Progressive Unsupervised Deep Transfer Learning for Forest Mapping in Satellite Image

Nouman Ahmed[1], Sudipan Saha[2], Muhammad Shahzad[1,2],
Muhammad Moazam Fraz[1], Xiao Xiang Zhu[2,3]
[1]School of Electrical Engineering and Computer Science (SEECS),
National University of Sciences and Technology (NUST), Islamabad, Pakistan
[2]Data Science in Earth Observation, Department of Aerospace and Geodesy,
Technical University of Munich (TUM), Munich, Germany
[3]Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling, Germany
{nahmed.bese16seecs;moazam.fraz}@seecs.edu.pk, {sudipan.saha;muhammad.shahzad}@tum.de,
xiaoxiang.zhu@dlr.de

Abstract

Automated forest mapping is important to understand our forests that play a key role in ecological system. However, efforts towards forest mapping is impeded by difficulty to collect labeled forest images that show large intra-class variation. Recently unsupervised learning has shown promising capability when exploiting limited labeled data. Motivated by this, we propose a progressive unsupervised deep transfer learning method for forest mapping. The proposed method exploits a pre-trained model that is subsequently fine-tuned over the target forest domain. We propose two different fine-tuning mechanism, one works in a totally unsupervised setting by jointly learning the parameters of CNN and the k-means based cluster assignments of the resulting features and the other one works in a semi-supervised setting by exploiting the extracted k-nearest neighbor based pseudo labels. The proposed progressive scheme is evaluated on publicly available EuroSAT dataset using the relevant base model trained on BigEarthNet labels. The results show that the proposed method greatly improves the forest regions classification accuracy as compared to the unsupervised baseline, nearly approaching the supervised classification approach.

1. Introduction

Forest mapping is an important process that helps to measure the deforestation and quantify its impact on the global climate. Fueled by the launch of many satellites by different space agencies, optical/radar sensors mounted over satellites provide the ability to map forest cover change both on local and global scales. In this context, Convolutional Neural Networks (CNNs) have recently become the defacto method for image classification [1]. However, forest images show large intra-class variation and often show significant spectral resemblance to non-forest images. Thus, annotating forest images requires domain expertise, impeding the acquisition of large scale labeled datasets and application of supervised methods in forest mapping [2].

Transfer learning is often adopted in data scarce situations by adopting a pre-trained model for another task [3]. Several CNN models pre-trained on large-scale ImageNet dataset [4] are available, which are generally used for transfer learning based image classification tasks [5]. However, ImageNet contains images of natural objects pertaining to everyday life. The learned feature representation are thus specific to such objects and do not adapt or generalize well for different domain data (e.g., satellite images showing forest) owing to the contrasting target distribution. For this purpose, to build an effective transfer learning based pipeline for forest classification, it is essential to train the CNNs on large annotated dataset of diverse satellite imagery. However inter-dataset variation is generally quite prominent in remote sensing [6], making mere transfer learning insufficient for forest mapping on new datasets. Moving beyond mere transfer learning, unsupervised learning can potentially tackle with the issues pertaining to supervised methods by extracting implicit patterns directly from the unlabeled input data. Techniques employing clustering, dimensionality reduction and density estimation have been widely used in computer vision [7] and remote sensing applications [8, 9]. Recently, few works [10, 11, 12] have shown the possibility of adapting the unsupervised methods based on clustering to deep models [13].

Such adaptation together with clustering allows extracting distinctive visual features that can be used for unsupervised learning of deep models.

Inspired by such strategies [7, 10, 11], this paper formulates the problem of forest mapping in an combined transfer learning and unsupervised learning paradigm by proposing a progressive learning framework enabling transfer of previously learned representations to the unlabeled data. For this purpose, a base CNN is first trained to extract useful features relevant to the target domain distribution that can later be used to fine-tune either in a totally unsupervised setting by jointly learning the parameters of CNN and the k-means based cluster assignments of the resulting features or in a one-shot semi-supervised scenario by exploiting the extracted k-nearest neighbor based pseudo labels. In either of the settings, no additional labeled data is required to learn domain-specific representations which makes them suitable for the problem of forest classification with scarce annotations. The main contributions proposed in this work are two-fold:

- First, a progressive unsupervised algorithm for forest mapping has been proposed which performs iterative CNN learning using the features extracted either over unlabeled data only or using both the available labeled data as well as the unlabeled data. The former setting enables simultaneous learning of the network together with the cluster assignments while the latter adopts a dynamic sampling approach to exploit the unlabelled data. To the best of our knowledge, such a cascaded formation in both the settings has not been adopted in the remote sensing forest classification.

- Secondly, the proposed unsupervised approach in both the settings is evaluated on images with various spectral modalities, i.e., RGB, RGB + Near Infrared (NIR) and five commonly used vegetation indices. The presented results demonstrates that the results improve after employing the proposed iterative framework with semi-supervised learning performing a slightly better than the totally unsupervised scenario with both approaching to supervised learning accuracy.

2. Related Works

A substantial amount of work has been done to classify remote sensing imagery in a supervised manner [14, 15] for different applications including forest classification and mapping [16, 17, 18]. However, considering relevance to our work, in this Section we detail the works on semi-supervised classification and unsupervised Classification.

Semi-supervised classification. Wu et al. [19] made use of the fact that hyperspectral images contain addtional spectral information to build a self-training classification system which made use of clustering and some spectral constraints to regulate the process. Negri et al. [20] compared Semi-supervised Support Vector Machive (S3VM) [21] and Expectation Maximization (EM) [22] for semi-supervised classification of remote sensing imagery by using data labelled by Fuzzy C-Means with high level of confidence. Meher [23] proposed a semi-supervised method with Granular Neural Networks (GNNs) [24] as the base classifier because of its decreased complexity in comparison to CNNs and further enhanced the model with fuzzy granulation of features using class belonging information and selection of granulated features using neighborhood rough sets.

Unsupervised Classification. Due to unavailability of labeled remote sensing data, recent works have focused on developing unsupervised and self-supervised methods in the computer vision literature. Many such methods use some pre-text task, e.g., image rotation [25] to learn semantic feature in an unsupervised way. Another effective approach towards this is deep clustering that jointly learns the parameters of the model and cluster assignment of input features [26]. Some methods are based on concept of contrastive learning [27]. Following the trend, a few unsupervised methods have been developed in the remote sensing literature [28, 3, 29]. Saha *et. al.* [28] use mutli-temporal image ordering as pre-text task for self-supervised learning. [29] proposed an unsupervised learning algorithm to cluster hybrid polarimetric SAR images, and dual-polarized SAR using the VGG16 [30] model.

The aforementioned unsupervised techniques are able to exploit the unlabeled data, however they are not designed to exploit forest remote sensing data. Most pre-text tasks like rotation are not effective in case of forest remote sensing data. Moreover, mere deep clustering may not be always sufficient to learning discriminative features from limited forest images. Another challenging aspect is that forest data (and associated classes) may vary from region to region. To address these issues, the paper proposes a scheme that utilizes the advantages of previously learned representations (pre-trained model weights) in conjunction with unsupervised techniques to progressively fine-tune network architecture using unlabelled data. The details of the proposed progressive and unsupervised learning strategy to learn new weights using unseen data is presented in the next Section.

3. Proposed Methodology

Let us assume that we have a set of unlabeled forest images $\mathcal{X} = \{x_1, x_2, ..., x_N\}$. Our objective is to assign each image in $\mathcal{X}$ to either ω_f (forest) or ω_{nf} (non-forest). We assume that a relevant base CNN model θ is available trained on another related but different labeled dataset $\mathcal{X}^l = \{x_1^l, x_2^l, ..., x_M^l\}$. The proposed forest region classification method can work under two different scenarios:

1. Using only the pre-trained network model θ, an iterative progressive unsupervised learning framework is employed.

2. Using the pre-trained network model θ along with its training dataset $\mathcal{X}^l$, a semi-supervised learning strategy is used.

Both the learning strategies later adopt an iterative progressive training procedure that includes populating the training set with the pseudo labels assigned to the available unlabeled images. The strategy to assign these pseudo labels to the unseen data is however different. When no labeled data is available, then we make use of the unsupervised clustering procedure to group images based on extracted features and later assign the cluster ID sequence numbers as the pseudo labels to the unlabeled images. Similarly, for the semi-supervised case, the extracted features from the available labeled images are used to assign the pseudo labels via *k*-nearest neighbors to the unlabeled images. In both cases, the unlabeled data with the assigned pseudo labels are further refined by incorporating certain heuristics constraints to ensure robust and progressive model training. The whole procedure is iteratively performed where in each iteration, the training sample set increases with more and more selected pseudo labels assigned to it enabling self-paced learning till convergence. Figure 1 presents the working procedure of the proposed forest region classification method.

3.1. Model Initialization

The proposed methodology includes a base shallow CNN θ trained on a relevant remote sensing labeled dataset $\mathcal{X}^l$ to learn weights w_t. The whole idea of using such a pre-trained model is so that information can be transfer learned (features and weights learned from the labeled dataset) to the proposed unsupervised forest classification technique which is further trained on new and unlabeled/unseen dataset. Thus, if a labeled remote sensing dataset $\mathcal{X}^l$ is given, the idea of model initialization is to fine tune any generic feature extractor like VGG16 [31] or even a custom shallow network where the last layer of the model is a fully connected layer having softMax activation and the output neurons equals to the number of forest classifications, i.e., two in our case (ω_f and ω_{nf}).

3.2. Unsupervised Model Training

After the model has been successfully initialized using $\mathcal{X}^l$, it is used for inference and and fine-tuning on the unlabeled dataset which includes feature extraction, pseudo labels assignment and reliable images selection to be explained in the subsequent sections.

3.2.1 Feature Extraction

In this step, the pre-trained model θ is used to extract the features from the labeled dataset $\mathcal{X}^l$ and the unlabeled dataset $\mathcal{X}$ in order to perform clustering (unsupervised) and *k*NN(semi-supervised). The output of the max pooling layer, prior to the fully connected layer in the shallow network, is considered as the features of the input image i.e, the fully connected layer of the network is removed and the max pooling layer is used as output to obtain the features of the input image. This can be further elaborated by the following equation:

$$f_j = \theta(x_j, w_t) \tag{1}$$

where x_j represents an image and extracted feature vectors are represented by f_j. Features are extracted for all the input images in both $\mathcal{X}$ and $\mathcal{X}^l$. In every subsequent iteration, features are extracted again but with the newly fine-tuned model which is considered more accurate and will yield more robust and useful features.

3.2.2 Pseudo labels Assignment

After successfully extracting the features from the labeled and unlabeled datasets, we populate the training set by assigning pseudo labels to the available unlabeled images. The procedure for assigning these labels depends on the availability of labelled data and is mentioned below:

Only θ is used

Even though θ is trained on $\mathcal{X}^l$, in many practical settings $\mathcal{X}^l$ may not be available for future tuning. As an example, such situation may arise in partnership between two organizations, where the organization owning $\mathcal{X}^l$ is not authorized to share the images.

In such case, after the features have been successfully extracted from the input images in $\mathcal{X}$ using θ, they are fed to an unsupervised clustering algorithm that in principle groups the similar images together in order to generate a relatively more refined training sample. In our case, the features are grouped into two clusters i.e, forest and non forest. In order to achieve that, the extracted features for all unlabeled images are fed into the standard *k*-means algorithm that clusters the similar features together. This enables us to formulate the basis for progressive training in a sense that the clustered group of images are assigned the same pseudo-labels which when iteratively used for training improves the accuracy of model and thus the feature extraction in each subsequent iteration. In this case, *k*-means allows us to extract two cluster centroids c_1 (for forest images) and c_2 (for non-forest images). *k*-means attempts to minimize

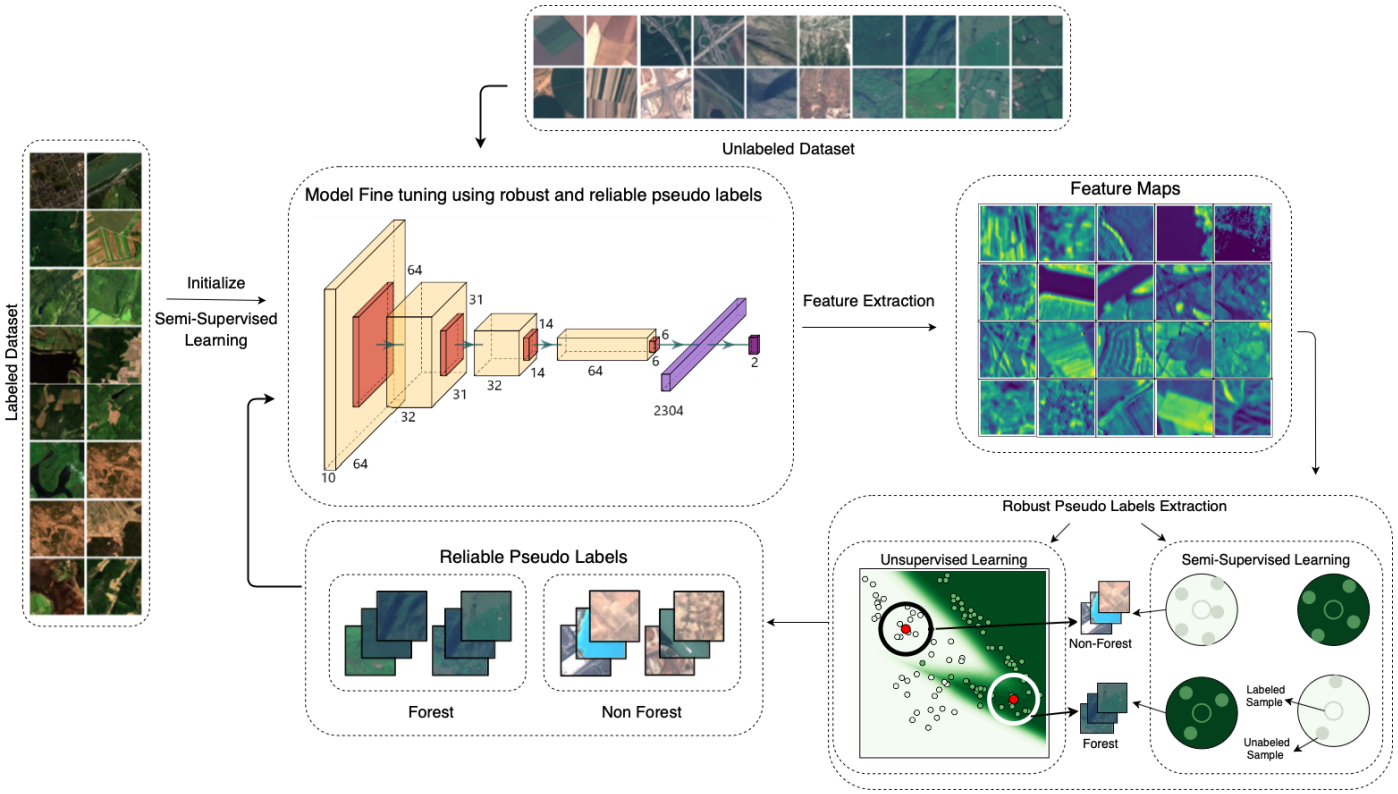

Figure 1: Illustration of the proposed progressive loop. The above CNN model is initialized using a relevant labeled dataset and the model is used to extract features from the labeled and the unlabeled dataset. In case of no labeled data, pseudo labels are assigned using unsupervised clustering which groups images based on the extracted features and assign the cluster ID as the pseudo label. For semi-supervised learning, the extracted features from the labelled dataset are used to assign pseudo labels using k-nearest neighbors to the unlabelled images. Both cases employ certain heuristics constraints to enable robust selection of pseudo labels. Lastly, model is fine-tuned using these labels till convergence.

the following optimization function:

$$C \leftarrow \arg\min_{c_j} \sum_{i=1}^{n} \sum_{j=1}^{2} | f_i - c_j |^2 \qquad (2)$$

As it is known that k-means cluster similar images together but due to various reasons many images can get wrong cluster assignments. This can happen because the model is not yet fully trained on the unlabeled dataset $\mathcal{X}$ and thus is not perfectly able to extract the desired features. These wrong cluster assignments can make the proposed formulation more susceptible to error. All this points to refining the obtained clustering results in order to improve the accuracy of the model.

Both θ and $\mathcal{X}^l$ are available

If the labeled data $\mathcal{X}^l$ is available during fine-tuning, we adopt the k Nearest Neighbors (kNN) classifier for the label estimation. For our purposes, the kNN classifier in the feature space may be a better choice, since similar input data always have similar feature representations. The kNN classifier assigns the label of each unlabeled image by its nearest labeled neighbors in the feature space. We define the confidence of label estimation as the distance between the unlabeled data and its nearest labeled neighbor. For the candidates selection, we select some of top reliable pseudo labeled data according to their label estimation confidence.

Formally, we define the dissimilarity cost (label estimation confidence) for each x_j in unlabelled dataset as:

$$d_j \leftarrow \min_{x_i} \| \theta(x_i^l, w_t) - \theta(x_j, w_t) \| \qquad (3)$$

where x_j represents the input unlabeled image, x_i^l represents an image in the labeled dataset $\mathcal{X}^l$, θ represents the relevant base model with learned weights w_t which outputs the extracted feature vectors. The cost is the l_2 distance between the unlabeled image x_j and the labeled image x_i. This cost acts as a criterion for confidence estimation for our pseudo-labeled data. The higher the cost, lower the confidence.

3.2.3 Reliable images Selection

In both the techniques, we apply certain heuristics to make sure we fine-tune the model with the most accurate pseudo-labels. For unsupervised technique, similarity is calculated between all the features and the cluster centroids and only those features having similarity higher than a pre-defined threshold are selected as reliable samples. The cosine similarity [11] can easily be calculated by the following formula:

$$\frac{f_i}{|f_i|} \cdot \frac{c_j}{|c_j|} > \mathcal{T} \tag{4}$$

The above equation defines cosine similarity as the dot product and the reliable images are selected if their similarity to the centroids is above a threshold, $\mathcal{T}$.

For semi-supervised technique, we select reliable samples based on the dissimilarity cost defined in Eq: (3). The size of the reliable samples is calculated as : $m_t = m_{t-1} + p.N$. Here, $p \in (0, 1)$ represents the ***enlarging factor*** which is the factor by which we increase our reliable samples at each iteration step t and N is the size of the unlabeled dataset. At each iteration step t, we select m_t nearest unlabelled images for all the labelled images.

3.2.4 Model training and optimization

After reliable images are selected, the training set now contains available images r_p from $p = 1$ to R, where R represents the number of reliable images. In case of semi-supervised learning, the reliable set also contains the labeled images. So, $R = M + n_r$, where M represents the number of labelled images and n_r represents the reliable images selected after applying kNN. The corresponding labels y_p which are the pseudo-labels from the cluster assignments on the extracted features(unsupervised) or from the kNN(semi-supervised). The model θ defined above is used as the base CNN model with max pooling and final classification layer as SoftMax activation. Categorical cross entropy is used for the loss function $\mathcal{L}$. The number of epochs are determined dynamically according the sample size of reliable images at each iteration, i.e, higher the reliable images, higher the epochs and vice versa in order to prevent overfitting and underfitting. In each iteration, the fine-tuning of the model is achieved using the reliable training set of images with the following optimization equation:

$$w_t^* \leftarrow \min_{w_t} \sum_{p=1}^{R} \mathcal{L}(y_p, \theta(r_p, w_t)) \tag{5}$$

where w_t^* represents the fine-tuned weights which are updated in each iteration and r_p and y_p represent the reliable training image and its corresponding pseudo label respectively. This procedure is performed iteratively resulting in larger reliable training samples as a result of more robust pseudo label assignment. The iterative process stops when model is converged i.e, the sample size of reliable images remains the same in subsequent iterations (unsupervised) or all the unlabeled images have been incorporated (semi-supervised).

3.2.5 Final inference

After the model θ has converged it is used to assign all images in $\mathcal{X}$ (or any related set of images) to either ω_f (forest) or ω_{nf} (non-forest).

4. Datasets

As described in Section 3, two datasets are employed in the different phases of proposed method: the labeled dataset $\mathcal{X}^l$ and the unlabeled dataset $\mathcal{X}$.

For labeled dataset $\mathcal{X}^l$, the publicly available BigEarthNet dataset is used [32]. It contains 590,326 multi-label remote sensing images with 43 land cover classes out of which 3 belong to forests. Since, the work in this paper is focused on binary forest classification, the dataset is converted to single-label binary data by selecting images having just forest and non forest labels. This resulted in a dataset of 59,701 images with 29,701 forest images and 30,000 non-forest images. This is divided into 44,731, 4,970 and 10,000 images for training, validation and testing sets respectively. Please note here that the testing set is not used for the self-supervised learning strategy but is only used as a sanity check while training the relevant base model.

For unlabeled dataset $\mathcal{X}$, the publicly available dataset, EuroSAT is used [33]. It contains 27,000 images having 10 land cover classes. 5,970 images are selected containing 3,000 forest and 2,970 non-forest images. The non-forest images were uniformly distributed among the other 9 classes. These are further divided into 4,970 images for unsupervised tuning and 1,000 for testing. Sample images from both datasets are given in Figure 2.

4.1. Implementation Details

For the initialization on a relevant dataset, a shallow CNN with 3 Convolution layers, a fully connected layer and a SoftMax classification layer with two output neurons is used. Max Pooling and Batch Normalization is applied after every layer. Adam optimizer with batch size of 16 and an initial learning rate of 0.001 is used. Decaying learning rate is used with validation loss being monitored in subsequent epochs. The training and validation losses for initialization on BigEarthNet on all image variations i.e, RGB, RGB + NIR and vegetation indices are shown in Fig 3.

For RGB, RGB + NIR and vegetation indices, all the bands of images in BigEarthNet had a size 120x120 (all

(a) BigEarthNet (b) EuroSAT

Figure 2: Sample Images from BigEarthNet and EuroSAT datasets

Vegetation Index	Formula
Normalized Difference Vegetation Index (NDVI)	$\frac{(NIR-R)}{(NIR+R)}$
Green Leaf Index (GLI)	$\frac{(G-R)+(G-B)}{(2*G)+R+B}$
Difference Vegetation Index (DVI)	$NIR-R$
Green Difference Vegetation Index (GDVI)	$NIR-G$
Ratio Vegetation Index (RVI)	$\frac{R}{NIR}$

Table 1: Equations of vegetation indices. RGB represent the Red, Green and Blue channels respectively and NIR represent the Near Infrared channel

bands were upsampled to match the dimension of RGB bands) and the bands of EuroSAT were upsampled from 64x64 to match the model's input. Standard normalization of zero mean and standard deviation of 1 is applied.

In case of semi-supervised scheme, k is set to 3 in kNN classification for assigning labels to the unlabeled data. The *enlarging factor*(p) is set to 0.1 during the iterations. For the reliable image selection after k-means clustering, a threshold of 0.85 is set after calculating cosine similarity. During the progressive learning, the number of epochs were selected in direct proportion of reliable images in order to prevent overfitting (in case of small reliable sample set) and underfitting (in case of large reliable sample set). This fine tuning process is repeated until the model is converged i.e, the number of reliable images remains the same in successive iterations (in case of unsupervised learning) or all the unlabeled images have been incorporated (in case of semi-supervised learning).

4.2. Results

The proposed model is first evaluated on the RGB bands of remote sensing imagery of forest regions. Both, unsupervised and semi-supervised progressive models, achieved an F1 Score of 0.86 as compared to that of 0.95 achieved under supervised training. Near-Infrared (NIR) band is added to the RGB bands for further evaluation. The F1 Score achieved by the semi-supervised method was 0.91, whereas unsupervised method achieved 0.89. In comparison, the supervised model achieved an F1 score of 0.98.

The proposed model was further evaluated on 5 vegetation indices, derived from the spectral bands, (see Table 1). This resulted in an increase in the overall F1 score achieved by both, unsupervised and semi-supervised methods, which was 0.91 and 0.93 respectively as compared to 0.96 achieved under supervision.

The results are summarized in Table 2. It can be seen that as the number of bands are increased the accuracy increases. Also, the best results were achieved using vegetation indices. This is in accordance with the fact that vegetation indices better represent the information regarding forests than raw pixels.

Fig 4 and 5 show some of the images from EuroSAT dataset that were correctly classified as forest and non forest respectively by the model after progressive unsupervised learning. The prediction clearly shows that the model is able to distinguish between just vegetation (or pastures) and forest regions and also able to classify forest images even though roads/water bodies pass through them.

5. Discussion

The proposed algorithm adopts a progressive unsupervised and semi-supervised formulation to tackle the problem of classification of forest regions in satellite images. Some of the design patterns regarding the proposed methodology are discussed in this section. Firstly, the base architecture consists of just 3 convolutional layers along with a fully connected and a classification layer. The reason behind choosing this shallow network instead of lets say VGG16[31], ResNet18 [34] or Inception[35] was that the model in the proposed approach is highly memory efficient as compared to these and in this specific case, a deep architecture is not suitable. The proposed methodology was tried using a ResNet18 too. Although the supervised training resulted in similar accuracies, the model was unable to improve after the progressive unsupervised training. The reason being that the number of robust and reliable pseudo labels at the start were quite low and a deep architecture tends to overfit on them and fails to generalize to the whole dataset.

Secondly, the proposed model was initialized on a relevant dataset. The other two options were to initialize it randomly or on a general purpose dataset like ImageNet. The

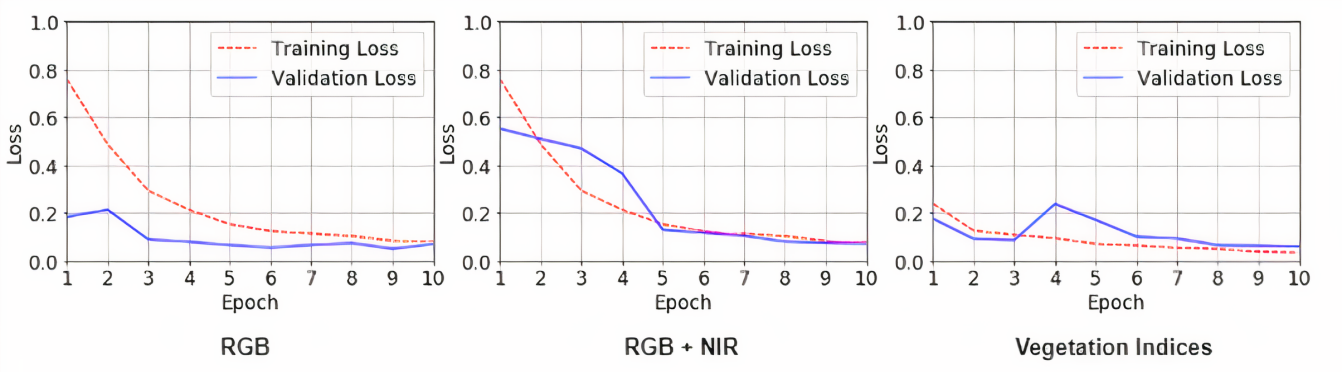

Figure 3: Training and validation losses

		Images with various spectral signatures/modalities					
		RGB		RGB + NIR		Indices	
		F1	Acc.	F1	Acc.	F1	Acc.
Models	**Base Model (without adaptation)**	0.66	0.66	0.79	0.80	0.87	0.87
	Semi-Supervised Learning	0.86	0.86	0.91	0.91	0.93	0.93
	Progressive Unsupervised Learning	0.85	0.86	0.89	0.89	0.91	0.91
	Supervised	0.94	0.94	0.98	0.98	0.96	0.96

Table 2: Comparison of F1 Score and Accuracy achieved by the Relevant Base Model(model trained on a relevant dataset), relevant base model fine tuned using Progressive Semi-Supervised and Unsupervised Learning and also the Supervised Model on the unlabelled dataset, EuroSAT. As shown, the proposed method improves the accuracy of the base model in each case and eventually approaches supervised model. Results are given for images with various spectral signatures/modalities which include RGB, RGB + NIR and Vegetation Indices

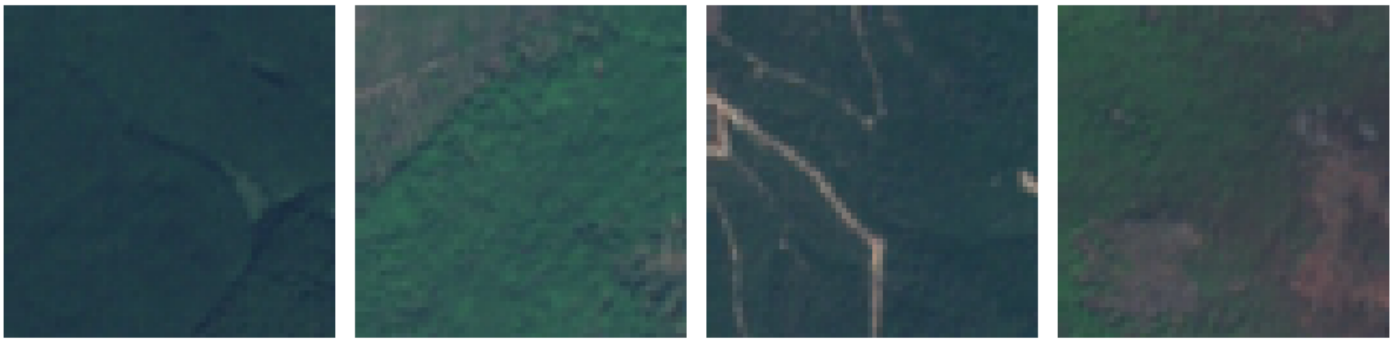

Figure 4: Forest Regions correctly classified by the proposed unsupervised progressive model

randomly initialized model will extract inaccurate features and thus form random clusters. The model was not initialized using ImageNet due to the difference between features in normal day-to-day images and the satellite images containing several spectral bands other than just RGB[36]. This will lead to extraction of features that are not useful for remote sensing forest classification.

For semi-supervised learning, two hyper-parameters were of extreme importance. First, the value k in k Nearest Neighbors was set to 3 as it resulted in the most correct pseudo-labels. We tried with other values of k that yielded comparatively poor result. Secondly, the *enlarging*

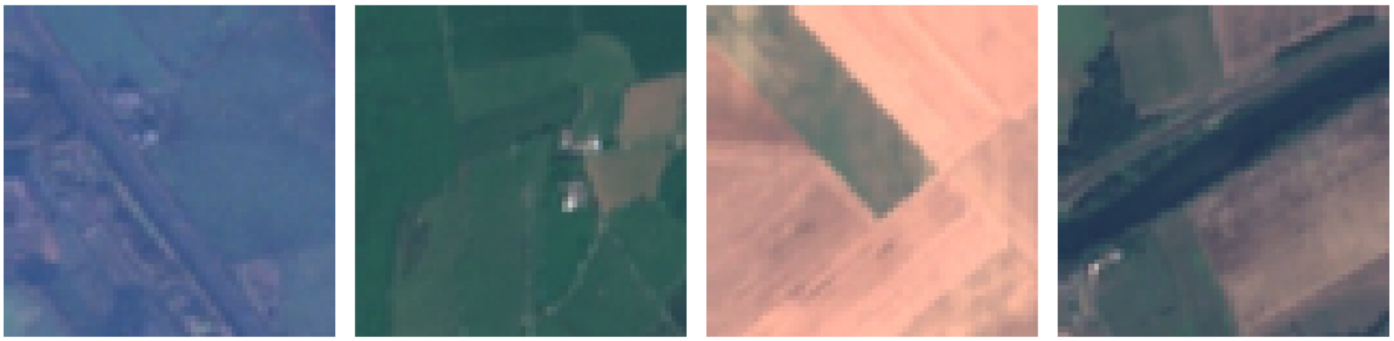

Figure 5: Non-Forest Regions correctly classified by proposed unsupervised progressive model

Ground Truth: Non-Forest
Predicted Label: Forest

Ground Truth: Forest
Predicted Label: Non-Forest

Figure 6: Images wrongly classified by the proposed method

factor(p) was set to 0.1 which increases the reliable unlabelled images by around 500 with each iteration (total number of unlabelled images were 4975). We can either set p to a very large or very small value. A large value urges m_t to increase rapidly, resulting in unreliable pseudo-labels. A small value means m_t progressively enlarges with a small change in each step, with large computation time.

Lastly, two of the images wrongly classified by the model after proposed progressive unsupervised learning are shown in Fig 6, one that is wrongly classified as forest and the other wrongly classified as non-forest. As it shows, the images are spectrally quite similar to each other. For the proposed method, though deep model implicitly captures the textual semantics, One improvement can be incorporating explicit textual semantics along with these.

6. Conclusion

This paper presents a progressive unsupervised deep learning based approach for forest mapping. The crux of the idea is to initialize a base model on a relevant dataset and subsequently transfer the learned information on to a deep unsupervised progressive scheme which is then trained using unlabeled dataset. The approach is generic and the results have been reported using variety of three different spectral (imaging) modalities. Although the proposed approach yields high accuracy but still there are different aspects for improvement. One such improvement will be to further enhance the unsupervised strategy by replacing the k-means clustering and reliable selection procedure with a single more robust clustering scheme incorporating textural semantics of forests. This may not only aid in improving the convergence of the progressive scheme but may also potentially lead towards completely self paced learning bypassing the need of any relevant base model. Our work must not be seen as a competitor to completely supervised forest mapping methods, rather as a complementary to them.

7. Acknowledgements

The authors would like to thank German Academic Exchange Service (DAAD) for supporting this work through Deutsch-Pakistanische Forschungskooperationen. 2j ab 19 grant number: 57459164.

References

[1] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. *CoRR*, abs/1403.6382, 2014. 1

[2] Zhong Zheng, Jinfei Wang, Bo Shan, Yongjun He, Chunhua Liao, Yanghua Gao, and Shiqi Yang. A new model for transfer learning-based mapping of burn severity. *Remote Sensing*, 12(4):708, 2020. 1

[3] Sudipan Saha, Francesca Bovolo, and Lorenzo Bruzzone. Unsupervised deep change vector analysis for multiple-change detection in vhr images. *IEEE Transactions on Geoscience and Remote Sensing*, 57(6):3677–3693, 2019. 1, 2

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 1

[5] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *CoRR*, abs/1703.00121, 2017. 1

[6] Jakob Gawlikowski, Sudipan Saha, Anna Kruspe, and Xiao Xiang Zhu. Out-of-distribution detection in satellite image classification. *arXiv preprint arXiv:2104.05442*, 2021. 1

[7] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2

[8] M. Shahzad and X. X. Zhu. Automatic detection and reconstruction of 2-d/3-d building shapes from spaceborne tomosar point clouds. *IEEE Transactions on Geoscience and Remote Sensing*, 54(3):1292–1310, 2016. 1

[9] M. Shahzad and X. X. Zhu. Robust reconstruction of building facades for large areas using spaceborne tomosar point clouds. *IEEE Transactions on Geoscience and Remote Sensing*, 53(2):752–769, 2015. 1

[10] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. *CoRR*, abs/1807.05520, 2018. 1, 2

[11] Hehe Fan, Liang Zheng, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. *CoRR*, abs/1705.10444, 2017. 1, 2, 5

[12] R.M.S. Bashir, M. Shahzad, and M.M. Fraz. Vr-proud: Vehicle re-identification using progressive unsupervised deep architecture. *Pattern Recognition*, 90:52 – 65, 2019. 1

[13] Sudipan Saha, Lichao Mou, Chunping Qiu, Xiao Xiang Zhu, Francesca Bovolo, and Lorenzo Bruzzone. Unsupervised deep joint segmentation of multitemporal high-resolution images. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12):8780–8792, 2020. 1

[14] Eka Miranda, Achmad Mutiara, Emastuti, and Wahyu Wibowo. Classification of land cover from sentinel-2 imagery using supervised classification technique (preliminary study). pages 69–74, 09 2018. 2

[15] Meysam Majidi Nezhad, Azim Heydari, Lorenzo Fusilli, and Giovanni Laneve. Land cover classification by using sentinel-2 images: A case study in the city of rome. 04 2019. 2

[16] Andrew Mellor, Andrew Haywood, Simon Jones, and Phil Wilkes. Forest classification using random forests with multisource remote sensing and ancillary gis data. 01 2012. 2

[17] Angel Fernandez-Carrillo, David Fuente, Freddy Rivas-Gonzalez, and Antonio Franco-Nieto. An automatic sentinel-2 forest types classification over the roncal valley, navarre: Spain. page 58, 10 2019. 2

[18] Suman Sinha, Laxmi Kant Sharma, and Mahendra Singh Nathawat. Improved land-use/land-cover classification of semi-arid deciduous forest landscape using thermal remote sensing. *The Egyptian Journal of Remote Sensing and Space Science*, 18(2):217 – 233, 2015. 2

[19] Yue Wu, Guifeng Mu, Can Qin, Qiguang Miao, Wenping Ma, and Xiangrong Zhang. Semi-supervised hyperspectral image classification via spatial-regulated self-training. *Remote Sensing*, 12(1), 2020. 2

[20] Rogério Negri, Sidnei Sant'Anna, and Luciano Dutra. Semi-supervised remote sensing image classification methods assessment. pages 2939–2942, 07 2011. 2

[21] Kristin P. Bennett and Ayhan Demiriz. Semi-supervised support vector machines. In *Proceedings of the 11th International Conference on Neural Information Processing Systems*, NIPS'98, page 368–374, Cambridge, MA, USA, 1998. MIT Press. 2

[22] T. K. Moon. The expectation-maximization algorithm. *IEEE Signal Processing Magazine*, 13(6):47–60, 1996. 2

[23] Saroj K. Meher. Semisupervised self-learning granular neural networks for remote sensing image classification. *Applied Soft Computing*, 83:105655, 2019. 2

[24] Yan-Qing Zhang. *Granular Neural Network*, pages 4402–4411. Springer New York, New York, NY, 2009. 2

[25] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 2

[26] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018. 2

[27] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 2

[28] Sudipan Saha, Francesca Bovolo, and Lorenzo Bruzzone. Change detection in image time-series using unsupervised lstm. *IEEE Geoscience and Remote Sensing Letters*, 2020. 2

[29] A. Chatterjee, J. Saha, J. Mukherjee, S. Aikat, and A. Misra. Unsupervised land cover classification of hybrid and dual-polarized images using deep convolutional neural network. *IEEE Geoscience and Remote Sensing Letters*, pages 1–5, 2020. 2

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. 2

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 3, 6

[32] Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. *CoRR*, abs/1902.06148, 2019. 5

[33] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *CoRR*, abs/1709.00029, 2017. 5

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 6

[35] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014. 6

[36] Jia Song, Shaohua Gao, Yunqiang Zhu, and Chenyan Ma. A survey of remote sensing image classification based on cnns. *Big Earth Data*, 3(3):232–254, 2019. 7