INTER-REGION TRANSFER LEARNING FOR LAND USE LAND COVER CLASSIFICATION

J. Siddamsetty[1], M. Stricker[1,2], M. Charfuelan[1], M. Nuske[1] and A. Dengel[1,2]

[1] German Research Center for Artificial Intelligence (DFKI GmbH), Kaiserslautern, Germany
[2] Rhineland-Palatinate Technical University of Kaiserslautern-Landau, Kaiserslautern, Germany
(jayanth.siddamsetty, marco.stricker, marcela.charfuelan_oliva, marlon.nuske, andreas.dengel)@dfki.de

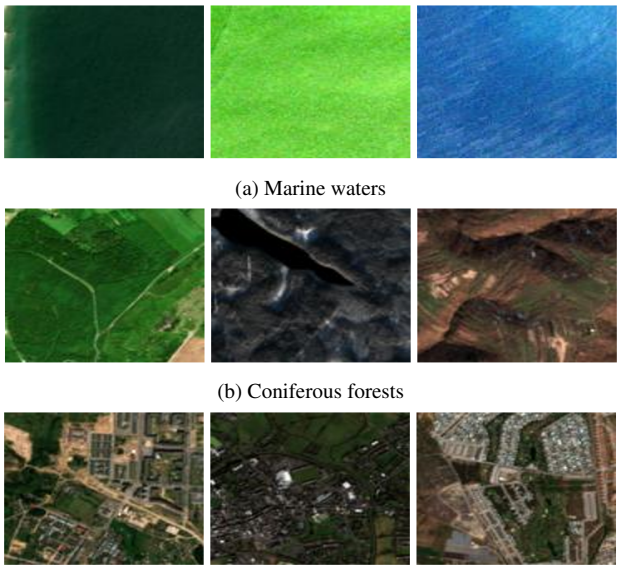
KEY WORDS: Land Use Land Cover (LULC), Remote Sensing, Transfer Learning

ABSTRACT:

Land use land cover (LULC) classification is an essential task in Earth Observation (EO) as it helps in monitoring long-term developments, detecting changes and analysing their environmental impacts. Due to advancements in remote sensing, there is an abundance of open data available but annotating this data is expensive. As a result, many research works in EO create a labelled dataset for one selected region and perform a corresponding regional analysis. By employing transfer learning, we can reuse these labelled datasets for different regions and thereby minimize the manual annotation costs. However, there are some open questions: to what extent can the features learned in one region be transferred to another? Does a larger pre-training dataset mean better transfer learning performance? How can we estimate the transfer learning performance? To answer these questions, we divide a large EO dataset called BigEarthNet into sub-datasets by region and perform region to region transfer learning. We find that the models trained on one region do not perform well on another region. We applied transfer learning techniques and showed that the class imbalance can hinder learning. If the source region has additional classes which are dominant in the source region or has fewer images for the classes dominant in the target region, transfer learning can have negative impacts on the model performance in the target region. We also demonstrate the use of chi-squared distance in selecting an appropriate source region for transfer learning.

1. INTRODUCTION

Land use land cover (LULC) classification is the task of categorizing the Earth's surface according to a set of land use and land cover classes. While land use classes describe terrain altered by humans such as 'Agricultural land', land cover classes describe natural terrain such as 'Forest'. Accurate LULC maps are important because land use change has an immense ecological impact. The LULC maps become outdated with the continuous change in the use of land and therefore need to be updated often. Using deep learning methods, it is possible to update the land cover maps automatically. Due to the advancement in the field of remote sensing, the task of acquiring images has become easier and the cost of acquisition has also been reduced, resulting in an abundance of data. The Copernicus programme of the European Commission has a series of satellites in the Sentinel mission which provide Earth Observation (EO) data for free under a full and open data policy. EO data are global and therefore cover a variety of landscapes. However, annotating this amount and variety of data manually is expensive. Transfer learning can reduce the amount of manual labelling needed. The technique transfers the weights of the model learned on a large dataset, called the source domain, to a target domain where only a few or no labelled images are available.



(a) Marine waters

(b) Coniferous forests

(c) Continuous urban fabric

Figure 1. Visualization of images from the same class appearing different in different regions

In this work, we perform inter-region transfer learning, where we train a classification model using a labelled dataset from one region and use it for inference or fine-tuning the model using a dataset from a different region with fewer labelled examples. The motivation behind this is bi-fold: First, a huge amount of data is available, but it is expensive to annotate. Transfer learning enables us to apply a trained model for a 'source' region on a different 'target' region with a minimal need for annotating data. Second, recent research in remote sensing is region-specific. For example, Jardim et al. perform LULC classification on images belonging to the Caatinga biome region of Northeast Brazil (Jardim et al., 2022), Costa et al. conduct land cover mapping on images from Portugal (Costa et al., 2022), Yuan et al. identify land use hot spots in Kazakhstan and Mongolia (Yuan et al., 2022) and several other research works are carried out on region-specific datasets (Zong et al., 2020) (Silva-Perez et al., 2020). It is however not investigated whether the same models can be used for

inference in different regions, which limits the applicability of the developed methods. The transfer of models to new regions is challenging, since the dominant LULC classes may differ from one region to another, or the target region may not have labelled data samples at all. In fact, even the same land use class might appear differently depending on the region, as we can see in Figure 1. It is therefore important to analyse in more detail to what extent models trained on one region are transferable to another.

The main contributions of our work are:

- We perform a thorough analysis of different region-to-region transfer learning scenarios.
- We compare fine-tuning and linear probing techniques.
- We use the chi-squared distance between the class distributions to estimate the performance of transfer learning.
- We investigate the potential of unsupervised domain adaptation techniques between countries.

We begin with a discussion about related work in Section 2 followed by a description of the BigEarthNet dataset and our derivations in Section 3. Next, in Section 4, we discuss the classification model used for this work. Later, in Section 5 we present the experiments and results. Finally, in Section 6, we discuss domain adaptation experiments and then conclude this work in Section 7.

2. RELATED WORK

LULC classification is a difficult task because of the challenges in remote sensing images. Cheng et al. (Cheng et al., 2020) have discussed these challenges in detail: they highlight a large intra-class diversity, high inter-class similarity, the big variance of object scales, and the coexistence of multiple ground objects belonging to different classes. LULC classifiers need good feature representations of the remote sensing image for correct classification. Before the usage of deep learning, these image features were hand-crafted. However, these descriptors have limited capabilities in representing an image, require expert knowledge, and are time-consuming to design. Deep learning methods, on the other hand, have the ability to learn the image features that can be used for different tasks. A major breakthrough in this direction was achieved by Krizhevsky et al. (Krizhevsky et al., 2012) who won the ImageNet classification challenge in 2012 by using Convolutional Neural Networks (CNNs) as feature extractors. Because of their ability to learn good features, CNNs were also used for remote sensing image classification. Vali et al. (Vali et al., 2020) published a survey of deep learning methods applied to LULC classification. They also highlight the need for large, labelled datasets. Additionally, Phiri et al. (Phiri et al., 2020) published a survey of LULC classification methods applied to Sentinel-2 imagery, which further classifies these methods as pixel-level and object-level methods. This survey shows the rise of object-level methods due to the limitations of pixel-level methods.

Before the emergence of large-scale remote sensing datasets, Penatti et al. (Penatti et al., 2015) have shown that CNNs trained to classify everyday objects could generalize well in classifying aerial images and also outperformed several visual descriptors. Castelluccio et al. (Castelluccio et al., 2015) also employed pre-trained CNNs and fine-tuned the weights of several layers of the CNN to adapt the model to classify remote sensing images. They showed that this approach is better than randomly initializing the weights of the CNN.

An additional challenge in remote sensing image datasets is the typically large Ground Sampling Distance (GSD), which represents the distance between adjacent pixel locations on the ground. For example, Sentinel-2 images have a GSD of 10 metres. Thus, a lot of information might be present inside each image and a single label may not be sufficient to annotate the whole image if it contains instances of many classes in a given image. Therefore, it is necessary to have more than one label per image for the model to be able to learn meaningful representations of different classes. The BigEarthNet dataset (Sumbul et al., 2019) is a large-scale remote sensing dataset which annotates images with multiple labels when there are instances of several classes in the image. In such cases, learning the pattern of class co-occurrence may also help improve the classification. For example, the presence of marine waters in the image increases the probability of a beach in the image. Huang et al. (Huang et al., 2021) have proposed an explicit network to infer label relations. They use the label co-occurrence matrix as an input to this explicit network, which is then used with the features extracted by the main CNN network for classification. Sumbul et al. (Sumbul and Demİr, 2020) on the other hand, divided each image into non-overlapping patches to generate local descriptors. Using a novel multi-attention strategy, they defined global descriptors for each image which will be used for classification.

The size of the BigEarthNet dataset allows applying transfer learning to remote sensing. Stojnic et al. (Stojnić and Risojević, 2021) used the remote sensing benchmark datasets to perform a self-supervised task as a pre-training task and then fine-tuned the model for remote sensing image classification. They showed that pre-training with multi-spectral remote sensing datasets is better than pre-training on a computer vision benchmark dataset like ImageNet (Russakovsky et al., 2014). Interestingly, they have also shown that pre-training with the ImageNet dataset was better than pre-training with only RGB images of the remote-sensing dataset. Walsh et al. (Walsh et al., 2021) used the BigEarthNet dataset for pre-training a ResNet classifier (He et al., 2016) and used the classifier as an encoder in a U-Net architecture to generate land cover maps for Ireland. Transfer learning between datasets has been beneficial in such cases. In this work, we investigate the transfer learning performance between different combinations of source and target regions. The general practice in transfer learning is that the models are first trained on a large dataset and then adapted to the task using a smaller dataset. Bigger pre-training datasets imply longer training time. If pre-training with big datasets does not significantly outperform pre-training with small datasets, their use should be reconsidered. We, therefore, perform an empirical analysis of data quantity and class distribution similarity on the application of transfer learning.

Fuller et al. (Fuller et al., 2022), performed similar experiments to investigate how pre-trained vision transformers perform on data not seen during training. They pre-train vision transformer models using around 1.3 million remote sensing images in an unsupervised way. They divide an image classification dataset into a set of datasets, each of them

containing images from different biomes. The pre-trained transformer is fine-tuned using images from one of the biomes and tested on images from a different biome that are not seen during the training. In contrast to that, we do not perform any unsupervised pre-training on such a large corpus of data. Instead, we investigate the performance of the model when trained first on images from a 'source' region and tested under different settings using the images from the 'target' region. Compared to the previously mentioned study, we provide more detailed reasoning in terms of class distributions and chi-squared distances between them to explain our results, which will help in choosing a better pre-training dataset.

3. DATASET

BigEarthNet-S2 (Sumbul et al., 2019) referred to as BigEarthNet in this paper is a multi-spectral dataset composed of images of 12 bands captured by the Sentinel-2 satellites. The dataset consists of 590,326 images captured over 10 countries during different seasons between June 2017 and May 2018. Each image is annotated with one or more LULC classes provided by the CORINE Land cover database of the year 2018. We use the nomenclature of 19 classes that were defined by the dataset creators based on Sentinel-2 image properties. For this work, we consider only the RGB bands, which have a spatial resolution of 10 m per pixel.

Country	Images
Finland	174,943
Portugal	99,174
Serbia	71,724
Lithuania	55,513
Ireland	49,197
Austria	44,240
Belgium	11124
Switzerland	4,940
Luxembourg	3,910
Kosovo	1,737

Table 1. Number of images in each country of the BigEarthNet dataset.

To facilitate the experiments, we divided the dataset into 10 smaller datasets, with each of them containing images from one country. Table 1 shows the number of images from each country. We use an early version of the large-scale data management tool developed by Aksoy et al. (Aksoy et al., 2022) to query labels using geometrical shapes. To generate a training, validation and testing split for each country, we define an intersection of the original training split for the whole dataset and the images belonging to a country. For example, to get the training split for Portugal, we use all images that are in the training set provided by the BigEarthNet dataset which belong to Portugal.

Figure 2 illustrates the distributions of classes in Portugal, Serbia, Finland, and Kosovo. It is interesting to notice the skewed distribution in the Finland dataset. Classes like 'Permanent crops'(3), 'Natural grassland and sparsely vegetated areas'(11), and 'Beaches, dunes, sands'(14) have only two, three, and nine samples respectively. Additionally, the number of samples in a few other classes amounts to only a fraction compared to the class "Coniferous forest"(9) with 83660 images. This problem is defined as long-tail learning and is a prevalent problem in the machine-learning community. Zhang et al. (Zhang et al., 2021) have presented a systematic survey that deals with this problem.

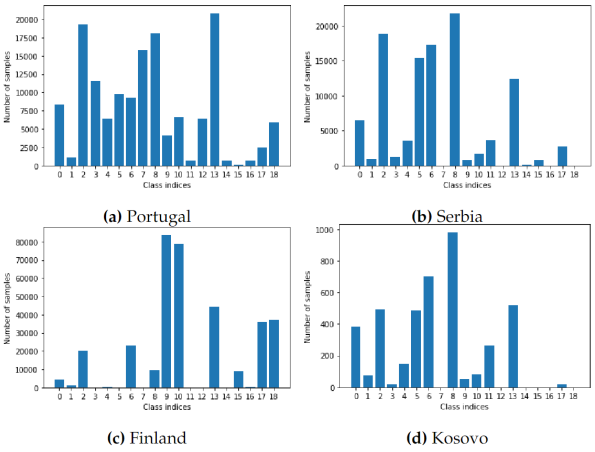


Figure 2. Different countries have a different distribution of samples per class. The x-axis of the plots represents the class indices. The class corresponding to these indices is listed in Table 3 in the Appendix

To help investigate the effects of datasets having a similar or different class distribution, we calculated the Chi-square distance between two distributions using Equation 1. The Chi-square distance matrix is plotted in Figure 3.

$$\mathcal{X}^2 = \frac{1}{2}\sum_{i=1}^{n}\frac{(x_i - y_i)^2}{(x_i + y_i)} \tag{1}$$

where $x = x_0, x_1, ..., x_{18}$, $y = y_0, y_1, ..., y_{18}$, are the vectors that contain the ratio of each class in the corresponding dataset: $\sum x_i = 1$, and $\sum y_i = 1$. A chi-square value of 0 means that the two class distributions match perfectly, while 1 means that the two datasets share no common classes.

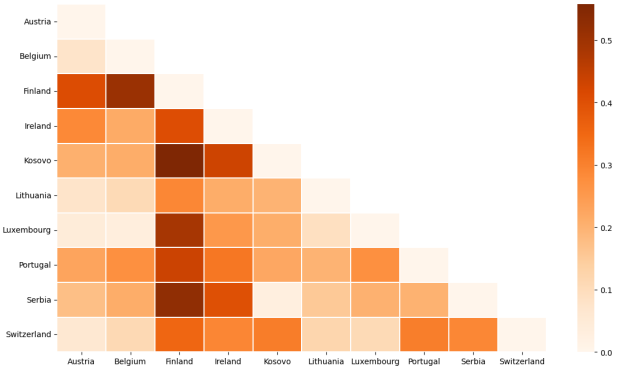


Figure 3. Chi-square distances between the class distribution of all countries contained in the BigEarthNet dataset. We see that the class distribution for Finland significantly differs from all other class distributions.

4. CLASSIFICATION MODEL AND MOTIVATION FOR TRANSFER LEARNING

In multi-label image classification, it is important to learn the features of the image and also to establish the relationship between the labels. For example, let us consider an aerial image of a ship sailing in an ocean. For the model, it is important to learn the features of the ship and the ocean and also to learn

that the probability of a ship class is higher when ocean class is predicted.

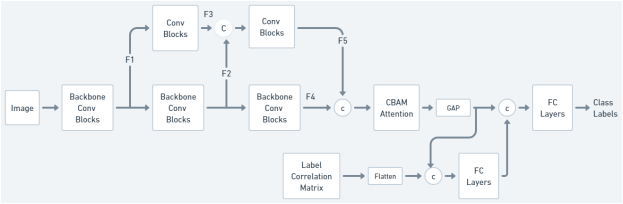


Figure 4. Sketch of the network architecture used for our experiments. We use the network architecture proposed in Huang et al., but replace the backbone with the EfficientNet-B5 architecture.

To find a good feature extractor, we evaluated the performance of AlexNet (Krizhevsky et al., 2012), variants of ResNet (He et al., 2016) and EfficientNet (Tan and Le, 2019) classifiers, and we found that the EfficientNet classifiers performed better than the others. We chose the EfficientNet-B5 variant to balance between performance and model size. We then replace the VGG16 (Simonyan and Zisserman, 2014) backbone used in the architecture proposed by Huang et al. (Huang et al., 2021) with the EfficientNet-B5 backbone, as shown in Figure 4. To exploit the features learned with different receptive fields, the features from different stages of the backbone are extracted. Features $F1$ are extracted from the network when the spatial size of the features is equal to half of that of the input. Features $F2$ are extracted from the network when the spatial size of the features is equal to one-fourth of that of the input. $F1$ is then passed through a block convolution layer, followed by batch normalization and a rectified linear unit (ReLU) to match the size of $F2$. The output of this block is defined as $F3$. $F2$ and $F3$ are then concatenated and passed through two sets of convolution layers, batch normalization and a rectified linear unit (ReLU) to match the size of the final features from the backbone network. These features are then passed through a Convolutional Block Attention Module (CBAM) (Woo et al., 2018) and a pooling layer. To help the model learn the correlations between the labels, the correlation matrix is separately calculated and passed through fully connected layers. The output of these layers is concatenated with the pooled features and used for classification.

For all the experiments presented in this paper, we used the Stochastic Gradient Descent optimizer with a learning rate of 0.0571. We determined this learning rate using a range test implemented in the Learning Rate Finder proposed by Smith et al.(Smith, 2015). This implementation uses a base learning rate and increases the learning rate by some factor for each mini-batch until the loss gets worse. Then we select a learning rate slightly lower than the rate at which the loss gets worse.

The following classification metrics are calculated for each class to evaluate the performance of the classification:

$$Precision, P = \frac{TP}{TP + FP} \tag{2}$$

$$Recall, R = \frac{TP}{TP + FN} \tag{3}$$

$$\mathrm{F_2\ Score} = \frac{5}{\frac{4}{P} + \frac{1}{R}} \tag{4}$$

where TP: True Positives, TN: True Negatives, FP: False Positives, and FN: False Negatives of the classification.

To adapt to the multi-class setting, these metrics can be averaged in different ways. In this work, three ways of averaging the metrics are used. They are:

- Micro averaging: calculates metrics globally by counting the total true positives, false negatives and false positives;
- Macro averaging: calculates metrics for each class and computes their unweighted mean; and,
- Samples averaging: calculates metrics for each instance, and computes their average.

Hamming Loss (L_H), Coverage (*cov*), and Label Ranking Average Precision (*lrap*) are also calculated to evaluate the performance of the classification. Given y, the true labels, and $\hat{y}$, the predicted labels:

$$L_H(y, \hat{y}) = \frac{1}{n * m} \sum_{i=0}^{n-1} \sum_{j=0}^{m-1} 1(\hat{y}_{ij} \neq y_{ij}) \tag{5}$$

$$cov(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} \max_{j:y_{ij}=1} rank_{ij} \tag{6}$$

$$lrap(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{-1} \frac{1}{\|y_i\|_0} \sum_{j:y_{ij}=1} \frac{\mathcal{L}_{ij}}{rank_{ij}} \tag{7}$$

where n is the number of samples, m is the number of classes, y_{ij} represents the prediction score for class i and sample j and,

$$rank_{ij} = \left|\left\{k : \hat{y}_{ik} \geq \hat{y}_{ij}\right\}\right|$$

$$\mathcal{L}_{ij} = \left\{k : y_{ik} = 1, \hat{y}_{ik} \geq \hat{y}_{ij}\right\}$$

, $|\cdot|$ computes the set cardinality and the $\|\cdot\|_0$ l_0 norm.

	MAML SiB_RGB	Our Model
f_2 micro (%)	49.8	**61.7**
f_2 samples (%)	56.0	**66**
f_2 macro (%)	35.1	**45.1**
recall micro (%)	45.6	**59**
recall samples (%)	53.7	**65.5**
recall macro (%)	33.6	**42.8**
hamming loss (%)	**4.8**	9
coverage	**7.1**	12.96
lrap (%)	**80.0**	62.2
Parameters in millions	**1.1**	29.27

Table 2. Comparison of the classification results with the baseline provided with the BigEarthNet dataset for RGB images (MAML_SiB_RGB)

Our LULC classification model as shown in Table 2 outperforms the baseline model MAML SiB_RGB by Sumbul et al. (Sumbul and Demİr, 2020) in terms of f_2 scores and recall. Therefore, we are considering this network for further experiments in this work.

To investigate the importance of images from a region during training, we trained the classification model on datasets from Kosovo, Finland and Portugal and the entire dataset without images from Kosovo (referred to as full \kosovo) and evaluated the four models using the test split from the Kosovo dataset. We have plotted the f_2 scores of the classification in Figure 5. As we can see, the results were significantly better when the model

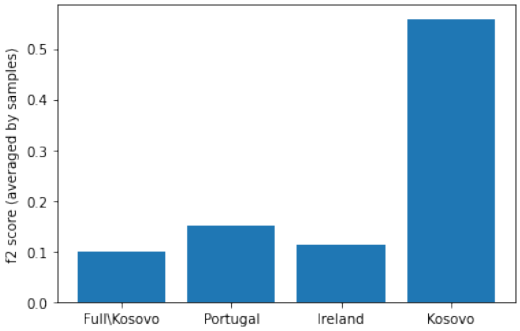


Figure 5. Comparison of supervised training with different region-to-region transfer learning scenarios. In all cases, the testing dataset contains only samples from Kosovo. We compare the f_2 score for different training datasets averaged by samples. For the first bar "Full\Kosovo" the training dataset contains all training samples from all countries except Kosovo. For the second bar training samples from Portugal are used and for the third bar training samples from Ireland are used. The last bar represents regular supervised training with training data from Kosovo.

was trained and tested using the Kosovo dataset, even though the training split of Kosovo dataset is much smaller compared to the other datasets. This indicates that it is important for the model to see images from a particular region during training to classify the images from that region. It also suggests that the images from different regions cannot be treated as equal, even though they are captured similarly using the same sensors and belong to the same class. This indicates that a LULC classification model trained using images from one region does not generalize well. To address this challenge, we are exploring inter-region transfer learning. Our results indicate that the similarity of labels and images is crucial for the performance of region-to-region transfer learning. In the following section, we perform a more thorough experimental analysis of this statement.

5. TRANSFER LEARNING EXPERIMENTS AND RESULTS

To apply transfer learning, we pre-train the model using a large 'source' dataset and then use a few samples from the 'target' dataset to update all or some weights of the model. For our experiments, we use the following set of source regions. These country combinations have been selected to ensure a variety of different dataset sizes and different Chi-square distances to the target domains.

1. IALi: Ireland + Austria + Lithuania
2. IALiSe: Ireland + Austria + Lithuania + Serbia
3. IALiSeP: Ireland + Austria + Lithuania + Serbia + Portugal
4. IALiSeF: Ireland + Austria + Lithuania + Serbia + Finland
5. IALiSePF: Ireland + Austria + Lithuania + Serbia + Portugal + Finland
6. BLuSw: Belgium + Luxembourg + Switzerland
7. KLuSw: Kosovo + Luxembourg + Switzerland

We choose two target datasets, the Kosovo dataset, which is a dataset of a landlocked country with 1737 images and the Belgium dataset, which is around 5 times larger than the Kosovo dataset and also contains images with classes like 'marine waters'. For both these datasets, we perform supervised training, which we use as the baseline for comparison. For the rest of the experiments, we train the model on each of the combinations mentioned above and evaluate them in the following three ways:

- Direct testing: Test the classification performance on the testing dataset of the target country;
- Fine-tuning: Initialize the model with the pre-trained weights and train all the layers in the model using the training dataset of the target country and then test the classification performance on the testing dataset of the target country,
- Linear probing: Initialize the model with the pre-trained weights and train only the fully connected layers with the training dataset of the target country, and test the classification performance on the testing dataset of the target country.

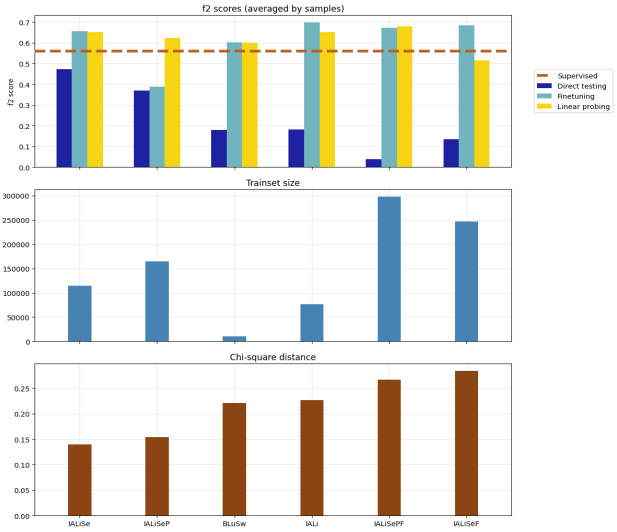


Figure 6. Comparison of the performance of different transfer learning scenarios with the Kosovo dataset as the target dataset. In the first row, the f_2 scores averaged by samples are plotted for supervised training and the evaluation methods that are explained in Section 5. In the second row, the size of the different combinations of training data is shown. In the last row, we show the Chi-square distance of the corresponding combination from the Kosovo dataset. The abbreviations in the x-axis are explained in Section 5.

The results of transfer learning experiments on the Kosovo dataset are plotted in Figure 6. In the following, we analyse the results of the transfer learning approaches.

5.1 Direct testing

In congruence with the results shown in Figure 5, we can see in Figure 6 that the classification performance is significantly lower than the supervised learning baseline when the model has not seen samples from the target dataset during training. We see a sharp decline when we add Portugal and Finland to

the training data, for example in the combinations 'IALiSeF' and 'IALiSePF'. To understand this decline, it helps to see how the classes are distributed in these datasets. While we are significantly increasing the size of the training dataset, we are also making the class distribution increasingly different from the target class distribution, as can be seen by the increased Chi-squared distances in Figure 6. These results confirm our intuition that direct testing works well for similar class distributions between the source and target domains.

In Figure 2, we can see the class distribution of the Finland dataset. Even though the dataset is more than 100 times larger than the Kosovo dataset, the class imbalance tends to make the model biased towards those classes that are dominant in the Finland dataset, which explains the worst performance with 'IALiSeF' and 'IALiSePF' as the training dataset. It is also noticeable that the performance of the model is better when trained on the combination 'IALiSe' and 'BLuSw' as compared to training on 'IALiSeF', although the 'IALiSeF' dataset is much larger. The best performance is indeed achieved by using 'IALiSe' as the training dataset, which has the least Chi-squared distance.

5.2 Fine-tuning

We can see from the plots in Figure 6 that fine-tuning the model with the training dataset from the target country has improved the performance in most cases compared to the supervised training with the target dataset. We see an exception when the model was pre-trained with the combination 'IALiSeP'. Portugal is the only dataset which has samples for the class 'Agro-Forestry Areas' and around 31% of the images are annotated with this class. Our inference is that the model found a local minimum during the training on the source dataset, and the number of samples in the target domain is not enough for generalization in the target domain; thus, the performance during the test time in the target domain is poor.

It must also be noted that pre-training with the smallest combination 'BLuSw' and the other combinations 'IALi' and 'IALiSe' yields similar performance compared to pre-training on datasets of much larger scales. We assign this to the fact that the difference between the class distribution of Kosovo and these combinations is small. The results confirm our statement that a similar class distribution leads to better performance.

5.3 Linear probing

After linear probing, the model's performance is better than the supervised learning in all cases, except when Finland is added to the combination.

As we see in Figure 2, the Finland dataset contains a disproportionately large number of samples for a small set of classes, while other classes are almost completely missing. This limits the model's ability to learn good representations of these classes and makes the model biased towards frequently occurring classes. This effect is not very evident during fine-tuning because the model was able to learn features related to the missing classes while fine-tuning, whereas in the linear probing phase, we only update the parameters of the fully connected layers. It is also evident that the pre-training combinations 'BLuSw','IALi' and 'IALiSe' yields similar performance compared to pre-training on datasets of much larger scales, consistently confirming our interpretation.

From the results, we can conclude that directly using a model which is trained on a dataset from a certain region may not work well for inference in a new region. It is also observed that a few examples from a similar class distribution are similarly helpful as many examples from a different class distribution. Finally, we note that in our experiments, linear probing was more robust than fine-tuning, with smaller drops in performance for source combinations that did not work well.

6. DOMAIN ADAPTATION

Domain adaptation is a subclass of methods under transfer learning, which learns a good feature extractor when there is a shift in distribution between the training and testing sets. We discussed in the previous sections how similar class distribution of the source and the target domain is crucial for the performance of region-to-region transfer learning. While the classical methods limit themselves to using weights from the pre-training on the source domain, domain adaptation methods can also learn in an unsupervised setting where there are no labels in the target domain. In this section, we explore two such methods for LULC classification.

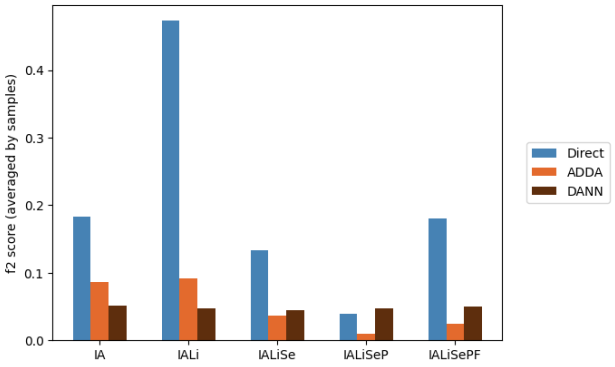


Figure 7. Results of Domain Adaptation on Kosovo dataset. We plot the f_2 scores averaged by samples for several different combinations of source countries. We compare the results of two advanced domain adaptation techniques (ADDA and DANN) with the results obtained by directly testing the model in the target region, as indicated in the legend. The abbreviations are explained in Section 5.

Ganin et al. (Ganin et al., 2016) proposed a framework called 'Domain-Adversarial Training of Neural Networks (DANN)' to minimize the difference between the feature representations of source and target domain using adversarial training. The objective is to minimize the loss of the label predictor which predicts the class of the image and to maximize the loss of the domain classifier which classifies the image as source domain or target domain. A gradient reversal layer is used to negate the gradients flowing from the domain classifier, making it possible to train the entire network using standard backpropagation. This allows the model to learn features that are invariant of the domain yet discriminative for the main learning task. We train our model with this framework to investigate if it can be successfully applied to our region-to-region transfer learning goal.

Tzeng et al. (Tzeng et al., 2017) proposed another adversarial framework called 'Adversarial Discriminative Domain Adaptation (ADDA)' which trains a feature extractor

in two stages. In the first stage, a feature extractor for the source domain is learned to perform well on the main learning task with a label predictor. Then the feature extractor is treated like a Generator and a domain classifier is considered as a discriminator. The Generator is trained with inverted labels such that the discriminator cannot distinguish between source and target domains. The Generator is then used with the label predictor to make predictions in the target domain.

We apply both frameworks discussed above to the LULC classification task. In Figure 7 we have plotted the results, and we see that the unsupervised domain adaptation methods performed poorly in the target region. Xia et al. (Xia et al., 2022) performed region-to-region domain adaptation, and they showed that most adversarial training methods did not perform well due to the diversity in the data and also highlighted the large content and style gap. They also showed that model performance decreased significantly for land cover class as opposed to land use classes. Our finding confirms this statement and therefore motivates the investigation of domain adaptation techniques specific to earth observation application in future work.

7. CONCLUSIONS AND OUTLOOK

Land Use Land Cover classification is an essential task in Earth Observation as it helps us understand land use, which has huge ecological effects. Since the terrain changes from region to region, we investigate the applicability of classification models on regions whose images are not seen during training. We find that (a) a classification model trained in one region would not be useful for inference in a new region because of the decline in the classification performance. (b) the chi-squared distance between the class distribution can be used as an indicator to estimate the performance of the model in the new region and the suitability of the dataset as a source domain for transfer learning (c) when we pre-train the model on a dataset with a significant number of samples from a new class, the model may be trapped in a local minimum and may not be able to generalize well during fine-tuning of the model, rendering the transfer learning ineffective, and (d) when probing the linear layers of the model, pre-training with datasets that have similar class distribution requires fewer samples during pre-training compared to a big dataset with a different class distribution. The results of this study have also shown that a significantly big pre-training dataset does not necessarily mean equivalently better classification performance. We also highlight that the similarity between the class distribution of the source and the target dataset can lead to better transfer learning results.

We also applied adversarial domain adaptation methods that learn domain invariant features to improve classification performance in the target domain. We observe that the classification performance is worse than inference in the target domain without transfer learning. This may be due to the challenges of remote sensing images discussed earlier. We, therefore, encourage the development of domain adaptation techniques that are specific to EO and consider the specifics of remote sensing images as well as the geography of the region under consideration. Furthermore, content-based image similarity methods for selecting a pre-training dataset can be explored, as solely depending on the class distribution might fail in cases where datasets have similar classes but varying representations. A similar analysis must be performed to see how well we can transfer the learning between different tasks, such as regression for yield prediction, for example. This would be a significant contribution to leveraging the potential of a large amount of freely available EO data.

ACKNOWLEDGEMENTS

This project was partly funded through the ESA InCubed Programme (`https://incubed.esa.int/`) as part of the project AI4EO Solution Factory (`https://www.ai4eo-solution-factory.de/`).

REFERENCES

Aksoy, A., Dushev, P., Tzirita Zacharatou, E., Hemsen, H., Charfuelan, M., Quiané-Ruiz, J.-A., Demir, B., Markl, V., 2022. Satellite image search in AgoraEO. *Proceedings of the VLDB Endowment*, 15, 3646-3649.

Castelluccio, M., Poggi, G., Sansone, C., Verdoliva, L., 2015. Land Use Classification in Remote Sensing Images by Convolutional Neural Networks. *CoRR*, abs/1508.00092. http://arxiv.org/abs/1508.00092.

Cheng, G., Xie, X., Han, J., Guo, L., Xia, G., 2020. Remote Sensing Image Scene Classification Meets Deep Learning: Challenges, Methods, Benchmarks, and Opportunities. *CoRR*, abs/2005.01094. https://arxiv.org/abs/2005.01094.

Costa, H., Benevides, P., Moreira, F. D., Moraes, D., Caetano, M., 2022. Spatially Stratified and Multi-Stage Approach for National Land Cover Mapping Based on Sentinel-2 Data and Expert Knowledge. *Remote Sensing*, 14(8). https://www.mdpi.com/2072-4292/14/8/1865.

Fuller, A., Millard, K., Green, J. R., 2022. Transfer Learning with Pretrained Remote Sensing Transformers. *arXiv preprint*. https://arxiv.org/abs/2209.14969.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., Lempitsky, V., 2016. Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research*, 17(59), 1–35. http://jmlr.org/papers/v17/15-239.html.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778.

Huang, R., Zheng, F., Huang, W., 2021. Multilabel Remote Sensing Image Annotation With Multiscale Attention and Label Correlation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 6951-6961.

Jardim, A. M. d. R. F., Araújo Júnior, G. d. N., Silva, M. V. d., Santos, A. d., Silva, J. L. B. d., Pandorfi, H., Oliveira-Júnior, J. F. d., Teixeira, A. H. d. C., Teodoro, P. E., de Lima, J. L. M. P., Silva Junior, C. A. d., Souza, L. S. B. d., Silva, E. A., Silva, T. G. F. d., 2022. Using Remote Sensing to Quantify the Joint Effects of Climate and Land Use/Land Cover Changes on the Caatinga Biome of Northeast Brazilian. *Remote Sensing*, 14(8). https://www.mdpi.com/2072-4292/14/8/1911.

Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. F. Pereira, C. Burges, L. Bottou, K. Weinberger (eds), *Advances in Neural Information Processing Systems*, 25, Curran Associates, Inc.

Penatti, O. A. B., Nogueira, K., dos Santos, J. A., 2015. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 44–51.

Phiri, D., Simwanda, M., Salekin, S., Nyirenda, V. R., Murayama, Y., Ranagalage, M., 2020. Sentinel-2 Data for Land Cover/Use Mapping: A Review. *Remote Sensing*, 12(14). https://www.mdpi.com/2072-4292/12/14/2291.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. S., Berg, A. C., Fei-Fei, L., 2014. ImageNet Large Scale Visual Recognition Challenge. *CoRR*, abs/1409.0575. http://arxiv.org/abs/1409.0575.

Silva-Perez, C., Marino, A., Cameron, I., 2020. Monitoring Agricultural Fields Using Sentinel-1 and Temperature Data in Peru: Case Study of Asparagus (Asparagus officinalis L.). *Remote Sensing*, 12(12). https://www.mdpi.com/2072-4292/12/12/1993.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition.

Smith, L. N., 2015. No More Pesky Learning Rate Guessing Games. *CoRR*, abs/1506.01186. http://arxiv.org/abs/1506.01186.

Stojnić, V., Risojević, V., 2021. Self-supervised learning of remote sensing scene representations using contrastive multiview coding. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1182–1191.

Sumbul, G., Charfuelan, M., Demir, B., Markl, V., 2019. BigEarthNet: A Large-Scale Benchmark Archive for Remote Sensing Image Understanding. 5901–5904.

Sumbul, G., Demİr, B., 2020. A Deep Multi-Attention Driven Approach for Multi-Label Remote Sensing Image Classification. *IEEE Access*, 8, 95934-95946.

Tan, M., Le, Q., 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 97, PMLR, 6105–6114.

Tzeng, E., Hoffman, J., Saenko, K., Darrell, T., 2017. Adversarial discriminative domain adaptation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society, Los Alamitos, CA, USA, 2962–2971.

Vali, A., Comai, S., Matteucci, M., 2020. Deep Learning for Land Use and Land Cover Classification Based on Hyperspectral and Multispectral Earth Observation Data: A Review. *Remote Sensing*, 12(15). https://www.mdpi.com/2072-4292/12/15/2495.

Walsh, E., Bessardon, G., Gleeson, E., Ulmas, P., 2021. Using machine learning to produce a very high resolution land-cover map for Ireland. *Advances in Science and Research*, 18, 65–87. https://asr.copernicus.org/articles/18/65/2021/.

Woo, S., Park, J., Lee, J.-Y., Kweon, I., 2018. *CBAM: Convolutional Block Attention Module: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VII*. 3–19.

Xia, J., Yokoya, N., Adriano, B., Broni-Bediako, C., 2022. Openearthmap: A benchmark dataset for global high-resolution land cover mapping.

Yuan, J., Chen, J., Sciusco, P., Kolluru, V., Saraf, S., John, R., Ochirbat, B., 2022. Land Use Hotspots of the Two Largest Landlocked Countries: Kazakhstan and Mongolia. *Remote Sensing*, 14(8). https://www.mdpi.com/2072-4292/14/8/1805.

Zhang, Y., Kang, B., Hooi, B., Yan, S., Feng, J., 2021. Deep Long-Tailed Learning: A Survey. *CoRR*, abs/2110.04596. https://arxiv.org/abs/2110.04596.

Zong, L., He, S., Lian, J., Bie, Q., Wang, X., Dong, J., Xie, Y., 2020. Detailed Mapping of Urban Land Use Based on Multi-Source Data: A Case Study of Lanzhou. *Remote Sensing*, 12(12). https://www.mdpi.com/2072-4292/12/12/1987.

APPENDIX

Experimental setup

All the experiments on the classification, fine-tuning and linear probing were performed on NVIDIA V100 GPUs with 16 GB VRAM. The training was run for 100 epochs, stopping early if the validation loss did not decrease for 20 consecutive epochs. Consequently, fine-tuning and linear probing was also run for a maximum of 100 epochs with the same early stopping conditions. We experimented with a smaller learning rate of 0.00571 during fine-tuning of an experiment and found out that it was worse compared to the learning rate found using the LR finder as described by Leslie N. Smith (Smith, 2015). The model has 29.27 million parameters and during the linear probing experiments, only 219.62 thousand parameters are updated. Due to numerous experiments, this work does not focus on hyperparameter tuning.

LULC Classes in BigEarthNet

Index	Class Name
0	Urban fabric
1	Industrial or commercial units
2	Arable land
3	Permanent crops
4	Pastures
5	Complex cultivation patterns
6	Land principally occupied by agriculture, with significant areas of natural vegetation
7	Agro-forestry areas
8	Broad-leaved forest
9	Coniferous forest
10	Mixed forest
11	Natural grassland and sparsely vegetated areas
12	Moors, heathland and sclerophyllous vegetation
13	Transitional woodland, shrub
14	Beaches, dunes, sands
15	Inland wetlands
16	Coastal wetlands
17	Inland waters
18	Marine waters

Table 3. The classes in BigEarthNet dataset