remote sensing

MDPI

Article

Few-Shot Learning for Crop Mapping from Satellite Image Time Series

Sina Mohammadi, Mariana Belgiu and Alfred Stein *

Department of Earth Observation Science (EOS), Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, 7522 NH Enschede, The Netherlands; s.mohammadi@utwente.nl (S.M.); m.belgiu@utwente.nl (M.B.)

* Correspondence: a.stein@utwente.nl

Abstract: Recently, deep learning methods have achieved promising crop mapping results. Yet, their classification performance is constrained by the scarcity of labeled samples. Therefore, the development of methods capable of exploiting label-rich environments to classify crops in label-scarce environments using only a few labeled samples per class is required. Few-shot learning (FSL) methods have achieved this goal in computer vision for natural images, but they remain largely unexplored in crop mapping from time series data. In order to address this gap, we adapted eight FSL methods to map infrequent crops cultivated in the selected study areas from France and a large diversity of crops from a complex agricultural area situated in Ghana. The FSL methods are commonly evaluated using class-balanced unlabeled sets from the target domain data (query sets), leading to overestimated classification results. This is unrealistic since these sets can have an arbitrary number of samples per class. In our work, we used the Dirichlet distribution to model the class proportions in few-shot query sets as random variables. We demonstrated that transductive information maximization based on α-divergence (α-TIM) performs better than the competing methods, including dynamic time warping (DTW), which is commonly used to tackle the lack of labeled samples. α-TIM achieved, for example, a macro F1-score of 59.6% in Ghana in a 24-way 20-shot setting (i.e., 20 labeled samples from each of the 24 crop types) and a macro F1-score of 75.9% in a seven-way 20-shot setting in France, outperforming the second best-performing methods by 2.7% and 5.7%, respectively. Moreover, α-TIM outperformed a baseline deep learning model, highlighting the benefits of effectively integrating the query sets into the learning process.

Keywords: crop mapping; few-shot learning; deep learning; transfer learning; time series

check for updates

Citation: Mohammadi, S.; Belgiu, M.; Stein, A. Few-Shot Learning for Crop Mapping from Satellite Image Time Series. *Remote Sens.* **2024**, *16*, 1026. https://doi.org/10.3390/rs16061026

Academic Editor: Jianxi Huang

Received: 28 January 2024
Revised: 10 March 2024
Accepted: 11 March 2024
Published: 14 March 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Accurate information on cultivated crops plays a crucial role in crop growth monitoring, yield estimation, and food security [1,2]. Traditionally, crop information is collected through field surveys. Since these surveys are labor-intensive and costly, they are conducted infrequently and over limited geographic areas [3]. Consequently, there is a pressing need for the advancement of automated and accurate methods for crop mapping.

In recent years, deep learning-based methods, including recurrent neural networks (RNNs), convolutional neural networks (CNNs), fully convolutional networks (FCNs), and self-attention networks (or Transformers) have demonstrated remarkable performance in crop mapping [4–11]. This success can be explained by their ability to learn very complex patterns from underlying time series data [12]. However, deep learning models require a large volume of labeled samples for training [13]. This limitation hinders their application in numerous regions across the globe, particularly sub-Saharan countries, where the availability of labeled crop samples is scarce.

Recently, few-shot learning (FSL) methods have been proposed to address this challenge in natural images [14,15]. These methods are able to generalize knowledge from a

source domain with abundant labeled samples to a target domain where a limited amount of labeled samples is available. In remote sensing applications, FSL methods have been successfully used for aerial scene classification [16], land cover classification [17], and dwelling extraction [18]. Nevertheless, to the best of our knowledge, there is only one study that applied FSL to crop mapping from satellite image time series [19]. However, it employed a single FSL method, namely model agnostic meta-learning (MAML) [15], to address a simplified binary crop-type classification problem with similar classes in the source and target domain. In real-world agricultural scenarios, the complexity is considerably higher with a high diversity of crop types. In our work, for the first time, we adapt eight FSL methods for crop mapping within highly diverse agricultural regions.

The FSL methods typically assume that the unlabeled sets from the target domain data, referred to as query sets, are class-balanced [20]. This leads to overestimated classification results. In reality, however, the query sets can contain different class proportions. In order to enable realistic evaluations of FSL methods, inspired by the work of Veilleux et al. [20], we employ the Dirichlet distribution to model the proportions of the classes in few-shot query sets as random variables. Originally, FSL methods were designed for non-temporal natural images. These methods are based on backbone networks with 2D CNNs used across the spatial dimensions. In order to learn the sequential relationship of image time series data essential for crop mapping, we design a backbone network based on 1D convolution and employ it across the temporal dimension. Finally, to handle the variability in lengths of time series data with the capability of capturing information across multiple levels of granularity, we add a temporal pyramid pooling (TPP) layer to the output of our backbone network. This layer generates fixed-length feature representations.

We designed realistic experimental settings by defining two crop mapping scenarios. In both scenarios, the goal was to utilize the available large labeled datasets from the source domains to transfer knowledge to label-scarce environments called target domains. In these scenarios, the target domains comprise various classes, where some or all do not align with the classes present in the source domains. Scenario 1 focuses on crop mapping in a study area from Ghana, a country with widespread malnutrition and extreme food insecurity [21]. Despite the critical importance of agriculture in this country, the current food systems are poorly understood due to the lack of timely information on cultivated crops [21]. In our research, mapping crop types in the selected study area from this country faces challenges not only because of limited labeled samples but also because of small and irregular fields in complex landscapes, as well as high cloud coverage [21]. Scenario 2 is dedicated to mapping infrequent crop types in several study areas from France. Previous crop mapping studies focused mainly on major crop types [22]. Although infrequent crop types cover only a small portion of the total cropped land, they play a vital role in cropping systems by boosting crop diversity, which contributes to food and nutritional security, particularly in the presence of climate change [22].

The eight implemented FSL methods have been tested for the two above-described scenarios using the FewCrop benchmark that combines three publicly available datasets, namely Panoptic Agricultural Satellite TIme Series (PASTIS) [23], ZueriCrop [24], and Ghana [21].

The main contributions of this study can be summarized as follows:

- In order to address the challenge of crop-type mapping in label-scarce environments, we apply eight FSL methods to crop mapping. To adapt the FSL methods, which were designed for natural images, to crop mapping, we employed a realistic query sampling strategy based on the Dirichlet distribution and 1D CNN with a TPP layer. The TPP layer enables the use of varying-length time series while capturing information at different levels of granularity.
- We reveal that effectively integrating the query sets into the learning process using an unsupervised loss function improves the performance.
- The literature on few-shot crop type mapping is very limited. In order to inspire further research in this direction and encourage the research community to design and

implement new few-shot crop mapping methods, we released the full implementation code and instructions on setting up a benchmark for few-shot crop mapping, which can be found here: https://github.com/Sina-Mohammadi/FewShotCrop (accessed on 9 March 2024). We hope that researchers can benefit from our work in advancing FSL for crop mapping.

2. The Method

2.1. Few-Shot Learning Problem Formulation

In FSL, we have access to a source domain dataset with a large number of labeled samples, referred to as meta-training or the base dataset [13]. FSL aims to obtain a model that is capable of leveraging the knowledge learned from this dataset to recognize target domain classes with a few labeled samples. In FSL, episodic sampling is used to construct a series of randomly sampled few-shot tasks (episodes), each with a few labeled samples. N_S labeled samples are sampled from randomly sampled K different classes to create each of the K-way N_S-shot tasks. This set of labeled samples, denoted as S, is referred to as the support set, with a size of $|\mathcal{S}| = N_S \cdot K$. Moreover, each task contains a query set, denoted by Q, created by randomly sampling N_Q unlabeled samples from each of the K classes, and therefore, its size is $|\mathcal{Q}| = N_Q \cdot K$.

In the few-shot settings, a model is first trained on a large labeled source dataset with base classes (base training). Subsequently, in the meta-testing stage, model generalization is evaluated on the unlabeled query sets after being adapted to the tasks at hand using only the support sets (inductive FSL) or using both the support sets and the query sets (transductive FSL).

According to the base training implementation, FSL methods can be categorized into episodic (or meta-learning) methods and non-episodic methods. In episodic methods, the model is trained on randomly sampled few-shot tasks constructed from the base dataset. In non-episodic methods, the model is trained following the standard supervised training by minimizing cross-entropy on the classes of the base dataset. The episodic methods incorporated in our framework are MAML [15], Prototypical Networks [14], and MetaOptNet [25], and the non-episodic methods are Baseline [26], SimpleShot [27], Entropy-min [28], transductive information maximization (TIM) [13], and α-TIM [20].

The FSL methods are categorized as transductive and inductive methods based on whether the model has access to the query set (Q) or not.

2.2. Transductive Methods

In the transductive setting, in addition to the support set, the model also has access to the query set. While adapting to a task, the model can improve the performance by integrating the query set into the learning process.

We provide a detailed visual illustration of our framework based on the TIM and α-TIM methods in Figure 1. For the remaining methods, we provide descriptions and refer the readers to our code implementation and the original papers for additional details. Below, we explain the four transductive FSL methods incorporated in our framework.

TIM: For a few-shot task at the meta-testing stage, TIM [13] aims to improve the performance on the query set by maximizing the mutual information between the query features and their label predictions. Following Boudiaf et al. [13], in the base training stage, we trained an embedding function, i.e., feature extractor, denoted as f_θ, and a classifier, denoted as g_s, on the source domain (base dataset) classes using supervised loss and non-episodic training (Figure 1).

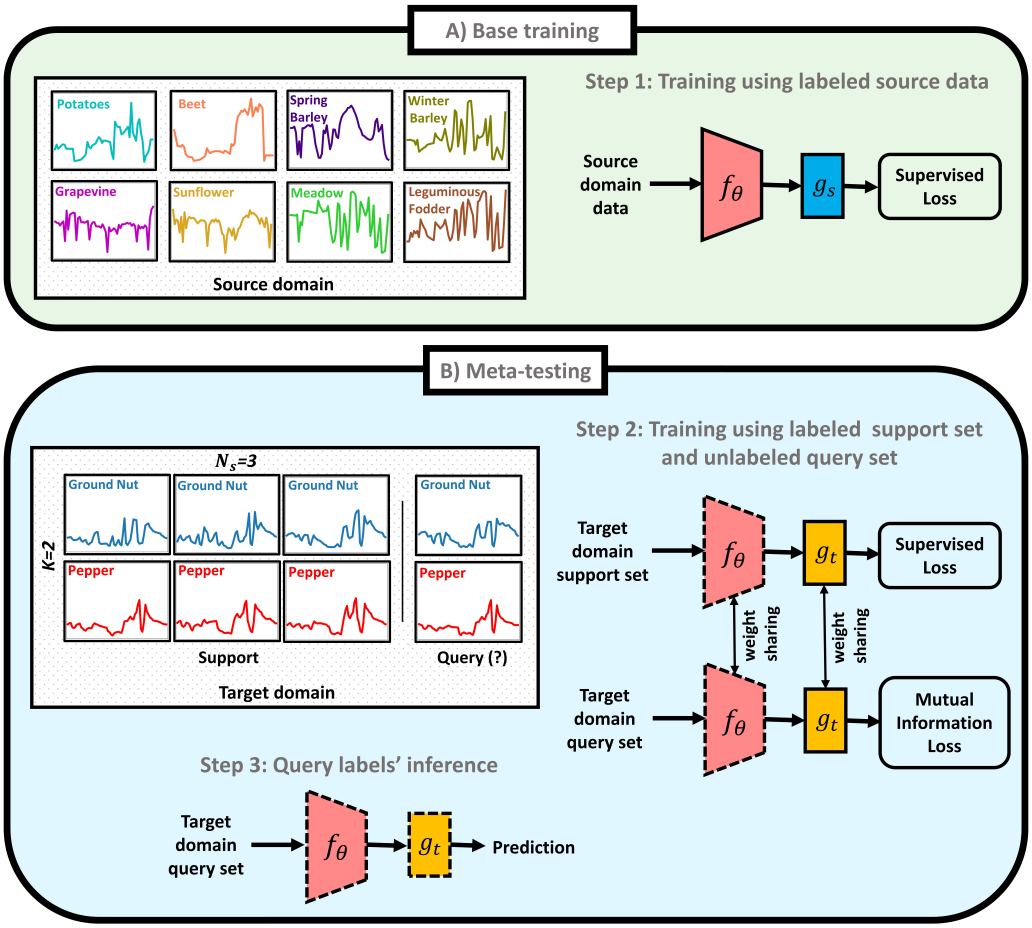


Figure 1. The framework is composed of two stages: (**A**) Base training and (**B**) meta-testing. This figure shows the framework based on the TIM and α-TIM methods. Note that the solid lines indicate that the network is being trained, whereas the dashed lines indicate that the network was already trained during a previous step and is fixed in the current step.

The feature extractor, which maps the input data to a d-dimensional feature space, is composed of a backbone network and a temporal pyramid pooling (TPP) layer, as shown in Figure 2. The backbone network learns to generate increasingly abstract and discriminative features from the input data, which are essential for understanding and discriminating between different patterns in the time series. In order to do so, our backbone network uses four 1D convolutional layers, each followed by a batch normalization layer and a relu activation. In order to deal with varying-length time series, we propose employing a TPP layer, drawing inspiration from pyramid pooling in computer vision and word spotting [29,30]. We adopt this layer on top of the feature maps generated by the backbone network to learn fixed-dimensional feature representations. As shown in Figure 2, the backbone network generates features with 128 channels (corresponding to 128 convolution filters in the last layer) and length L. The value of L depends on the length of the input time series. The key idea of TPP is to divide the feature map obtained from the backbone network into multiple pyramid levels with different numbers of bins. Our TPP layer has three pyramid levels with 1, 4, and 16 bins, following the number of bins used in He et al. [29]. At level 1 with one bin, the entire feature map is treated as a single segment (1 bin). By performing global average pooling over this segment, a feature representation with a size of 1×128 is obtained. At level 2, the feature map is divided into four bins. Each bin undergoes pooling, resulting in a feature representation with a size of 4×128. Similarly, at level 3, after performing pooling over each of the 16 different bins, a feature representation with a size of 16×128 is obtained. Subsequently, the resulting fixed-size feature representations from all levels are concatenated together to form the final fixed-

length feature representation with a size of 21 × 128 (=2688), which encodes information from different temporal scales. TPP enables dealing with varying-length time series used as input to our framework while capturing information at different levels of granularity.

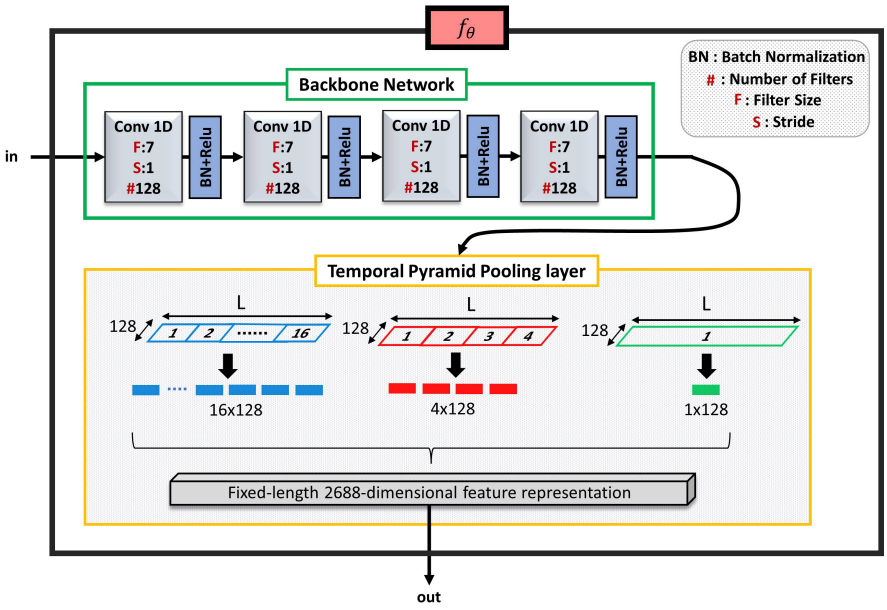


Figure 2. The architecture of our feature extractor is composed of a backbone network and a temporal pyramid pooling layer. Note that we use this architecture as the feature extractor for all the employed FSL methods.

After completing the base training, at every meta-testing task, f_θ, is frozen, and a new classifier, g_t, is trained on the support set using supervised loss and on the unlabeled query set using mutual information loss, which is an unsupervised loss function. Formally, a soft classifier with a weight matrix $W := [w_1, \ldots, w_K] \in \mathbb{R}^{K \times d}$ is defined, for which the posterior distribution (over labels given features) and marginal distribution (over query labels) are denoted as

$$p_{ik} = \mathbb{P}(Y = k \mid X = x_i; W, \theta) \text{ and } \widehat{p}_k = \mathbb{P}(Y_{\mathcal{Q}} = k; W, \theta) \tag{1}$$

where X and $Y \in \{1, \ldots, K\}$ are the random variables associated with the raw features (within $\mathcal{S} \cup \mathcal{Q}$) and labels, respectively, and p_{ik} and $\widehat{p}_k$ are given by

$$p_{ik} \propto \exp\Big(-\frac{\tau}{2}\|w_k - z_i\|^2\Big), \quad \text{and} \quad \widehat{p}_k = \frac{1}{|\mathcal{Q}|}\sum_{i \in \mathcal{Q}} p_{ik} \tag{2}$$

Here, τ is a temperature parameter, and $z_i = \frac{f_\theta(x_i)}{\|f_\theta(x_i)\|_2}$ shows the L2-normalized embedded features. The weighted mutual information between the query features and their label predictions is defined as

$$\widehat{\mathcal{I}}(X_{\mathcal{Q}}; Y_{\mathcal{Q}}) := \widehat{\mathcal{H}}(Y_{\mathcal{Q}}) - \gamma \widehat{\mathcal{H}}(Y_{\mathcal{Q}} \mid X_{\mathcal{Q}}) \tag{3}$$

where $\gamma \geq 0$. Note that for $\gamma = 1$, $\widehat{\mathcal{I}}(X_{\mathcal{Q}}; Y_{\mathcal{Q}})$ is equivalent to the standard mutual information. $\widehat{\mathcal{H}}(Y_{\mathcal{Q}})$ and $\widehat{\mathcal{H}}(Y_{\mathcal{Q}} \mid X_{\mathcal{Q}})$ denote the label-marginal entropy and conditional entropy of labels given the raw query features, respectively. They are obtained as

$$\widehat{\mathcal{H}}(Y_{\mathcal{Q}}) = -\sum_{k=1}^{K} \widehat{p}_k \log \widehat{p}_k \tag{4}$$

$$\widehat{\mathcal{H}}(Y_{\mathcal{Q}} \mid X_{\mathcal{Q}}) = -\frac{1}{|\mathcal{Q}|} \sum_{i \in \mathcal{Q}} \sum_{k=1}^{K} p_{ik} \log(p_{ik}) \tag{5}$$

At the meta-testing stage, the overall loss function is composed of (i) supervised loss, denoted as $\mathcal{L}_{\text{Supervised}}$ and defined over the support samples, and (ii) mutual information loss, denoted as $\mathcal{L}_{\text{MI}}$ and defined over the query samples. The overall loss, dubbed transductive information maximization (TIM) loss, is formulated as

$$\min_{\mathrm{W}} \lambda \cdot \mathcal{L}_{\text{Supervised}} + \mathcal{L}_{\text{MI}} \tag{6}$$

where $\mathcal{L}_{\text{MI}} = -\widehat{\mathcal{I}}(X_{\mathcal{Q}}; Y_{\mathcal{Q}})$, and $\mathcal{L}_{\text{Supervised}}$ is the standard cross-entropy loss, obtained as

$$\mathcal{L}_{\text{Supervised}} = -\frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \sum_{k=1}^{K} y_{ik} \log(p_{ik}) \tag{7}$$

where y_{ik} is the k-th element of the one-hot encoded label of the i-th support sample.

In the framework based on TIM, in the meta-testing stage, the mutual information between the query features and their label predictions is maximized, while the cross-entropy loss on the support set is minimized. TIM assumes that the query sets are class-balanced, which is unrealistic in most crop mapping applications.

α-TIM: Veilleux et al. [20] proposed a generalization of mutual information loss based on α-divergences to deal with class-imbalanced query sets in few-shot tasks at the meta-testing stage. They proposed an effective extension of Shannon mutual information by replacing the Shannon entropy used in Equations (4) and (5) with the α-entropy:

$$\widehat{\mathcal{I}}_{\alpha}(X_{\mathcal{Q}}; Y_{\mathcal{Q}}) = \widehat{\mathcal{H}}_{\alpha}(Y_{\mathcal{Q}}) - \widehat{\mathcal{H}}_{\alpha}(Y_{\mathcal{Q}} \mid X_{\mathcal{Q}}) = \frac{1}{\alpha - 1} \left(\frac{1}{|\mathcal{Q}|} \sum_{i \in \mathcal{Q}} \sum_{k=1}^{K} p_{ik}^{\alpha} - \sum_{k=1}^{K} \widehat{p}_{k}^{\alpha} \right) \tag{8}$$

Therefore, α-TIM loss is formulated as

$$\min_{\mathrm{W}} \lambda \cdot \mathcal{L}_{\text{Supervised}} + \mathcal{L}_{\alpha\text{-MI}} \tag{9}$$

where $\mathcal{L}_{\alpha\text{-MI}} = -\widehat{\mathcal{I}}_{\alpha}(X_{\mathcal{Q}}; Y_{\mathcal{Q}})$, and $\mathcal{L}_{\text{Supervised}}$ is the standard cross-entropy loss, which is calculated as per Equation (7).

Entropy-min: Dhillon et al. [28] proposed to minimize the entropy of the softmax predictions of the query set in each meta-testing few-shot task. In other words, if $\widehat{\mathcal{H}}(Y_{\mathcal{Q}})$ (label-marginal entropy) is removed from Equation (3), the loss function in Equation (6) turns into the loss function employed by Entropy-min. Therefore, the framework based on Entropy-min is similar to Figure 1, but with a different unsupervised loss function.

MAML: Introduced by Finn et al. [15], MAML aims at learning a global initialization of parameters such that a small number of gradient steps with a few support samples from a new task will produce large improvements on that task. MAML uses episodic training and comprises a meta-learner and several base learners. The base learners focus on acquiring quick knowledge for a single task, whereas the meta-learner aims to gradually accumulate general knowledge that can be shared across different tasks. MAML uses transductive batch normalization to take advantage of the statistics of the query set of a given few-shot task.

2.3. Inductive Methods

In the inductive setting, it is assumed that the model has access to the support set, whereas no information from the query set of the target dataset is provided. Below, we explain the four inductive FSL methods incorporated in our framework.

Prototypical Networks: Snell et al. [14] introduced a meta-learning method that uses episodic training. In order to incorporate this method into our framework, we employ the feature extractor to obtain the embedded representation of the input data. During the base training stage, the weights of the feature extractor are optimized across tasks such that the distance between the mean vector of embedded support samples of each class (class prototypes) and the query samples of the corresponding class is minimized. Subsequently, at every meta-testing task, the classification is performed based on the distance between the query sample and the class prototypes.

Baseline: Following a baseline deep learning method employed by Chen et al. [26], we train the feature extractor f_θ and a classifier g_s on the base dataset classes using non-episodic training. Subsequently, in the meta-testing stage, we freeze f_θ and train a new classifier, g_t, on the support set of the sampled task. In other words, the framework based on Baseline is similar to TIM and α-TIM (Figure 1), with the difference that it does not employ the mutual information loss on the query set in the meta-testing stage.

SimpleShot: SimpleShot [27] uses the same base training procedure as the Baseline, Entropy-min, TIM, and α-TIM methods but a different meta-testing stage. At this stage, centering and L2 normalization on the embedded representation of the input data are applied, and nearest neighbor classification is performed by using Euclidean distance as the distance measure.

MetaOptNet: Introduced by Lee et al. [25], MetaOptnet is a meta-learning method that investigated linear support vector machine (SVM) as the base learner for MAML.

We evaluate the aforementioned FSL methods using not only the balanced query sampling strategy but also the realistic query sampling strategy by means of the Dirichlet distribution, which was inspired by Veilleux et al. [20]. In standard few-shot settings, it is typically assumed that the proportion of the query samples assigned to a particular class k in a few-shot task is fixed and known a priori: $p_k = 1/K$, for all k and all few-shot tasks [20]. In order to relax this impractical assumption, we employ a Dirichlet distribution representing the proportions of the classes within few-shot query sets as random variables [20]. We denote the Dirichlet distribution parameter, often referred to as the alpha parameter (or concentration parameter), as $\boldsymbol{a}$. Assuming that $\boldsymbol{a} = a \cdot \mathbb{1}_K$, where $\mathbb{1}_K$ is the K-dimensional vector with all components equal to 1, a determines the concentration of the Dirichlet distribution. A common way to depict the Dirichlet distribution is by focusing on the scenario where $K = 3$, as it allows for visualization in two dimensions. In Figure 3, we present the Dirichlet density for $K = 3$, with two-simplex support, represented with an equilateral triangle. This triangle's vertices correspond to the probability vectors (1, 0, 0), (0, 1, 0), and (0, 0, 1). Various concentration parameters are showcased, including 0.5, 2, 5, and 15. Higher a values correspond to more concentrated distributions, with the category proportions being more evenly distributed, i.e., all categories are likely to have similar probabilities. Notably, when the concentration parameter approaches infinity, only uniform distribution, i.e., the point in the middle of the simplex, could occur as the marginal distribution of the classes. Therefore, this case reflects the balanced sampling setting with perfectly balanced tasks. Lower a values, especially $a < 1$, make the distribution sparser, with a few categories more likely to dominate with higher probabilities.

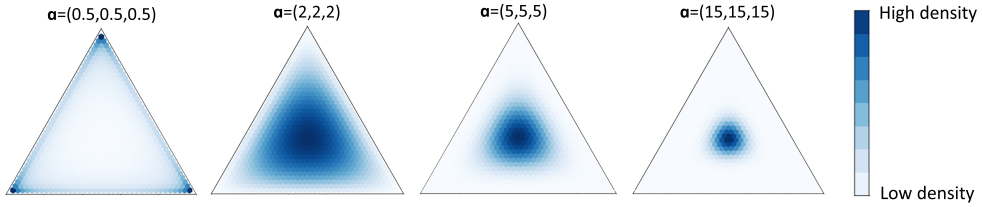


Figure 3. Dirichlet density function for $K = 3$, with different choices of parameter vector $\boldsymbol{a}$.

3. The Study Areas and the FewCrop Benchmark

3.1. The Study Areas

In this study, we use three datasets, namely the PASTIS [23], the ZueriCrop [24], and the Ghana datasets [21]. These datasets were collected from France, Switzerland, and Ghana, respectively (See Figure 4).

Covering more than 4000 km^2, the PASTIS dataset was collected from four different regions of France with varying climates and crop distributions. This dataset is composed of 2433 patches obtained from the Sentinel-2 satellite, with a spatial size of 128 pixels by 128 pixels. Acquired between September 2018 and November 2019, each patch contains between 38 and 61 time steps. This dataset is made up of five folds for the sake of five-fold cross-validation. It includes 124,422 parcels, containing 18 crop types and a background class that corresponds to non-agricultural land uses.

Covering over a 50 by 48 km area, the ZueriCrop dataset [24] was acquired using the Sentinel-2 satellite between January 2019 and December 2019 in the Swiss Cantons of Zurich and Thurgau. The dataset is organized in five folds to facilitate five-fold cross-validation. All patches in the dataset have 71 time steps, and the spatial size 24 pixels by 24 pixels. The dataset encompasses 116,000 field instances with 48 crop types.

The Ghana dataset [21] consists of sparse ground truth labels of crop fields in northern Ghana, covering 8937 fields with 24 crop types. The dataset includes time series of satellite imagery from Sentinel-1, Sentinel-2, and PlanetScope satellites. In this study, we use Sentinel-2 images, which were collected throughout the entirety of 2016 and include 4040 patches, with spatial size 64 pixels by 64 pixels, and time steps that range in number from 11 to 67.

As suggested by the authors, the cloudy patches were not excluded from these datasets since an adequate method should be robust to cloudy acquisitions. In fact, deep learning algorithms have been experimentally proven to be robust to cloud cover [10].

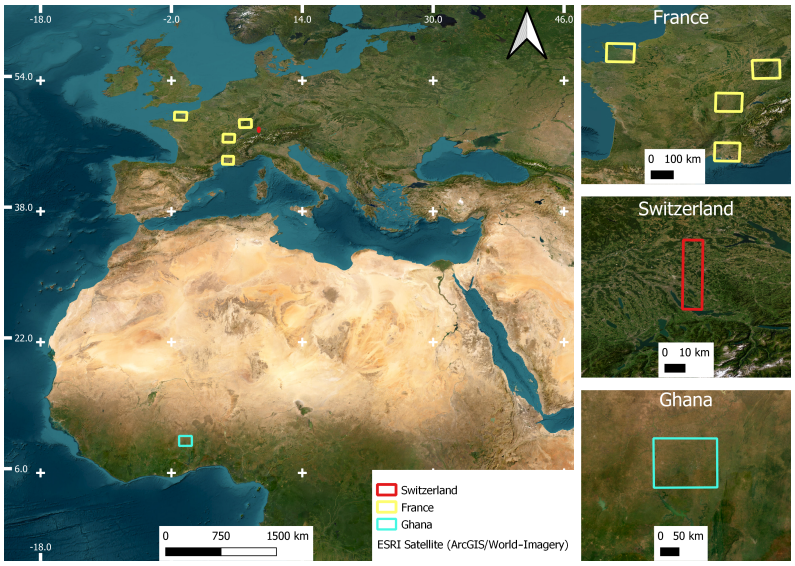


Figure 4. Location of the study areas in the PASTIS, ZueriCrop, and Ghana datasets, collected from France, Switzerland, and Ghana, respectively.

3.2. The FewCrop Benchmark

In this section, we introduce the benchmark prepared for testing the few-shot crop mapping methods, namely FewCrop. This benchmark is made up of the aforementioned publicly available datasets. We defined two real-world scenarios in our benchmark. For Scenario 1, we employed the first fold of the PASTIS dataset as our label-rich environment and aimed to map the crop types of the Ghana dataset. In this scenario, we used the third fold of the ZueriCrop dataset as our validation dataset for the early stopping of the training

process. For Scenario 2, we used the seven major classes of the first fold of the PASTIS dataset as our label-rich environment and aimed to map the seven infrequent classes of the same fold. We used the remaining five classes of this fold as the validation set for early stopping. The class distributions of the three datasets used in our study are shown in Figure 5. The selected major and infrequent classes are shown in the upper left section of this figure.

In Scenario 1, the label-rich and label-scarce environments come from different geographic areas with different climatic conditions and agricultural practices. The larger variation among the environments in Scenario 1 compared to Scenario 2 can hamper the exploitation of the knowledge learned from the label-rich environment in the label-scarce environment due to the smaller similarity between the two environments. In our study, we will investigate to what extent the knowledge learned from the label-rich environment can help classify the crops in the label-scarce environment.

The number of observations and their dates in the image time series in the PASTIS and Ghana datasets are different across different patches within each dataset. In order to handle this issue, we preprocessed the data using linear interpolation, the effectiveness of which was demonstrated by Xu et al. [4]. By doing so, we filled in the gaps caused by missing acquisitions and obtained 61 time steps for all the patches of the first fold of the PASTIS dataset and 67 time steps for all the patches of the Ghana dataset. The ZueriCrop dataset, as provided by its creators, already contains a consistent number of time steps for all the patches, which is 71 time steps.

As mentioned above, the PASTIS, ZueriCrop, and Ghana datasets consist of patches sized at 128 pixels by 128 pixels, 32 pixels by 32 pixels, and 64 pixels by 64 pixels, respectively. These patches are accompanied by labels of the same dimensions, as each patch is associated with its corresponding pixel-level label. We reshaped all the patches and their labels in the three datasets and converted them into pixel-based datasets since our framework works with pixels rather than patches. The number of samples (pixels) of the source, validation, and target datasets in both scenarios are reported in Table 1. All pixels have a 10 m resolution and contain nine Sentinel-2 bands, namely Blue, Green, Red, Red Edge 1, Red Edge 2, Red Edge 3, Near Infra-Red (NIR), Short Wave Infra-Red (SWIR) 1, and SWIR 2.

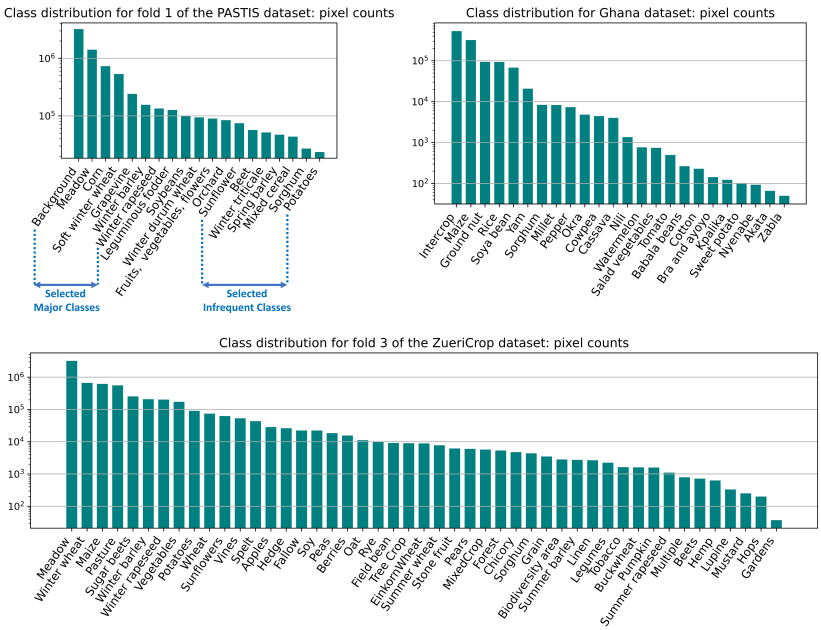


Figure 5. Class distributions of the three datasets used in our study (at log scale). Note that in the Ghana dataset, Kpalika, Nyenabe, Zabla, Bra, and Ayoyo are local names corresponding to the scientific names Jatropha curcas, Mucuna pruriens, Ziziphus mauritiana, Hibiscus sabdariffa, and Corchorus olitorius, respectively.

Table 1. The number of samples (pixels) of the source, validation, and target datasets in the two scenarios.

	Scenario 1	**Scenario 2**
Source dataset	7,254,310	6,434,782
Validation dataset	6,465,945	494,954
Target dataset	1,164,896	324,574

4. Experiments

Our framework is implemented in PyTorch 1.12.1 using a machine with NVIDIA RTX A4000 GPU. The models were trained using the Adam optimizer [31] with an initial learning rate of 1×10^{-4} and cosine decay. In this study, we did not fine-tune the hyperparameters to simplify the experiments. Regarding the concentration parameter a, it is essential to achieve a good trade-off between two cases: a very high a value corresponds to balanced sampling and a very low a value corresponds to extremely imbalanced sampling (i.e., only one class is present in the task). Since both cases are unlikely to happen in realistic scenarios, the parameter a should be carefully selected to mimic realistic class proportions. In line with this consideration, Veilleux et al. [20] recommended setting it to 2 to mimic realistic class proportions.

Dynamic time warping (DTW) [32] is used for time series classification by aligning two time series to find the best match. DTW computes a distance matrix, finds the optimal alignment path, and classifies new time series based on their similarity to known examples, making it effective for handling time series data that may have temporal distortions [32]. DTW has been proven to be a powerful method to deal with the scarcity of labeled samples as well as distorted temporal profiles caused by the presence of clouds [33]. We compared the results of the FSL methods with DTW to demonstrate their effectiveness when tackling these challenges. Since the datasets contain nine spectral bands, we employed a multi-dimensional implementation of DTW [34], leveraging GPU-accelerated code provided by Maghoumi et al. [35] and Maghoumi et al. [36].

Base training for the episodic methods: We performed base training for this group of methods using episodic training with 100,000 randomly sampled tasks. In order to perform early stopping during base training, we evaluated the methods after every 1000 iterations using 10,000 randomly sampled episodes from our validation data.

Base training for the non-episodic methods: For these methods, we trained a single model for 100 epochs using standard supervised training via cross-entropy loss with a batch size of 512. During base training, we evaluated the model after each epoch using 10,000 randomly sampled episodes from our validation data to perform early stopping.

Meta-testing: We performed meta-testing on 30,000 randomly sampled tasks and reported the macro F1-score averaged over these tasks. The query set in each task contains 75, 105, and 360 samples in five-way, seven-way, and 24-way settings. Note that in the balanced sampling setting, the query set contains 15 samples per class, whereas it has an arbitrary number of samples per class in the realistic sampling setting implemented by the Dirichlet distribution.

4.1. Performance Comparison Using the Standard Five-Way Setting

In this section, we conduct experiments for both introduced scenarios using the most common setting in few-shot classification, i.e., the five-way setting. We compare the results of eight FSL methods incorporated in our framework with each other and with multi-dimensional DTW in terms of the macro F1-score (Table 2). We evaluate these methods under the settings of the realistic query set sampling and the balanced query set sampling.

In order to facilitate an easy comparison, we calculate the mean of the scores obtained in all five-way settings for each method. Table 2 shows that under the realistic sampling strategy, our framework based on α-TIM outperforms the other methods in Scenario 1 in all five-way settings (1-shot, 5-shot, 10-shot, and 20-shot). Specifically, it achieves a mean score

of 62.0%, which is superior to the next best-performing methods, DTW (58.1%) and Baseline (56.5%). In Scenario 2, α-TIM outperforms the other methods in 10-shot and 20-shot settings. In the 1-shot setting, however, the best-performing methods are SimpleShot and TIM, and in the 5-shot setting, those are Baseline and SimpleShot. Nevertheless, α-TIM achieves a mean score of 67.7%, outperforming the other methods, including SimpleShot (66.6%), Baseline (66.0%), Entropy-min (65.6%), and TIM (59.1%), which are the next best-performing methods. Therefore, α-TIM showed superior overall performance in both scenarios.

In realistic sampling, the sensitivity of the F1 measure to the imbalanced query sets results in a decrease in the performance of most methods. The numbers in parentheses in Table 2 show that if the experiments are conducted under balanced sampling, the scores increase in almost all cases. Additionally, it is anticipated that the transductive methods will exhibit an inherent sensitivity to the proportion of the query sets as they use these sets in the learning process, which distinguishes them from the inductive methods. This issue is particularly pronounced in the TIM method, where the performance difference between the two sampling settings is larger.

Table 2. Comparison of the FSL methods and DTW in terms of macro F1-score averaged over 30,000 tasks using the five-way setting. All values are reported in percentages. The numbers in black show the performance when the query sets are sampled following the realistic sampling setting using a Dirichlet distribution with $\boldsymbol{a} = 2 \cdot \mathbb{1}_K$. The numbers in parentheses show the performance change obtained when the query sets are sampled following the balanced sampling setting. A blue arrow (↑) indicates a performance improvement, whereas a red arrow (↓) indicates a performance drop.

		Scenario 1: Ghana (Five-way)			
	Method	**1-shot**	**5-shot**	**10-shot**	**20-shot**
Induct.	Protonet	41.5 (↑ 2.7)	56.4 (↑ 3.2)	61.6 (↑ 3.2)	63.8 (↑ 2.9)
	MetaOptNet	39.9 (↑ 2.7)	54.4 (↑ 3.5)	61.9 (↑ 3.4)	67.7 (↑ 3.3)
	Baseline	42.5 (↑ 2.7)	57.3 (↑ 2.8)	61.4 (↑ 2.8)	64.7 (↑ 2.9)
	SimpleShot	42.0 (↑ 2.9)	56.8 (↑ 2.9)	60.5 (↑ 2.9)	63.0 (↑ 2.9)
Transduct.	MAML	38.9 (↑ 3.2)	57.8 (↑ 3.9)	63.9 (↑ 3.8)	66.9 (↑ 3.7)
	Entropy-min	38.8 (↑ 1.2)	57.5 (↑ 0.3)	61.0 (↑ 1.2)	62.8 (↓ 0.1)
	TIM	42.9 (↑ 10.5)	52.7 (↑ 13.8)	56.5 (↑ 15.3)	59.3 (↑ 16.7)
	α-TIM	**43.4** (↑ 3.7)	**60.4** (↑ 1.9)	**68.8** (↑ 2.1)	**75.2** (↑ 2.3)
	DTW	40.3 (↑ 3.0)	56.9 (↑ 3.3)	64.2 (↑ 3.3)	70.9 (↑ 3.1)
		Scenario 2: Infrequent Crop Types of France (Five-way)			
	Method	**1-shot**	**5-shot**	**10-shot**	**20-shot**
Induct.	Protonet	47.0 (↑ 3.0)	63.5 (↑ 3.9)	69.8 (↑ 4.0)	73.1 (↑ 4.0)
	MetaOptNet	43.5 (↑ 2.9)	55.3 (↑ 3.9)	66.3 (↑ 4.0)	71.1 (↑ 4.0)
	Baseline	46.2 (↑ 3.1)	**68.4** (↑ 3.8)	73.3 (↑ 3.9)	76.1 (↑ 3.7)
	SimpleShot	**48.8** (↑ 3.4)	**68.4** (↑ 3.8)	73.2 (↑ 3.9)	75.9 (↑ 3.7)
Transduct.	MAML	39.5 (↑ 3.4)	60.0 (↑ 5.0)	67.2 (↑ 5.1)	71.2 (↑ 5.2)
	Entropy-min	46.3 (↑ 4.1)	67.6 (↑ 6.8)	72.9 (↑ 5.5)	75.6 (↑ 4.7)
	TIM	**48.8** (↑ 12.0)	60.0 (↑ 16.7)	62.8 (↑ 17.6)	64.7 (↑ 18.3)
	α-TIM	48.1 (↑ 3.9)	66.5 (↑ 1.0)	**75.5** (↑ 2.1)	**80.8** (↑ 2.3)
	DTW	26.2 (↑ 2.2)	39.3 (↑ 3.4)	46.3 (↑ 3.8)	54.7 (↑ 4.3)

4.2. Comparison of the Classification Performance Obtained Using the Number of Classes in the Target Datasets as the Number of Ways

FSL methods are usually evaluated in a five-way setting. Nonetheless, this artificial setting cannot fully provide a realistic assessment of FSL methods. This is because, in a realistic scenario, such as crop mapping, there is an arbitrary and possibly larger number of classes in the target dataset. In order to mitigate this issue, we also evaluated the performance of the methods in more challenging 24-way and seven-way settings for the first and second scenarios, respectively, in terms of the macro F1-score (Table 3).

Table 3. Comparison of the FSL methods and DTW in terms of the macro F1-score averaged over 30,000 tasks using the number of classes in the target datasets as the number of ways. All values are reported in percentages. The numbers in black show the performance when the query sets are sampled following the realistic sampling setting using a Dirichlet distribution with $\boldsymbol{a} = 2 \cdot \mathbb{1}_K$. The numbers in parentheses show the performance change obtained when the query sets are sampled following the balanced sampling setting. A blue arrow (↑) indicates a performance improvement, whereas a red arrow (↓) indicates a performance drop. '-' indicates that the result was computationally intractable to obtain.

		Scenario 1: Ghana (24-way)			
	Method	**1-shot**	**5-shot**	**10-shot**	**20-shot**
Induct.	Protonet	29.0 (↑ 2.0)	38.9 (↑ 2.4)	42.5 (↑ 2.3)	43.6 (↑ 2.3)
	MetaOptNet	26.3 (↑ 2.1)	35.8 (↑ 3.2)	-	-
	Baseline	**29.4** (↑ 2.2)	39.6 (↑ 2.2)	42.1 (↑ 2.2)	44.5 (↑ 2.4)
	SimpleShot	28.9 (↑ 2.2)	39.1 (↑ 2.2)	41.2 (↑ 2.2)	42.6 (↑ 2.2)
Transduct.	MAML	-	-	-	-
	Entropy-min	27.4 (↑ 1.7)	37.9 (↑ 1.2)	39.1 (↑ 1.2)	39.6 (↑ 1.3)
	TIM	28.0 (↑ 7.1)	38.1 (↑ 10.0)	41.5 (↑ 11.3)	44.0 (↑ 12.5)
	α-TIM	28.8 (↑ 1.0)	**44.8** (↑ 2.2)	**52.6** (↑ 2.5)	**59.6** (↑ 2.7)
	DTW	27.6 (↑ 2.3)	42.7 (↑ 2.9)	49.9 (↑ 3.1)	56.9 (↑ 3.1)
		Scenario 2: Infrequent Crop Types of France (Seven-way)			
	Method	**1-shot**	**5-shot**	**10-shot**	**20-shot**
Induct.	Protonet	39.3 (↑ 3.0)	56.0 (↑ 4.0)	63.2 (↑ 4.2)	67.0 (↑ 4.3)
	MetaOptNet	35.9 (↑ 2.6)	49.0 (↑ 3.9)	61.1 (↑ 4.0)	66.3 (↑ 4.1)
	Baseline	38.8 (↑ 2.7)	**61.4** (↑ 4.1)	67.0 (↑ 4.2)	70.2 (↑ 4.1)
	SimpleShot	41.0 (↑ 3.1)	**61.4** (↑ 4.2)	66.9 (↑ 4.2)	69.8 (↑ 4.2)
Transduct.	MAML	-	-	-	-
	Entropy-min	36.9 (↑ 4.3)	60.4 (↑ 7.2)	66.1 (↑ 6.2)	68.9 (↓ 0.1)
	TIM	**41.4** (↑ 9.5)	54.4 (↑ 15.2)	57.9 (↑ 16.3)	60.3 (↑ 17.3)
	α-TIM	37.4 (↑ 2.5)	58.4 (↑ 1.4)	**69.4** (↑ 2.2)	**75.9** (↑ 2.6)
	DTW	21.0 (↑ 1.8)	33.0 (↑ 3.0)	40.1 (↑ 3.6)	48.9 (↑ 4.3)

In Scenario 1, α-TIM achieved a mean score of 46.5%, outperforming the second and the third best-performing methods, namely DTW (44.3%) and Baseline (38.9%). In Scenario 2, α-TIM (60.3%) outperformed SimpleShot (59.8%) and Baseline (59.4%), being the second and the third best-performing methods, respectively.

Note that the episodic methods (Protonet, MetaOptNet, and MAML) in Table 3 were trained using the standard five-way setting because training them with larger numbers of ways in the base training is computationally intractable due to memory constraints, which was also discussed by Chen et al. [26]. Obtaining results for MetaOptNet in the 10-shot and 20-shot settings was also not possible due to memory constraints. Furthermore, no results could be obtained for MAML since it learns the initialization for the classifier and, thus, can only be applied to the same number of classes in the target dataset.

In order to report detailed prediction results in both scenarios, we provide confusion matrices of the best-performing method, namely α-TIM, in the 20-shot 24-way setting for Scenario 1 and in the 20-shot seven-way setting for Scenario 2 (Figures 6 and 7). As seen from these confusion matrices, the method is challenged when classifying some crop types, such as groundnut, maize, soyabean, yam, and intercrop, in Scenario 1, with a recall score below 30%. In scenario 2, the method correctly classifies a large portion of pixels, obtaining a recall score higher than 80% for spring barley, sunflower, beet, winter triticale, and sorghum and obtaining a recall score of 75.4% for potatoes and 66.6% for mixed cereal.

The qualitative comparisons of the non-episodic FSL methods and DTW for three patches containing a subset of crops from each scenario are presented in Figure 8. This figure shows that α-TIM yields fewer misclassified pixels in both scenarios compared to the competing methods. Note that we only used the results obtained by the non-episodic methods for visualization due to the aforementioned constraints of the episodic methods.

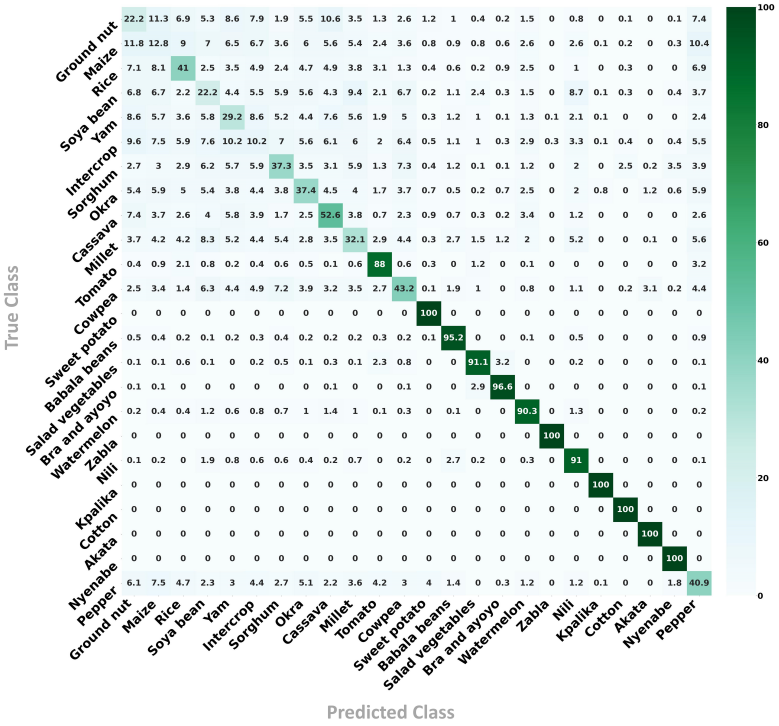


Figure 6. Confusion matrix of α-TIM in the 20-shot 24-way setting for Scenario 1, following the realistic sampling setting using a Dirichlet distribution with $\boldsymbol{a} = 2 \cdot \mathbb{1}_K$. Each row is normalized such that the sum of its numbers is equal to 100. Thus, the percentage in diagonal boxes is the recall (producer's accuracy) of each class.

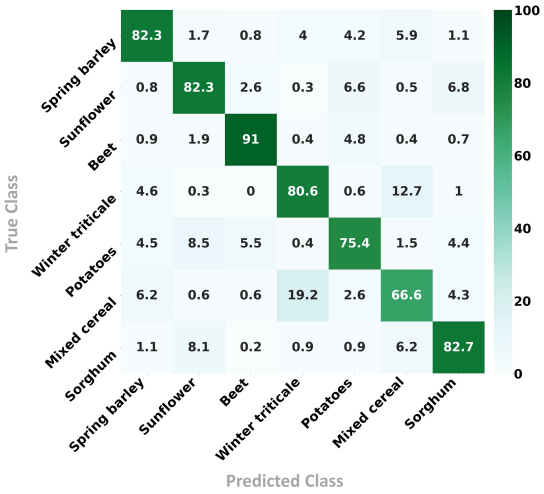


Figure 7. Confusion matrix of α-TIM in the 20-shot seven-way setting for Scenario 2, following the realistic sampling setting using a Dirichlet distribution with $\boldsymbol{a} = 2 \cdot \mathbb{1}_K$. Each row is normalized such that the sum of its numbers is equal to 100. Thus, the percentage in diagonal boxes is the recall (producer's accuracy) of each class.

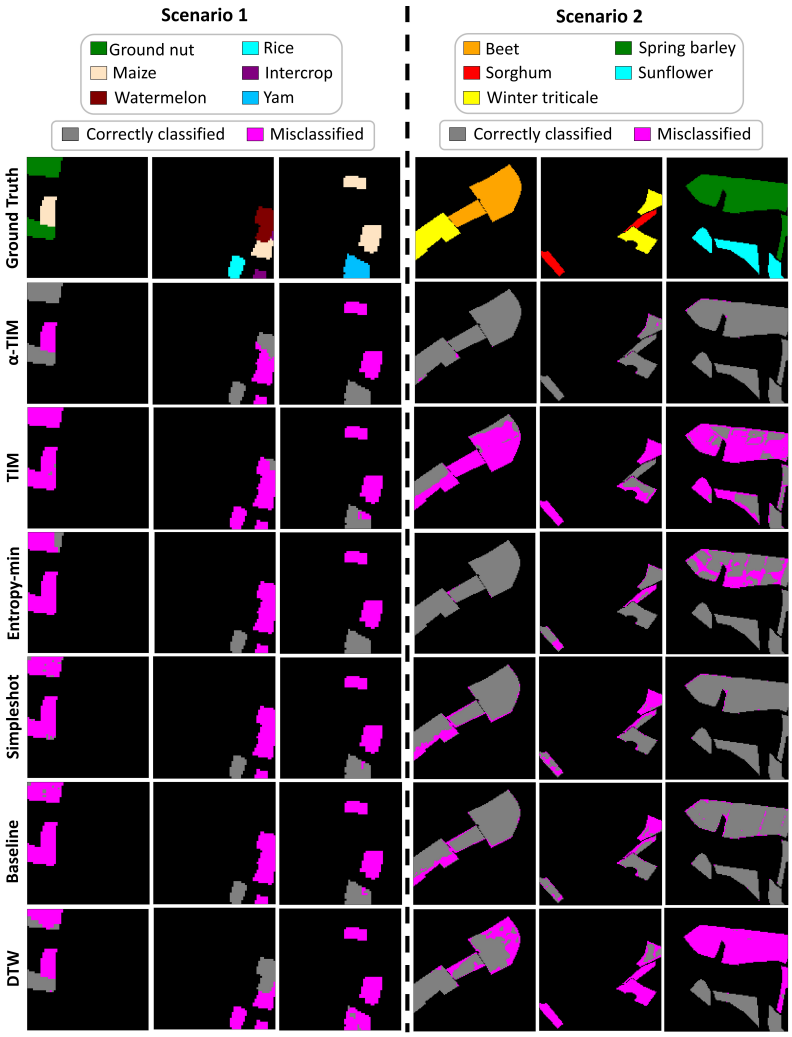


Figure 8. Visual results for three different samples from each scenario. The results were obtained by averaging the predictions of the methods trained on 5,000 randomly selected support sets using the 20-shot 24-way and 20-shot seven-way settings at the meta-testing stage for Scenario 1 and Scenario 2, respectively.

4.3. Results Obtained When Removing Base Training on the Source Domain

Table 4 reports the results obtained when removing base training on the source domain in terms of the macro F1-score. This is done by performing a meta-testing stage on the models with randomly initialized weights rather than the models pre-trained on the source domain. The performance drops strongly in most cases when removing base training on the source data (note the numbers with ↑).

Table 4. Classification results obtained when removing base training on the source domain in terms of the macro F1-score averaged over 30,000 tasks. All values are reported in percentages. Query sets are sampled following the realistic sampling setting using a Dirichlet distribution with $\boldsymbol{a} = 2 \cdot \mathbb{1}_K$. The numbers in parentheses show the performance change obtained when the base training is performed. A blue arrow (↑) indicates a performance improvement, whereas a red arrow (↓) indicates a performance drop. In the cases where there is no arrow, there is no change in the performance. '-' indicates that the result was computationally intractable to obtain.

		Scenario 1: Ghana			
		Five-way		**24-way**	
	Method	**5-shot**	**20-shot**	**5-shot**	**20-shot**
Induct.	Protonet	47.6 (↑ 8.8)	49.5 (↑ 14.3)	29.4 (↑ 9.5)	29.2 (↑ 14.4)
	MetaOptNet	57.5 (↓ 3.1)	72.1 (↓ 4.4)	43.7 (↓ 7.9)	-
	Baseline	48.0 (↑ 9.3)	51.0 (↑ 13.7)	29.1 (↑ 10.5)	29.3 (↑ 15.2)
	SimpleShot	47.1 (↑ 9.7)	50.5 (↑ 12.5)	28.8 (↑ 10.3)	28.4 (↑ 14.2)
Transduct.	MAML	34.0 (↑ 23.8)	35.1 (↑ 31.8)	-	-
	Entropy-min	57.5	61.4 (↑ 1.4)	37.6 (↑ 0.3)	39.1 (↑ 0.5)
	TIM	51.7 (↑ 1.0)	56.4 (↑ 2.9)	33.6 (↑ 4.5)	35.4 (↑ 8.6)
	α-TIM	56.1 (↑ 4.3)	67.4 (↑ 7.8)	39.0 (↑ 5.8)	43.9 (↑ 15.7)
		Scenario 2: Infrequent Crop Types of France			
		Five-way		**Seven-way**	
	Method	**5-shot**	**20-shot**	**5-shot**	**20-shot**
Induct.	Protonet	37.5 (↑ 26.0)	46.2 (↑ 26.9)	30.7 (↑ 25.3)	35.4 (↑ 31.6)
	MetaOptNet	47.1 (↑ 8.2)	67.0 (↑ 4.1)	42.2 (↑ 6.8)	60.6 (↑ 5.7)
	Baseline	37.0 (↑ 31.4)	43.7 (↑ 32.4)	29.9 (↑ 31.5)	35.7 (↑ 34.5)
	SimpleShot	38.5 (↑ 29.9)	44.6 (↑ 31.3)	31.4 (↑ 30.0)	36.7 (↑ 33.1)
Transduct.	MAML	36.9 (↑ 23.1)	45.9 (↑ 25.3)	-	-
	Entropy-min	61.0 (↑ 6.6)	68.1 (↑ 7.5)	55.3 (↑ 5.1)	63.1 (↑ 5.8)
	TIM	40.9 (↑ 19.1)	50.5 (↑ 14.2)	34.7 (↑ 19.7)	43.8 (↑ 16.5)
	α-TIM	46.7 (↑19.8)	64.4 (↑ 16.4)	40.4 (↑ 18.0)	57.1 (↑ 18.8)

4.4. Classification Results Obtained When Removing the Temporal Pyramid Pooling Layer

We perform an ablation analysis to investigate the advantage of the TPP layer. In order to do so, we replace it with a global average pooling layer, which is equivalent to using only Level 1 of the TPP layer. We report the p-values of a two-sample t-test to investigate whether the TPP layer has a significant effect on the results. The results presented in Table 5 show that the TPP layer significantly improves the results (as reflected in the p-values) compared to using a global average pooling layer as a replacement for the TPP layer.

Table 5. Investigating the effect of the TPP layer on the results of the three selected FSL methods. The numbers in black show the performance of the framework when the TPP layer is replaced with a global average pooling layer. The numbers in parentheses show the performance change obtained when the TPP layer is used. A blue arrow (↑) indicates a performance improvement, while a red arrow (↓) indicates a performance drop. The performance is measured in terms of the macro F1-score averaged over 30,000 tasks. All values are reported in percentages. Query sets are sampled following the realistic setting using a Dirichlet distribution with $\boldsymbol{a} = 2 \cdot \mathbb{1}_K$. The p-values of two-sample t-tests are reported to investigate whether the TPP layer has a significant effect on the results.

	Scenario 1: Ghana (24-way)			
Method	**1-shot**	**5-shot**	**10-shot**	**20-shot**
Entropy-min	19.5 (↑ 7.9)	20.3 (↑ 17.6)	23.7 (↑ 15.4)	22.7 (↑ 16.9)
TIM	28.4 (↓ 0.4)	34.1 (↑ 4.0)	35.6 (↑ 5.9)	36.8 (↑ 7.2)
α-TIM	26.7 (↑ 2.1)	39.0 (↑ 5.8)	41.5 (↑ 11.1)	43.5 (↑ 16.1)
p-value	<0.05			
	Scenario 2: Infrequent Crop Types of France (Seven-way)			
Method	**1-shot**	**5-shot**	**10-shot**	**20-shot**
Entropy-min	27.4 (↑ 9.5)	35.9 (↑ 24.5)	37.3 (↑ 28.8)	37.6 (↑ 31.3)
TIM	37.4 (↑ 4.0)	50.3 (↑ 4.1)	54.2 (↑ 3.7)	56.7 (↑ 3.6)
α-TIM	35.7 (↑ 1.7)	55.9 (↑ 2.5)	62.1 (↑ 7.3)	66.3 (↑ 9.6)
p-value	<0.05			

5. Discussion

Our study addressed a critical research question of crop type mapping studies, namely, how to handle the problem of labeled sample scarcity in the target study areas. This is a relevant question for many regions across the world where large labeled datasets are not available.

5.1. Interpretation of the Reported Classification Results

α-TIM achieved the best results, emphasizing the importance of effectively exploiting the query sets in the learning process. Additionally, comparing α-TIM to Baseline demonstrates that taking advantage of the underlying information in the query set can mitigate the limited performance of a standard baseline deep learning model when few labeled samples are available in the region of interest.

TIM performed significantly better than Entropy-min in the balanced sampling setting (Tables 2 and 3) because TIM minimizes not only conditional entropy but also label-marginal entropy. In contrast, Entropy-min focuses solely on minimizing conditional entropy, which encourages the model to generate confident predictions with high separability between different classes and unambiguous cluster assignments [13]. This can lead to trivial, single-class solutions, i.e., mapping all samples to a single class [13]. By minimizing the label-marginal entropy, TIM encourages the marginal distribution of labels to be uniform to avoid such solutions. However, in a realistic scenario where the distribution of the classes is not uniform, the performance of TIM decreases significantly. In contrast to TIM, the mutual information loss of α-TIM can handle imbalanced query sets effectively.

Comparing the results presented in Tables 2 and 3 demonstrates that the evaluated methods are challenged when the number of ways increases. Their performance drops in these scenarios because of the inter-class homogeneity and intra-class heterogeneity of complex agricultural areas [37,38].

In Scenario 2, all FSL methods outperformed DTW. Yet, in Scenario 1, only α-TIM outperfromed DTW (see Tables 2 and 3). These results could be explained by the capability of DTW to successfully align distorted temporal profiles caused by cloud coverage (Belgiu et al. [33]), which is higher in Scenario 1 compared to Scenario 2.

We showed the importance of base training for achieving high classification results (Table 4). This was also highlighted in transfer learning for crop mapping studies [39]. By removing the base training, the performance of the FSL methods in Scenario 1 dropped less compared to Scenario 2 (Table 4). The larger difference between the source and target datasets in Scenario 1 compared to Scenario 2 might explain these results. Table 4 shows MetaOptNet was not able to benefit from the source dataset for optimal model parameter initialization (numbers with ↓). This can be explained by the sensitivity of the method to the hyperparameters tuning [40], which is beyond the scope of this study.

The confusion matrix shown in Figure 6 demonstrates that even the performance of the best method (α-TIM) is seriously challenged when classifying some crop types in Scenario 1, characterized by a highly diverse agricultural landscape, i.e., 24 crop types, as well as high cloud cover [21]. The experiments conducted by Rustowicz et al. [21] using the same dataset also led to similar classification performance, which emphasizes the challenges of crop mapping in areas dominated by high cloud cover and complex landscapes in the smallholder setting. The best-classified crops in Ghana are nyenabe, akata, cotton, kpalika, zabla, sweet potatoe, babala beans, salad vegetables, bra and ayoyo, watermelon, and nili, with a recall score of at least 90%. The least accurate results (less than 50%) were obtained for ground nut, maize, rice, soya bean, yam, intercrop, sorghum, okra, millet, cowpea, and pepper. The confusion matrix (Figure 6) shows that the classifier often confuses these crop types with each other. In order to delve into the underlying reasons, we present the temporal profiles of the normalized difference vegetation index (NDVI) of these crop types in Figure 9. These profiles highlight the difficulty in distinguishing between these crops due to their similar growth patterns. Specifically, the wide standard deviation ranges and high overlap between them from different crop types indicate both the challenge of inter-class homogeneity and intra-class heterogeneity. The two challenges impact the classifier's performance significantly.

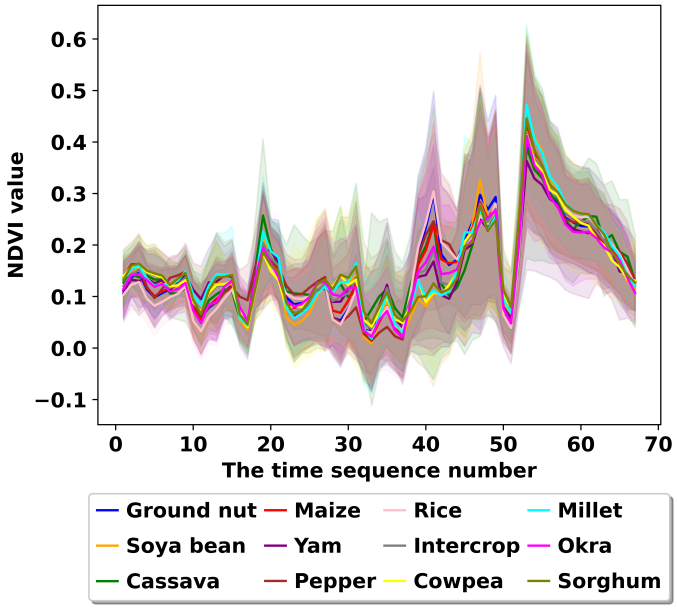


Figure 9. The temporal NDVI profiles of challenging crop types cultivated in the investigated study area from Ghana.

Additionally, we used the t-SNE visualization of a subset of crop samples to investigate the factors contributing to high performance in certain classes and lower performance in others. We selected four crop types with low performance, namely intercrop, maize, ground nut, and rice, which are the dominant crop types in this scenario. The confusion matrix (Figure 6) highlights that the classification model tends to confuse some of these crop types with each other and with other crops, notably yam, soya bean, pepper, and

cassava. For comparison, we also included a set of crop types—tomato, salad vegetables, watermelon, kpalika, and cotton—that pose less of a classification challenge. The t-SNE visualization of a subset of samples from all these crop types (Figure 10) shows that tomato, salad vegetables, watermelon, kpalika, and cotton are better clustered when compared to the other chosen crop types, which are mixed. Therefore, tomato, salad vegetables, watermelon, kpalika, and cotton are easier to classify due to lower inter-class homogeneity and intra-class heterogeneity, whereas others are less distinguishable due to higher inter-class homogeneity and intra-class heterogeneity. This is in line with the results presented in the confusion matrix (Figure 6). In Scenario 2, most of the classes obtained a recall score higher than 80%, with the exception of mixed cereals (66.6%) and potatoes (75.4%) (Figure 7). Due to the similar challenges posed in the first scenario, there is a relatively high confusion between mixed cereals, winter triticale, and spring barley and between potatoes, sunflower, and beet. On the other hand, the beet class is less confused with the other crops, which explains its high classification performance.

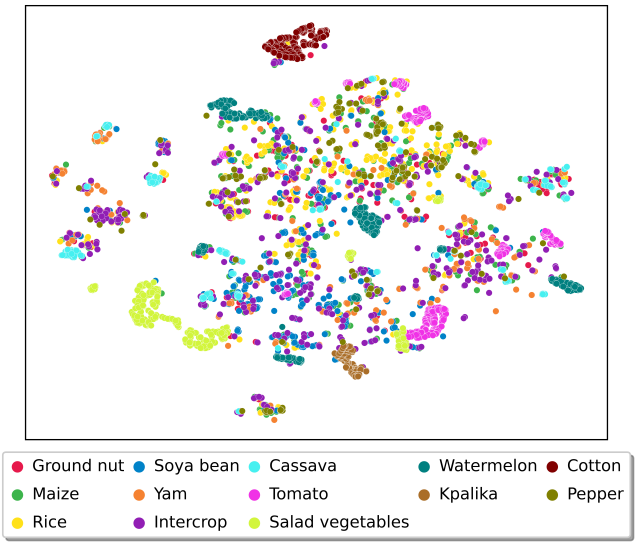


Figure 10. t-SNE visualization of a subset of samples from crops cultivated in the investigated study area from Ghana.

Finally, as evident from the p-values reported in Table 5, the TPP layer significantly improves the results when used instead of a global average pooling layer. Considering that patterns of interest can appear at different scales, the inclusion of the TPP layer empowers the network to capture information across multiple levels of granularity. This underscores the significance of employing methods tailored for crop mapping.

5.2. Limitations and Future Work

As evident in the results presented in the confusion matrices (Figures 6 and 7), the classification model tends to confuse some classes with each other due to the problem of high inter-class homogeneity and intra-class heterogeneity. This issue is especially problematic in the first scenario, where the model assigns a large portion of samples to the wrong classes. Therefore, future works should focus on developing methods capable of mitigating this challenge. For example, the cross-entropy loss used in FSL methods can be replaced with more advanced supervised loss functions, such as supervised contrastive loss [41], which is able to learn more discriminative feature representations. Furthermore, attention mechanisms such as Squeeze and Excitation Networks [42] or Transformers [43,44] can also be implemented to give the model the ability to focus on the most important features of each crop type.

Following the procedure employed by FSL methods for natural images, we first performed base training on the source domain and subsequently performed meta-testing

on the target dataset. Nevertheless, this procedure might limit the transferability of the model from the source domain to the target domain because it does not explicitly consider the domain shift between the two domains. This challenge was also discussed in Islam et al. [45], where the authors demonstrated that using both labeled source data and unlabeled target data during training provides a common embedding for both domains, resulting in better feature representations. Therefore, future studies might focus on integrating unsupervised domain adaptation methods [46,47] in their FSL framework to address the challenge of high discrepancies between the temporal-spectral characteristics of crops from different regions. These methods use both labeled source data and unlabeled target data to explicitly reduce the domain shift and learn common embedding. They can be particularly helpful in Scenario 1 for achieving more satisfying results in regional-scale crop mapping.

Pretraining on the label-rich source dataset that is more similar to the target dataset could lead to better results [3]. This was also observed when comparing the impact of pretraining on the results of Scenario 1 versus Scenario 2 (Table 4). In order to identify the most similar label-rich source dataset to the target dataset, the degree of similarity between the two domains can be assessed using growing degree day (GDD) [48], as employed in Wang et al. [3], or Earth mover's distance (EMD) [49], presented in Oh et al. [50].

As seen in Tables 2 and 3, the performance improves as the number of shots increases. This implies that as we label more samples from the target study area, we can expect to achieve a greater improvement in performance. However, labeling randomly selected samples can be suboptimal as it does not take into account their informativeness [51]. The performance in the label-scarce region might be improved through the strategic labeling of a few additional samples. This can be achieved using active learning strategies, which focus on selecting the most informative samples for model training to be labeled, with the goal of maximizing model performance while minimizing the amount of labeled data required [51–53].

The FSL methods incorporated in our framework include many hyperparameters. In order to simplify the experiments and stay within the scope of the study, the default hyperparameters used in the FSL studies were employed. Hyperparameter tuning remains a task for the future studies. Another possible future work is to replace our backbone network with different architectures employed in the crop mapping literature, such as LSTM-based [4] or self-attention-based architectures [10]. Future studies should also investigate the applicability of other FSL methods to crop mapping, such as work by Rodriguez et al. [54] and Wang et al. [55].

Notwithstanding these limitations, the present study proved the potential of FSL for crop type mapping and it can serve as a basis for future works dedicated to exploring and advancing these methods to alleviate the challenges in label-scarce regions further.

6. Conclusions

We implemented and evaluated four inductive and four transductive FSL methods to address the challenge of crop mapping in label-scarce and complex agriculture environments. We evaluated the performances of the FSL methods realistically by defining two real-world crop mapping scenarios and using the Dirichlet distribution to model the class proportions as random variables in few-shot query sets. Our experiments on classifying a large number of crop types in Ghana and infrequent crop types in France showed that the α-TIM method outperformed the other evaluated methods. For example, in a 24-way 20-shot setting in Ghana, it achieved a Macro F1-score of 59.6%, outperforming the second-best method by 2.7%. In a seven-way 20-shot setting in classifying infrequent crop types of France, it achieved a macro F1-score of 75.9%, surpassing the second-best method by 5.7%. The results demonstrated that α-TIM outperformed DTW, which has commonly been used in previous studies to deal with the lack of labeled samples. Moreover, the superior performance of α-TIM compared to a baseline deep learning model suggests exploiting the potential of the query set to improve crop mapping performance. Removing base

training on the large labeled source dataset resulted in huge performance drops, which highlights the importance of reusing the knowledge of the models trained in areas where abundant labels are available. We hope that our study, together with the provided code and benchmark, will open up new avenues for researchers to advance FSL for crop mapping.

Author Contributions: Conceptualization, S.M.; methodology, S.M.; formal analysis, S.M., M.B. and A.S.; writing—original draft preparation, S.M. and M.B.; writing—review and editing, S.M., M.B. and A.S.; software, S.M.; validation, S.M.; visualization, S.M. and M.B.; supervision, M.B. and A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The implementation code of the evaluated methods and the instructions on setting up the benchmark dataset are available at: https://github.com/Sina-Mohammadi/FewCrop (accessed on 9 March 2024).

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

References

1. Ramankutty, N.; Mehrabi, Z.; Waha, K.; Jarvis, L.; Kremen, C.; Herrero, M.; Rieseberg, L.H. Trends in global agricultural land use: Implications for environmental health and food security. *Annu. Rev. Plant Biol.* **2018**, *69*, 789–815. [CrossRef]
2. Kussul, N.; Lemoine, G.; Gallego, F.J.; Skakun, S.V.; Lavreniuk, M.; Shelestov, A.Y. Parcel-based crop classification in Ukraine using Landsat-8 data and Sentinel-1A data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *9*, 2500–2508. [CrossRef]
3. Wang, S.; Azzari, G.; Lobell, D.B. Crop type mapping without field-level labels: Random forest transfer and unsupervised clustering techniques. *Remote Sens. Environ.* **2019**, *222*, 303–317. [CrossRef]
4. Xu, J.; Zhu, Y.; Zhong, R.; Lin, Z.; Xu, J.; Jiang, H.; Huang, J.; Li, H.; Lin, T. DeepCropMapping: A multi-temporal deep learning approach with improved spatial generalizability for dynamic corn and soybean mapping. *Remote Sens. Environ.* **2020**, *247*, 111946. [CrossRef]
5. Chen, B.; Zheng, H.; Wang, L.; Hellwich, O.; Chen, C.; Yang, L.; Liu, T.; Luo, G.; Bao, A.; Chen, X. A joint learning Im-BiLSTM model for incomplete time-series Sentinel-2A data imputation and crop classification. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *108*, 102762. [CrossRef]
6. Zhong, L.; Hu, L.; Zhou, H. Deep learning based multi-temporal crop classification. *Remote Sens. Environ.* **2019**, *221*, 430–443. [CrossRef]
7. Pelletier, C.; Webb, G.I.; Petitjean, F. Temporal convolutional neural network for the classification of satellite image time series. *Remote Sens.* **2019**, *11*, 523. [CrossRef]
8. Wang, L.; Wang, J.; Zhang, X.; Wang, L.; Qin, F. Deep segmentation and classification of complex crops using multi-feature satellite imagery. *Comput. Electron. Agric.* **2022**, *200*, 107249. [CrossRef]
9. Mohammadi, S.; Belgiu, M.; Stein, A. Improvement in crop mapping from satellite image time series by effectively supervising deep neural networks. *ISPRS J. Photogramm. Remote Sens.* **2023**, *198*, 272–283. [CrossRef]
10. Rußwurm, M.; Körner, M. Self-attention for raw optical satellite time series classification. *ISPRS J. Photogramm. Remote Sens.* **2020**, *169*, 421–435. [CrossRef]
11. Garnot, V.S.F.; Landrieu, L.; Chehata, N. Multi-modal temporal attention models for crop mapping from satellite time series. *ISPRS J. Photogramm. Remote Sens.* **2022**, *187*, 294–305. [CrossRef]
12. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
13. Boudiaf, M.; Ziko, I.; Rony, J.; Dolz, J.; Piantanida, P.; Ben Ayed, I. Information maximization for few-shot learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 2445–2457.
14. Snell, J.; Swersky, K.; Zemel, R. Prototypical networks for few-shot learning. In Proceedings of the 31st International Conference on Neural Information, Long Beach, CA, USA, 4–9 December 2017; pp. 4080–4090
15. Finn, C.; Abbeel, P.; Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 1126–1135.
16. Zhai, M.; Liu, H.; Sun, F. Lifelong learning for scene recognition in remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1472–1476. [CrossRef]
17. Rußwurm, M.; Wang, S.; Korner, M.; Lobell, D. Meta-learning for few-shot land cover classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 13–19 June 2020; pp. 200–201.
18. Gella, G.W.; Tiede, D.; Lang, S.; Wendit, L.; Gao, Y. Spatially transferable dwelling extraction from Multi-Sensor imagery in IDP/Refugee Settlements: A meta-Learning approach. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *117*, 103210. [CrossRef]
19. Tseng, G.; Kerner, H.; Nakalembe, C.; Becker-Reshef, I. Learning to predict crop type from heterogeneous sparse labels using meta-learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1111–1120.

20. Veilleux, O.; Boudiaf, M.; Piantanida, P.; Ben Ayed, I. Realistic evaluation of transductive few-shot learning. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 9290–9302.
21. Rustowicz, R.M.; Cheong, R.; Wang, L.; Ermon, S.; Burke, M.; Lobell, D. Semantic segmentation of crop type in Africa: A novel dataset and analysis of deep learning methods. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019; pp. 75–82.
22. Waldner, F.; Chen, Y.; Lawes, R.; Hochman, Z. Needle in a haystack: Mapping rare and infrequent crops using satellite imagery and data balancing methods. *Remote Sens. Environ.* **2019**, *233*, 111375. [CrossRef]
23. Garnot, V.S.F.; Landrieu, L. Panoptic Segmentation of Satellite Image Time Series with Convolutional Temporal Attention Networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 4872–4881.
24. Turkoglu, M.O.; D'Aronco, S.; Perich, G.; Liebisch, F.; Streit, C.; Schindler, K.; Wegner, J.D. Crop mapping from image time series: Deep learning with multi-scale label hierarchies. *Remote Sens. Environ.* **2021**, *264*, 112603. [CrossRef]
25. Lee, K.; Maji, S.; Ravichandran, A.; Soatto, S. Meta-learning with differentiable convex optimization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10657–10665.
26. Chen, W.Y.; Liu, Y.C.; Kira, Z.; Wang, Y.C.F.; Huang, J.B. A Closer Look at Few-shot Classification. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30April–3 May 2018.
27. Wang, Y.; Chao, W.L.; Weinberger, K.Q.; van der Maaten, L. SimpleShot: Revisiting nearest-neighbor classification for few-shot learning. *arXiv* **2019**, arXiv:1911.04623.
28. Dhillon, G.S.; Chaudhari, P.; Ravichandran, A.; Soatto, S. A Baseline for Few-Shot Image Classification. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
30. Sudholt, S.; Fink, G.A. Evaluating word string embeddings and loss functions for CNN-based word spotting. In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; Volume 1, pp. 493–498.
31. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations (ICLR), Banff, AB, Canada, 14–16 April 2014.
32. Sakoe, H.; Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoust. Speech, Signal Process.* **1978**, *26*, 43–49. [CrossRef]
33. Belgiu, M.; Csillik, O. Sentinel-2 cropland mapping using pixel-based and object-based time-weighted dynamic time warping analysis. *Remote Sens. Environ.* **2018**, *204*, 509–523. [CrossRef]
34. Ten Holt, G.A.; Reinders, M.J.; Hendriks, E.A. Multi-dimensional dynamic time warping for gesture recognition. In Proceedings of the Thirteenth Annual Conference of the Advanced School for Computing and Imaging, Heijen, The Netherlands, 13–15 June 2007; Volume 300, p. 1.
35. Maghoumi, M. Deep Recurrent Networks for Gesture Recognition and Synthesis. Ph.D. Thesis, University of Central Florida, Orlando, FL, USA, 2020.
36. Maghoumi, M.; Taranta, E.M.; LaViola, J. DeepNAG: Deep Non-Adversarial Gesture Generation. In Proceedings of the 26th International Conference on Intelligent User Interfaces, Station, TX, USA, 14–17 April 2021; pp. 213–223.
37. Hamidi, M.; Safari, A.; Homayouni, S. An auto-encoder based classifier for crop mapping from multitemporal multispectral imagery. *Int. J. Remote Sens.* **2021**, *42*, 986–1016. [CrossRef]
38. Zhang, P.; Hu, S.; Li, W.; Zhang, C. Parcel-level mapping of crops in a smallholder agricultural area: A case of central China using single-temporal VHSR imagery. *Comput. Electron. Agric.* **2020**, *175*, 105581. [CrossRef]
39. Nowakowski, A.; Mrziglod, J.; Spiller, D.; Bonifacio, R.; Ferrari, I.; Mathieu, P.P.; Garcia-Herranz, M.; Kim, D.H. Crop type mapping by using transfer learning. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *98*, 102313. [CrossRef]
40. Antoniou, A.; Edwards, H.; Storkey, A. How to train your MAML. In Proceedings of the Seventh International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
41. Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; Krishnan, D. Supervised Contrastive Learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 18661–18673.
42. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
43. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
44. Li, Z.; Chen, G.; Zhang, T. A CNN-transformer hybrid approach for crop classification using multitemporal multisensor images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 847–858. [CrossRef]
45. Islam, A.; Chen, C.F.R.; Panda, R.; Karlinsky, L.; Feris, R.; Radke, R.J. Dynamic distillation network for cross-domain few-shot recognition with unlabeled data. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 3584–3595.

46. Chen, C.; Xie, W.; Huang, W.; Rong, Y.; Ding, X.; Huang, Y.; Xu, T.; Huang, J. Progressive feature alignment for unsupervised domain adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 627–636.
47. Choi, J.; Jeong, M.; Kim, T.; Kim, C. Pseudo-Labeling Curriculum for Unsupervised Domain Adaptation. In Proceedings of the British Machine Vision Conference (BMVC), Cardiff, UK, 9–12 September 2019; Springer: Berlin/Heidelberg, Germany, 2019.
48. Zhong, L.; Gong, P.; Biging, G.S. Efficient corn and soybean mapping with temporal extendability: A multi-year experiment using Landsat imagery. *Remote Sens. Environ.* **2014**, *140*, 1–13. [CrossRef]
49. Rubner, Y.; Tomasi, C.; Guibas, L.J. A metric for distributions with applications to image databases. In Proceedings of the Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271), Bombay, India, 7 January 1998; pp. 59–66.
50. Oh, J.; Kim, S.; Ho, N.; Kim, J.H.; Song, H.; Yun, S.Y. Understanding Cross-Domain Few-Shot Learning Based on Domain Similarity and Few-Shot Difficulty. *arXiv* **2022**, arXiv:1508.04409.
51. Yoo, D.; Kweon, I.S. Learning loss for active learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 93–102.
52. Su, T. Active learning with prediction vector diversity for crop classification in western Inner Mongolia. *Multimed. Tools Appl.* **2023**, *82*, 15079–15112. [CrossRef]
53. Zhang, Z.; Pasolli, E.; Crawford, M.M. Crop Mapping through an Adaptive Multiview Active Learning Strategy. In Proceedings of the 2019 IEEE International Workshop on Metrology for Agriculture and Forestry (MetroAgriFor), Portici, Italy, 24–26 October 2019; pp. 307–311.
54. Rodríguez, P.; Laradji, I.; Drouin, A.; Lacoste, A. Embedding propagation: Smoother manifold for few-shot classification. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XXVI 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 121–138.
55. Wang, Y.; Zhang, L.; Yao, Y.; Fu, Y. How to trust unlabeled data? instance credibility inference for few-shot learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 6240–6253. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.