Semantic Segmentation of Canopy Coverage in Remote Sensing using Deep Learning

Student: Niamh Donnelly (17204509), Supervisor: Dr. Brian Mac Namee,

University College Dublin, Dublin 4, Ireland
niamh.donnelly1@ucdconnect.ie

Abstract. The following work outlines the current state-of-the art in the area of classification of earth observation data in remote sensing. First, the landscape of remote sensing is explored in terms of the three predominant classification types: pixel-based classification, object-based classification, and scene classification. Within these broad categories a review of traditional and contemporary approaches are discussed. Finally, a body of research is proposed for the exploration of Convolutional Neural Networks (CNNs) as an efficient approach to the semantic segmentation of tree coverage from earth observational imagery.

Keywords: Remote Sensing · Deep Learning · Canopy Coverage · Convolutional Neural Networks.

1 Introduction

Improvements in satellite technology and public availability of high resolution aerial imagery, has lead to considerable efforts towards harnessing the power of machine learning (ML) algorithms for land classification tasks. An important and relevant area where ML has already demonstrated utility relates to tree canopy coverage segmentation and classification. Remote sensing of canopy coverage in urban areas is essential to the task of monitoring environmental resources (i.e. air-quality, habitat of wildlife) and for large scale urban planning. This process was traditionally achieved by human manual annotation of tree areas from satellite imagery. However, advances in deep learning architectures has led to a move towards the development of automated systems. While notable advances in remote sensing have been achieved with deep learning architectures, the approach has rarely been used in the problem of landmass classification. This paper presents an in-depth review of existing studies relevant to this task in remote sensing. The report discusses several distinct approaches for classification that are commonly used in remote sensing, which include: pixel-based classification, object-based classification, and scene classification. Additionally, a review of traditional and contemporary classification methods in the domain specific to canopy coverage is explored. The final section provides a summary of the review, and identifies key areas of research interest. In particular, the study found high suitability of Convolutional Neural Network (CNN) architectures to the task of segmentation and classification of earth observation data obtained from satellite images.

2 Review of the Literature

Three distinct classification approaches currently exist in the area of remote sensing; pixel-based segmentation (involves the classification of each pixel in an image), object-based (the detection and classification of whole objects in an image), scene classification (the assignment of an entire scene to a thematic class).

2.1 Pixel-based Approaches

Pixel-based approaches are considered one of the more traditional methods of image classification in the remote sensing and geoscience fields. This approach is also referred to in machine learning literature as pixel-wise segmentation [1]. Unlike traditional classification, the methodology involves not only the detection of an object in an image, but subsequently the assignment or semantic labeling of each pixel to an object class [2][3]. Traditionally pixel-based approaches involve the use

of spectral features (i.e. information from single pixels) with shallow learning models (e.g. SVMs, KNNs). However, it is advantageous to include spatial information surrounding individual pixels. Recent work has made use of CNN architectures to obtain rich spectral and spatial information from image data. CNNs are a particular type of neural network designed to detect visual patterns directly from image pixels [4]. These approaches are discussed in detail in the following sections.

Spectral Features & Shallow Architectures Earth observation image data often consists of multiple spectral bands containing plentiful discriminative information. Each image pixel has a n-dimensional vector representation of spectral bands. Previous research has made use of spectral feature based representations with varying machine learning algorithms for earth observation data. Camps-Valls and Bruzzone provide a comparison of Support Vector Machines (SVMs) and other kernel based algorithms for hyperspectral (i.e. multiple spectral bands) image classification [5]. The study employs a coarse feature extraction to remove a number of spectral bands, ultimately observing that an SVM provides superior classification results taking into account robustness to noise, overall accuracy, computational cost and reduced training data. SVM classifiers have shown some of the highest success rates for pixel based classification of earth observation data and are one of the more common models used. Rahmani and Akbarizadeh [6] examine the accuracies of varying machine learning models including: kNN, Neural Network and SVM models on a hyperspectral earth observation dataset. Rahmani and Akbarizadeh also identified an SVM as the superior classifier in terms of computational efficiency and classification accuracy.

Spectral-Spatial Features While the use of spectral information can provide promising classification accuracies, recent approaches explore the use of spatial information in conjunction spectral features. Issues arise when frameworks are developed to label a single pixel without considering its surrounding context. A pixel label may pertain to short or even long range information. For example, classification of a green pixel to a either a tree, a meadow or other vegetation requires a wider contextual consideration. Pixel-based methods often suffer from what is known as the "salt and pepper effect" [7], whereby an area of the classification contains instances where single isolated pixels are labeled incorrectly and differently to their surroundings, resulting in the the segmented mapping appearing visually speckled. Consideration of surrounding pixels could potentially influence these isolated pixels to an accurate classification and result in stronger segmentation between spatial regions. Inclusion of spatial features is commonly achieved using a pixel wise patch-based approach [8] . This is implemented during model training by extracting a patch of pixels of fixed width and length and labeling the center pixels based on its surrounding neighbors contained within the patch area.

Convolutional Neural Network Approaches Traditional techniques amalgamate spatial aspects through extraction of spatial features (e.g. texture, shape) and use of shallow classifiers like SVMs. While these techniques show improvements in accuracy in comparison to exclusively spectral approaches [9], research has demonstrated that further performance enhancements are possible using convolutional neural networks. In particular, CNNs have been the architecture of choice for remote sensing image analysis. Approaches to segmentation via CNN approaches can be categorized into a number of architecture assemblies:

Pixel-based Vector Approaches Each pixel in a hyperspectral image has a visual representation of its spectral bands. Often, in remote sensing, hyperspectral data is fed to a CNN using this visual representation for each pixel. The input to the CNN is a pixel spectral vector with the number of spectral bands pertaining to its length. A study by Hu et al. [10] developed a CNN architecture using the Theano python programming library to classify landmass objects (e.g. trees, gravel, meadows) at a pixel level in this fashion. This study compared the proposed 5 layer CNN (one convolutional and one fully connected layer) with a SVM model and four CNNs with architectures of varying complexity, including the well established LeNet-5 [11]. Comparison of their custom CNNs with other established architectures of a higher complexity provided lower accuracies. Authors postulate this was due to an inadequate number of training samples, which is a common occurrence when developing CNNs. In comparison to the SVM, the proposed model showed an average of 2% accuracy improvements over all three datasets.

Patch-based Approaches In a patch-based approach, each pixel is classified in context to a patch of surrounding image pixels. This procedure can be efficiently implemented via CNN architectures whereby the CNN takes as input a patch of an image, and generates a classified output patch. The output patch is smaller in size and centered around the input patch. By implementing a patch-based system, Mnih and Hinton [8] observed high accuracies and a reduction in the salt-and-pepper appearance/noise for classification of aerial imagery using CNNs. Another study by Zhao et al. [12], employed a patch-based CNN approach for classification of earth observation data and observed higher accuracy rates in comparison to engineered features and a SVM classifier.

Fully Convolutional Networks (FCNs) Fully Convolutional Netowrks gained much popularity in the area of machine learning following a study by Shelhammer et al. [3] for their use in semantic image segmentation. The architecture involves the replacement of fully connected layers with convolutional layers. Shelhammer's research adapted pre-existing established models (VGG-16 [13], AlexNet [14], Google LeNet [15]) into fully connected networks. In order to obtain a classification per pixel, the output is scaled to the size of the input layer through the use of a final "upsampling layer". The upsampling process often results in very coarse prediction mappings due to the the rapid increase in size from the reduced sizes in pooling layers. This study addressed this nuance by adding "skip connections"; this process creates paths from early layers to deeper layers, thus re-introduces finer pixel layers to the deeper layers, which are more likely to bare the negative effects of pooling. This process compensates for the coarseness of the final prediction layer. Ultimately, the adapted fully connected VGG network out-performed other adapted architectures in terms of accuracy. A study by Maggiori et al. [16] that examined this approach for remote-sensing image classification observed higher accuracy and less computational expenditure in comparison to the CNN patch-based approach.

It is not uncommon to develop architectures to profit from a combination of both handcrafted features and high-level feature representations accessible via CNN models. Employing both of these techniques, Paisitkriangkrai et al. [17] developed a framework for pixel-based automated annotation of vegetation in urban areas. The study identified a number of discriminative pixel-level features, including Normalized Difference Vegetation Index (NDVI) which determines the density of green on a patch of land. A random forest classifier is trained on features to provide output probabilities for target classes. A FCN is used to attain high level feature representations and combined with a logistic regression classifier to provide outputs class probabilities. Results of both approaches are combined to provide an overall classification. Furthermore, a pixel-level Conditional Random Field (CRF) post-processing step was implemented to improve accuracy. CRF works as a segmentation technique by enforcing globally-consistent labeling and improving fragmented and marginal regions. Results indicated that CRF achieved increased classification accuracy. The framework observed an overall accuracy improvement in comparison to isolated implementations of the CNN and random forest. These observations indicate that while CNNs alone tend to outperform traditional methods, further improvements in performance can be achieved using a combination of CNNs and relevant hand-crafted feature representations.

Convolutional Encoder-Decoder Networks To further improve the prediction results and alleviate the effects of pooling layers, Badrinarayanan et al. introduced the SegNet architecture [18]. SegNet amalgamates a convolutional encoder and decoder network. The encoder is represented by the VGG-16 CNN architecture without the fully connected layers. The pooling layers of the encoder gradually down-sample and decrease the size of the input. Next, the decoder network replaces pooling layers with upsample layers to gradually increase the convolutional layer size back to the scale of the original image. This study observed that in most parameter layouts, the encoder-decoder architecture out performs FCNs, either in terms of accuracy or memory allocation. U-Net architectures [19] are an another well established encoder-decoder approach, which includes an added nuance of skip connections between coinciding encoder and decoder layers. A recent study by Aich et al. [20] found these out performed SegNet for semantic segmentation of land types. In the field of remote sensing, a recent study by Volpi and Tuia [21] implemented a custom convolutional encoder-decoder architecture for dense semantic labeling of earth observation data. Similar to Maggiori et al. , Volpi and Tuia draw a comparison to patch based architectures, with marginally higher results achieved by implementation of the convolutional encoder-decoder technique.

Dilated Convolutions An alternative method of maintaining resolution is to replace pooling layers in the CNN architecture with dilated convolution layers. Dilated convolution layers involve the dispersal of alignment of filter weights by a dilation factor. Increasing this factor increases the sparsity of alignment of weights, thus increasing the kernel size. By repeatedly increasing the dilation factor for each layer the convolutional area is expanded. This has the effect of negating loss in resolution in pixels. Interestingly, a study by Hamaguchi et al. [22] examined this approach for segmentation of small objects from earth observation data and observed poor performance accuracies when repeatedly increasing the dilation factor after each layer. It was postulated that failure to aggregate local features resulted from the high sparsity of the kernels associated with the increased dilation factor.

2.2 Object-based Approaches

In the field of remote sensing, especially when classifying objects in images (e.g. ship detection[23][24], building detection [25] and vehicle detection[26]), it is often beneficial to utilize approaches that incorporate a greater degree of abstraction than pixel-based alternatives. Object-based recognition of remote sensing data is considered a particularly difficult image classification problem [23]. This complexity is mostly attributed to the nature of remote sensing images. Any single image can contain large numbers of inherently small scale objects imposed on complex backgrounds. In comparison, natural image scenes are usually consumed by larger objects imposed on less intricate backgrounds. Machine learning techniques employed for object detection in remote sensing is an area of increasing interest. For any technique, the process of object-based classification in remote sensing is commonly categorized by two stages: object detection and object classification. Object detection identifies an area (referred to as the object proposal or region) in an image where an object is present, and object classification identifies its target class.

Object Detection As previously mentioned, object detection identifies if an object is present in an area of an image. Often, this detection process involves additional localization information. Techniques employed to classify with localization not only apply a class label to an image, but also identify a bounding box surrounding the object of interest, indicating its location.

Sliding Window Technique for Region Proposal In the area of remote sensing and the broader field of image categorization, object detection or region proposal may be modeled as a classification problem using sliding window detection [27]. This approach obtains patches/windows of fixed width and length at every location of an input image by sliding the window across the image. These patches are subsequently fed to an image classifier to obtain the probability of an object being present in an image, and potentially its localization points. This technique is effective, however, it is computationally expensive to extract patches at every location of an image. In addition the method incurs an added redundancy as many patches will not contain an object at all. This downfall is particularly common in remote sensing where windows are frequently small in size in order to locate minute objects. To avoid searching the entire image, it is often beneficial to propose an area of interest which could potentially contain an object.

Sliding Window Advancements Region proposal is commonly implemented in the area of remote sensing through the use of a number of techniques. Diao et al. [28] examined the application of a novel coarse object location method for detecting the proposal areas of airplanes in satellite images. This study implemented an algorithm to calculate the geometric centre of possible objects based on saliency mapping. Additionally, overlapping windows that contained the same object were fused into one window, further reducing the number of windows necessary. Once a proposed region was identified, a deep belief network provided the probability of the presence of an airplane in the area of interest. Considering performance and efficiency metrics, Diao et al. observed the detection method used in the study achieved superior results to that of a sliding window approach.

Another interesting tool for object proposal is the selective search sliding window approach. Selective search was first introduced to the field of computer vision by Uijlings et al. [29]. It works through identifying a number of proposal windows (which potentially contain an object) via a graph-based segmentation algorithm. A parameter in the segmentation algorithm determines

the number of possible object proposals. Interestingly, Cheng et al. observe a direct relationship between the aforementioned parameter and detection performance [30]. Increasing the parameter pertaining to the allowed number of object proposals results in an increased computational cost however also improves detection. The authors provide an optimal point of tradeoff between quantity of object proposals and accuracy, where performance plateaus.

Object Classification Object classification is an important aspect of remote sensing, and is deployed after the region of object proposal has been identified. Early attempts at object detection and classification in the field of remote sensing focused on algorithms to identify lower-level feature representations of image data. Examples of these approaches include Bag-of-Words and Histogram of Oriented Gradients. However, in recent years, the state-of-the-art in the field has begun to move towards deep-learning approaches.

Bag-of-Words (BoW) Using a bag-of-words (BoW) approach, image features are represented as unordered descriptive words. An image is then assigned a vector/histogram of occurrence counts for each word. Xu et al. [31] use a bag-of-words (BoW) based model for object-based classification in land-use of high spatial resolution aerial images. BoW feature vectors are extracted from the aerial data and fed to an SVM model for classification. Results showed using Bow features obtained higher accuracies than using hand-crafted low-level features, for example spectral and texture attributes.

Histogram of Oriented Gradients (HOG) Low-level feature extraction techniques include the histogram of oriented gradients (HoG), which represents objects based on gradient intensities and orientations in spatial regions [32]. This technique has been applied in the area of remote sensing with success particularly in the representation of edges and shapes of objects. Tuermer et al. [33] observed high accuracies applying HoG features with the Adaboost classifier for vehicle detection. However, the framework encountered misclassification with similar shaped objects or occurrences where vehicles are partially concealed. Zhang et al. [34] apply a rotation invariant variation on the HoG feature representation by evaluating the dominant orientation of a region. Objects are rotated to according to the dominant orientation and the HoG feature is encoded in the revised orientation.

Deep Learning Approaches Although the feature representation methods have been successful in the past, more recent approaches make use of deep learning models. Deep convolutional neural networks are particularly popular for object detection in both the remote sensing arena and the computer vision community due to high success rates. CNNs can learn rich features automatically without the need for specific domain knowledge and have showed improvement in object-based detection with respect to other feature learning techniques, like BoW [35] [30]. While CNNs provide an element of translation invariance, they occasionally have difficulty in effectively dealing with the problem of object rotation variations [36]. Remote sensing data is particularly influenced by rotational variances in comparison to natural scenes. In natural images, objects are mostly upright, however, remote sensing images are often taken from multiple orientations. Cheng et al. [30] propose a novel framework through adoption of a rotation-invariant CNN (RICNN) model for object detection in Very High Resolution (VHR) optical remote sensing images. Rotational invariance is introduced through data augmentation by introducing rotated variations of objects in the training set. Next, introducing a regularization constraint term enforces these training examples to share features with their rotated counterparts. Cheng et al. [30] found the RICNN showed higher performance rates in comparison to the aforementioned BoW methods and CNN, without the rotation regularization approaches. Interestingly, the authors observed detection accuracy of basketball and tennis courts were particularly low. Authors postulate this was due to pooling layers reducing or even removing critical discriminative properties such as straight lines or arcs common to these classes.

Scale variations of objects are a common object-based detection complexity explored in computer vision. Often in remote sensing, image objects (e.g. buildings, vehicles, vegetation) appear at different scales. Chen et al. [26] made considerable advancements in tackling the complexity of scale variances through the use of novel a hybrid deep CNN (HDNN). The HDNN divides maps of the final CNN layer into different sizes through varying filter sizes, thus enabling the HDNN to extract variable-scale features. Results showed the HDNN outperforms a deep CNN with static filter sizes for vehicle detection.

Apart from CNNs, Deep Belief Networks (DBNs) also have shown improvements in comparison to lower level feature extraction methods such as those described in previous sections. A DBN is a neural network consisting of stacked restricted boltzmann machines, which are stochastic undirected graphical models with a hidden layer and an input layer. Effectively, the output of one layer is the input to the next boltzmann machine on the stack. Chen et al. [37] found that using a DBNs for aircraft detection outperform the traditional method based on HOG.

2.3 Scene Classification

Scene classification involves the semantic labeling of images according to a discrete set of land cover and land use classes (e.g. residential, agriculture, urban areas) dependent on the image contents. If the required function of an algorithm is to categorize a scene for a given image, pixel-based or object-based methods can falter in the ability to account for surrounding context. For example, if classifying a scene is dependent on object based methods, a building may belong to both a residential area and an urban area. Thus, these variations and differing structural patterns make scene categorization an inherently difficult task. The associated challenges along with the availability of high resolution earth observation images in recent years have spurred a surge of interest in scene classification methods.

Traditional Shallow Methods Scene classification involves the extraction of features from an area of an image and subsequently the classification of the sectioned scene. Early methods in the field made use of handcrafted features [38][39][40], requiring a level of expert domain knowledge to provide features such as texture, color gradients, spatial, and spectral information. Handcrafted feature representations used in scene classification also include the HoG and BoW methods discussed previously in the object-based section of this report. Scenes are commonly representative of multiple interrelating descriptive features. Zhao et al. [41] derive a multi-feature based classification model for high spatial resolution remote sensing imagery. The study employees local spectral features, local structural features and global textural features. Local spectral features are acquired as vectors representations of spectral bands (specifically the mean and standard deviation of spectral bands of each scene patch). The Scale-invariant feature transform (SIFT) [42] is used to describe local structural information for each scene patch. The global textural features of each scene patch are represented using the shape-based invariant texture index (SITI). The authors identified that using a combination of these three feature types in conjunction with an SVM classifier provided marginally higher accuracy results for scene classification in comparison to isolated instances of feature types. While these methods have been shown to be reasonably effective in use, like many other areas in remote sensing, attention in recent years has turned to exploitation of rich high level features through the use of CNNs.

Supervised Convolutional Neural Networks Supervised CNN deep learning architectures have shown encouraging results in the areas of scene recognition. Zhang et al. [43] benchmark a proposed Gradient Boosting Random Convolutional Network Framework (GBRCN) with number of established CNN architectures on different scene labeled earth observation data sets. While the proposed CNN showed the highest accuracy on all datasets, very few alternative CNN architectures tested provided an accuracy below 90%. Although CNNs achieve state-of-the-art performance, in order to benefit from a deep architecture with multiple convolutional layers, a substantial quantity of data is required for sufficient training. Often the availability of large labelled datasets is limited (particularly in the domain of scene classification). Thus, recent research efforts focus on the use of pre-trained CNN architectures whereby, a CNN is trained on an arbitrary image dataset and subsequently features extracted from intermittent layers to form global feature representations for some classification model. This strategy benefits from the fact that initial layers of CNNs tend to be generic low-level detectors (i.e. edges, corners), which are less tied to the final application, thus can have relevance to other detection uses. Examining the use of pre-trained networks in this manner, Marmanis et al. [44] used the open-source UC Merced Land Use (UCML) dataset [45] to benchmark the use of CNN features descriptors in comparison with engineered feature methods (for example, BOW, Salient Unsupervised Learning [46], spatial pyramid matching kernel (SPMK) [47]). Results indicate higher levels of accuracy when implementing pre-trained CNN features in comparison to alternative methods.

Pre-trained Convolutional Neural Networks An alternative method involves the fine-tuning of a pre-trained CNN. That is, particular layers of the pre-trained CNN are modified to adapt to the dataset of interest. Usually the early layers of the CNN are untouched and latter layer parameters (i.e. kernel weights) are tuned/re-learned to encode specific attributes relevant to the dataset of interest. Recent research has compared the performance of the feature extraction approach, and the use of fine-tuned pre-trained CNNs. A study by Nogueira et al. [48] analyzed performance of these two approaches in comparison with that of a CNN without any form of pre-training. Performance of six differing CNN architectures, with varying complexity, were examined on a number of benchmark remote sensing datasets. A number of approaches were examined: extraction of CNNs last fully connected layer as descriptive feature inputs to an SVM model, fine-tuned versions of CNN architectures, fully-trained CNN (i.e. without pre-training). The authors observed that for all combinations, fine-tuned networks mostly outperformed all other approaches for the varying CNN architectures and benchmark data sets. Furthermore, the addition of the descriptive features to this approach marginally increased performance. Fully trained networks produced lowest overall performance, the reason suggested by Nogueira et al. being the limiting size of the dataset for convergence.

2.4 Remote Sensing for Canopy Coverage Identification

Forest canopy cover is identified as the percentage cover of tree canopy in a give area, including only trees/shrubs and ignoring any other forms of vegetation [49]. It encompasses the vertical projection of the tree outline to the earth surface. Prior to accessibility of remote sensing data, the detection process of canopy coverage was achieved by ground-based techniques, whereby human surveyors were required to manually identify areas of tree coverage using hand-held devices. Another commonly employed traditional technique made use of surveyors to manually mark areas of tree coverage from satellite images. Government agencies in the UK previously implemented the latter technique through a crowd-sourcing methodology to roughly estimate the percentage of tree canopy cover in urban areas [50]. The initiative procured the use of voluntary human participants to manually label trees for a sample of earth observation data of the London city area. The percentage cover of the these areas was then extrapolated to provide an approximate calculation for other areas of the city not included in the sample subset.

The nature of the task of estimating the area of tree cover is one which could benefit from machine learning capabilities to generalise to sample data. Stojanova et al. [51] explored a broad set of machine learning models to predict canopy coverage from land satellite imagery. Models implemented in the study include decision trees, random forests, bagging techniques, and tree-based ensembles. Random forest models outperform in terms of root-mean-square deviation (RMSE) for canopy cover, with most models showing competitive performance rates on a benchmark dataset. While remote sensing areas of object-based, pixel-based, and scene classification have experienced high performance rates spurred by advancements in deep learning architectures, canopy coverage evaluation using these techniques is a relatively undiscovered area of interest. Nevertheless', a number of exploratory studies do exist in the literature. Guirado et al. [52] drew a comparative study between the use of traditional classifiers (e.g. SVM, KNN) and a CNN classification approach to determine the presence of tree shrubs in extracted patches of satellite image obtained via Google Earth. The authors obtained a 100% accuracy score on test data with the use of a ResNet CNN architecture for object based detection, outperforming object-based image analysis with SVM (88%), KNN (88%) and random forest (91.78%) classifiers. While other studies identify land types containing tree coverage through scene classification of forestry vegetation,[53] few focus solely on perfecting classifiers aimed for the task of identifying percentage of canopy coverage. The specific area of remote sensing is still uncharted in research pertaining to the pixel-wise semantic segmentation of tree cover.

2.5 Summary

A review of the literature indicates that in domain of remote sensing, classifier selection is largely depend on the required application or task. Moreover, CNN approaches have overtaken traditional classifiers in virtually every aspect of remote sensing. While CNNs have been used in the specific domain of canopy coverage, the application has predominantly used a object-based approach to

classify individual trees. The use of CNN architectures associated with pixel-wise segmentation strategies is still a largely unexplored area. These pixel-wise segmentation specific CNN architectures (i.e. FCN, encoder-decoder CNN, dilated convolutions) have been favored for land-use classification competitions, including the renowned DeepGlobe competition, with variations of UNet and Segnet holding top positions on the current leader-boards [54]. This provides a strong indication that the task of canopy segmentation could potentially also benefit from similar approaches.

3 Research Specification

Thesis Contribution Convolutional neural networks are currently considered state-of-the-art in the field of computer vision. CNNs have consistently achieved the highest benchmark results on pixel-wise semantic segmentation datasets like PASCAL VOC benchmark datasets [55] which contain a collection of everyday scenes and objects. Remote sensing imagery provides new challenges for segmentation algorithms. Remote visual sensing involves manipulation of very large volumes of data, often consisting of numerous image types which require models capable of high precision accuracy to identify small scale image features at a pixel level. Furthermore, earth observation imagery assumes an overhead aerial viewpoint, which is absent of depth related attributes associated with the everyday scene datasets. The primary contribution of the proposed body of work is to evaluate the ability of CNN architectures to perform pixel wise semantic segmentation in the area of remote sensing. Precisely, I aim to develop a comparison of different CNN approaches for pixel-based automatic detection of canopy coverage from earth observation data.

Proposed Framework Approaches The canopy coverage dataset provided by a 3rd party organization is accessed through the Google Earth Engine (EE) service [56]. Google EE is a cloud-based platform housing catalogs of geospatial datasets and satellite imagery and provides access through python and Java APIs. The framework of this study proposes to employ the python API to parse requests to earth engine servers to access earth observation imagery with the potential to host image data remotely.

In the area of pixel-wise semantic segmentation, a number of CNN framework approaches exist in the literature (see section 2.1). This research aims to evaluate two of these approaches in their applicability of pixel-wise automatic detection and segmentation of canopy coverage.

Approach 1: FCN Initial feasibility testing will be performed to explore variations of established CNN architectures using the FCN method applied by Shelhamer et al. [3], whereby fully connected layers are replaced by convolutional layers and a final upsample layer. Initially, network assembly of the FCN will replicate that of the VGG-16 CNN architecture, following that of a similar method employed by Fu et al. [57]. The VGG-16 network is one of the most influential CNN architecture and is known for its simplicity and depth owing to its high success rates. In a similar fashion to Shelhamer et al., the proposed research will also test the potential of other architectures (AlexNet and GoogleLeNet described in earlier sections) into fully convolutional networks.

Approach 2: Encoder-Decoder Architectures Recall the encoder-decoder approach aims to overcome the pitfalls associated with FCNs by replacing the final upsampling layer with a decoding network. This research aims to examine the efficiency of this method in comparison to an FCN approach. Experimentation will leverage the use of a number of existing approaches, including SegNet and U-Net that have shown previous successes in the areas of semantic segmentation of earth observation data.

Implementation and Evaluation Development of architectures will use Keras [58] and TensorFlow [59] libraries using the python programming language. Furthermore, the architecture will potentially benefit from a of pre-trained VGG-16 network for the encoder network of the SegNet and the FCN. Following the initial training stages, architectures for each approach showing highest accuracy on a subset of the data will undergo fine tuning and an evaluation comparison on the final test set. The applicability of post-processing steps (e.g. CRF [60]) on the final models will be determined. Finally, the models will be evaluated in comparison to a previously implemented

approach leveraging an SVM with engineered features and using benchmark datasets provided be the DeepGlobe competition [54].

A graphical illustration of the planned research, described using a Gantt chart, is presented in Appendix A.

References

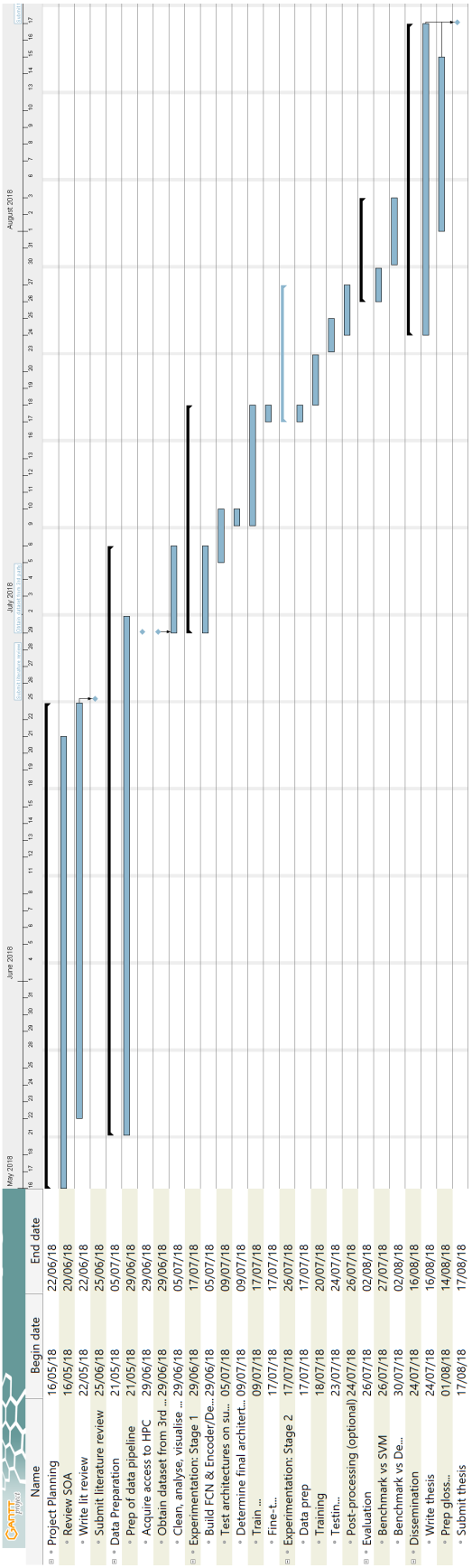
1. L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, pp. 834–848, April 2018.
2. H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1520–1528, Dec 2015.
3. E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 640–651, April 2017.
4. Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, pp. 2278–2324, Nov 1998.
5. G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, pp. 1351–1362, June 2005.
6. M. Rahmani and G. Akbarizadeh, "Unsupervised feature learning based on sparse coding and spectral clustering for segmentation of synthetic aperture radar images," *IET Computer Vision*, vol. 9, no. 5, pp. 629–638, 2015.
7. C. Liu, L. Hong, S. Chu, and J. Chen, "A svm ensemble approach combining pixel-based and object-based features for the classification of high resolution remotely sensed imagery," in *2014 Third International Workshop on Earth Observation and Remote Sensing Applications (EORSA)*, pp. 140–144, June 2014.
8. V. Mnih and G. E. Hinton, "Learning to label aerial images from noisy data," in *Proceedings of the 29th International conference on machine learning (ICML-12)*, pp. 567–574, 2012.
9. Y. Wei, Y. Zhou, and H. Li, "Spectral-spatial response for hyperspectral image classication," *Remote Sensing*, vol. 9, no. 3, 2017.
10. W. Hu, Y. Huang, W. Li, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *Journal of Sensors*, 2015.
11. Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, pp. 2278–2324, Nov 1998.
12. J. Zhao, W. Guo, S. Cui, Z. Zhang, and W. Yu, "Convolutional neural network for sar image classification at patch level," in *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 945–948, July 2016.
13. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.
14. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, pp. 84–90, May 2017.
15. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *CoRR*, vol. abs/1409.4842, 2014.
16. E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Convolutional neural networks for large-scale remote-sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, pp. 645–657, Feb 2017.
17. S. Paisitkriangkrai, J. Sherrah, P. Janney, and A. V.-D. Hengel, "Effective semantic pixel labelling with convolutional networks and conditional random fields," in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 36–43, June 2015.
18. V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 2481–2495, Dec 2017.
19. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015.
20. S. Aich, W. van der Kamp, and I. Stavness, "Semantic binary segmentation using convolutional networks without decoders," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.
21. M. Volpi and D. Tuia, "Dense semantic labeling of subdecimeter resolution images with convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, pp. 881–893, Feb 2017.
22. R. Hamaguchi, A. Fujita, K. Nemoto, T. Imaizumi, and S. Hikosaka, "Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1442–1450, March 2018.

23. Z. Li, D. Yang, and Z. Chen, "Multi-layer sparse coding based ship detection for remote sensing images," in *2015 IEEE International Conference on Information Reuse and Integration*, pp. 122–125, Aug 2015.
24. X. Yang, H. Sun, K. Fu, J. Yang, X. Sun, M. Yan, and Z. Guo, "Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks," *Remote Sensing*, vol. 10, no. 1, 2018.
25. C. Benedek, X. Descombes, and J. Zerubia, "Building detection in a single remotely sensed image with a point process of rectangles," in *2010 20th International Conference on Pattern Recognition*, pp. 1417–1420, Aug 2010.
26. X. Chen, S. Xiang, C. L. Liu, and C. H. Pan, "Vehicle detection in satellite images by hybrid deep convolutional neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, pp. 1797–1801, Oct 2014.
27. H. Sun, X. Sun, H. Wang, Y. Li, and X. Li, "Automatic target detection in high-resolution remote sensing images using spatial sparse coding bag-of-words model," *IEEE Geoscience and Remote Sensing Letters*, vol. 9, pp. 109–113, Jan 2012.
28. W. Diao, X. Sun, X. Zheng, F. Dou, H. Wang, and K. Fu, "Efficient saliency-based object detection in remote sensing images using deep belief networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, pp. 137–141, Feb 2016.
29. J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, pp. 154–171, Sep 2013.
30. G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, pp. 7405–7415, Dec 2016.
31. S. Xu, T. Fang, D. Li, and S. Wang, "Object classification of aerial images with bag-of-visual words," *IEEE Geoscience and Remote Sensing Letters*, vol. 7, pp. 366–370, April 2010.
32. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 886–893 vol. 1, June 2005.
33. S. Tuermer, F. Kurz, P. Reinartz, and U. Stilla, "Airborne vehicle detection in dense urban areas using hog features and disparity maps," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 6, pp. 2327–2337, Dec 2013.
34. W. Zhang, X. Sun, K. Fu, C. Wang, and H. Wang, "Object detection in high-resolution remote sensing images using rotation invariant parts based model," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, pp. 74–78, Jan 2014.
35. E. Okafor, P. Pawara, F. Karaaba, O. Surinta, V. Codreanu, L. Schomaker, and M. Wiering, "Comparative study between deep learning and bag of visual words for wild-animal recognition," in *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–8, Dec 2016.
36. G. Cheng, P. Zhou, and J. Han, "Rifd-cnn: Rotation-invariant and fisher discriminative convolutional neural networks for object detection," in *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pp. 2884–2893, IEEE, 2016.
37. X. Chen, S. Xiang, C. L. Liu, and C. H. Pan, "Aircraft detection by deep belief nets," in *2013 2nd IAPR Asian Conference on Pattern Recognition*, pp. 54–58, Nov 2013.
38. G. Cheng, J. Han, L. Guo, Z. Liu, S. Bu, and J. Ren, "Effective and efficient midlevel visual elements-oriented land-use classification using vhr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, pp. 4238–4249, Aug 2015.
39. Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '10, (New York, NY, USA), pp. 270–279, ACM, 2010.
40. J. A. dos Santos Otávio Augusto Bizetto Penatti Ricardo da Silva Torres, "Evaluating the potential of texture and color descriptors f remote sensing image retrieval and classification," 2009.
41. B. Zhao, Y. Zhong, G. S. Xia, and L. Zhang, "Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, pp. 2108–2123, April 2016.
42. D. G. Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, vol. 2, pp. 1150–1157, Ieee, 1999.
43. F. Zhang, B. Du, and L. Zhang, "Scene classification via a gradient boosting random convolutional network framework," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, pp. 1793–1802, March 2016.
44. D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Deep learning earth observation classification using imagenet pretrained networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, pp. 105–109, Jan 2016.
45. Y. Yang and S. Newsam, "Uc merced land use dataset,"

46. F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, pp. 2175–2184, April 2015.
47. S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, pp. 2169–2178, 2006.
48. K. Nogueira, O. A. Penatti, and J. A. dos Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recognition*, vol. 61, pp. 539 – 556, 2017.
49. S. Jennings, N. Brown, and D. Sheil, "Assessing forest canopies and understorey illumination: canopy closure, canopy cover and other measures," *Forestry: An International Journal of Forest Research*, vol. 72, no. 1, pp. 59–74, 1999.
50. "Measuring the tree canopy cover in london: An analysis using aerial imagery," Tech. Rep. MSU-CSE-06-2, London SE1 2AA, September 2015.
51. D. Stojanova, P. Panov, V. Gjorgjioski, A. Kobler, and S. Deroski, "Estimating vegetation height and canopy cover from remotely sensed data with machine learning," *Ecological Informatics*, vol. 5, no. 4, pp. 256 – 266, 2010.
52. E. Guirado, S. Tabik, D. Alcaraz-Segura, J. Cabello, and F. Herrera, "Deep-learning versus obia for scattered shrub detection with google earth imagery: Ziziphus lotus as case study," *Remote Sensing*, vol. 9, no. 12, p. 1220, 2017. Copyright - Copyright MDPI AG 2017; Last updated - 2018-01-18; SubjectsTermNotLitGenreText - Ziziphus lotus.
53. S. Basu, S. Ganguly, S. Mukhopadhyay, R. DiBiano, M. Karki, and R. Nemani, "Deepsat: A learning framework for satellite imagery," in *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '15, (New York, NY, USA), pp. 37:1–37:10, ACM, 2015.
54. I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, "Deepglobe 2018: A challenge to parse the earth through satellite images," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.
55. M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, pp. 303–338, June 2010.
56. Google, "Google earth engine."
57. G. Fu, C. Liu, R. Zhou, T. Sun, and Q. Zhang, "Classification for high resolution remote sensing imagery using a fully convolutional network," *Remote Sensing*, vol. 9, no. 5, p. 498, 2017. Copyright - Copyright MDPI AG 2017; Last updated - 2017-09-22.
58. F. Chollet *et al.*, "Keras." https://keras.io, 2015.
59. M. Abadi and A. A. and, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software available from tensorflow.org.
60. F. Liu, G. Lin, and C. Shen, "Crf learning with cnn features for image segmentation," *Pattern Recognition*, vol. 48, no. 10, pp. 2983 – 2992, 2015. Discriminative Feature Learning from Big Data for Visual Recognition.

A Appendix I

A.1 Project Gantt Chart



Name	Begin date	End date
Project Planning	16/05/18	22/06/18
Review SOA	16/05/18	20/06/18
Write lit review	22/05/18	22/06/18
Submit literature review	25/06/18	25/06/18
Data Preparation	21/05/18	05/07/18
Prep of data pipeline	21/05/18	29/06/18
Acquire access to HPC	29/06/18	29/06/18
Obtain dataset from 3rd ...	29/06/18	29/06/18
Clean, analyse, visualise ...	29/06/18	05/07/18
Experimentation: Stage 1	29/06/18	17/07/18
Build FCN & Encoder/De...	29/06/18	05/07/18
Test architectures on su...	05/07/18	09/07/18
Determine final architert...	09/07/18	09/07/18
Train ...	09/07/18	17/07/18
Fine-t...	17/07/18	17/07/18
Experimentation: Stage 2	17/07/18	26/07/18
Data prep	17/07/18	17/07/18
Training	18/07/18	20/07/18
Testin...	23/07/18	24/07/18
Post-processing (optional)	24/07/18	26/07/18
Evaluation	26/07/18	02/08/18
Benchmark vs SVM	26/07/18	27/07/18
Benchmark vs De...	30/07/18	02/08/18
Dissemination	24/07/18	16/08/18
Write thesis	24/07/18	16/08/18
Prep gloss...	01/08/18	14/08/18
Submit thesis	17/08/18	17/08/18