Meta-Learning for Few-Shot Land Cover Classification

Marc Rußwurm[1,*,†], Sherrie Wang[2,3,*], Marco Körner[1], and David Lobell[2]

[1]Technical University of Munich, Chair of Remote Sensing Technology
[2]Stanford University, Center on Food Security and the Environment
[3]Stanford University, Institute for Computational and Mathematical Engineering

Abstract

The representations of the Earth's surface vary from one geographic region to another. For instance, the appearance of urban areas differs between continents, and seasonality influences the appearance of vegetation. To capture the diversity within a single category, such as urban or vegetation, requires a large model capacity and, consequently, large datasets. In this work, we propose a different perspective and view this diversity as an inductive transfer learning problem where few data samples from one region allow a model to adapt to an unseen region. We evaluate the model-agnostic meta-learning (MAML) algorithm on classification and segmentation tasks using globally and regionally distributed datasets. We find that few-shot model adaptation outperforms pre-training with regular gradient descent and fine-tuning on the (1) Sen12MS dataset and (2) DeepGlobe dataset when the source domain and target domain differ. This indicates that model optimization with meta-learning may benefit tasks in the Earth sciences whose data show a high degree of diversity from region to region, while traditional gradient-based supervised learning remains suitable in the absence of a feature or label shift.

1. Introduction

A growing constellation of satellites, combined with cloud computing and deep learning, offers an objective and scalable way to monitor global issues from deforestation and wildfires to urban development and road flooding [15, 6, 4, 24]. For many of these prediction problems, the bottleneck to making accurate and timely predictions has shifted away from satellite imagery availability or data processing limits and toward a lack of ground truth labels [34, 31, 25]. At the same time, these tasks share characteristics in remotely sensed imagery—such as ground sampling distance, seasonality, and spectral characteristics—no matter where on Earth they are taken. This raises the question of whether prediction in label-scarce regions could be improved if each model were to benefit from knowledge contained in all the datasets, rather than solving the same prediction problem across different geographies or time slices with independent models trained on small disjoint datasets.

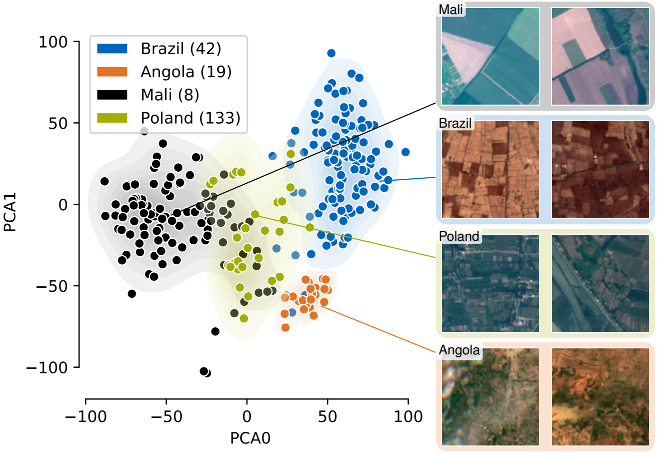


Figure 1: A principal component analysis (PCA) on VGG-16 [26] features of cropland images from different countries. Representations of the same class vary geographically; applying models trained on one geography to another would violate the assumption in traditional supervised learning that train and test distributions are equal. Model-agnostic meta learning provides a framework for inductive transfer learning that adapts the model to a new region with few data samples.

The concept of using knowledge gained while solving one problem to aid the solving of another is known in machine learning as **transfer learning** [17]. Transferring

[*]equal contribution; † this work was conducted during a research stay at the Lobell Lab and supported by a fellowship within the IFI programme of the German Academic Exchange Service (DAAD).

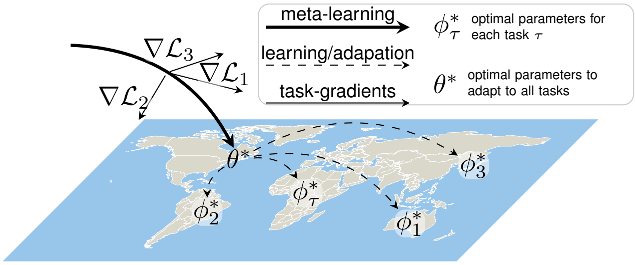


Figure 2: The model-agnostic meta learning (MAML) algorithm [8] finds initial weights θ from which a model can adapt to a new geographic region τ with few data samples.

knowledge between tasks or domains is successful when the problems are different but related [29]. We argue that the diverse nature of representations on the Earth's surface is a prime example of different-but-related tasks. We illustrate this in Fig. 1 using representations of cropland from four different countries. Croplands across the world are distinct from each other, yet they share characteristics. Transfer learning allows models to both adapt to each distribution individually and share knowledge across regions: countries like Angola and Mali, for which smaller labeled datasets are available, could then benefit from larger labeled datasets from countries like Brazil and Poland.

Thus far, transfer learning on remote sensing data has largely focused on fine-tuning pre-trained models and performing domain adaptation (Section 2). In this work, we explore **meta-learning**, in which models not only learn from data to perform tasks but *learn how to learn* to perform tasks through experiencing tasks on a variety of datasets. In particular, we use model-agnostic meta-learning (MAML) for the problem of inductive transfer-learning, where the generalization is induced by a few labeled examples in the target domain [17]. A schematic of MAML is shown in Fig. 2 and the algorithm is described in Section 3.2.

Our main contributions are (1) demonstrating that remote sensing tasks across geographies can be restructured as one meta-learning problem and (2) evaluating MAML for few-shot classification and segmentation of multi-spectral and high-resolution remote sensing images; specifically, the well-cited benchmark datasets Sen12MS and DeepGlobe.

2. Related Work

Transfer learning can be divided into subcategories depending on the amount of labeled data available in the source and target domains. Our work is focused on the scenario in which ample labels exist in the source domain, but few exist in the target domain. We summarize the related remote sensing methodology accordingly.

In such a setting, one common transfer learning technique is pre-training a neural network on ImageNet and fine-tuning [14] it on an application-specific dataset. For high-resolution remotely sensed imagery, these include airplane detection [5], high-resolution land cover classification [28], and disaster mapping [9]. Xie et al. (2016) [32] extended this concept by swapping ImageNet for the proxy task of night-light prediction that allowed them to estimate poverty in African regions with a limited number of labeled poverty data points. These approaches require a significant amount of problem design, such as the choice of proxy datasets or model and which parameters to fine-tune, and, thus, usually focus on a limited number of hand-selected tasks.

A second class of methods using deep learning for label-scarce tasks in remote sensing has focused on developing novel network architectures or loss functions to make learning more label-efficient. So far, these methods have focused on optical [11], SAR [21], and hyperspectral image classification [12]. While they decrease the number of labels required for any optical, SAR, or hyperspectral task, these methods do not explicitly endeavor to transfer knowledge from a data-rich geography to a data-poor one.

Non-deep learning methods for domain adaptation were summarized by Tuia et al. (2016) [29] and include selecting invariant features, adapting data distributions, and adapting classifiers via semi-supervised learning. For the most part, such methods generalize only across small regions rather than worldwide, while sometimes requiring a feature space in which inputs can be modeled as a mixture of Gaussians or some other predefined distribution.

Lastly, meta-learning is beginning to be explored for remote sensing applications. Alajaji and Alhichri (2020) [1] describe preliminary results of MAML on few-shot UC Merced, OPTIMAL-31, and AID RS classification, though again not with a focus on cross-geography generalization.

3. Meta-learning

Meta-learning [22] considers a large number of related tasks $\tau \in \mathcal{T} = \{\tau^{(1)}, \ldots, \tau^{(N)}\}$ to arrive at a predictive function that can perform well on unseen tasks τ after seeing a few data samples. Even though meta-learning has been a topic in machine learning for decades [22, 3], it has recently gained popularity for few-shot problems [30, 27, 19] and has been re-introduced under a "model agnostic" framework [8] with rapid developments in the field [18, 16, 2].

3.1. Terminology and Definitions

Meta-learning introduces a set of terms that may be new to some readers, so we clarify them in this section.

A **task** τ is comprised of a **support** dataset D_{support} to adjust the model parameters to the specific task and a **query** dataset D_{query} to evaluate the performance. Each dataset is comprised of inputs $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$ and corresponding labels $\{y_1, y_2, \ldots, y_m\}$ from a data distribution. A $\boldsymbol{k}$-

Algorithm 1: Regular Gradient Descent

```
$p(\mathcal{D})$: distribution over data points;
$\alpha$: step size hyperparameters;
randomly initialize $\phi$;
repeat
    sample $D \sim p(\mathcal{D})$;
    evaluate $\boldsymbol{g} = \nabla\mathcal{L}(f_\phi, D)$;
    update parameters $\phi \leftarrow \phi - \alpha\boldsymbol{g}$;
until convergence;
```

Algorithm 2: Model-Agnostic Meta-Learning

```
$p(\mathcal{T})$: distribution over tasks;
$\alpha, \beta$: step size hyperparameters;
randomly initialize $\theta$;
repeat
    sample batch of tasks $\tau \sim p(\mathcal{T})$;
    foreach $\tau_i \in \tau$ do
        initialize $\phi_i$ with $\theta$;
        sample $\{D_\text{support}, D_\text{query}\} \sim p(\tau_i)$;
        evaluate $\boldsymbol{g} = \nabla_{\phi_i}\mathcal{L}_{\tau_i}(f_{\phi_i}, D_\text{support})$;
        adapt parameters $\phi_i \leftarrow \phi_i - \alpha\boldsymbol{g}$;
        evaluate test loss $\mathcal{L}_{\tau_i}(f_{\phi_i}, D_\text{query})$ ;
    end
    update $\theta \leftarrow \theta - \beta \sum_{\tau_i \sim p(\tau)} \nabla_\theta \mathcal{L}_{\tau_i}(f_{\phi_i}, D^{\tau_i}_\text{query})$;
until convergence;
```

shot, n**-way** classification task aims to distinguish between n classes and is trained on k examples per class. Each task is drawn from a distribution over tasks $\tau \sim p(\mathcal{T})$ to yield a set of tasks $\{\tau^{(1)}, \tau^{(2)}, \dots, \tau^{(N)}\}$. The meta-learner *learns how to learn* by training and evaluating on the **meta-training set**. Meta-learning hyperparameters are tuned on the **meta-validation set**. The **meta-test set** measures generalization on new, unseen tasks.

3.2. Model-Agnostic Meta Learning (MAML)

Neural network parameters ϕ are usually initialized randomly and optimized iteratively via gradient descent to perform well on a single dataset, as shown in Algorithm 1. Model-agnostic meta-learning (MAML) extends gradient descent by optimizing for a model initialization θ that leads to good performance on a set of related tasks $\{\tau^{(1)}, \tau^{(2)}, \dots, \tau^{(N)}\}$. We contrast the regular gradient descent with the MAML optimization algorithm in Algorithms 1 and 2. Meta-training is divided into an inner loop and an outer loop. In the inner loop, networks initialized with θ are updated to each task via t steps of gradient descent on D_support of each task. This results in models with parameters ϕ_i adapted to each task $\tau^{(i)}$. The outer loop updates θ based on the performance of ϕ_i on D_query of the meta-training batch. In so doing, MAML requires second-order gradient calculations. The algorithm looks for a better θ until convergence, upon which the generalization error is computed on unseen meta-test tasks.

4. Datasets

We evaluate model-agnostic meta-learning on two public remote sensing datasets that cover optical and radar data at medium and very high resolution.

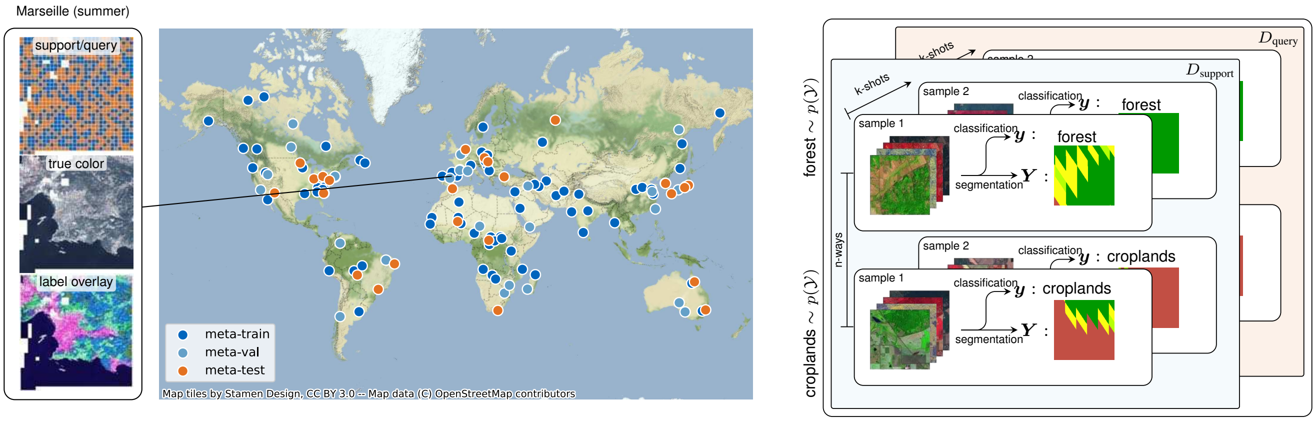
4.1. Sentinel-1/2 Multi-Spectral (Sen12MS) Dataset

The *Sentinel-1/2 Multi-Spectral (Sen12MS)* [23] dataset is a novel globally distributed satellite image classification and segmentation dataset. It contains 280 662 Sentinel 2 (optical) and Sentinel 1 (radar) tiles from 125 distinct regions at four different seasons. The optical and radar images were resampled to 10 m ground sampling distance and span 256 × 256 px in height and width. The original dataset uses tile-overlaps of 50%. For this work, we removed the overlap to ensure independence of support and query datasets, which yielded 200 306 128 × 128 px tiles. We show true color examples and principal component embeddings on VGG-16 features of four distinct regions in Fig. 1. Each image tile is accompanied by a land cover label with a comparatively coarse resolution of 500 m from the MODIS Land Cover product MCD12Q1 V6 upsampled to 10 m. In this work, we use the Sen12MS dataset for classification and assign the most common pixel-level label to the image tile. We use the simplified label-scheme of International Geosphere Biosphere Programme (IGBP) categories [13] with 10 distinct classes, consistent with the IEEE Data Fusion Contest 2020 [33]. In Fig. 3a, the 125 globally distributed regions are shown separated into meta-train, meta-validation, and meta-test sets. Each region contains between 196 and 850 tiles with a region-specific class distribution. We also show an overview of all tiles of the region 131 (Marseille) from the summer season true-color and labels. The individual 128 × 128 px tiles are randomly assigned to the support or query partition of each region. The objective is to classify each tile with its most frequent label class. Figure 3b illustrates this on an example of a 2-shot 2-way task. In this case, task datasets D^τ_{query} and $D^\tau_{support}$ contain $k = 2$ randomly chosen tile-label pairs of $n = 2$ distinct classes chosen from the available classes in the region.

4.2. DeepGlobe Land Cover Segmentation Dataset

The DeepGlobe Challenge [7] was introduced at CVPR 2018 to advance state-of-the-art satellite image analysis. Here, we used the land cover segmentation data to explore the use of MAML on high-resolution satellite imagery.

The DeepGlobe land cover segmentation dataset is comprised of very high resolution (0.5 m) DigitalGlobe Vivid+



(a) The 125 regions of the Sen12MS dataset. The 25 meta-test regions have been selected based on the hold-out set of the Data Fusion Contest 2020 [33]. The 75 meta-train and 25 meta-val have been randomly randomly partitioned.

(b) Example of a Sen12MS 2-way-2-shot task from region 87 and in the summer season. Ways determines the number of classes per task while shot the number of samples per class.

Figure 3: The Sen12MS dataset [23] is a public remote sensing dataset of 128 globally distributed regions and four distinct seasons. In this work, we sample tasks (b) from the dataset that include samples from one region and season aiming at adapting a deep learning model to one specific region.

images of dimension 2448 × 2448 px with three RGB channels. In total, there are 803 training images, each with human-annotated semantic segmentation labels covering seven land cover classes: urban, agriculture, rangeland, forest, water, barren, and unknown. For the competition, 171 validation images and 172 test images were also provided. However, since they do not have corresponding labels, we did not include them in the following experiments. Across the training images, the most common class is agriculture (58 % of pixels), followed by forest (11 %), urban (11 %), rangeland (8 %), barren (8 %), water (3 %), and unknown (0.05 %).

We divided the DeepGlobe training set into three meta-datasets: a meta-training set on which to train MAML, a meta-validation set on which to tune MAML hyperparameters, and a meta-test set on which to evaluate generalization (Fig. 4a). Ideally, we would evaluate whether meta-learned models generalize better to new geographic regions. However, the DeepGlobe Land Cover dataset does not tag images with latitude and longitude. In the absence of geographic information, we split the images in two ways:

1. At random, *i.e.* the 803 images were sampled uniformly at random into a 500-image meta-train, a 150-image meta-val, and a 153-image meta-test set.

2. Using unsupervised clustering on features extracted from a pre-trained network. DeepGlobe images were fed into a VGG-16 network pre-trained on ImageNet, and for each image, a 4096-dimensional vector was extracted from the first layer in the classifier. We used k-means to assign the images into 6 clusters and the 6 clusters were divided at random into the meta-train, meta-val, and meta-test sets. The resulting datasets contained 454, 166, and 183 images, respectively.

Figure 6a visualizes the distributions of image features for the meta-train, meta-val, and meta-test sets under these two splitting methodologies. The results across the two splits will illuminate the settings under which MAML improves upon pre-training and training from scratch.

Each image was further divided into 16 sub-images, each of dimension 612 × 612 px (Fig. 4b). Eight sub-images were placed in the support set and 8 in the query set. At meta-train time, k shots of 306 × 306 px tiles were sampled from the support set and q queries were sampled from the query set. At meta-test time, the entire query set was fed into the model as 32 tiles to compute metrics (Fig. 4c).

Put succinctly, our DeepGlobe experiments explore whether a model can learn to segment a large region (1.2 km × 1.2 km) of high resolution satellite imagery after seeing only a small labeled section (153 m × 153 m) of it.

5. Models

Model-agnostic meta-learning is an optimization algorithm that uses gradient descent and can be employed for any neural network architecture. In this work, we chose two popular models for image classification and segmentation.

5.1. CNN Classification Model

Following other meta-learning approaches [8, 30], we used a straightforward CNN architecture for the Sen12MS classification objective. The network consisted of 7 stacked layers with 64 convolutional 3 × 3 px kernels followed by batch normalization [10], ReLU activation function, and max-pooling of size 2. The input tensor $\boldsymbol{X} \in \mathbb{R}^{128\times128\times15}$ of joint Sentinel-2 and Sentinel-1 bands is projected to a

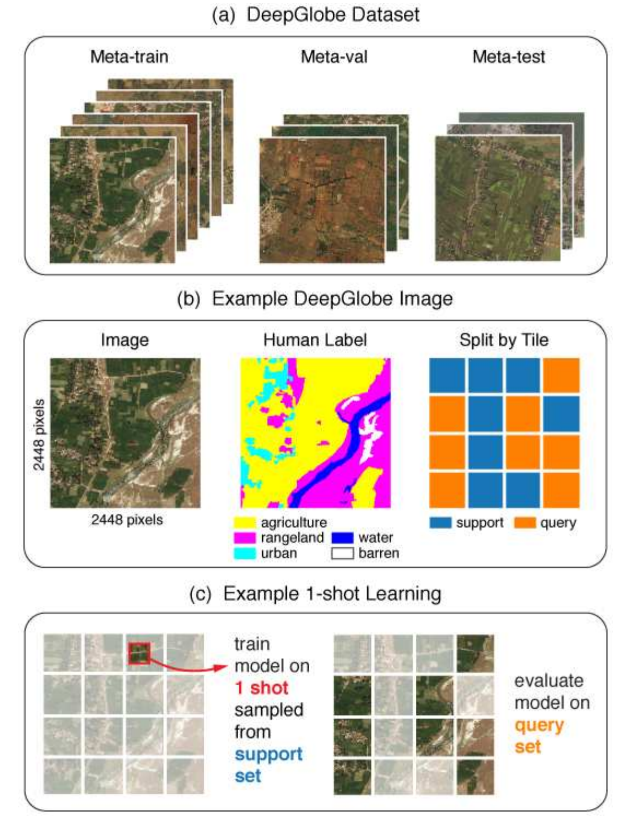


Figure 4: The DeepGlobe dataset contains high resolution RGB satellite imagery with land cover labels segmented by humans. To repurpose DeepGlobe for meta-learning, we (a) split the images into meta-train, meta-val, and meta-test sets. Then (b) each image was split into 16 sub-images, 8 of which were placed in the support set and 8 in the query set. Under such a setup, (c) we trained models on the meta-train set to segment the queries after seeing k shots from the support.

64-dimensional feature vector that maps to the output vector $\boldsymbol{y} \in \mathbb{R}^{10}$ for each of the classes.

5.2. U-Net Segmentation Model

For the DeepGlobe segmentation task, we employed the popular U-Net [20] architecture. It is a fully-convolutional segmentation model with skip connections between encoder and decoder. We used four downsampling and upsampling layers so that the input tensor is projected to a hidden representation, which is then added to intermediate hidden states from the encoder (skip connections) while being upsampled and convolved to an output tensor whereupon each pixel represents one class label.

6. Experiments

We experimentally evaluated the classification and segmentation performance of deep learning models with the same architecture trained with regular gradient descent (pretrained) Algorithm 1 and MAML Algorithm 2.

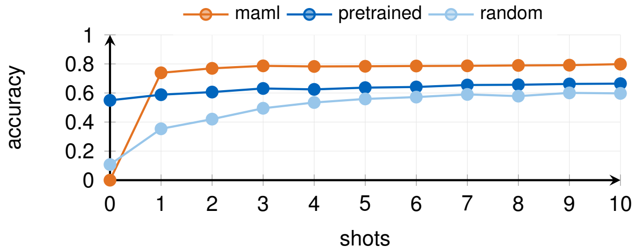


Figure 5: Classification results on Sen12MS. Regular pre-training with gradient descent leads to good zero-shot performance, while models trained with the model-agnostic meta learning algorithms outperform regular pretraining and the randomly initialized baseline clearly throughout all ten seen examples from a unseen region.

6.1. Sen12MS Classification

We assumed that data from the meta-train regions were readily available, but at most ten image-label pairs per class can be seen from the meta-test regions. This corresponds to a 4-way 10-shot classification scenario with four randomly selected classes from one region. It reflects use-case of interest to this work, where labeled data is available in some regions but not in others.

We trained the classification models with MAML on 4-way 2-shot datasets from the meta-train regions. We treated each sub-dataset D_{support} and D_{query} as a single batch of $N = k \cdot n = 8$ samples.

Baselines. We compared the *MAML-trained* model with a model that was pre-trained on all available data from the meta-train regions using regular gradient descent Algorithm 1. We *pre-trained* this model with the same 4-way 2-shot batches as MAML but used the combined support and query sets for training. This resulted in a batch size of 16 image label pairs. Finally, we also considered the scenario of having no additional data from meta-train regions. Here, we initialized the model randomly without any prior training, and train on each task's support set from scratch; we refer to this baseline as the *random* model.

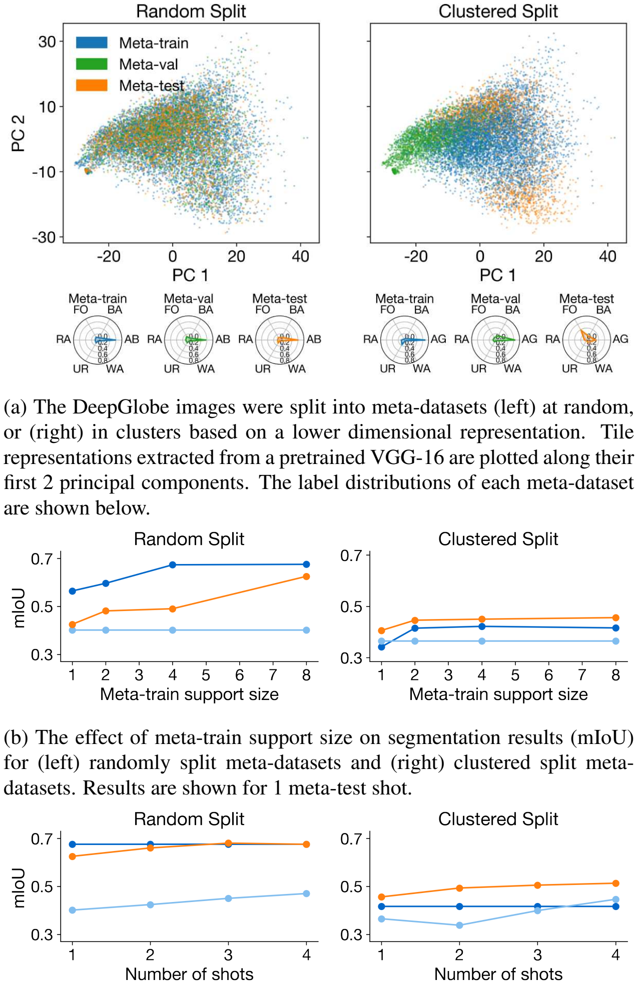
Evaluation. With the three initial CNN model parameterizations, *i.e. MAML-trained*, *pretrained*, and *random*, we evaluated the ability to adapt to new unseen meta-test regions based on at most ten data samples. For this, we sampled 100 4-way 10-shot tasks from the meta-test regions. We fine-tuned the models on subsets of D_{support}, while we report performance metrics on D_{query} on all ten examples per class. The number of samples seen from D_{support} was varied incrementally from zero-shot to 10-shot. Zero-shot represents no fine-tuning and shows the performance that can be obtained solely based on data from the meta-train regions. Training on batches of 1-shot to 10-shot provides increasingly more data from the target region to the models. The meta-val regions were used to determine a suitable step size $\alpha \in \{0.001, 0.0025, \ldots, 0.5, 0.75, 1\}$ and gradient steps on the same data batch $n \in \{1, 2, 5, 10, 50, 100\}$ for fine-tuning the pre-trained model. We evaluated these hyperparameters via grid search for each shot independently.

Classification Results. In Section 6.1, we report the accuracy scores for an increasing number of shots. The zero-shot case, without any adaptation to the particular meta-test region, shows that the regular pre-trained model performed best with 55 % accuracy and a kappa score of 0.47. Without any adaptation on the target-region, MAML predictions are low in accuracy, which highlights a distinct difference between meta-learning and pre-training. However, when a single data sample from the meta-test region is provided (1-shot), the MAML-trained model (74 % accuracy, 0.68 kappa score) outperforms the pre-trained model (59 % accuracy, 0.51 kappa score) by a large margin. The pre-trained model only shows a comparatively slight increase in accuracy (54 % to 66 %) throughout all seen examples while the MAML-trained model scores 80 % accuracy and 0.76 kappa score with all 10 shots.

6.2. DeepGlobe Land Cover Segmentation

Our second experiment demonstrates the use of MAML on the DeepGlobe land cover segmentation dataset. Each DeepGlobe image was considered its own task and we trained a U-Net via MAML to segment the query set of an image after being shown k shots from the support set. The experiments were designed to investigate the effect on the generalization of (1) meta-training label quantity (number of support and query sub-images), (2) meta-test label quantity (number of shots), and (3) distributional shift between meta-train and meta-test sets (random split versus clustered split of meta-datasets). The number of labeled sub-images in the support and query sets was varied to be $m \in \{1, 2, 4, 8\}$ and the number of shots used to adapt the U-Net was in the range $k \in \{1, 2, 3, 4, 5\}$. Hyperparameters, such as the number of epochs to meta-train MAML or train a model from scratch, were selected using performance on the meta-validation set.

Baselines. Similar to the Sen12MS evaluation, we compared MAML to two baselines: (1) a U-Net pre-trained on the meta-training set and fine-tuned on k shots in each meta-test task, and (2) randomly initialized U-Nets trained independently from scratch on k shots in each meta-test task. To make comparisons fair, we showed the pre-trained model the same amount of data as seen by MAML. If MAML was meta-trained on m support tiles and m query tiles and adapted using k shots, the baseline U-Net was pre-trained on $2m$ tiles per image and fine-tuned on k shots per meta-test tile, and the randomly initialized model was trained on k shots. The U-Net architecture was shared among MAML and both baselines.



(a) The DeepGlobe images were split into meta-datasets (left) at random, or (right) in clusters based on a lower dimensional representation. Tile representations extracted from a pretrained VGG-16 are plotted along their first 2 principal components. The label distributions of each meta-dataset are shown below.

(b) The effect of meta-train support size on segmentation results (mIoU) for (left) randomly split meta-datasets and (right) clustered split meta-datasets. Results are shown for 1 meta-test shot.

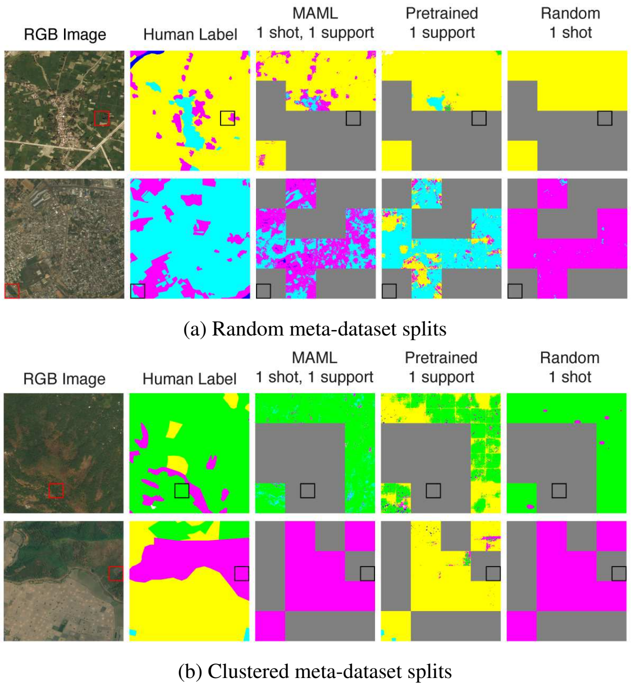
(c) The effect of number of adaptation shots on segmentation results. Results are shown for a support size of 8.

Figure 6: Segmentation results on DeepGlobe, with two ways of splitting the images into meta-datasets.

Evaluation. The performance of all models was evaluated on the query tiles of an unseen meta-test set of images. The location of the k shots for each meta-test image was sampled at random from its support set and fixed across all models for direct comparison. The models were evaluated by means of pixel-wise accuracy and the mean intersection over union (mIoU) score across the meta-test queries. For elaboration on the formula used to compute mIoU, please refer to the DeepGlobe publication [7].

Random Meta-Dataset Split Results. When the meta-datasets were randomly split, the pre-trained model performed better than MAML and the randomly initialized model. This was especially true at smaller meta-training set sizes (Fig. 6b). In other words, MAML requires a large set of meta-training tasks in order to perform well on new tasks. As the number of shots seen by the meta-learner increases, MAML catches up to the pre-trained model (Fig. 6c). In these experiments, we did not observe fine-tuning of the pre-trained model to improve its performance.

Figure 7a visualizes the predictions of MAML and the



(a) Random meta-dataset splits

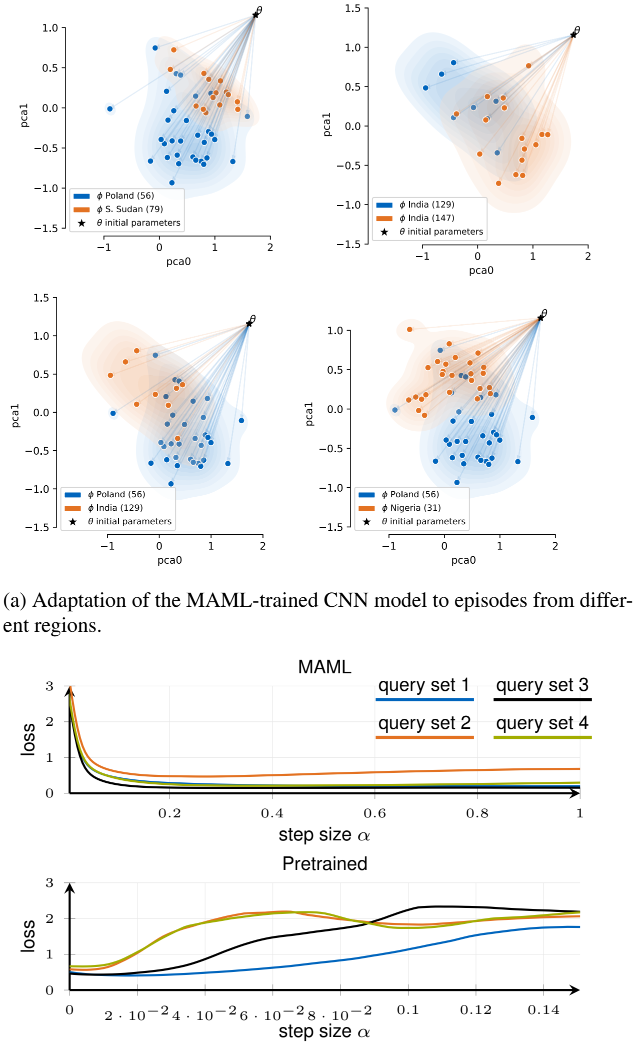
(b) Clustered meta-dataset splits

Figure 7: Example segmentation predictions by MAML, a pre-trained U-Net, and U-Nets trained from scratch.

baselines on 1-shot learning for two images: one where MAML performs well and one where it fails. MAML appears heavily influenced by the choice of the 1 shot, while the pre-trained model is biased toward predicting agriculture (the most common class). The model trained from scratch is even more heavily influenced by the choice of the 1 shot, as this is the only data it sees during training.

The success of pre-training can be attributed to the complete overlap of meta-train and meta-test distributions, seen in Fig. 6a. In the setting where $p(X, y)$ are identical in the source domain and target domain, a model trained on the source domain transfers perfectly to the target domain. These results also expose MAML's weaknesses when meta-train size is small: it is not able to retain information about land cover types as effectively as a straightforward supervised model.

Clustered Meta-Dataset Split Results. When the meta-datasets were split along clusters, the meta-train and meta-test distributions overlapped less (Fig. 6a) but could still be considered to arise from the same data-generating distribution. Whereas the meta-train set contains mostly agriculture pixels, the meta-test set contains predominantly forest. Figures 6b and 6c show, first and foremost, that this meta-test set is more difficult than the randomly split meta-test set for all three models. However, MAML is able to adapt to this distributional shift more successfully than the pre-trained model. Example segmentations shown in Fig. 7b reveal that MAML's flexibility to adaptation can again be both helpful and detrimental: helpful when the 1 shot is representative of the image, but detrimental when it is not. We see that the pre-trained model carries its bias toward agriculture into its meta-test set predictions, whereas MAML does not appear to retain a strong enough prior to recognize agriculture without being provided a shot containing that class.



(a) Adaptation of the MAML-trained CNN model to episodes from different regions.

(b) 1D Loss surface on multiple query sets along the gradient of one support set.

Figure 8: The adapted weights ϕ_τ for task τ vary from region to region (a). The loss surface along the direction of initial weights θ to ϕ_τ (b) is more convex and allows larger gradient step sizes for model-agnostic meta learning compared to regular pretraining.

6.3. Visualization of Model Adaptation

In the introduction and Fig. 1, we showed the regional diversity of representations on the Earth's surface using PCA on pre-trained VGG-16 image features. In Section 3 and Fig. 2, we assumed that a neural network would achieve optimal performance with a different set of weights ϕ^* for each geographic region. In this experiment, we empirically confirmed this hypothesis with two evaluations on meta-test regions of the Sen12MS dataset. In Section 6.3.1, we visualize the adapted MAML weights for two distinct geographies. Then in Section 6.3.2 we compare the loss surfaces of a MAML-trained and a pre-trained model along one adaptation trajectory. The two evaluations are meant to provide the reader with some intuition of what MAML is doing in different regions and how this differs from pre-training.

6.3.1 Region-wise Adaptation

We studied the adaptation of MAML-model parameters θ trained on 2-shot 4-way tasks of Section 6.1. We sampled 1000 1-shot 4-way classification tasks from the meta-test regions for the four most common classes (forests, grassland, savanna, urban) and split these into a support and query partition at ratio of 4:1. For each training task, we evaluated the gradient and adapted the model using gradient descent with step size 0.75 to new parameters for each task $\phi_\mathcal{T}$. We visualized this adaptation by flattening all model parameters to a 231 818-dimensional vector and using PCA to map the parameters to the first two principal components. We colored this embedding by region and drew lines from the initial weights θ to the adapted task-specific weights $\phi_\mathcal{T}$ in Fig. 8. The adapted model-weights differ from region to region in embedding space, as can be seen in the examples of Poland and South Sudan. This empirically shows that a different set of model parameters is optimal for two different regions.

6.3.2 Loss Surface along Support Gradient

Next, for four example query tasks, we evaluated the loss along a line from the initial parameters θ to task-adapted parameters ϕ with the MAML-trained model and the pre-trained model. For this evaluation, we selected one support set and four query sets from the same region and season. The gradient $\boldsymbol{g}$ was evaluated on the support set, and different model weights ϕ_α were obtained along the gradient direction using $\phi_\alpha = \theta + \alpha \boldsymbol{g}$ with different step sizes $\alpha_{\text{MAML}} \in [0, 1]$, $\alpha_{\text{pre}} \in [0, 0.15]$ proportional to the optimal step sizes for MAML and pretrained model. We calculated the query loss using the model f_{ϕ_α} for each of the four queries at different step sizes α. This draws a one-dimensional slice of the loss-surface along the gradient direction determined by the support set. In Fig. 8b, we show this loss surface for the MAML-trained model and the pre-trained classification model. Without adaptation, at $\alpha = 0$, the MAML-trained model evaluates a high loss compared to the pre-trained model. This is consistent with the comparatively poor zero-shot results from Fig. 5. With increasing step size, however, we observe that the MAML loss decreases consistently while the pre-trained loss remains similar or increased for larger step sizes. The MAML-trained model achieves low loss in a large range of step sizes from 0.1 to 1 for all query sets, while a narrow range of step sizes between 0 and 0.05 lead to better accuracies on some tasks from the pre-trained model initialization.

In general, the loss surface of the MAML-trained model follows a convex curve for all of the test examples, while the loss surface of the pre-trained model is non-convex with local minima. This experiment illustrates the difference between meta-learning and pre-training: the two methods lead to very different model parameters. The loss surface of a meta-learned model is smooth and convex in the gradient direction of a novel task — in other words, when the MAML algorithm optimizes for an initialization θ that can adapt well to new tasks, it seeks out these smooth, convex regions in the loss landscape. By contrast, for the pre-trained model, there are cases like query sets 2 and 4 in which it appears beneficial not to adapt to the specific task's region.

7. Discussion and Conclusion

In this work, we evaluated the model-agnostic meta-learning (MAML) algorithm for few-shot problems in land cover classification to adapt deep learning models to individual regions with few data examples. Existing models use regular gradient descent to pre-train a model on a large body of data and use this pre-trained model as an initialization for datasets with fewer examples. We compared these two approaches on land cover classification on the Sen12MS dataset of optical and radar images from globally distributed regions and the DeepGlobe dataset with very high-resolution imagery in few regions. The results on Sen12MS in Section 6.1 demonstrate that MAML-optimization can outperform regular gradient descent and pre-training of models when the dataset includes a distinct regional diversity. The DeepGlobe results in Section 6.2 illustrate the advantage MAML offers when the source domain differs from the target domain in transfer learning but also highlight MAML's weaknesses in retaining prior knowledge and under-performing in ideal (identical source and target domain) settings. In Section 6.3, we evaluated the loss surfaces for pre-trained and MAML-trained models and showed that the loss surface was more convex for MAML-trained models when adapting to new unseen data.

We believe that the meta-learning framework can lead deep learning in Earth observation to a new direction: away from finding incrementally better model architectures for specific use-cases and toward unifying strategies that more closely reflect the reality on the Earth's surface. Much work remains to be done to improve MAML performance by retaining stronger priors on land cover classes, as well as to explore other meta-learning paradigms (*e.g.* prototypical networks).

References

[1] D. Alajaji and H. Alhichri. Few shot scene classification in remote sensing using meta-agnostic machine. In *2020 6th Conference on Data Science and Machine Learning Applications (CDMA)*, pages 77–80, 2020.

[2] Antreas Antoniou, Harrison Edwards, and Amos Storkey. How to train your MAML. *arXiv preprint arXiv:1810.09502*, 2018.

[3] Yoshua Bengio, Samy Bengio, and Jocelyn Cloutier. *Learning a synaptic learning rule*. Université de Montréal, Département d'informatique et de recherche ..., 1990.

[4] Basudeb Bhatta. *Analysis of urban growth and sprawl from remote sensing data*. Springer Science & Business Media, 2010.

[5] Zhong Chen, Ting Zhang, and Chao Ouyang. End-to-end airplane detection using transfer learning in remote sensing images. *Remote Sensing*, 10(1):139, 2018.

[6] Emilio Chuvieco. *Remote sensing of large wildfires: in the European Mediterranean Basin*. Springer Science & Business Media, 2012.

[7] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raska. DeepGlobe 2018: A challenge to parse the earth through satellite images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 172–17209. IEEE, 2018.

[8] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning (ICML)*, pages 1126–1135, 2017.

[9] Ananya Gupta, Elisabeth Welburn, Simon Watson, and Hujun Yin. Post disaster mapping with semantic change detection in satellite imagery. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

[10] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[11] Neal Jean, Sherrie Wang, Anshul Samar, George Azzari, David Lobell, and Stefano Ermon. Tile2vec: Unsupervised representation learning for spatially distributed data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3967–3974, 2019.

[12] B. Liu, X. Yu, A. Yu, P. Zhang, G. Wan, and R. Wang. Deep few-shot learning for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 57(4):2290–2304, 2019.

[13] Thomas R Loveland and AS Belward. The IGBP-DIS global 1km land cover data set, discover: first results. *International Journal of Remote Sensing*, 18(15):3289–3295, 1997.

[14] Dimitrios Marmanis, Mihai Datcu, Thomas Esch, and Uwe Stilla. Deep learning earth observation classification using ImageNet pretrained networks. *IEEE Geoscience and Remote Sensing Letters*, 13(1):105–109, 2015.

[15] Stephen D McCracken, Eduardo S Brondizio, Donald Nelson, Emilio F Moran, Andrea D Siqueira, and Carlos Rodriguez-Pedraza. Remote sensing and GIS at farm property level: Demography and deforestation in the Brazilian Amazon. *Photogrammetric Engineering and Remote Sensing*, 65:1311–1320, 1999.

[16] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

[17] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[18] Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. In *Advances in Neural Information Processing Systems*, pages 113–124, 2019.

[19] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International Conference on Learning Representations (ICLR)*, 2017.

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[21] Mohammad Rostami, Soheil Kolouri, Eric Eaton, and Kyungnam Kim. Deep transfer learning for few-shot SAR image classification. *Remote Sensing*, 11(11), 2019.

[22] Jürgen Schmidhuber. *Evolutionary principles in self-referential learning, or on learning how to learn: The meta-meta-... hook*. Diplomarbeit, Technische Universität München, München, 1987.

[23] Michael Schmitt, Lloyd Haydn Hughes, Chunping Qiu, and Xiao Xiang Zhu. SEN12MS–A curated dataset of georeferenced multispectral Sentinel-1/2 imagery for deep learning and data fusion. *arXiv preprint arXiv:1906.07789*, 2019.

[24] E Schnebele, N Waters, et al. Road assessment after flood events using non-authoritative data. *Natural Hazards and Earth System Sciences*, 14(4):1007, 2014.

[25] Grant J Scott, Matthew R England, William A Starms, Richard A Marcum, and Curt H Davis. Training deep convolutional neural networks for land–cover classification of high-resolution imagery. *IEEE Geoscience and Remote Sensing Letters*, 14(4):549–553, 2017.

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[27] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087, 2017.

[28] Xin-Yi Tong, Gui-Song Xia, Qikai Lu, Huanfeng Shen, Shengyang Li, Shucheng You, and Liangpei Zhang. Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sensing of Environment*, 237:111322, 2020.

[29] Devis Tuia, Claudio Persello, and Lorenzo Bruzzone. Domain adaptation for the classification of remote sensing data: An overview of recent advances. *IEEE Geoscience and Remote Sensing Magazine*, 4(2):41–57, 2016.

[30] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016.

[31] Sherrie Wang, William Chen, Sang Michael Xie, George Azzari, and David B Lobell. Weakly supervised deep learning for segmentation of remote sensing imagery. *Remote Sensing*, 12(2):207, 2020.

[32] Michael Xie, Neal Jean, Marshall Burke, David Lobell, and Stefano Ermon. Transfer learning from deep features for remote sensing and poverty mapping. In *AAAI Conference on Artificial Intelligence*, AAAI'16, pages 3929–3935. AAAI Press, 2016.

[33] Naoto Yokoya, Pedram Ghamisi, Ronny Hänsch, and Michael Schmitt. 2020 IEEE GRSS Data Fusion Contest: Global land cover mapping with weak supervision. *IEEE Geosci. Remote Sens. Mag.*, 2020. in press.

[34] Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36, 2017.