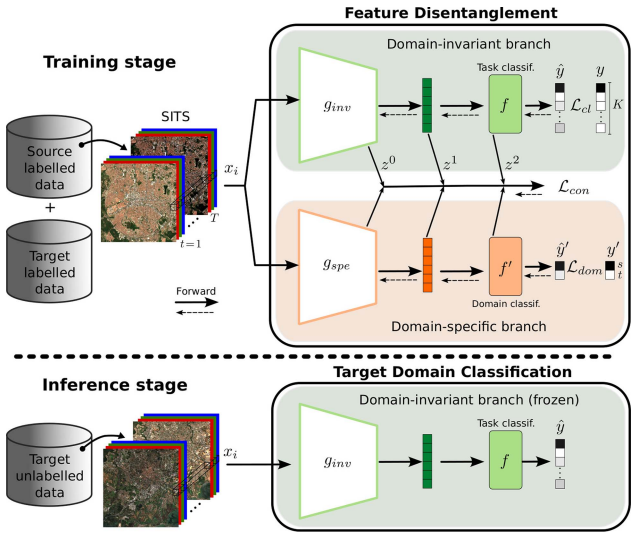


Fig. 5. Architecture of the proposed dual-branch network used in the training stage and composed of two independent branches which disentangle the domain-invariant information (top branch) from domain-specific information (bottom branch). Class ($\mathcal{L}_{\text{cl}}$) and domain ($\mathcal{L}_{\text{dom}}$) discrimination losses used respectively on the top and bottom branches, while a multilevel contrastive loss ($\mathcal{L}_{\text{con}}$) is used to intermediate features at different depths from both branches. At inference time, only the domain-invariant encoder is used for classifying the target domain.

*c) Contrastive Learning:* To further decouple the two separate branches, we employ contrastive loss which has shown promising results for feature disentanglement in previous works [45]. The general objective of contrastive learning approaches is to learn data representations by comparing and contrasting similar and dissimilar information in terms of pairwise comparisons. More precisely, pair of similar elements are denoted as positive pairs and pair of dissimilar elements are referenced as negative pairs where similarity is defined based to some criteria. The goal is to ensure that, in the manifold generated by the representation learnt by the neural network, positive pairs are close to each other and negative pairs are

located far apart. In our case, since we have access to both the class labels and domain labels, we leverage the supervised contrastive loss proposed by [46], where the positive pairs are given by all samples sharing the same label. However, here, the use of the supervised contrastive loss is not trivial since we have two separate label spaces (class and domain labels). To address this issue, we adopt a mixed label space $\mathcal{Y}_{\text{mix}}$ composed of $3K$ classes, where the domain-invariant features are mapped onto the $K$ first labels while the last $2K$ are reserved to the domain-specific features—more specifically, $K$ for the source domain and $K$ for target domain. This leads to $\mathcal{Y}_{\text{mix}} = \{1, \ldots, K, s1, \ldots, sK, t1, \ldots, tK\}$.[2]

For instance, given a sample $x_i^s$ coming from the source domain and associated with label $K$ (i.e., $y_i^s = K$), then its corresponding domain-specific embedding $g_{\text{spe}}(x_i^s)$ will be mapped to class $sK$, while its domain-invariant counterpart $g_{\text{inv}}(x_i^s)$ will be mapped to class $K$ in the mixed label space $\mathcal{Y}_{\text{mix}}$. Likewise, for a target sample $x_i^t$, its embeddings $g_{\text{spe}}(x_i^t)$ and $g_{\text{inv}}(x_i^t)$ are mapped to classes $tK$ and $K$, respectively. Notice that, even though the two samples come from different domains, their domain-invariant embeddings are purposely mapped to the same class $K$ in $\mathcal{Y}_{\text{mix}}$.

Denoting $z$ the extracted embeddings (features) $g_{\text{inv}}(x)$ and $g_{\text{spe}}(x)$, we consider an augmented batch $I$ of size $2B$ containing both $g_{\text{inv}}(x_i)$ and $g_{\text{spe}}(x_i)$ features for each $i \in \{1, \ldots, B\}$ in the original batch. The resulting supervised contrastive loss is defined as follows:

$$\mathcal{L}_{\text{con}} = -\sum_{i \in I} \frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{a \in I \setminus \{i\}} \exp(z_i \cdot z_a / \tau)} \quad (3)$$

where $P(i) := \{p \in I \setminus \{i\} : y_p = y_i\}$ with cardinality $|P(i)|$ is the set of *positive* examples w.r.t. the current *anchor* $i \in I := \{1, \ldots, 2B\}$ and $\tau \in \mathbb{R}^+$ is a scalar temperature parameter. Therefore, the goal of this loss is to push together, in the feature space, the embeddings corresponding to the same category (positive examples) while repelling them from the rest (negative examples). The positive examples here correspond to those sharing the same class in the mixed label space $\mathcal{Y}_{\text{mix}}$ defined above, in other words, those that share simultaneously the same class, among the K existing ones, and the same category type, among three possibilities: source or target domain (for domain-specific features), or domain-invariant features—recalling that domain-invariant features from different samples of the same class are matched together regardless of the their provenance (source or target domain). Finally, the disentanglement between domain-invariant and domain-specific features is achieved through the complementary action of the different loss functions. Specifically, the optimization terms enforce the following: i) domain-specific features should contain information about their respective domains; ii) domain-invariant features should be discriminative for the downstream classification task, and iii) contrastive learning structures the geometric manifold where domain-specific and domain-invariant features are pushed away from each other making them orthogonal thus, ensuring that these two groups of features carry complementary information to some extent.

### D. Multilevel Supervision

To further enforce the feature disentanglement, we propose to perform contrastive learning not only at the level of the encoder's output, but at multiple depths within the network architecture.

Then, the loss function described in (3) is actually also applied to intermediate features at different depths of the network. For that matter, we denote $\mathcal{L}_{\text{con}}^l$ the contrastive loss (3) applied to the intermediate features $z^l$ at depth $l$, as depicted in Fig. 5. Specifically, in our case, we use three levels of supervision with: $l = 0$ for the encoder's last internal layer; $l = 1$ for the encoder's output features; $l = 2$ for the output of the classifier's first fully connected layer. Note that $\mathcal{L}_{\text{con}}^l$ applies exclusively to features at depth $l$, which share the same space dimension, and, thus, features at different depths are never mixed together.

### E. Model Summary and Training

The resulting loss is given by

$$\mathcal{L} = \mathcal{L}_{ce} + \mathcal{L}_{\text{dom}} + \textstyle\sum_l \mathcal{L}_{\text{con}}^l \quad (4)$$

Empirically, we observed that weighting the different losses did not have a conclusive impact on the final model performance. For this reason, we used an unweighted sum of the three terms as described above.

As backbone, we leverage a widely popular architecture for the analysis of SITS data, specifically the temporal convolutional neural network (TempCNN) proposed by [47]. However, any other recent time series encoders could be used [48], [49]. The choice of TempCNN as the backbone is due to its simplicity, which provides us with some assurances that the behavior we analyze will be primarily related to the proposed framework, rather than being influenced by the use of a sophisticated and advanced backbone.

The TempCNN encoder consists of three 1-D convolutional layers with 64 channels each. The classifier is composed of a fully connected layer with 256 hidden units, batch normalization, and ReLU activation, followed by a linear output layer with Softmax activation.

## IV. Experiments

In this section, we report and discuss the experimental evaluation carried out on the study sites presented in Section II. Our objective is to evaluate the performance of *REFeD* across various dimensions. First, we undertake a quantitative assessment comparing the performance of *REFeD* against baselines and competing approaches. Second, we conduct a qualitative examination of the LC maps generated by *REFeD*. Finally, we inspect the internal representations learned by our model, visually comparing them to those of some of the top-performing competitors.[3]

[2]More generally, we define $\mathcal{Y}_{\text{mix}} = \mathcal{Y} \cup (\mathcal{Y} \times \mathcal{Y}')$ with a total of $|\mathcal{Y}_{\text{mix}}| = (|\mathcal{Y}'|+1)|\mathcal{Y}|$ classes.

[3]The code associated to this article is available at: https://github.com/cassiofragadantas/SDA_LULC.

### A. Baseline Methods

With the goal to assess the performance of *REFeD* w.r.t. baselines and strategies coming from Semi-Supervised Domain Adaptation and Domain Generalization literature, we consider.

- *Only Source:* This strategy trains a model only considering source data and, then, the obtained classifier is directly deployed on the target data. The main purpose of this method is to have an empirical estimate about the distribution shift between source and target domains. We implement this baseline considering Random Forest, XGBoost, TempCNN [47], and ATCNN [50] (an extension of temporal convolutional neural network that integrates attention on the temporal dimension) as classifiers.
- *Only Target:* This strategy trains a model only considering the target labeled data and, then, the obtained classifier is employed to classify the remaining target samples. The main purpose of this method is to provide the reference performances without the use of historical or out-of-year data. We implement this baseline considering random forest, XGBoost, TempCNN [47], and ATCNN [50] as classifiers.
- *Source+Target:* This strategy trains a model over both source and target labeled data. Then, the obtained classifier is employed to classify the remaining target samples. Here, the data coming from different distributions are mixed together constituting a multidomain training dataset. We implement this baseline considering Random Forest, XGBoost, TempCNN [47], and ATCNN [50] as classifiers.
- *Fine Tuning:* This strategy trains a model over the labeled source domain and, then, the resulting model is fine tuned on target labeled samples. This is an alternative way to combine both source and target data. We implement this baseline considering both TempCNN [47] and ATCNN [50] as backbone approaches.
- The domain adversarial neural network (DANN) method originally introduced in [20]. This is a well-known and largely employed UDA approach that exploits gradient reversal layer with the aim to obtain data representations that are invariant to the particular domain they come from.
- The conditional adversarial domain adaptation (CDAN) approach [51]. This method extends DANN by conditioning the domain discriminator on the classification output.
- *Sourcerer [35]:* This recent SSDA approach has been proposed to cope with the analysis of SITS data for the downstream task of LULC mapping. Sourcerer is a bayesian-inspired, deep learning-based framework, that internally exploits the TempCNN model as backbone, similarly to *REFeD*. The technique leverages a deep learning model trained on a source domain and then fine-tunes the model on the available target domain via a regularizing term that automatically adjusts the degree to which the model weights are modified to fit the target data.
- *POEM [52]:* This recent Domain Generalization approach learns domain-invariant and domain-specific representations with a similar dual-branch architecture as adopted by our framework. It enforces polarization via orthogonality constraints. This approach was primarily introduced for image classification. In order to transfer it on SITS data, also for this competitor we employ the TempCNN architecture as a backbone.

### B. Experimental Settings

For all the competing approaches, labeled source data are entirely employed while, for the target domain, data are split into three parts: training, validation, and test sets following a proportion of 50%, 20%, and 30% of the original target dataset, respectively. Regarding the *Only Source* baseline, the model is trained considering only the labeled source data. Concerning the DANN and CDAN domain adaptation methods, the target training set is leveraged to set up the adversarial learning stage. Furthermore, with the aim to avoid possible spatial bias in the evaluation procedure [53], we impose that all the pixels belonging to the same object will be exclusively associated with one of the data partitions (training, validation, or test). The splitting procedure is repeated five times and the average results are reported.

Concerning the evaluation tasks, according to the data presented in Section II, we set up two transfer tasks per benchmark. Each transfer task is denoted as ($\mathcal{D}_s$ + $\mathcal{D}_t$ $\rightarrow$ $\mathcal{D}_t$) where the right arrow indicates the transfer direction from the combined source/target labeled training dataset ($\mathcal{D}_s$ + $\mathcal{D}_t$) to the test target ($\mathcal{D}_t$) dataset. For the *Koumbia* study site, we consider as transfer tasks (2020 + 2021 $\rightarrow$ 2021) and (2021 + 2020 $\rightarrow$ 2020) and for the *Centre-Val de Loire* study site, we consider the transfer tasks (2018 + 2021 $\rightarrow$ 2021) and (2021 + 2018 $\rightarrow$ 2018).

The values of the SITS benchmarks were scaled per year and per band considering the 2nd and 98th percentile of the data distribution as minimum and maximum values. The assessment of the model performances was done considering the following metrics: *Weighted F1-score* (simply indicated with *F1-score*) and *Accuracy* (global precision).

*Implementation details:* For the neural network approaches, the training stage has been conducted for 200 epochs. For methods based on fine-tuning, we used 100 epochs for the initial training and 100 epochs for the fine-tuning stage. For all methods, we adopt a learning rate of $10^{-4}$, the AdamW [54] optimizer, and a batch size of 256. Regarding *REFeD*, based on recent literature on contrastive learning [55], we set the temperature hyperparameter $\tau$ to 0.07 and we consider a batch size of 512 since it has been noted that contrastive loss benefits from larger batch sizes. The drop out value is set to 50%. Considering Random Forest classifiers, we optimize the model via the tuning of one parameter: the number of trees in the forest. We vary this parameter in the range {100, 200, 300, 400, 500}. The optimization of this parameter is based on the validation set. Experiments are carried out on a workstation with a dual Intel (R) Xeon (R) CPU E5-2667v4 (@3.20GHz) with 256 GB of RAM and TITAN X (Pascal) GPU. All the deep learning methods are implemented using the *Pytorch* deep learning library. All the models run on a single GPU. Random Forest is implemented using the Python *Scikit-learn* library [56] and run on CPU.

TABLE III
OVERALL PERFORMANCES (F1 SCORES) IN THE *KOUMBIA* STUDY SITE

| **Strategy** | **Method** | 2020 + 2021 → 2021 | | 2021 + 2020 → 2020 | |
|---|---|---|---|---|---|
| | | **F1-Score** | **Accuracy** | **F1-Score** | **Accuracy** |
| *Only Source* | RF | 72.40 ± 3.94 | 72.25 ± 4.08 | 71.98 ± 3.54 | 71.81 ± 3.63 |
| | XGBoost | 61.91 ± 3.35 | 61.20 ± 3.60 | 59.24 ± 3.54 | 58.55 ± 3.45 |
| | TempCNN | 64.54 ± 4.32 | 65.25 ± 4.14 | 69.54 ± 4.88 | 69.95 ± 5.00 |
| | ATCNN | 68.15 ± 1.78 | 68.27 ± 1.54 | 66.58 ± 3.12 | 65.88 ± 2.97 |
| *Target domain* | RF | 77.56 ± 2.82 | 77.45 ± 2.90 | 76.78 ± 3.53 | 76.65 ± 3.54 |
| | XGBoost | 74.63 ± 3.93 | 74.76 ± 3.85 | 74.34 ± 3.76 | 74.48 ± 3.70 |
| | TempCNN | 76.59 ± 2.94 | 76.63 ± 2.91 | 75.54 ± 5.04 | 75.47 ± 5.09 |
| | ATCNN | 71.38 ± 2.10 | 71.54 ± 2.18 | 72.90 ± 3.89 | 72.83 ± 3.94 |
| *Source+Target* | RF | 77.49 ± 3.90 | 77.30 ± 3.93 | 78.42 ± 4.02 | 78.32 ± 4.03 |
| | XGBoost | 74.89 ± 4.37 | 75.01 ± 4.32 | 77.04 ± 3.30 | 77.14 ± 3.29 |
| | TempCNN | 78.60 ± 2.94 | 78.48 ± 3.03 | 78.95 ± 4.27 | 78.92 ± 4.27 |
| | ATCNN | 74.46 ± 2.90 | 74.42 ± 2.92 | 75.20 ± 2.61 | 75.18 ± 2.51 |
| *Fine Tuning* | TempCNN | 78.04 ± 2.75 | 78.00 ± 2.78 | 78.56 ± 3.63 | 78.55 ± 3.64 |
| | ATCNN | 72.30 ± 2.11 | 72.36 ± 2.20 | 72.64 ± 4.11 | 72.63 ± 4.08 |
| *UDA approaches* | DANN | 75.87 ± 4.87 | 75.93 ± 4.81 | 77.33 ± 3.92 | 77.20 ± 4.04 |
| | CDANN | 74.82 ± 4.49 | 74.89 ± 4.32 | 78.46 ± 2.30 | 78.54 ± 2.22 |
| SSDA approach | Sourcerer | 76.72 ± 2.63 | 76.54 ± 2.64 | 76.34 ± 3.94 | 76.24 ± 3.99 |
| DG approach | POEM | 78.35 ± 3.43 | 78.27 ± 3.41 | 78.41 ± 4.63 | 78.35 ± 4.66 |
| Proposed Method | *REFeD* | **79.23** ± 3.33 | **79.17** ± 3.27 | **82.15** ± 3.65 | **82.09** ± 3.67 |

The bold values indicate the best results among competitors.

TABLE IV
OVERALL PERFORMANCES (F1 SCORES) IN THE *CENTRE-VAL DE LOIRE* STUDY SITE

| **Strategy** | **Method** | 2018 + 2021 → 2021 | | 2021 + 2018 → 2018 | |
|---|---|---|---|---|---|
| | | **F1-Score** | **Accuracy** | **F1-Score** | **Accuracy** |
| *Only Source* | RF | 69.66 ± 1.64 | 72.35 ± 1.07 | 59.31 ± 3.19 | 62.36 ± 2.56 |
| | XGBoost | 47.86 ± 0.72 | 43.29 ± 0.83 | 35.76 ± 1.42 | 31.40 ± 1.28 |
| | TempCNN | 47.96 ± 1.23 | 46.26 ± 0.76 | 43.10 ± 2.46 | 41.68 ± 3.31 |
| | ATCNN | 52.89 ± 2.15 | 47.22 ± 2.95 | 44.13 ± 1.77 | 37.84 ± 1.27 |
| *Target domain* | RF | 79.95 ± 2.45 | 80.70 ± 1.78 | 70.63 ± 3.05 | 71.86 ± 2.55 |
| | XGBoost | 80.33 ± 0.81 | 79.10 ± 1.44 | 71.93 ± 2.58 | 70.50 ± 3.17 |
| | TempCNN | 82.60 ± 1.15 | 80.61 ± 1.41 | 73.39 ± 3.12 | 71.46 ± 3.46 |
| | ATCNN | 80.39 ± 1.37 | 79.43 ± 1.93 | 70.11 ± 1.38 | 68.19 ± 1.36 |
| *Source+Target* | RF | 79.00 ± 2.24 | 79.83 ± 1.60 | 68.94 ± 2.66 | 70.45 ± 2.31 |
| | XGBoost | 79.81 ± 0.96 | 79.10 ± 1.27 | 69.69 ± 1.77 | 68.11 ± 1.81 |
| | TempCNN | 83.46 ± 1.40 | 81.54 ± 2.07 | 75.17 ± 3.61 | 74.17 ± 3.71 |
| | ATCNN | 78.45 ± 1.06 | 77.39 ± 1.05 | 68.45 ± 1.20 | 67.08 ± 1.31 |
| *Fine Tuning* | TempCNN | 82.94 ± 2.04 | 83.51 ± 1.75 | 74.43 ± 3.78 | 75.70 ± 3.29 |
| | ATCNN | 80.72 ± 0.81 | 79.80 ± 1.25 | 69.13 ± 1.46 | 67.99 ± 1.93 |
| *UDA approaches* | DANN | 52.84 ± 3.47 | 56.17 ± 3.78 | 52.67 ± 2.35 | 53.33 ± 1.91 |
| | CDAN | 47.29 ± 0.32 | 52.72 ± 0.28 | 42.81 ± 0.76 | 44.07 ± 1.10 |
| SSDA approach | Sourcerer | 82.76 ± 1.69 | 83.49 ± 1.29 | 74.02 ± 3.36 | 74.93 ± 2.90 |
| DG approach | POEM | 82.89 ± 1.71 | 83.51 ± 1.31 | 74.28 ± 3.11 | 75.14 ± 2.80 |
| Proposed Method | *REFeD* | **84.45** ± 1.61 | **84.86** ± 1.35 | **77.60** ± 3.15 | **78.02** ± 2.90 |

The bold values indicate the best results among competitors.

### C. Comparison With Competing Methods

Tables III and IV summarize the results obtained for the two study areas, *Koumbia* and *Centre-Val de Loire* respectively by reporting the average F1-score and the accuracy considering the different combination of methods and strategies. As expected, regardless of the dataset, for RF, XGBoost, TempCNN and ATCNNclassifiers, the lowest accuracy is obtained when only source-labeled data are considered, while the highest classification results are obtained when training data from both domains are used.

TABLE V
PER-CLASS AVERAGE F1 SCORES FOR *KOUMBIA*: (A) SCENARIO (2020 + 2021 → 2021), (B) SCENARIO (2020 + 2021 → 2020)

(a)

| Class | TempCNN (Source+Target) | Sourcerer | POEM | REFeD |
|---|---|---|---|---|
| Cereals | 74.06 | 70.79 | 72.48 | **75.73** |
| Cotton | 75.77 | 72.79 | 76.04 | **77.87** |
| Oleaginous | 64.18 | 60.22 | 64.24 | **66.96** |
| Grassland | **78.58** | 75.62 | 78.11 | 77.61 |
| Shrubland | **78.90** | 76.47 | 78.32 | 77.90 |
| Forest | 84.27 | 85.06 | 84.78 | **86.05** |
| Bare soil | **83.45** | 80.11 | 82.83 | 83.04 |
| Water | **100.0** | **100.0** | **100.0** | **100.0** |

(b)

| Class | TempCNN (Source+Target) | Sourcerer | POEM | REFeD |
|---|---|---|---|---|
| Cereals | 77.48 | 74.25 | 76.39 | **79.38** |
| Cotton | 78.10 | 75.48 | 77.18 | **82.96** |
| Oleaginous | 72.91 | 69.93 | 71.69 | **75.14** |
| Grassland | 85.22 | 82.68 | 84.14 | **86.38** |
| Shrubland | 77.31 | 74.64 | 77.29 | **81.38** |
| Forest | 77.03 | 74.97 | 77.33 | **81.94** |
| Bare soil | **79.55** | 77.61 | 76.78 | 78.71 |
| Water | **100.0** | **100.0** | **100.0** | **100.0** |

The bold values indicate the best results among competitors.

TABLE VI
PER-CLASS AVERAGE F1 SCORES FOR *CENTRE-VAL DE LOIRE*: (A) SCENARIO (2018 + 2021 → 2021), (B) SCENARIO (2021 + 2018 → 2018)

(a)

| Class | TempCNN (Source+Target) | Sourcerer | POEM | REFeD |
|---|---|---|---|---|
| Soft wheat | 91.35 | 91.53 | 90.52 | **91.91** |
| Maize | 94.91 | 92.72 | 94.73 | **96.14** |
| Barley | 94.94 | 94.71 | 94.75 | **95.58** |
| Other cereals | 85.55 | 85.84 | 84.36 | **87.54** |
| Oleaginous | 86.76 | **88.56** | 86.87 | 86.43 |
| Winter Fallows | 73.82 | 71.98 | 73.22 | **73.87** |
| Leguminous | 55.26 | 52.26 | 54.93 | **61.63** |
| Fodder | 74.31 | 73.04 | 72.11 | **75.82** |
| Meadow | **74.35** | 71.33 | 73.92 | 73.48 |
| Other crops | **84.16** | 83.88 | 84.06 | 83.92 |

(b)

| Class | TempCNN (Source+Target) | Sourcerer | POEM | REFeD |
|---|---|---|---|---|
| Soft wheat | 85.55 | 85.23 | 84.21 | **85.80** |
| Maize | 85.65 | 85.48 | 86.23 | **87.04** |
| Barley | 94.28 | 94.34 | 93.63 | **94.63** |
| Other cereals | 73.35 | 73.87 | 71.35 | **73.79** |
| Oleaginous | 72.49 | 70.01 | 73.02 | **76.98** |
| Winter Fallows | 74.79 | 74.24 | 73.78 | **75.13** |
| Leguminous | 54.77 | 46.51 | 53.05 | **65.15** |
| Fodder | 70.47 | 68.24 | 66.91 | **72.07** |
| Meadow | 72.99 | 71.62 | 71.34 | **73.04** |
| Other crops | 57.40 | 57.76 | 57.66 | **64.06** |

The bold values indicate the best results among competitors.

By focusing our attention on the TempCNN architecture, the F1-scores obtained using only the source-labeled data are 64.54 and 69.54 in *Koumbia* on EO data acquired in 2021 and 2020, respectively, and 47.96 and 43.10 in *Centre-Val de Loire* on EO data acquired in 2021 and 2018, respectively. Although valuable, the historical reference data may not be completely representative of the recently acquired EO data. This fact is further supported by the performance of the UDA competitors (DANN and CDAN), which only achieve on par results to the *Only Source* baseline and are outperformed by all other approaches. Moreover, the class statistical distributions of SITS acquired over different years can severely shift. Using only the target-labeled data the obtained accuracy increases, i.e., F1 scores of 76.59 and 75.54 in *Koumbia* on EO data acquired in 2021 and 2020, respectively, and 82.60 and 73.39 in *Centre-Val de Loire* on EO data acquired in 2021 and 2018, respectively. It is worth noting that the importance of using labeled data from the target domain is even more visible in the *Centre-Val de Loire* dataset since the source and target domains are different from both the spatial and temporal viewpoints. The joint use of source and target labeled data has a positive impact on the classification performances, leading to F1 scores of 78.60 and 78.95 in *Koumbia* on EO data acquired in 2021 and 2020, respectively, and 83.46 and 75.17 in *Centre-Val de Loire* on EO data acquired in 2021 and 2018, respectively.

These results further improve when using SSDA and DG methods, which better combine source and target information. However, the highest classification accuracy is obtained by the proposed approach *REFeD* which achieves F1 scores of 79.23 and 82.15 in *Koumbia* on EO data acquired in 2021 and 2020, respectively, and 84.45 and 77.60 in *Centre-Val de Loire* on EO data acquired in 2021 and 2018, respectively.

Per-class performances are detailed in Table V(a) and (b) for the *Koumbia* study site and in Table VI(a) and (b) and *Centre-Val de Loire*, respectively. Concerning the *Koumbia* benchmark, Table V(a) and (b), we can observe that, no matter the transfer task, *REFeD* achieves the best performances in terms of F1-Score on all the agricultural LC classes. This is of particular interest since such classes are characterized by strong shifts from one cultural year to another one due to crop rotations related to the underlying agricultural practices. A notable improvement, related to the proposed method, can also be noted on the *Forest* LC class, especially for the transfer task (2021 + 2020 → 2020). Regarding all the other LC classes, *REFeD* achieves comparable results w.r.t. to all the other competing methods with, for all these cases, less than a point of difference in terms of F1-score.

Regarding the *Centre-Val de Loire* benchmark, Table VI(a) and (b), we can note that, *REFeD* obtains a systematic improvement, in terms of F1-Score, for the family of cereal classes (*Soft wheat*, *Maize*, *Barley* and *Other cereals*), regardless of

TABLE VII
TRAINING TIME OF THE DIFFERENT COMPETING APPROACHES FOR THE *KOUMBIA* STUDY SITE UNDER THE TRANSFER TASK (2020 + 2021 → 2021)

| Method | Training Time |
|---|---|
| RF | 77m |
| XGBoost | 12m |
| TempCNN | 20m |
| ATCNN | 20m |
| DANN | 16m |
| CDANN | 23m |
| Sourcerer | 35m |
| POEM | 36m |
| *REFeD* | 40m |

For *RF, XGBoost, TempCNN* and *ATCNN* we consider the *source+target* training strategy to be comparable with all the other approaches.

TABLE VIII
ABLATION ANALYSIS OF *REFED* OVER THE *CENTRE-VAL DE LOIRE* STUDY SITE

| | *Centre-Val de Loire* | |
|---|---|---|
| | 2018 + 2021 → 2021 | 2021 + 2018 → 2018 |
| w/o $\mathcal{L}_{dom}$ | 83.63 ± 1.60 | 76.62 ± 2.79 |
| w/o $\sum_l \mathcal{L}^l_{con}$ | 83.15 ± 1.45 | 74.13 ± 2.59 |
| w/o Multi-level sup. | 83.69 ± 1.93 | 75.97 ± 2.35 |
| *REFeD* | **84.45** ± 1.61 | **77.60** ± 3.15 |

We consider both transfer tasks: (2018 + 2021→ 2021) and (2021 + 2018→ 2018). The bold values indicate the best results among competitors.

the transfer task. Notably, the most significant enhancement is observed in the *Leguminous* crop class, where *REFeD* attains an improvement between 6 and 10 points of F1-Score compared to the best competing approach. This is even interesting since the *Leguminous* class is the most underrepresented crop type in the considered benchmark in terms of number of samples. This point further underscores the quality of the proposed approach demonstrating its ability to handle scenarios characterized by significant class imbalances, a common situation in real-world applications. Finally, still regarding the *Centre-Val de Loire* benchmark, we can underline that for the transfer task (2021 + 2018 → 2018), *REFeD* achieves the best performances for all the crop type classes.

Table VII reports the training time of the various competing approaches involved in the experimental evaluation. All the methods require between 12 min and 1.5 h to complete the training stage, with *REFeD* taking approximately 40 min to learn its internal parameters. The Random Forest method has the highest training time, likely because it does not leverage GPU computation, unlike the other approaches. Given that a LC classification model typically needs to be trained once per season (or year), the observed training times remain more than reasonable and align well with the constraints associated with the downstream application.

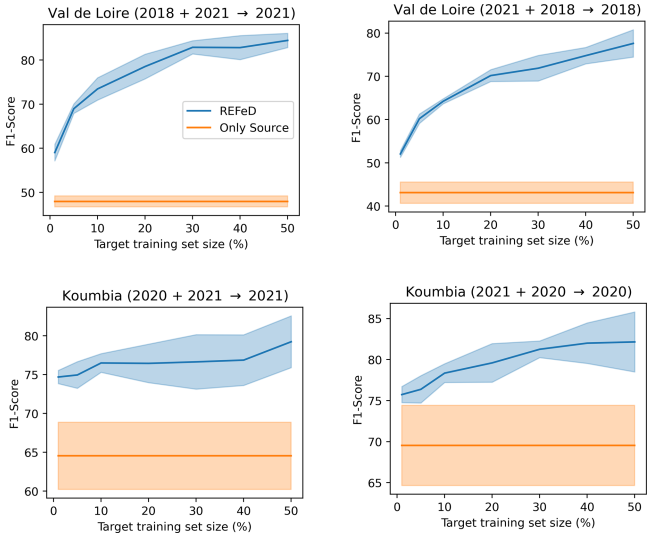


Fig. 6. Average performance of the proposed *REFeD* approach (solid blue line) and corresponding standard deviation (shaded blue area) when varying the target training set size from 1% to 50% of the available data. As a baseline, the performance using only source data (i.e., no target data) is shown in orange.

### D. Further Analysis of the Proposed Approach

*Ablation study:* In Table VIII, we evaluate the importance of each component of the proposed approach by reporting the method's performance when removing each component individually, namely: 1) the domain classifier, 2) the contrastive loss, and 3) the multilevel supervision. In the latter scenario, the contrastive loss is applied only to the output features (those at depth $l = 1$, i.e., $z^1$ in Fig. 5). All three tested variants achieve worse results than the full architecture, which corroborates the effectiveness and complementary nature of these components in enhancing the overall performance. This suggests that each component plays a crucial role in tackling different aspects of the problem, and their combined effect is necessary to fully leverage the strengths of the proposed approach.

*Reduced target training set:* In Fig. 6, we evaluate the performance of our proposed approach in a regime of limited target data availability. Naturally, performance improves as more target data becomes available. However, it is noteworthy that even when exploiting a very limited amount of target data (only 1% of the full dataset), *REFeD* already attains remarkably superior results than relying solely on source data.

### E. Visual Analysis

In this part of the experimental assessment, we provide qualitative analyses to further evaluate the behavior of *REFeD* considering the *Koumbia* site, in the transfer task (2021 + 2020 → 2020). To this end, in addition to our framework, we also consider the top-performing competitors: *POEM*, *Sourcerer* and *TempCNN* (Source + Target). We first inspect some extracts from the obtained LC maps and, then we visually examine the internal representations learned by the different methods by means of the t-SNE [57] dimensionality reduction technique.

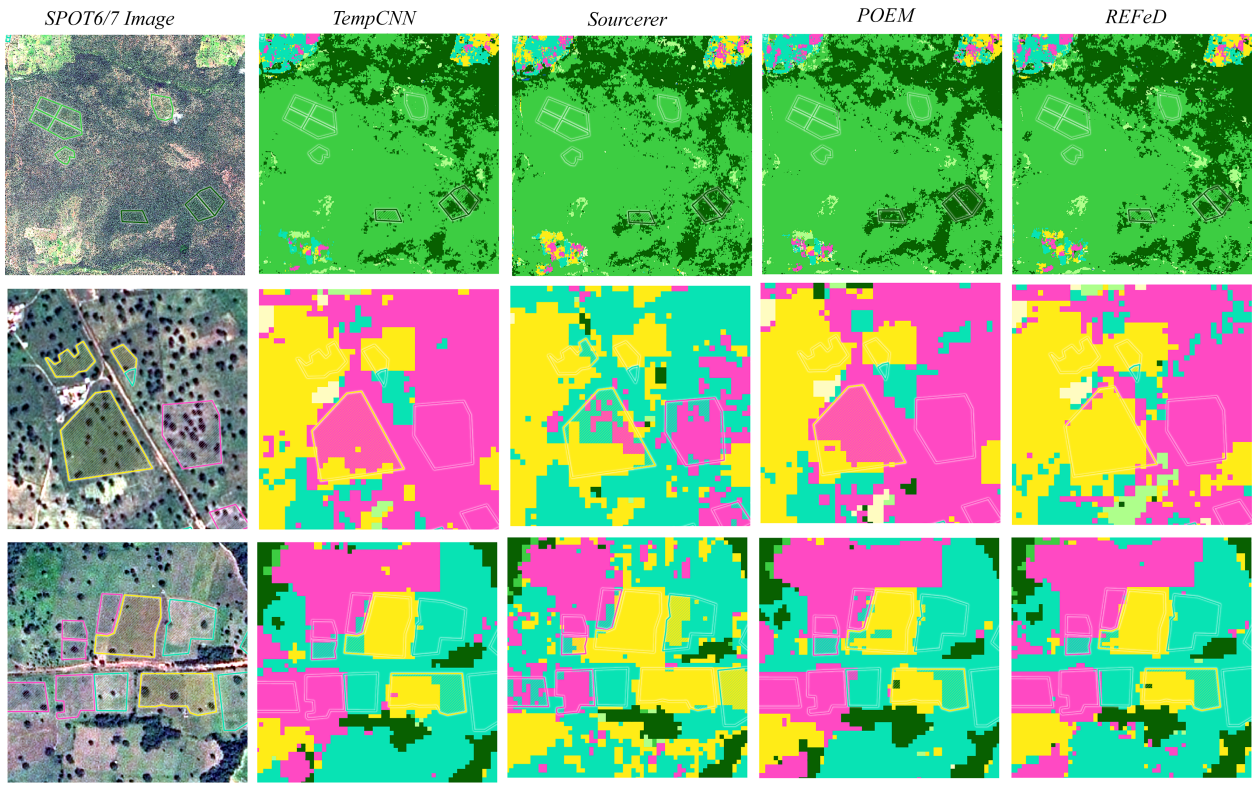


Fig. 7. Extracts from the provided LC maps per method. Ground truth areas outlined over the extracts using the same color codes of Fig. 1.

*1) Land Cover (LC) Maps:* In order to give a further insight into the performances of the proposed method, we performed a qualitative analysis of the LC maps provided by each of the competing methods. In Fig. 7, we report some examples for the (2021 + 2020 → 2020) transfer scenario over the *Koumbia* study site. As easily observable, the *Sourcerer* method tend to generate much noisier maps with respect to the competitors, which seems to be the main factor limiting its global performances. The other methods provide maps of comparable spatial characteristics, with *REFeD* significantly outperforming the competitors on classes related to natural vegetation with different densities (like the *Forest* class on the first row—bottom/right of the clip). More occasionally, *REFeD* also seems to retrieve the correct crop class to whole fields within the cropland which are entirely misclassified by other methods (e.g., second row, for the big *Cereal* field in the middle). Otherwise, the systematic improvement of *REFeD* over its best competitors mainly occurs at the finest scales, like for the fields in the example on the last row of Fig. 7, where most of the "holes" generated by *TempCNN* and *POEM* appear as properly filled.

*2) Visualization of Internal Model Representations:* In this last stage of our experimental evaluation, we provide a visual inspection of the internal feature representation learned by *REFeD*, *POEM*, *Sourcerer*, and *TempCNN* (Source + Target) on the *Koumbia* study site. To this end, we randomly chose 50 samples per LC class from the target domain and we extracted the corresponding feature representation per method. Subsequently, we applied t-SNE [57] to reduce the feature dimensionality for visualization purposes. Results are depicted in Fig. 8. We can

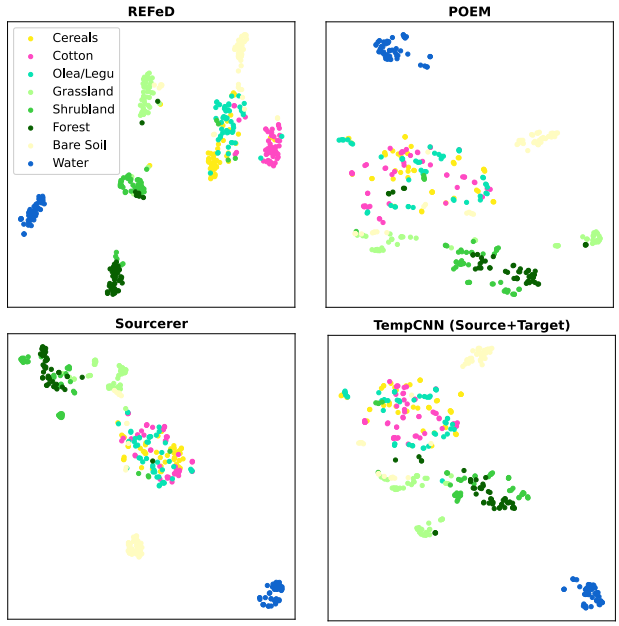


Fig. 8. t-SNE results for the proposed approach (top left) and three different baselines: POEM (top right), Sourcerer (bottom left), and TempCNN (Source+Target) (bottom right) on the *Koumbia* study.

note that all the methods well separate samples coming from the $Water$ and $Bare\ soil$ classes from the rest of the data. However, while competing approaches clearly mix together samples from all other LC classes, *REFeD* partially alleviates

clutter issues on the remaining classes providing a better visual behavior in terms of cluster structure, on the considered subset of target data. This can be noted, for instance, regarding both the agricultural (*Cereals*, *Cotton*, and *Oleaginous/Leguminous*) and natural vegetation (*Grassland*, *Shrubland*, and *Forest*) classes. Overall, the visualization of internal features representation is coherent with the quantitative as well as qualitative findings we previously discussed.

## V. Conclusion

In this work we have presented *REFeD*, a novel deep learning framework that enhances the accuracy of the current LC mapping process by combining together EO and reference historical and recent data with the aim to give value again to overlooked reference data under a data-centric perspective. *REFeD* is based on a dual-branch network, consisting of two encoders with identical architectures but unshared parameters. It relies on contrastive learning to disentangle invariant and specific per-domain features to recover the intrinsic information related to the downstream LC mapping task. Furthermore, *REFeD* is equipped with an effective supervision scheme where feature disentanglement is further enforced via multiple levels of supervision.

From the results obtained, it turned out that the use of historical reference data alone is not sufficient to perform an update of the existing LC maps, even when considering the same study area. As expected, the use of labeled data coming from the target year leads to improvements in classification accuracy, however, *REFeD* achieves the highest classification accuracy leading to the possibility of enhancing the considered LC mapping task by exploiting auxiliary reference labeled data (i.e., historical reference data on the same study area or a related one). In addition, the results obtained with *REFeD*, on study areas featured by diverse and contrasted landscapes, highlight its added value compared to both: i) the direct combination of source and labeled data and ii) recent competing methods from related literature. Most importantly, *REFeD* systematically outperforms models that only exploit target data paving the way to the reuse of historical and/or overlooked reference data for the LC mapping task taking as input SITS data.

*Limitations and future work:* In situations characterized by significant imbalance in data sample volumes between different domains (years), our approach, like any other machine learning framework dealing with data from multiple domain distributions, may struggle to effectively leverage historical knowledge to enhance the land cover mapping process. This issue has not been investigated in our current research, but it could be the focus of a dedicated study. Such a study would explore this dimension in depth, aiming to determine the minimum historical data volume necessary to improve the classification performance considering, at least, all the frameworks evaluated in this research. Further analysis could focus on characterizing the performance of the model under different degrees of distribution shift to assess its robustness. A straightforward approach would be to assume that the temporal gap between historical and current data is directly proportional to the degree of distribution shift between remotely sensed data. However, this assumption may be inaccurate because, depending on the specific pair of years considered in the case study, remotely sensed data from closer years may have a higher distribution shift than any data from any pair of years, regardless of the temporal gap. To address this point, we first need to define a method for estimating the magnitude and nature of distribution shifts. This is a particularly challenging task, but one that could have significant benefits in general areas of transfer learning. Once a measure of distribution shift has been established, we could then analyze the performance of our approach in scenarios with increasing degrees of distribution shift. A current limitation of our framework is the fact that it can only handle a single year of historical data. Future extensions could include adapting *REFeD* to scenarios where multiple year of historical ground truth data is available, further advancing the effective reuse of overlooked and/or neglected efforts related to past resource-intensive field campaigns. To this purpose, a straightforward extension should consist into directly adding data coming from other years into the learning process increasing the cardinality of the $\mathcal{Y}_{\text{mix}}$ set. While this can be a cost-free solution, it could struggle to scale up as the number of available years of historical ground truth data increases. Another solution could be model the problem as a kind of semisupervised multisource domain adaptation process where the goal would be to extract an invariant representation across all the data domains.

In addition, we can contemplate the use of complementary remote sensing data in a multisource setting where the domains are described by different sensors (e.g., Sentinel-2, Sentinel-1, and Landsat). Finally, another interesting enhancement for the current framework would be the capacity to leverage additional unlabeled data that might be available from the target domain.

## References


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

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

[3] D. Ienco, R. Gaetano, C. Dupaquier, and P. Maurel, "Land cover classification via multitemporal spatial data by deep recurrent neural networks," *IEEE Geosc. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1685–1689, Oct. 2017.

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

[5] Z. Huang et al., "A spectral-temporal constrained deep learning method for tree species mapping of plantation forests using time series Sentinel-2 imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 204, pp. 397–420, 2023.

[6] A. Kavvada et al., "Towards delivering on the sustainable development goals using earth observations," *Remote Sens. Environ.*, vol. 247, 2020, Art. no. 111930.

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

[8] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Vaughan, "A theory of learning from different domains," *Mach. Learn.*, vol. 79, pp. 151–175, 2010. [Online]. Available: http://www.springerlink.com/content/q6qk230685577n52/

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

[10] J. Peng, Y. Huang, W. Sun, N. Chen, Y. Ning, and Q. Du, "Domain adaptation in remote sensing image classification: A survey," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 9842–9859, 2022.

[11] L. Bruzzone and M. Marconcini, "Domain adaptation problems: A DASVM classification technique and a circular validation strategy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 770–787, May 2010.

[12] B. Sun and K. Saenko, "Deep CORAL: Correlation alignment for deep domain adaptation," in *Proc. Comput. Vis. Workshop*, vol. 9915, 2016, pp. 443–450.

[13] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018. [Online]. Available: https://proceedings.neurips.cc/paper/2018/file/ab88b15733f543179858600245108dd8-Paper.pdf

[14] D. Tuia, C. Persello, and L. Bruzzone, "Recent advances in domain adaptation for the classification of remote sensing data," 2021, *arXiv:2104.07778*.

[15] J. Karhunen, T. Raiko, and K. Cho, "Unsupervised deep learning: A short review," in *Proc. Adv. Independent Compon. Anal. Learn. Mach.*, pp. 125–142, 2015.

[16] C. Yaras, K. Kassaw, B. Huang, K. Bradbury, and J. M. Malof, "Randomized histogram matching: A simple augmentation for unsupervised domain adaptation in overhead imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 1988–1998., 2024.

[17] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.

[18] A. Saboori and H. Ghassemian, "Robust transfer joint matching distributions in semi-supervised domain adaptation for hyperspectral images classification," *Int. J. Remote Sens.*, vol. 41, no. 23, pp. 9283–9307, 2020.

[19] X. Ma, T. Zhang, and C. Xu, "GCAN: Graph convolutional adversarial network for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8266–8276.

[20] Y. Ganin et al., "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2096–2030, 2016.

[21] M. B. Bejiga, F. Melgani, and P. Beraldini, "Domain adversarial neural networks for large-scale land cover classification," *Remote Sens.*, vol. 11, no. 10, 2019, Art. no. 1153.

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

[23] F. Painblanc, L. Chapel, N. Courty, C. Friguet, C. Pelletier, and R. Tavenard, "Match-and-deform: Time series domain adaptation through optimal transport and temporal alignment," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discov. Databases*, 2023, pp. 341–356.

[24] E. Capliez, D. Ienco, R. Gaetano, N. N. Baghdadi, and A. Hadj-Salah, "Temporal-domain adaptation for satellite image time-series land-cover mapping with adversarial learning and spatially aware self-training," *IEEE J. Sel. Topics Appl. Earth Observ. Remote. Sens.*, vol. 16, pp. 3645–3675, 2023.

[25] D. Tuia and G. Camps-Valls, "Kernel manifold alignment for domain adaptation," *PLoS One*, vol. 11, no. 2, 2016, Art. no. e0148655.

[26] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Proc. Comput. Vis. Pattern Recognit.*, 2012, pp. 2066–2073.

[27] M. Volpi, G. Camps-Valls, and D. Tuia, "Spectral alignment of multitemporal cross-sensor images with automated kernel canonical correlation analysis," *ISPRS J. Photogrammetry Remote Sens.*, vol. 107, pp. 50–63, 2015.

[28] J. Liu, W. Xuan, Y. Gan, Y. Zhan, J. Liu, and B. Du, "An end-to-end supervised domain adaptation framework for cross-domain change detection," *Pattern Recognit.*, vol. 132, 2022, Art. no. 108960.

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

[30] C. Paris, L. Orlandi, and L. Bruzzone, "An interactive strategy for the training set definition based on active self-paced learning implemented on a cloud-computing platform," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2021.

[31] C. Persello and L. Bruzzone, "Active learning for domain adaptation in the supervised classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 11, pp. 4468–4483, Nov. 2012.

[32] Y. Yu and H. Lin, "Semi-supervised domain adaptation with source label adaptation," in *Proc. CVPR*, 2023, pp. 24100–24109.

[33] C. Persello and L. Bruzzone, "Active and semisupervised learning for the classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 11, pp. 6937–6956, Nov. 2014.

[34] G. Jun and J. Ghosh, "Semisupervised learning of hyperspectral data with unknown land-cover classes," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 273–282, Jan. 2012.

[35] B. Lucas, C. Pelletier, D. F. Schmidt, G. I. Webb, and F. Petitjean, "A Bayesian-inspired, deep learning-based, semi-supervised domain adaptation technique for land cover mapping," *Mach. Learn.*, vol. 112, no. 6, pp. 1941–1973, 2023.

[36] D. Zha et al., "Data-centric artificial intelligence: A survey," *CoRR*, vol. abs/2303.10158, pp. 1–39, 2023.

[37] R. Roscher et al., "Data-centric machine learning for geospatial remote sensing data," *CoRR*, vol. abs/2312.05327, pp. 1–48, 2023.

[38] K. Gao, A. Yu, X. You, C. Qiu, B. Liu, and F. Zhang, "Cross-domain multi-prototypes with contradictory structure learning for semi-supervised domain adaptation segmentation of remote sensing images," *Remote. Sens.*, vol. 15, no. 13, 2023, Art. no. 3398.

[39] S. Mohammadi, M. Belgiu, and A. Stein, "Improvement in crop mapping from satellite image time series by effectively supervising deep neural networks," *ISPRS J. Photogrammetry Remote Sens.*, vol. 198, pp. 272–283, 2023.

[40] J. Inglada et al., "Assessment of an operational system for crop type map production using high temporal and spatial resolution satellite optical imagery," *Remote Sens.*, vol. 7, no. 9, pp. 12356–12379, 2015.

[41] A. Jolivot et al., "Harmonized in situ datasets for agricultural land use mapping and monitoring in tropical countries," *Earth Syst. Sci. Data*, vol. 13, no. 12, pp. 5951–5967, 2021.

[42] M. Schneider, T. Schelte, F. Schmitz, and M. Körner, "EuroCrops: The largest harmonized open crop dataset across the European Union," *Sci. Data*, vol. 10, no. 1, Sep. 2023, Art. no. 612.

[43] L. Bruzzone and M. Marconcini, "Domain adaptation problems: A DASVM classification technique and a circular validation strategy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 770–787, May 2009.

[44] J. N. Kundu et al., "Towards inheritable models for open-set domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12376–12385.

[45] H. Chen, Q. Zhang, Z. Huang, H. Wang, and J. Zhao, "Towards domain-specific features disentanglement for domain generalization," *CoRR*, vol. abs/2310.03007, 2023.

[46] P. Khosla et al., in *NeurIPS*, 2020.

[47] C. Pelletier, G. I. Webb, and F. Petitjean, "Temporal convolutional neural network for the classification of satellite image time series," *Remote. Sens.*, vol. 11, no. 5, 2019, Art. no. 523.

[48] H. I. Fawaz et al., "Inceptiontime: Finding alexnet for time series classification," *Data Min. Knowl. Discov.*, vol. 34, no. 6, pp. 1936–1962, 2020.

[49] N. M. Foumani, C. W. Tan, G. I. Webb, and M. Salehi, "Improving position encoding of transformers for multivariate time series classification," *Data Min. Knowl. Discov.*, vol. 38, no. 1, pp. 22–48, 2024.

[50] N. N. Navnath, K. Chandrasekaran, A. Stateczny, V. M. Sundaram, and P. Panneer, "Spatiotemporal assessment of satellite image time series for land cover classification using deep learning techniques: A case study of Reunion island, France," *Remote. Sens.*, vol. 14, no. 20, 2022, Art. no. 5232.

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

[52] S. Jo and S. W. Yoon, "POEM: Polarization of embeddings for domain-invariant representations," in *Proc. AAAI*, B. Williams, Y. Chen, and J. Neville, Eds., 2023, pp. 8150–8158.

[53] N. Karasiak, J.-F. Dejoux, C. Monteil, and D. Sheeren, "Spatial dependence between training and test sets: Another pitfall of classification accuracy assessment in remote sensing," *Mach. Learn.*, pp. 2715–2740, 2021.

[54] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. 7th Int. Conf. Learn. Representations*, New Orleans, LA, USA, May 2019. OpenReview.net, 2019, pp. 1–18.

[55] C. Chen et al., "Why do we need large batchsizes in contrastive learning? A gradient-bias perspective," in *Proc. NeurIPS*, 2022, pp. 33860–33875.

[56] F. Pedregosa et al., "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[57] L. van der Maaten and G. E. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.