Masters Program in Geospatial Technologies

SEMI-AUTOMATIC CLASSIFICATION OF TREE SPECIES USING A COMBINATION OF RGB DRONE IMAGERY AND MASK RCNN:
Case study of the Highveld region in Eswatini.

Yan-Liang Lin

Dissertation submitted in partial fulfilment of the requirements
for the Degree of *Master of Science in Geospatial Technologies*

NOVA IMS Information Management School
Universitat Jaume I
Westfälische Wilhelms-Universität Münster

Semi-automatic classification of tree species using a combination of RGB drone imagery and Mask-RCNN: A case study of the Highveld region of Eswatini

Dissertation supervised by

Jan R.K. Lehmann, PhD

Remote Sensing and Spatial Modelling Research Group,

University of Münster

Münster, Germany

Dissertation co-supervised by

Joel Dinis Baptista Ferreria da Silva, PhD

Instituto Superior de Estatística e Gestão de Informação,

Universidade Nova de Lisboa

Lisbon, Portugal

Dissertation co-supervised by

Filiberto Pla Bañón, PhD

Institute of New Imaging Technologies,

Universitat Jaume I

Castellón de la Plana, Spain

February 2021

Declaration of Originality

I declare that the work described in this document is my own and not from someone else. All the assistance I have received from other people is duly acknowledged and all the sources (published or not published) are referenced.

This work has not been previously evaluated or submitted to NOVA Information Management School or elsewhere.

Lisbon, Portugal, February 2021

Yan-Liang Lin

Acknowledgments

I would like to thank my supervisor Prof. Jan Lehmann for his advice and encouragements during the thesis process, there are times when I was flagging and worried and he rallied me to produce better work and look forward not backwards.

I am very grateful for the guidance offered by my co-supervisors Prof. Joel da Silva and Prof. Filiberto Banon.

To my family and friends who encouraged, believed, and supported me during this process from start to end, I would like to express my eternal gratitude and love. Ma and Ba I finally made it! To the Lin babies, I have finally crossed this finish line. To my brother Brendon, thank you for being the rock and gnat in my ear, this wire has been pulled to the last.

To my little family abroad who I was adopted by and formed irreplaceable bonds with during this MSc. in Geospatial Technologies, you know who you are, you will forever hold a special place in my heart, a place that I did not even know I possessed.

SEMI-AUTOMATIC CLASSIFICATION OF TREE SPECIES USING A COMBINATION OF RGB DRONE IMAGERY AND MASK RCNN

A case study of the highveld region of Eswatini

Abstract

Tree species identification forms an integral part of biodiversity monitoring. Locating at-risk species and predicting their distribution is equally as important as tracing invasive alien plant species distributions. The high prevalence of the latter and their destructive impact on the environment is the focus for this thesis. In areas of the world where technology limitations are restrictive, an approach using low-cost, available RGB drone imagery is proposed to train advanced deep learning models to distinguish individual tree species; three dominant species (*Pinus elliotti, Eucalyptus grandis* and *Syzygium cordatum)* providing the bulk of sampling data, of which the first two are highly invasive in the region. This study explored the efficacy of utilizing Mask RCNN, an instance segmentation deep neural network, in identifying multiple classes of trees within the same image. In line with the low-cost approach, Google Colaboratory was utilized which drastically lowers the training time necessary and alleviates the need for high GPU systems. The model was trained on imagery from three study areas which were representative of three distinct landscapes: very dense forest, moderately dense forest with overlapping canopies, and open forest. The results indicate decent performance in open forest landscapes where overlapping tree crowns is infrequent with mean Average Precision of 0.71. On the contrary, in a dense forest landscape with many interlocking tree crowns, a mean Average Precision of 0.43 is highly indicative of the model's poor performance in such environments. The trained network was also observed to have higher confidence scores of detected objects within the open forest study areas as opposed to dense forest.

Keywords

Deep Learning

Mask RCNN

Invasive Alien Plant Species

Unmanned Aerial Vehicle

RGB Imagery

Acronyms

ANN	Artificial Neural Networks
CHM	Canopy Height Model
CNN	Convolutional Neural Network
DL	Deep Learning
FPN	Feature Pyramid Network
HS	Hyperspectral
IAPS	Invasive Alien Plant Species
NMS	Non-Max Suppression
RCNN	Region based Convolutional Neural Network
RGB	Red-Green-Blue
RPN	Region Proposal Network
ROI	Region of Interest
SGD	Stochastic Gradient Descent
UAV	Unmanned Aerial Vehicle
VIA	VGG Image Annotator
VGG	Visual Geometry Group

INDEX OF THE TEXT

INDEX OF TABLE OF FIGURES

1 Introduction

1.1 Contextual Background

Under the current critical climate change situation, biotic invasions by exotic plants are one of the most significant threats to vital ecosystem functioning(Kumar Rai & Singh, 2020). According to the International Union for Conservation of Nature (IUCN) the definition of Invasive Alien species are "...species that are introduced, accidentally or intentionally, outside of their natural geographic range and that become problematic"(IUCN, 2018). Apart from impacts on biodiversity, biological invasions affect environmental, socio-economic, and cultural changes on impacted areas.

Invasive plants across the world, although not as noticeable as natural disasters, contribute to losses and hardships faced by people that are proposed to be an order of magnitude higher than natural disasters(Ricciardi et al., 2011). They disrupt food chains and affect specific mutualisms between plants and animals which are felt in pollination and seed dispersal. Additionally, their invasive nature itself poses a threat to local biodiversity, often leading to catastrophic changes in local populations and in some cases, the extinction of indigenous species. Their incursion has been linked to land use change such as forest to grassland/savannah via alterations in the natural fire regimes(Pyšek et al., 2012). This has further implications for carbon sequestration and therefore cumulative impact on climate change. In light of the shifting climate and globalisation, the influence of Invasive Alien Plant Species(IAPS) is expanding rapidly in a positive feedback loop, with "transport, climate change and socio-economic change" being cited(Essl et al., 2020) as the main drivers for alien species invasion. IAPS further have repercussions for human health either directly by releasing certain toxins which can cause diseases or allergies, or indirectly as vectors for other pathogens to leverage in their propagation (*Lantana camara* for tsetse fly that causes sleeping sickness)(Kumar Rai & Singh, 2020).

In the Kingdom of Eswatini, the introduction of exotic plant species began with European colonisers bringing species such as Australian wattle for firewood and Eucalyptus and Pine for timber and building materials. These and many other introduced species have gained a foothold in Eswatini to such an extent that indigenous plants are outcompeted, and natural ecosystems heavily disrupted. A cascade effect follows their incursion, affecting aspects such as water availability, land use and cultivation, etc. Under the UN's Sustainable Development Goals, IAPS are highlighted under Goal 15, Life on Land as they pose a threat to ecosystem functioning.

Monitoring and detecting IAPS is the first step towards their control. In the latest report of the national strategy for the control of IAPS in Swaziland(Dlamini, 2020), a dedicated group is to be established to deal with the growing threat of IAPS. Yet it begs the question, how can we survey large plots of land for the presence of IAPS with limited technology at a low cost? Remote Sensing and Earth observation offer some of the most promising approaches to effective IAPS monitoring. With satellite data, airborne imagery and drone photography progressing rapidly and becoming increasingly available, people on the ground have many more tools at their disposal for plant monitoring and management.

The development of Unmanned Aerial Vehicles (UAVs) presents huge potential for advances in ecosystem monitoring. They drastically reduce surveying time and can take high resolution snapshots of areas that can be stored, combined, and then analysed with machine-learning, the use of UAV for the management of IAPS holds vast potential. There is growing interest in the combination of machine-learning with UAV imagery to build better classification models. The advent of deep learning, which utilizes not just the spectral information, but shapes and sizes of objects in imagery is at the forefront of these developments. Thus, this thesis proposes to use UAV imagery in combination with a deep learning algorithm (Mask RCNN) to build a semi-automatic classifier for detecting IAPS in Eswatini.

1.2 Problem Statement and Motivation

Plants and animals have been transported outside of their original ranges for centuries (accidentally or intentionally: alien/exotic), and a necessary definition for distinguishing between naturalised and invasive species was proposed by David M. Richardson. The definition of IAPS in this paper follows this concept, which in short states that naturalized plants are exotic plants that can sustain populations over many life cycles without human intervention. These essentially differ from invasive plants which are "Naturalized plants that produce reproductive offspring, often in very large numbers, at considerable distances from parents plants…"(Richardson et al., 2000).

Invasive species not only have effects on the natural environment and its wildlife, but they also exert pressure on agriculture and water security, with further effects on economy and livelihood of people. A study aimed at investigating the impact of IAPS on water flows in South Africa estimated that through the increased evaporation and transpiration losses that IAPS bring, an estimated 1.44-2.44 billion m^3 per year is lost in surface run-off in South Africa, and an estimated 193 million m^3 annually in Swaziland(Le Maitre et al., 2020). These are expected to increase with further invasions of IAPS and under existing shifting climatic conditions.

Many emaSwati are dependent on the land they live on, and the presence of IAPS reduces their ability to cultivate land, access to water and threatens their livelihoods. As invaders are foreign, the local people usually do not exploit them for their potential uses, they face no natural predators and may pose a threat to locals or wildlife if they are poisonous. Furthermore, invaders usually spread rapidly, taking up valuable land which becomes harder to cultivate or use and outcompetes local flora and fauna. All these factors indicate that IAPS in Southern Africa pose a threat to sustainability in the regions they invade. In Eswatini where a large portion of the economy is dominated by agriculture, the management of IAPS is important to improving living standards.

1.3 Research Questions

This work is aimed at assessing the use of the deep learning model Mask RCNN towards identifying and locating IAPS in natural environments. To fulfil this aim, the following research questions are specified:

a) What are the optimal hyperparameters to train the Mask RCNN model towards identifying multi-class trees in natural environments?
b) What forest classes are suitable for implementation of the Mask RCNN deep learning model for classification and localization of tree species?

1.4 Contribution

To the author's knowledge, this study is one of the first forays into exploring the use of high-resolution drone imagery in combination with instance segmentation to perform multi-class tree detection. The main contributions of this thesis consist of:

- Exploring feasibility of low-cost drone equipment to survey IAPS impacted ecosystems.
- Evaluating the performance of Mask RCNN for multi-class tree extraction in varying natural forest landscapes

1.5 Methodology

This thesis is structured into 4 stages namely, i) review and choice of method to detect IAPS; ii) data collection and processing; iii) Implementation and optimization of chosen method and iv) evaluation and performance comparison.

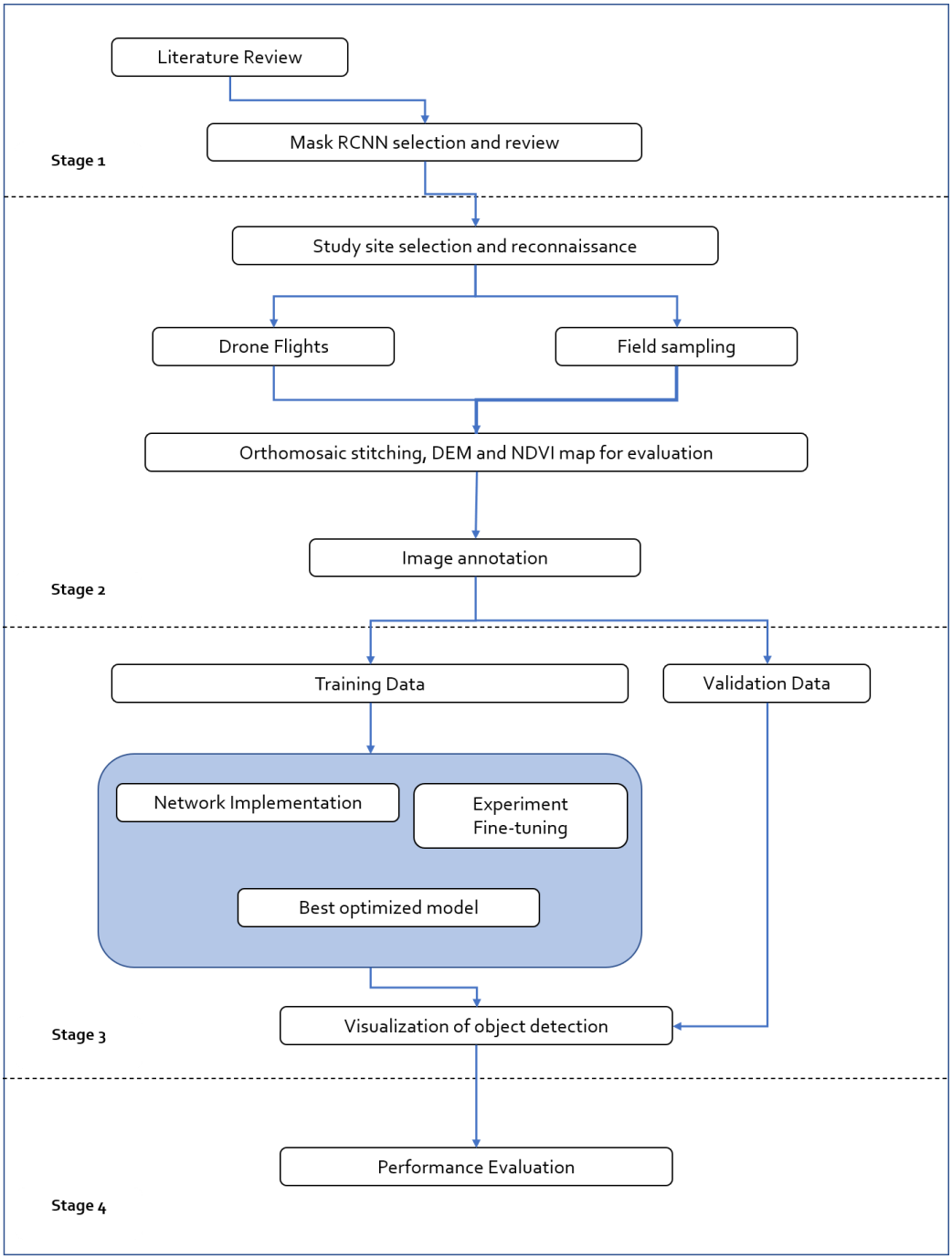

Figure 1: Thesis workflow

In the first stage, a review of the traditional approaches to IAPS monitoring was conducted. Given the available avenues of data collection and equipment, Mask-RCNN was selected as the most suitable architecture that could produce good object detection and classification results from high resolution RGB drone imagery. Mask-RCNN also outputs masks of the identified objects which would be vital towards the IAPS management process.

In the second stage, existing bastions of established IAPS were evaluated in the Highveld region in consultation with local plant ecologists. The Highveld Region of Eswatini where the capital is located has many riparian areas and the montane environment of Pine Valley is self-indicative of the extent of IAPS. As exotic species (gum, pine and wattle) grow exceptionally well in disturbed environments and outcompete local flora, they have established themselves heavily in this area which was selected. Data collection consisted of drone imagery acquired via drone pilots with guidance from the ecologists who also provided the ground truth data via in-situ field sampling between August and October. The images obtained were stitched together using DroneDeploy and visualised with QGIS to provide accurate field site maps of the study areas. Individual photos were then labelled with VGG Image annotator to produce labelled ground truth data to train the Mask RCNN models.

The third stage consisted of designing and implementing the Mask RCNN model with the dataset pre-processed from stage 2. After initial literature review and study, Mask RCNN was chosen as the suitable architecture to base the model on as the instance segmentation advantages can generate relevant output for IAPS management. Exploration of the best hyperparameters and backbone architecture for Mask RCNN was followed by extensive training of the model with optimal parameters with the prepared dataset.

In the final phase, the performance of the model was evaluated keeping in mind the two stages that Mask RCNN is composed of; namely the Region Proposal Network (RPN)

and the ensuing Classification, Bounding Box and Mask Generation that make up the second part of Mask RCNN. The metrics used were both qualitative with visualization of detection results using validation imagery, and quantitative in the form of calculations of average precision for evaluation and allowing a mean Average Precision (mAP) to be calculated for the dataset.

2 Literature Review

This chapter elucidates the reasoning behind the direction of this thesis and the considerations taken towards building new knowledge in remote sensing of invasive plants. The review begins with section 2.1 analysing the approaches previously trialled with satellite and airborne surveillance towards IAPS detection and monitoring. Following this, in section 2.2 a review of the advancements in Unmanned Aerial Vehicle (UAV) plant detection is expounded upon with the considerations towards low-cost drone surveillance and the new approaches that deep learning can leverage with only high-resolution RGB imagery.

2.1 Traditional approaches to Tree Species Identification

Remote sensing techniques serve as an effective way to monitor environmental problems such as IAPS. Satellites have the advantage of being equipped with advanced sensors and can capture vast swathes of land at once. However, the resolution of free satellite imagery is typically not high enough for species level distinction and individual location of IAPS especially in complex natural landscapes, which is the case for IAPS invasions.

The Sentinel-2 constellation has been investigated for the potential to identify forest species using spectral data with low accuracy after cross-validation (65%)(Immitzer et al., 2016). In a separate study investigating mangroves in coastal areas, (Wang et al., 2018) evaluated the use of Sentinel-2, Landsat 8 and Pleiades-1 in mapping mangrove extent and species. The results indicated good mapping of mangrove extents, but when it came to community level species identification, the findings had low-medium accuracy when using free satellite data from Sentinel and Landsat (70.95% and 68.57% respectively). Although Pleiades-1 imagery yielded decent accuracy (78.57%) it should be noted that the study area comprised of 17 different species, from which their Random Forest Model classified them into a total of 6 communities. This approach of grouping

species into community levels makes sense depending on the objective of the project and the morphological species variation.

Even when using higher resolution satellite data replete with more spectral information, (Abutaleb et al., 2020) found that when researchers used machine learning algorithms with Worldview-2 and Spot-7 imagery to map eucalyptus trees in Johannesburg city, the overall accuracies were 81.67% using Worldview-2 and 72.78% using SPOT-7 imagery and lower user accuracy of 73.77% and 60% respectively using the better performing Random Forest algorithm. Although imagery from Worldview-2, of higher spatial and spectral resolution performed better, the results for eucalyptus mapping was still below or at 80%. Understandably higher resolution imagery is only available commercially and this has proved to be both expensive and data heavy. These costs are significant obstacles to IAPS management and alternatives to IAPS monitoring have been discovered with the entrance of Unmanned Aerial Vehicles.

2.2 Unmanned Aerial Vehicles and Deep Learning

Alternately, drones are becoming more common and their incorporation for remote sensing due to their low-cost and rapid deployment in various situations makes them very attractive for areas that have technological constraints. UAVs can be equipped with increasingly complex sensors like hyperspectral (HS) and LiDAR to differentiate tree species and the combination of these sensors have yielded high accuracies. A study by (Nezami et al., 2020), compared the efficacy of Deep Learning Convolutional Neural Networks (DL CNN) with various datasets ranging from simple RGB channels, Hyperspectral and, a Canopy Height Model (CHM) separately and in a combination to classify trees into the 3 most common types of tree in a forest in Finland. Demonstrably high accuracies were obtained for each "species" using a 3D-CNN (between 99.6-94.8% producer's accuracy) that were also markedly better than classification with a Multi-Layer Perceptron (MLP) model. The trees were being classified into spruce, pine, and birch.

However, these sensors are usually rather expensive, and processing of this data requires a level of expertise and processing power that is not always available. Recent developments of drones and low-cost sensors are establishing an important place for UAV monitoring of IAPS. As most drones nowadays come with a standard high-resolution RGB camera, leveraging these cameras for IAPS identification is suitable for areas where IAPS are problematic but funding is not readily available for more advanced sensors. The limitations of RGB cameras in plant identification is slowly being overcome with the advent of computer vision. Deep Learning that uses Convolutional Neural Networks (CNNs) have been tested with high resolution drone imagery in segmenting plant species with successful results, better than typical pixel-based methods(Kattenborn, Eichel, et al., 2019). The authors mention the further exploration of similar CNN based tree segmentation either semantically, or using instance segmentation methods, such as Mask-RCNN.

Machine-learning methods can be applied to a plethora of imagery, ranging from simple RGB to Light Detection and Ranging (LiDAR to enhance and infer more information. Although advanced sensors have been found to generate highly accurate predictions of tree identification and spread when combined with machine learning(Underwood et al., 2003)(Naidoo et al., 2012), advances in convolutional neural networks in combination with simple RGB imagery has been found to have equal or even better predictions than standard machine learning algorithms. Researchers have found that convolutional layers successfully predict certain plant species using more than just spectral indices. Other factors such as shape and texture are successfully incorporated by deep learning to generate more accurate predictions on plant species.

Despite their ability to leverage more features of objects for segmentation and classification prediction,(Kattenborn, Eichel, et al., 2019) discussed how limitations exist for such methods. They propose further exploration into CNN models and how the performance of said models would in landscapes with varying vegetation patterns. In

increasing spatial resolution of UAV imagery, the trade-off lies in the total area coverage for plant monitoring decreasing. As drones must fly lower to get higher detail photos, the total area they can cover will then decrease. Mask RCNN was mentioned as one of the alternatives to semantic segmentation, by being able to detect individuals within object classes, more detail can be extracted and used for plant monitoring. The masks generated could also be used to provide estimates of plant cover on the ground. By combining raw RGB imagery with ancillary products such as elevation and Structure from Motion 3D data, accurate classification maps could be generated.

Given that this project intends to explore the use of drone RGB imagery for plant monitoring, the suggested Mask RCNN model is attractive for several reasons. Firstly, multiple CNN models have been shown to perform well in tree and plant cover detection. Most focus on binary classification methods run for separate species over various landscapes. An exploration into multi-class species detection is proposed that aims to identify multiple species in the same image. Mask RCNN is one of the few models that allows for instance segmentation, which differs from semantic segmentation in that individual objects are clearly delineated, and masks of these individuals generated. This is desirable for IAPS monitoring as it allows for counts of trees to be generated as well as the total plant cover.

3 Theoretical Background

The following chapter lays the theoretical foundation that form the backbone for the concepts utilized in this thesis. Primarily, it provides information on classical deep learning models used in Object Detection and Classification. Further it explains the framework and inner workings of the Mask RCNN model and expounds on both the benefits and drawbacks of using such a model.

3.1 Traditional Machine Learning approaches to plant monitoring

Much of the research into plant monitoring has utilized descriptive analysis and traditional machine learning methods such as Support Vector Machine (SVM) and Random Forest (RF) models to solve the problems of invasive tree location. Whilst these approaches are powerful and in some scenarios are superior to the chosen deep-learning tactic employed in this thesis, relying on these descriptive models has a pitfall when one considers that there still be some unmodelled variables that the engineers and scientists have failed to consider due to their hidden, complex, or non-intuitive nature. In comparison, predictive analysis, which forms the core of deep-learning models, seeks to minimise the error between the actual and predicted outcome. This is tackled by feeding the model a large set of training data which inherently possess certain patterns that will play a part in the model's computation of new patterns which are relevant to the problem at hand.

3.2 Artificial Neural Networks (ANN) and Deep Learning

ANNs were born from a deeper understanding of the human central nervous system and borrows from the biological framework of our brain. By mimicking this structure researchers can build more intuitive models for problem solving. One implication of an inter-connected structure of "neurons", is that Artificial Neural Networks eliminate the need for an all-inclusive understanding of all problem variables. This predictive approach has been shown to produce highly accurate models. It should be noted, however, that this may be sacrificing a deeper understanding of the variables that

contribute towards a pattern. Given that this study is not seeking to discover or explain new variables in invasive tree species distribution, ANNs are a valid methodology for this objective. The goal of this work is to explore whether deep learning can produce accurate and reliable IAPS output for better management without burdening itself with explaining the distribution on an ecological level. Stemming from the core ANN, Deep Learning delved to utilize more complex layers which can better capture the intricacies of the input data or phenomenon. Whilst Deep-Learning (DL) further branches into many separate models we will be focusing on Convolutional Neural Networks which have been recognised as a high performing model in the object detection field.

3.3 Convolutional Neural Networks (CNN)

CNNs are characterised by the addition of multiple hidden convolutional layers and filters, the hierarchy of which allows input data to be organized into smaller and simpler patterns from which more complex patterns arise. CNN research and application has made rapid progress in recent years, this boost can be largely attributed to 2 factors: the development and availability of high computing power which has sped up the training time and feasibility of such grand projects and secondly, the availability of large amounts of data to train these networks with. This section will explain the general framework of CNNs followed by a review of a few landmark models developed for object detection.

As an end-to-end learning framework, CNN models rely on updating weights between each layer which play a part in the model's ability to make more accurate predictions. The defining feature of a CNN are the convolutional layers, hidden layers between the input and output layer which are made up of learnable kernels (also called filters). These filters are applied over an image in a sliding-window fashion typically dealing with an *n x n* pixel size of the image. As this *n x n* sized filter slides across the image, the scalar product is calculated for each value in the kernel. The kernels extract features which are captured as output and fed-forward to the next layer.

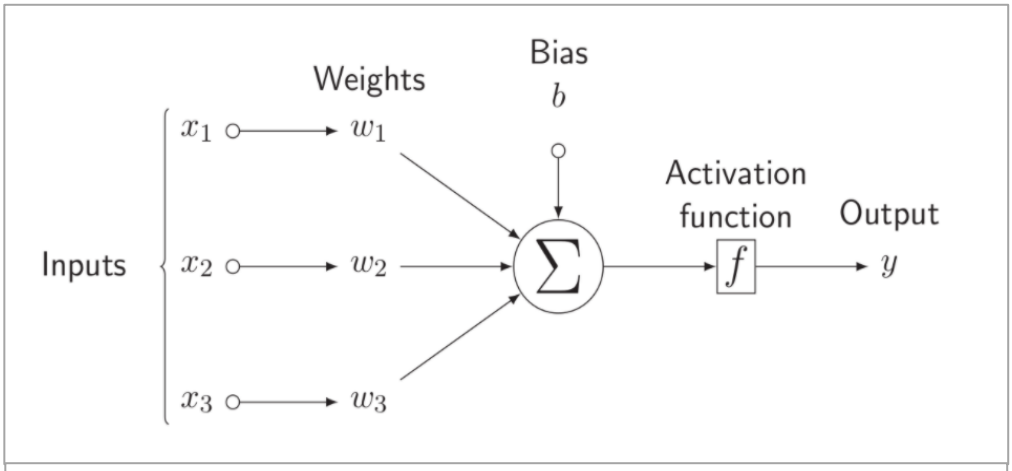

Figure 3.1: Fundamental building block of CNN, a fully connected neuron

The diagram illustrates how inputs pass to outputs, the input is multiplied by the weights and the bias value. Weights define the strength between the input and output. Stronger weights will influence more and vice versa. Additionally, the bias value is introduced as constants. They are not influenced by the training and updates that previous layers impart on the model, instead they serve to ensure that activation will be achieved even in the case that all inputs result in zero. The sum of all the weights as well as the bias are then passed to an activation function which produce the output.

The **activation function** that the sum of the values is passed into allows for non-linearity to be embedded into the model. There are various activation functions such as the Rectified Linear Unit (ReLU) and Sigmoid. With typical Sigmoid activation functions, there is a problem with a vanishing or exploding gradient which occurs with high

initialized weights or a vanishing gradient when the gradient tends to zero. To avoid this, ReLU activation function thresholds the negative values to zero and allows positive values to retain their original value.

A core part of Neural Networks is their ability to backpropagate. This is achieved by taking the weights that have been calculated across layers and feeding them backwards into previous layers which will then update their weights after each iteration.

Pooling is done to reduce the stress on computation that arises from such complex networks. By reducing the dimensions by taking the maximum value or average value of the pre-defined filter (e.g 3x3) and using those as the values for the neuron in the next layer, a pooling of the prior layer's values is performed.

When the outputs are calculated in layers, a loss value can be calculated. This is understood as the difference between the predicted value and the true value of the object. For example, if the final predicted value is 0.75 and the actual value is 1, the loss is then 0.25. By minimizing these losses after many iterations, the model can make more accurate predictions which are in effect, closer to the true value of the object. In this regard, optimization algorithms are applied so that these loss functions can be decreased, and the prediction capability maximized.

Stochastic Gradient Descent (SGD) with momentum

During training, the loss is calculated for each of our inputs. The gradient of that loss is then obtained with respect to each of the weights in the model. The addition of momentum remembers the updated change in weights at each iteration. The term momentum is borrowed from physics, where a particle accelerating in a direction, the predicted value, still travels even after no further acceleration or force is applied. This prevents the value from getting stuck in a local minimal and streamlines the learning in the direction of the actual value. By using SGD, the exponentially weighted average of the weights is calculated. This is done in conjunction with the learning rate, a pre-

defined hyper-parameter that indicates how big of a *step* that the model should take in the direction of its weights. Finally, the dot product of learning rate and gradient will be used when updating the weights during backpropagation as the model learns.

Backbone architecture

A powerful ability of CNNs is that after a framework or "backbone" has been decided, additional segments can be added to address different objectives such as object recognition, image segmentation and instance segmentation. Many backbones have been built and trialled on various datasets to further computer vision. Of these, the Resnet Family of backbones is utilized by our chosen Mask RCNN model. When deep learning models first appeared, researchers showed a problem regarding the number of layers involved. By comparing 20 layer and 56 layer networks, they found that the deeper layers resulted in higher training and test errors(He et al., 2016). ResNet architecture overcomes this problem by introducing a skip connection function between layers. This circumvents the errors that arise when the spatial resolution of an image changes (e.g. by a 3x3 convolution on a 32x32 image resulting in a 30x30 image). By adding the original identity to the final weight output, the original "identity" or value is preserved despite learning of the model pushing it further from the original.

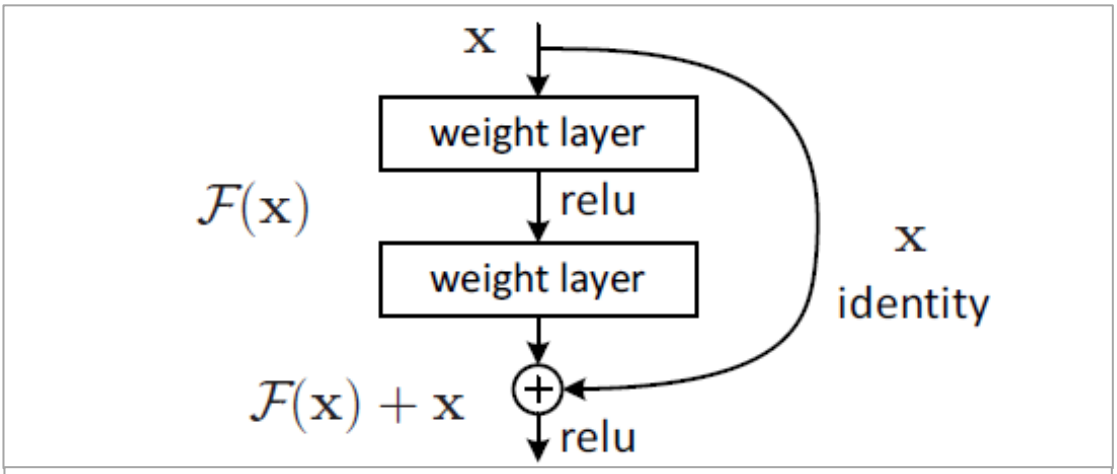

Figure 3.2: Skip connection for 2 layers introduced by ResNet (He at al. 2015)

ResNet50 and ResNet101 are two such models that adopt the deep residual learning approach mentioned above. They vary in complexity of the number of convolutional layers used with 3-layer block skip connections. Smaller ResNet models (ResNet34) use 2-layer block skip connections.

3.4 Region-Proposal Convolutional Neural Networks (RCNN)

In computer vision, object recognition is a term that is used to describe a collection of tasks that are used to identify objects from a digital image input. RCNNs were developed to better localize where individual objects were within an image which had multiple objects from the background. To better understand this, we can group these into 4 tasks: Image Classification, Object Localization, Object Detection and, Object Segmentation.

Image classification predicts the type or classes of object that are present in an image. The input being an image with a single object present and the output being a class label that has corresponding labels associated with them.

Object localization deals with locating an object within an image and proposing a bounding box that localizes where the object is in said image. The input is an image with one *or more* objects present and the output is the bounding box(es) generated around said objects.

Object detection is the cumulative product of the previously mentioned tasks, it detects the location of objects within an image and produces a bounding box as well as corresponding class label predictions for these proposed objects. Input is an image with one or more objects and its outputs bounding box(es) with a class label for each bounding box. From this output RCNNs branch into providing either semantic segmentation, all objects with the same class label were labelled collectively, or instance segmentation which defined each object as individuals within the image.

Object segmentation is performed after the detection of the object within an image and defines the areas taken up by recognized objects on a pixel-by-pixel basis. This is superior to the localization step because the object is not delineated by a rough bounding box but instead the edges are clearly defined and can be associated with the object more accurately. In the final chosen model, Mask RCNN explicitly describes a mask for each object detected.

3.2.1 RCNN

When object detection first arose, researchers built the first R-CNN. The following figure details how this was structured.

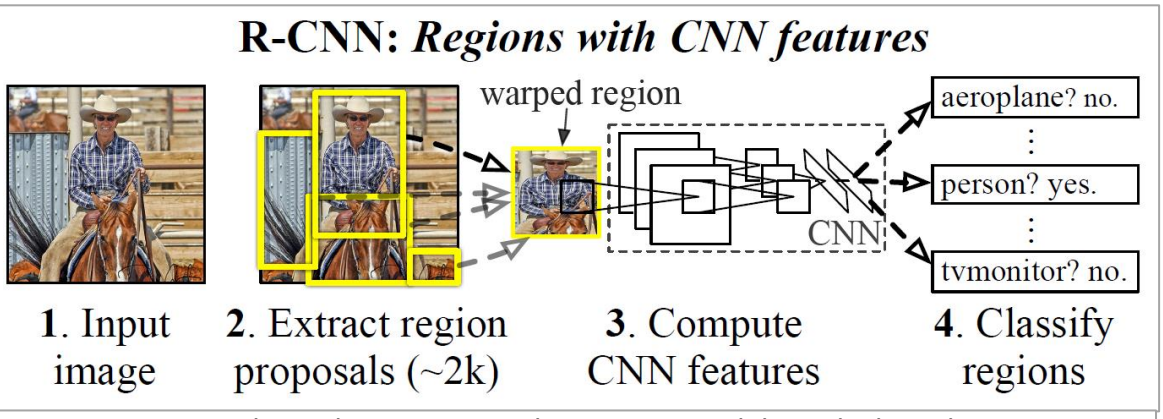

Figure 3.3: Object detection steps by a RCNN model (Girshick et al.,2016)

This can be summarized into a few steps:

1. Generate proposals for bounding boxes (bbox)
2. Pass the areas in the bounding boxes through a pre-trained model to determine the label of the object.
3. Pass the box through a linear regression model to improve the coordinates and tighten the boundaries once the object has been classified.

To produce the bounding boxes, a process called Selective Search is used where windows of different sizes are placed over the image and by grouping together pixels

from varying windows of the same texture, colour and intensity, final boxes are proposed. It should be noted that although this method was accurate, it involves a lot of computing power.

3.5 Faster RCNN

Scientists discovered that the selective search process was still a major bottleneck for the process. A workaround was proposed where instead of the region proposer relying on this method it would reuse the results from the first classification step performed by the CNN. Thus, the proposed regions would borrow feature maps that had already been generated and would not have to be run separately(Ren et al., 2017).

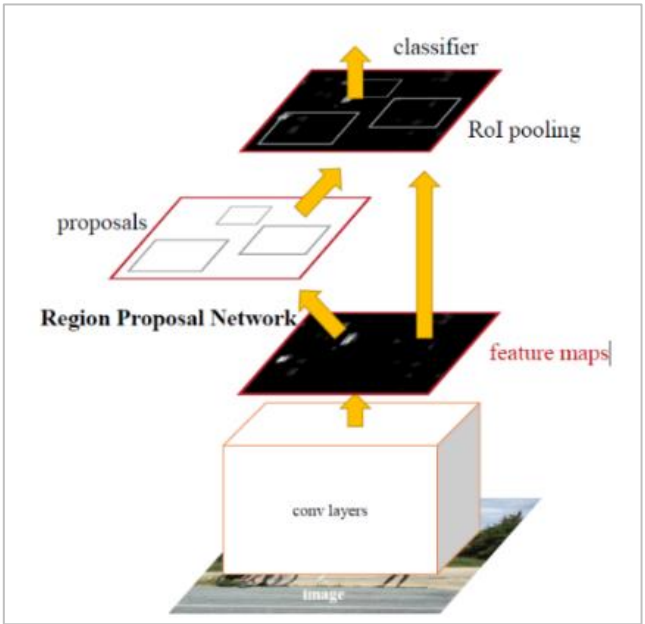

Figure 3.4: Region Proposal Network stage feeding Region of Interest layer (Ren et al. 2017)

3.6 Mask RCNN

Previous developments had optimized the runtime of RCNNs toward object detection. Although still heavy on computing power, there was a great improvement in the times needed to run the models. Up to this stage, the object localization was narrowed down to the bounding boxes generated by the RCNN models, researchers now moved

towards object segmentation. Various researchers attempted to build on the Faster-RCNN framework and improve segmentation, and of these Mask RCNN has proved to be the most viable model leading the field of object detection. The method proposed functioned in parallel with the classification and bounding box generation and created masks of the objects. This feature map is made up of binary output detailing whether the pixel was part of the object (1) or not (0). The hurdle of losing information from pooling features was overcome by their novel RoI Align method.

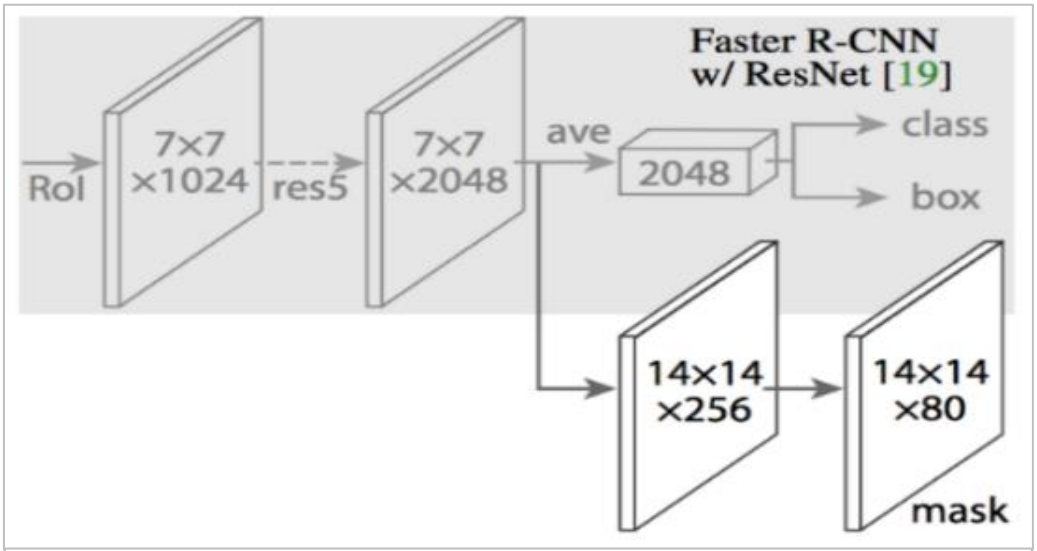

Figure 3.5: Mask RCNN architecture building on Faster RCNN (He at al. 2018)

Region of Interest Align solves the issue that arises from loss of information when feature maps are pooled in convolutional layers. As max pooling occurs, the dimensions of the feature map decrease, as bounding boxes are generated over only images, by reducing the size of the image, the similar transformation for the bounding box occurs and intuitively the aspect ratio for the objects bounding box may not match and typically the values are rounded because they are stored as integers. RoI Align enables these values to be stored as floats using bilinear interpolation which retains values while still undergoing the relevant transformations. The final output consists of the bounding box around predicted objects, the class label, and a mask of where it resides in the image. This thesis will focus on leveraging Mask RCNN to identify and generate masks for identified trees in Eswatini.

4 Study area and methodology

This chapter provides background into the study areas chosen for this project in the first section. Further, data acquisition, visualization and pre-processing are the focus of the second section and finally image annotation is described in the last section.

4. 1.1 Study area

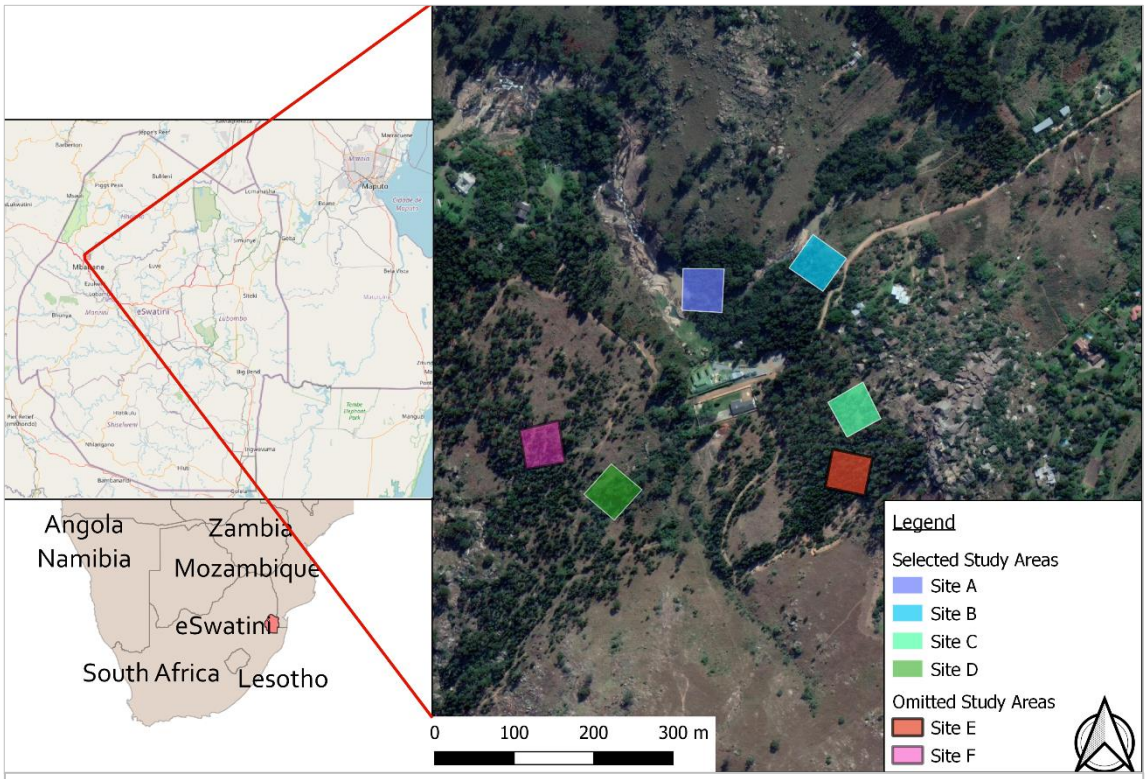

Figure 4.1: Study Area locations in Pine Valley, eSwatini

Eswatini is a landlocked country found between Mozambique and South Africa. There are 4 ecological regions, and this study focuses on the Highveld region, which is more mountainous, higher in elevation with cooler temperatures. The region's climate is characterised by mild winters when most rainfall is recorded and hot arid summers. The mountainous terrain dictates the course of runoff and the valleys and interlocking spurs between ranges are where the unique Afromontane forests (mist belt forest) can be

found. With increasing human settlement, these valleys and riparian areas have become gradually more disturbed which allows for invasion by IAPS.

Under consultation from local ecologists, the Pine Valley locality was explored for viable study areas using verification with Google Earth. Six sites were examined in total and after reconnaissance of the six by the ground truth team, 4 final sites were selected. This selection was based on the number of mature tree species present, the accessibility of the sites for field studies and the tree cover (open forest, moderately dense forest, and very dense forest). This study followed the definition of forest cover by (Forest Survey of India, 2013), which details that very dense forest is land with a tree canopy density of 70% or more, moderately dense is tree canopy density between 40 and 70% and open forest has more than 10% but less than 40% tree canopy density. The three final study areas are representative of these different landscapes and the plant ecologists helped to choose the study sites based on their classification of these landscapes into the three aforementioned forest classes.

4.1.2 Ground truth collection

As this study aims to identify trees, the dominant invasive species were established to be 1) Slash Pine (*Pinus elliottii),* 2) Black Wattle (*Acacia mearnsii),* 3) Red gum (*Eucalyptus grandis).* Additionally, dominant indigenous trees included: 4) Waterberry (*Syzigium Cordatum)* and 5) Fig (*Ficus bubu)*, with waterberry being the class in majority. Although this project undertakes to detect Invasive Alien Plant Species, the ability of the model to also classify commonly occurring indigenous species was investigated. Taking into consideration the model will be trained on drone imagery, it was decided that grouping indigenous trees into one class would only lower the classification ability especially since their morphology varied greatly between species. Morphological differences are hypothesized to assist in the model's learning.

Ground truth collection consisted of field visits to the sites and convenient random sampling whereby the tallest trees with the most exposed crowns were chosen. GPS

coordinates of trees were taken by laying a Garmin GPS instrument at the trunk. Upon inspection of these GPS coordinates overlaid with the orthoimages, it was found that some coordinates were offset from trees and this was accredited to poor GPS signal resulting from dense canopy cover. Site B was determined by ground crew to be difficult to traverse and collect data, so a plant specialist assisted to perform visual inspection of the stitched orthoimage to determine tree species from drone imagery.

4.1.3 UAV data collection

Each study area was 50m x 50m in dimension. The study areas were mapped with google earth and the kmz files were supplied to drone pilots so that they could plan appropriate grid missions with a mobile app drone route planner. Pix4D and DroneDeploy apps were used by the drone pilots. The first two study areas (A and B) were surveyed with a DJI Mavic Pro with a set altitude of 50meters above ground level. Unfortunately, when surveying Site C, the drone pilot crashed his drone into a tree. This resulted in the hiring of a second drone pilot to survey Sites C and D at an altitude of 56meters above ground level with a DJI Phantom 4. Both drones had similar camera specs of 12.3 Megapixels with a CMOS sensor of ½.3. The image sizes were 4000 x 3000 pixels, and the sensors were only capable of collecting RGB data. Site C drone imagery was collected near dusk and shadow was deemed to add too much noise to the imagery, thus it was omitted.

4.2.1 Image Processing

After drone imagery was collected, they were processed with DroneDeploy software to provide stitched orthomosaics, Digital Elevation Models and Visible Atmospherically Resistant Index (VARI) maps. VARI maps were used in lieu of NDVI maps because no NDVI sensors are available and it was deemed the next best visualization method to measure plant health. Orthomosaics of all 4 study sites are presented in the following figures.

Figure 4.2: Orthomosaic of Site A

Figure 4.3: Orthomosaic of Site B

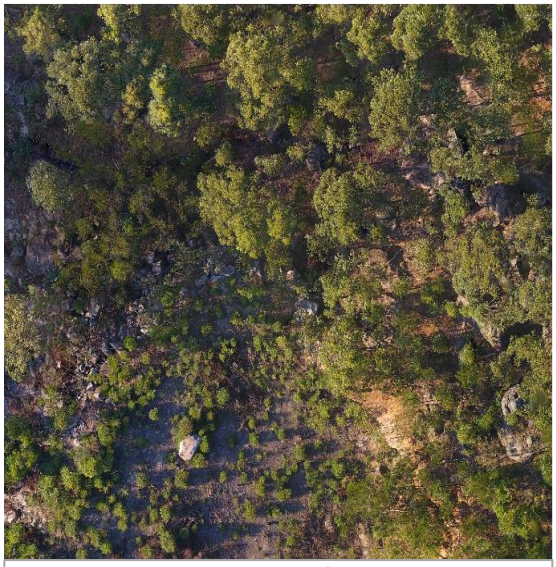

Figure 4.4: Orthomosaic of Site C (ommitted)

Figure 4.5: Orthomosaic of Site D

With assistance from the plant specialist, more ground truth via visual inspection was achieved by inspecting the orthoimages with the GPS coordinates overlaid in QGIS. This was further supplemented with inference drawn from the VARI and DEM map which identified the crowns of individual trees. The final ground truth results are thus a combination of field site visits and visual inspection aided by DEM and VARI models of the study areas.

Site	Pine	Wattle	Gum	Waterberry	Fig	Total
A	6	2	12	6	3	29
B	7	2	2	7	0	18
D	4	2	11	7	0	24
Total	17	6	25	20	3	71

Table 4.1: Species count for study sites (ground truth and visual inspection)

4.2.2 Image annotation

Once pre-processing was completed, the individual imagery was prepared for training in the Mask RCNN model. This involved annotating the individual photos with polygons designating the trees and class labels for each species. Instead of standard bounding boxes, it was decided that polygons detailing tree crowns would facilitate better learning as some of the imagery was of mixed canopy. The tree crown edges are more obvious to the human eye and this was leveraged to train the model. It should be noted that whilst manually delineating tree crowns is possible by visual inspection, in natural environments, the canopies of trees may overlap, especially in mixed forest environments. Frequently, the branches of one individual extend into the main tree crown of those adjacent as they compete for sunlight. Thus, while polygons are drawn over the effective tree crowns of species with ground truth, the actual discrepancy between tree crowns is typically blurred. The VGG image annotator (VIA) tool was used for this project after trialling LabelMe online annotator. VGG software runs as an offline application in a browser window which was superior to LabelMe that stores the data

online which leads to bottlenecks in the annotation because of load times in the browser. Images are loaded into VIA and class labels created in the region attribute section. The output is in JSON format with the x, y co-ordinates for the object saved as well as any labels that the user creates.

The total dataset of 121 images was split into train and validation sets with a proportion of 77.5%(92), and 22.5%(29) respectively. To balance the splitting more appropriately, this ratio of train and validation was applied to each study site separately and then all training imagery was agglomerated, and all validation grouped. This was to prevent validation data coming mostly from one study area and becoming imbalanced.

4.3 Setup on local system

Once all the data is prepared, a virtual environment was set up to train the model on local CPU. This was performed first to check if the annotations generated were suitable to the task and debug any problems that would occur when implementing the model. Using Anaconda and Jupyter Notebook, a virtual environment was setup to run the code from the github repository of the source Matterport Mask RCNN model(Abdulla, 2017). This code runs using Tensorflow 1.3 and Keras 2.0.8 and the appropriate packages were installed once the virtual environment was activated. One of the differences from the original code is that our model will be used to determine multi-class object detection, the 5 different species of trees. With online resources and forums, even with beginner level coding it was possible to alter the code to accommodate these changes through trial and error.

4.4 Implementation of Mask RCNN

4.4.1 Training on local system CPU

Initial trials were run on Local CPU to test training times and to check for any errors that would arise. The Mask RCNN implementation uses a MIT license and from this generosity, many users and programmers have adapted code to suit various objectives. Additionally, they have updated the base code from Matterport to function on TensorFlow versions > 2.x. Using later releases of TensorFlow will allow for better visualization and inference tracking with Tensorboard which we will explore later. Initially, a train, validation and test set from Study site A and B was used to see how long it would take to run with a smaller dataset. The default hyper-parameters were used (Table 4.2), the parameters of most importance being: learning rate, backbone, and pre-trained weights from benchmark datasets. The model can be trained from scratch which updates the weights for all layers. However, many similar works found that by freezing the training layers and using pre-trained MS Coco or ImageNet weights which are large-scale image datasets, the performances of the model were better able to learn and detect objects. This also serves to reduce run time of the model as it uses weights from previously learned objects to make inferences. Both these benchmark datasets allow the model to begin learning from an established checkpoint of machine learning instead of learning from scratch. Initial training with training images from Site A and B (78 images) took approximately 80 hours running on local CPU. This was deemed too time-intensive and a more appropriate method for training the model was explored.

4.4.2 Google Colaboratory (Google Colab)

Cloud computing services offered by Google have greatly assisted deep learning research. Dedicated Graphics Processing Units (GPU) and Tensor Processing Units (TPU) are available for data scientists either free of charge, or via a subscription for Google Colab Pro. This project utilized first Google Colab as a free service to check the capabilities and improvements in model training. After initial testing, a subscription to Google Colab was purchased for increased TPU RAM and longer runtime disconnection

timeout limits since prior experiments ran for 10+ hours. Google Colab can be linked to individual's Google Drive accounts, which is where the dataset containing train and validation images and their associated annotation files were stored. By using Google Colab, researchers are working in a virtual environment with many packages pre-installed such as TensorFlow and TensorBoard 2.4. This project mounted supporting python files made by contributors to the Mask-RCNN source code for Tensorflow 2.4. Default parameters were used in the first experiment which included all labelled training and validation images mounted from the google drive folder where they are stored.

Hyper parameters	**Learning Rate**	**Epochs**	**Steps p/epoch**	**Backbone**	**Optimizer**	**Momentum**	**Pre-trained weights**
Values	0.001	10	100	ResNet101	SGD w/momentum	0.9	MSCoco Dataset

Table 4.2: Default hyperparameters

Elapsed time for the first experiment was approximately 9.7 hours using TPU and High-RAM as runtime shape. Whilst still a very long runtime, the loss values calculated were observed to be fruitful and further experimentation was conducted with various changes to the hyper-parameters. Before elaborating on the changing loss values of the experiment, a short explanation of these values is given.
The following losses are output after each iteration, continuously updated as the model learns:

1. epoch_loss,
2. mrcnn_bbox_loss,
3. mrcnn_class_loss,
4. mrcnn_mask_loss,
5. rpn_bbox_loss
6. rpn_class_loss.

After every epoch, a validation check is performed with the validation set. The validation check uses the validation set of images and calculates the losses for the images based on the weights generated from the epoch on training data. A total of 50 validation steps is used to calculate these values after each epoch.

4.4.3 Loss metrics

The loss metrics listed above can be grouped into 3 categories, ones that relate to mrcnn, rpn and the overall epoch loss metric. Stage 1 of Mask RCNN involves a Region Proposal Network layer which precedes the multi-class object classifier, bounding box and mask generation neural networks. Understandably the RPN loss metrics will not include a mask loss segment.

Classification Loss

In the theoretical background section, the concept of loss was explained to be the difference between the actual value of the object and the predicted value. Therefore, classification loss reflects how much confidence the model has in predicting the object's class. Mrcnn_class_loss covers all the object classes, different from rpn_class_loss because the RPN stage only determines whether it is classifying foreground or background (object or not). Therefore, rpn_class_loss values are typically lower than mrcnn_class_loss. Less classes mean less chances for the classification to incorrectly make predictions.

Bounding box loss (bbox_loss)

The model will output a predicted bounding box and by comparing this to the labelled true bounding box provided, the difference in x and y coordinates (height and width) are computed. As a result of its regression loss nature, larger absolute differences between the true box and the predicted box are penalized and will result in higher bbox_loss values. Rpn_bbox_loss is the ability of the model to locate objects within the image and

mrcnn_bbox_loss refers to how well the model is at predicting the areas in an image corresponding to the different objects.

Mask Loss

In the second stage of Mask RCNN, masks are generated in parallel with the classifier and bbox networks. The model generates a binary mask for each class in the RoIs. Mask_loss is then calculated based on the difference in binary pixel classification of the mask (background, foreground) corresponding to its true class. This prevents it from being affected by class predictions.

The **epoch loss** can be understood as the sum of all the other losses generated during that iteration. Generally, this will gradually decrease as losses are minimized for each of the networks. Metrics that are prefixed by "val" indicate that the metrics are for the validation set.

5 Results

5.1 Experiment 1: Default values

Through inspection of the loss graphs at default hyperparameters, the performance of the model was evaluated. In the figures below (Figure 6.1.1 and 6.1.2), epoch loss is plotted on the secondary axis. All other loss metrics are plotted to the primary axis. Epoch loss descends to 0.6944 after 10 epochs and the values are not plateauing if we inspect the curve of loss over time. Regarding classification loss the RPN classifier has a simpler task compared with the multi-class mrcnn classifier. This is evident in the final values for RPN class loss (0.0352) and mrcnn class loss (0.1523). In object detection, epoch loss for the validation set close to the training set is most preferable. Otherwise, a case of underfitting or overfitting is occurring with the model predicting very well on only the training set but failing to generalise (underfitting) or performing badly at prediction in general (overfitting).

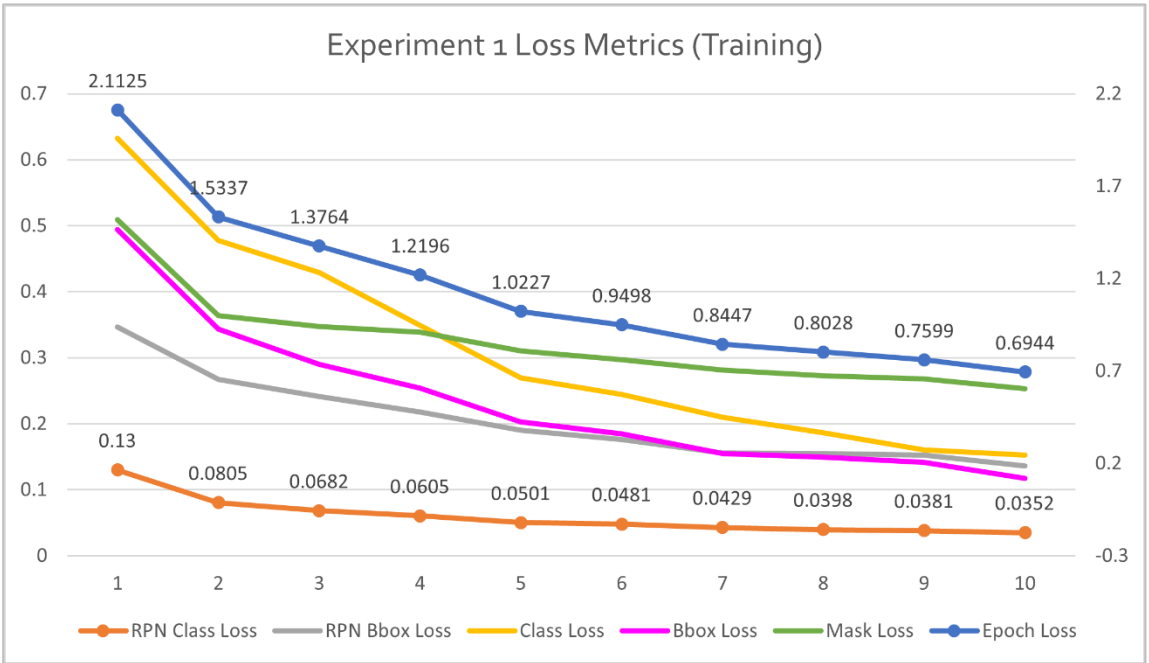

Figure 5.1.1: Loss metrics (primary axis) and epoch loss (secondary axis) for the training dataset using default hyperparameters for 10 epochs each of 100 steps

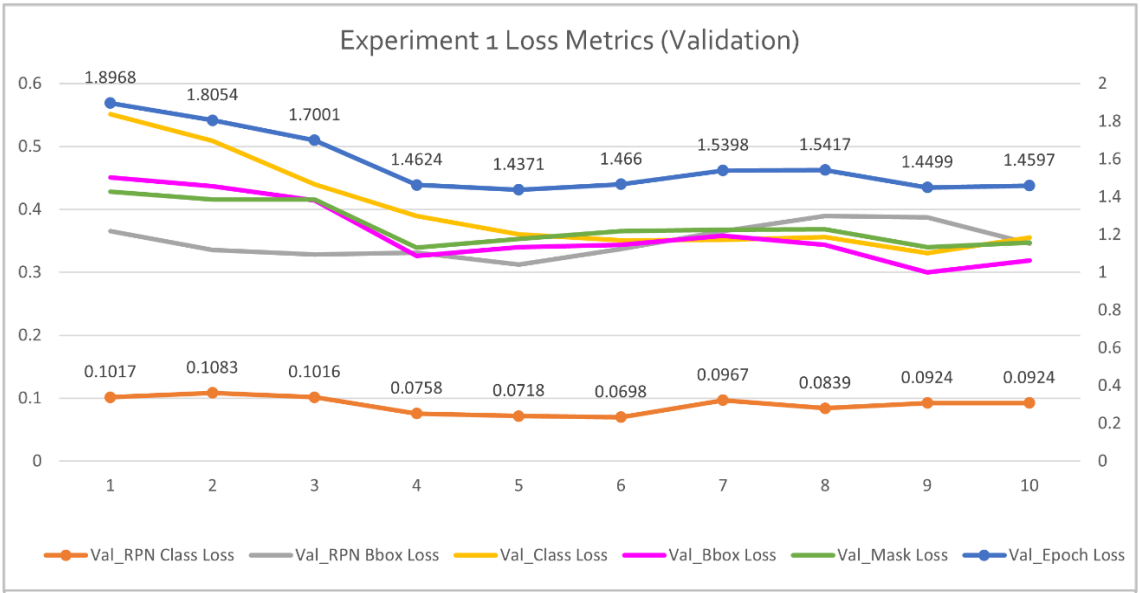

Figure 5.1.2: Loss metrics (primary axis) and epoch loss (secondary axis) for the validation dataset using default hyperparameters for 10 epochs each of 100 steps

Examining the bounding box loss metrics, both the validation and training sets had improved loss values for Mrcnn_bbox. RPN_bbox_loss descends to 0.1359 and the mrcnn_bbox_loss drops to 0.1169. Therefore, using default settings, the model has an improved ability to locate the precise location of the multi-class objects in the study area but the Region Proposal Network does not perform as well in locating Regions of Interest within the images.

Further experimentation was done to optimize the model's performance. Below is a table (Table 6.1) of the experiments.

Experiment	Learning Rate	Backbone	BenchmarkDataset	Epoch
1	0.001	ResNet101	Ms Coco	1-10
2	0.001	ResNet101	ImageNet	1-10
3	0.002	ResNet101	Ms Coco	1-10
4	0.002	ResNet101	ImageNet	1-10
5	0.004	ResNet101	Ms Coco	1-10
6	0.002	ResNet50	Ms Coco	1-10
7	0.004	ResNet50	Ms Coco	1-10
8	0.008	ResNet50	Ms Coco	1-10
9	0.01	ResNet50	Ms Coco	1-10
10	0.012	ResNet50	Ms Coco	1-10
11	0.008	ResNet50	Ms Coco	10-20
12	0.01	ResNet50	Ms Coco	10-20

Table 5.1: Experiments performed and associated hyper-parameters

5.2 Effect of benchmark datasets

Running further experiments (#2-4) whilst altering learning rate and pre-trained weights enabled us to compare which pre-trained weights from datasets yielded better results. The pre-trained weights available for Mask RCNN were the MS Coco dataset and the ImageNet dataset. Both these datasets were created to advance object recognition. MS Coco is a large-scale dataset with 91 object classes of everyday objects, and over 2.5 million labelled instances in 328,000 images. To further the field of object detection, the collaborators recognized that for computer vision to progress, machine learning must incorporate not just single object imagery, but be able to identify multiple objects in a scene. Thus, with multiple labelled objects in an image, the scope of computer vision is enhanced by bringing context to the task and function.

The alternative benchmark dataset, ImageNet, is a large-scale ontology of images. Although both projects are on-going, ImageNet included over 14 million images with over a million objects annotated and delineated with bounding boxes. The ontology of WordNet is utilized by ImageNet to add context and enable stepwise filtering of objects towards their final label. Keeping this project's objectives in mind, the MS Coco dataset is believed to be better at optimizing the model's performance as each drone image is a snapshot of the scenery below, made up of the landscape and populated by trees.

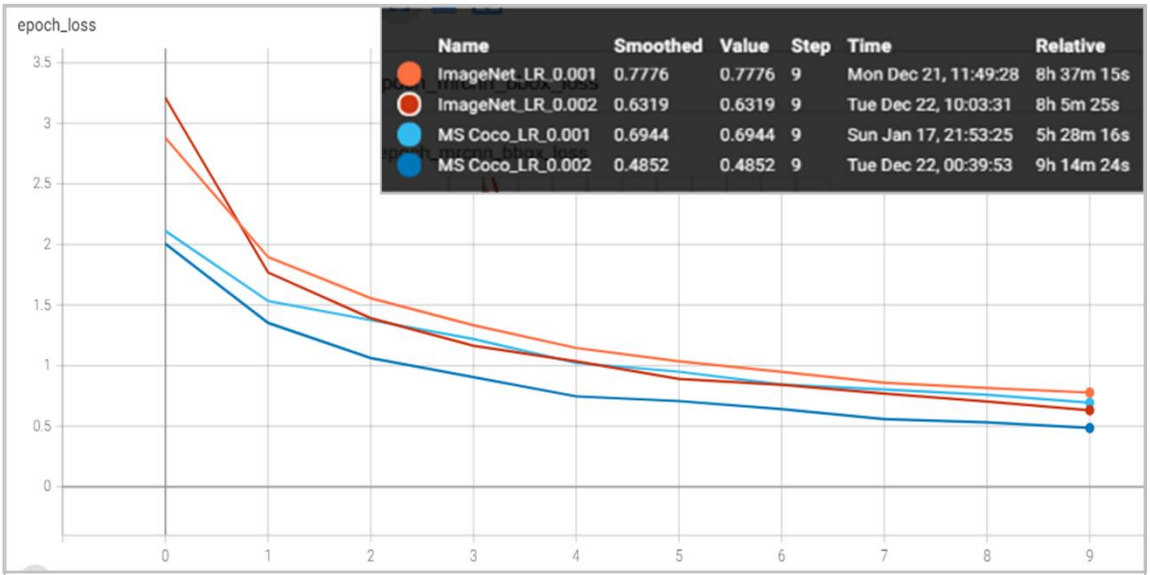

Figure 5.2.1: Comparison of Epoch loss with MS Coco and ImageNet pre-trained weights at increasing learning rates (0.001-0.002). Experiments 1(teal), 2(orange), 3(navy blue), and 4(red).

The graph above shows four experiments (Experiment 1-4) that trialled a learning rate of 0.001 and 0.002 with the weights from two pretrained datasets. The relative times for completion vary between the datasets, but this can be attributed to the random allocation of a TPU from Google Colab. Researchers are guaranteed a dedicated GPU or TPU but cannot choose any specific units. Further trials showed that the training times for each step varied between 22 seconds to 32 seconds.

The MS Coco dataset has better loss values when compared to ImageNet experiments. This improved performance does not apply to every loss metric, and the TensorBoard graphs for the specific loss metrics are shown in the annex. Upon closer inspection the ImageNet dataset has superior performance for all metrics save two: mrcnn_bbox_loss and mrcnn_mask_loss. Using ImageNet with a learning rate of **0.002** resulted in **mask loss** values of **0.3614** compared to 0.2162 using MS Coco at the same learning rate.

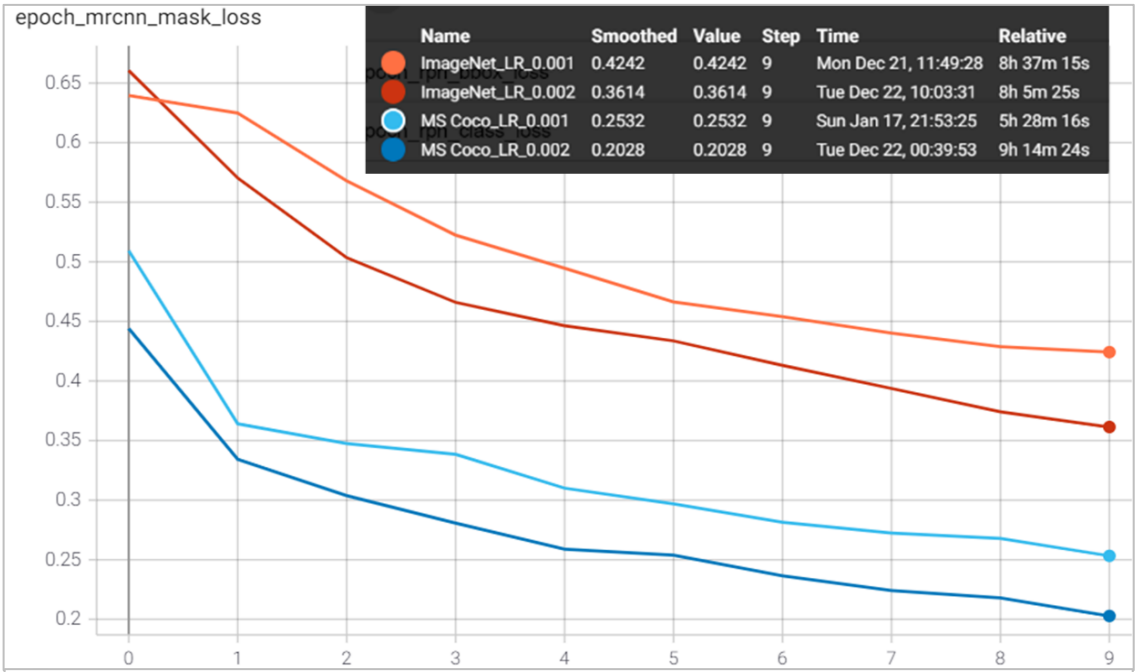

Figure 5.2.2: Comparison of mask loss for MS Coco and ImageNet pre-trained weights and increasing learning rate. Experiments 1(teal), 2(orange), 3(navy blue) and 4(red).

Thus, whilst there were more metrics that ImageNet performed better in, the improvements were minor. However, this does signal that ImageNet as a dataset may provide better results if the object detection models were not concerned with generating masks (i.e Faster-RCNN) because there are consistent improvements in classification loss and RPN bbox loss.

5.3 Effect of backbone architecture

Loss values for the training set were noticeably improved with ResNet101. With a learning rate of 0.002, ResNet101 reached an overall loss of 0.48 whilst ResNet50 only reached 0.51. At an increased learning rate of 0.004, ResNet101 again outperformed ResNet50 reaching 0.38 overall loss compared to 0.43.

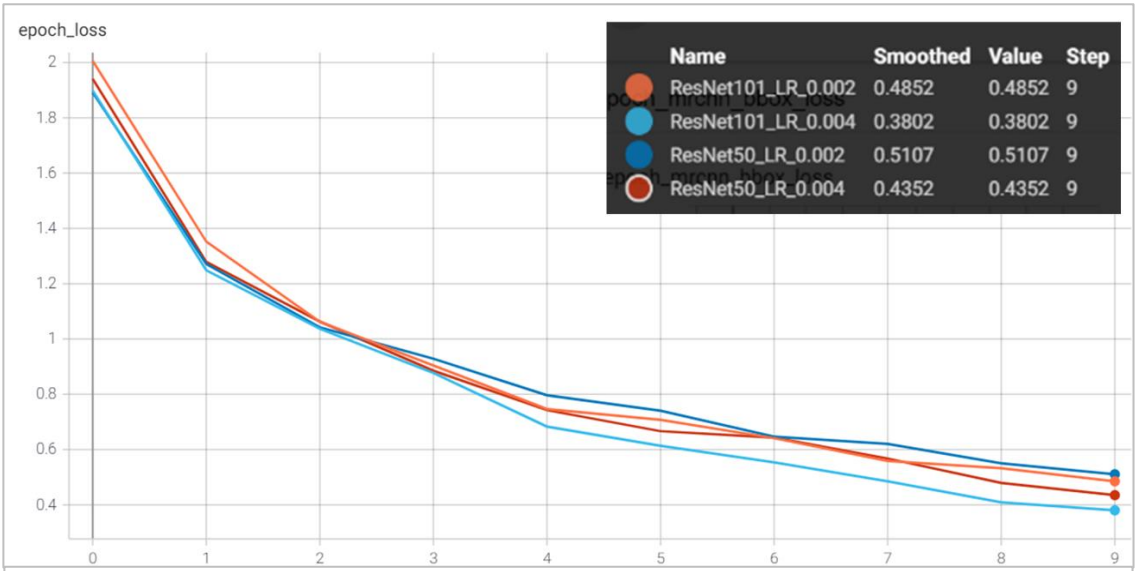

Figure 5.3.1: Comparison of epoch loss for ResNet50 and ResNet101 backbone and increasing learning rate. Experiments 3(orange), 5(teal), 6(navy blue) and 7(red)

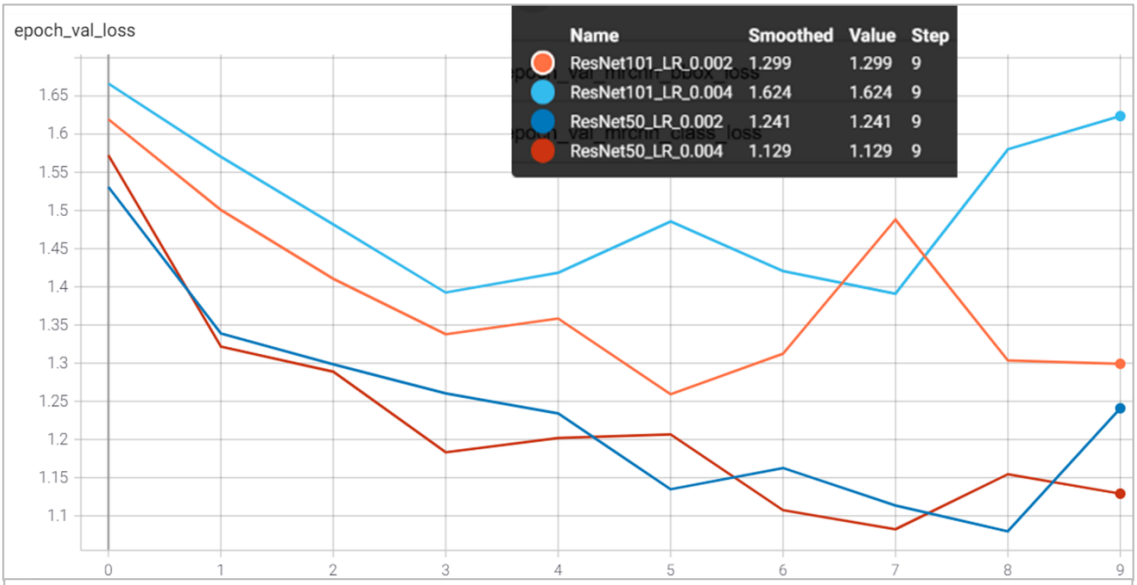

Figure 5.3.2: Comparison of epoch loss for validation dataset with ResNet50 and ResNet101 backbone and increasing learning rate. Experiments 3(orange), 5(teal), 6(navy blue) and 7(red)

The issue was that the validation losses resulting from using the ResNet101 backbone were observed to be much higher than the training losses. At a higher learning rate of 0.004, ResNet50 backbone had an epoch loss of 1.129 whilst ResNet101 logged a loss of 1.624, an even higher validation loss than its 0.002 model trained with the same backbone.

The metrics attributing to these higher validation loss values were found to be Mrcnn_class_loss and RPN_Bbox_loss. ResNet101 at the higher learning rate of 0.004 resulted in 0.4353 losses for mrcnn classification and 0.3947 loss for RPN object localization. These are indicative that the increased complexity of the backbone is performing well for the training dataset but struggled to apply the classification and detection to validation imagery. Thus, a case of overfitting was occurring with the deeper ResNet101 architecture.

There are four common ways to reduce overfitting: 1) adding data, 2) data augmentation, 3) reducing complexity of the model and, 4) dropout. This project could not afford to charter more drone flights and add ground truth surveys so there was not an option to add data. Regarding data augmentation, as the images included landscape views of mixed forest, rotation and cropping of the images was thought to be inapplicable since overlapping tree canopies may add more confusion to the classifier. Additionally, augmentation methods were outside of the scope of this thesis as it would require more time to adapt the original source code and dataset. Mask RCNN is easily implemented with the two types of ResNet, thus by reducing the complexity of the model to ResNet50, validation loss could be restricted from exploding. Finally, with this project's size of dataset, by including dropout to the model, valuable data and weights could be lost from the model so this option was disregarded.

5.4 Effect of Learning Rate (LR)

After running experiments with gradually increasing learning rates (0.002, 0.004, 0.008, 0.01, 0.012) it was found that a learning rate greater than 0.01 did not result in significant loss decrease. Loss began rising when learning rate was 0.012, coupled with increased validation loss. A learning rate of 0.008 and 0.01 had very similar overall loss, 0.3968 and 0.3918 respectively. The model trained on 0.01 had better validation loss (1.062) compared to a learning rate of 0.008 (1.236).

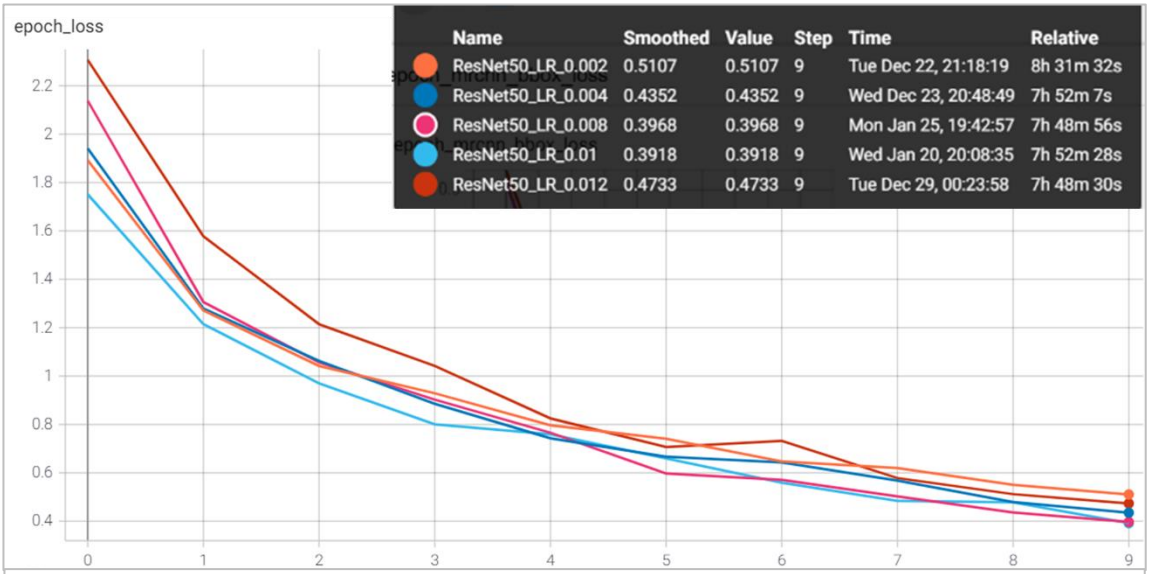

Figure 5.4.1: Comparison of learning rate (0.002-0.012) with ResNet50 backbone and pre-trained weights from the MS Coco dataset.

By comparing the epoch loss and validation epoch loss values of the two experiments, validation loss is significantly higher than the training loss. In ideal circumstances, if the model were learning substantially on the training set, its ability to generalize or predict on datasets it was not trained on or had not seen before (validation set) then the validation loss would be close to the values of the training loss. This is not the case for our model as we can see a general trend of decreasing loss for the training dataset as learning rate increases, but validation loss seems to plateau at a point or increase. Thus, further experimentation with these two learning rates was performed to discover how loss metrics for training and validation sets would develop.

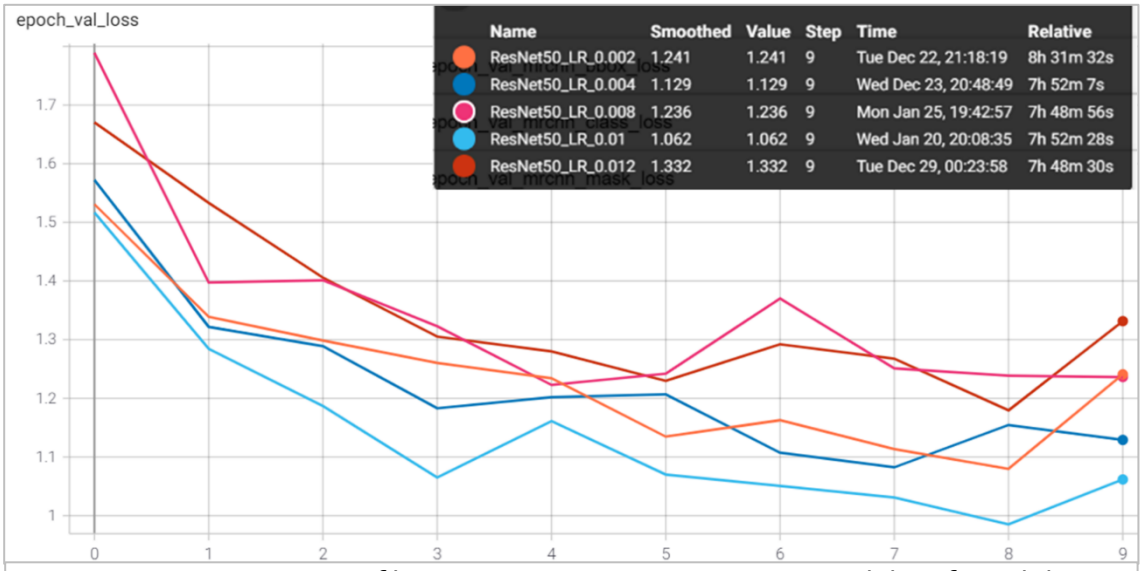

Figure 5.4.2: Comparison of learning rates (0.002-0.012) on epoch loss for validation dataset with ResNet50 backbone and Ms Coco. Experiments 6(orange), 7(navy blue), 8(pink), 9(teal) and 10(red).

5.5 Effects of optimal learning rates

At the end of each experiment, weights are updated and saved as h5 files after every epoch. Further training was performed using these weights to explore the performance. Experiment 11 uses weights from the previous experiment with a learning rate of 0.008 whilst Experiment 12 uses those with a LR of 0.01. Although the initial 10 epochs trained at a LR of 0.01 exhibited better performance than the alternative learning rates, what can be seen in Figure 6.5.1 is that at additional epochs, training loss descends to *0.2199* with a LR of 0.01 compared to *0.1455* at a LR of 0.008.

Experiment 11 had superior Mrcnn classification (*0.0033***)** compared to Experiment 12 (*0.0449***)**. This is indicative of the LR at 0.008 performing better at classifying individual objects in the 2nd stage. For Mrcnn bbox loss, Experiment 11 displayed superior loss values after training, reaching <0.01, compared to Experiment 12 which never decreased below 0.02. The main contributor to epoch loss for both experiments was mask Loss. A learning rate of 0.008 achieves better loss values than at higher learning rates. Epoch loss for experiment 11 increases to 1.12 but is still superior to experiment 12 at 1.52. Whilst the validation loss does increase, the associated training loss

significantly decreases such that this project decides to use the weights generated after training on 20 epochs using a learning rate of 0.008.

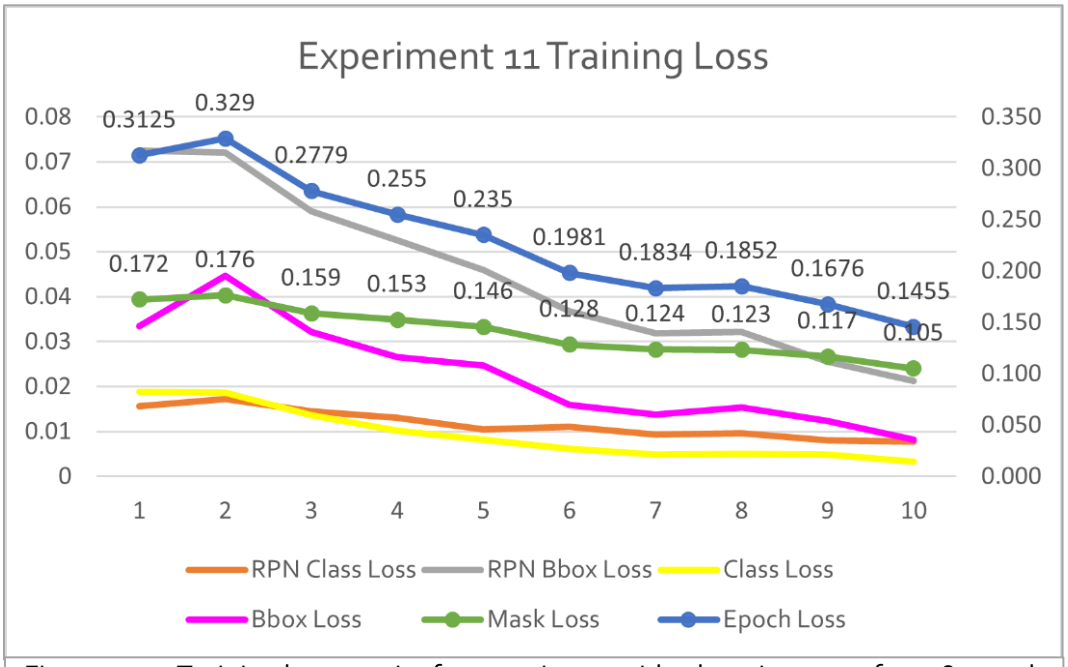

Figure 5.5.1: Training loss metrics for experiment with a learning rate of 0.008, epoch loss and mask loss shown on secondary axis. All other loss metrics on primary axis

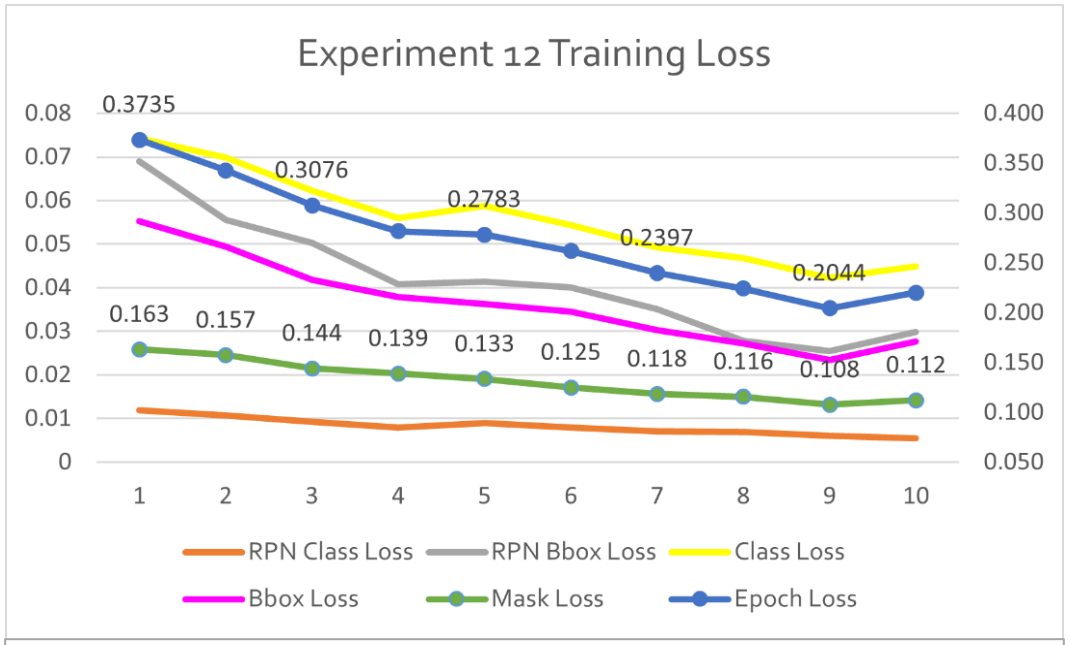

Figure 5.5.2: Training loss metrics for experiment with a learning rate of 0.01, epoch loss and mask loss plotted on secondary axis. All other loss metrics on primary axis

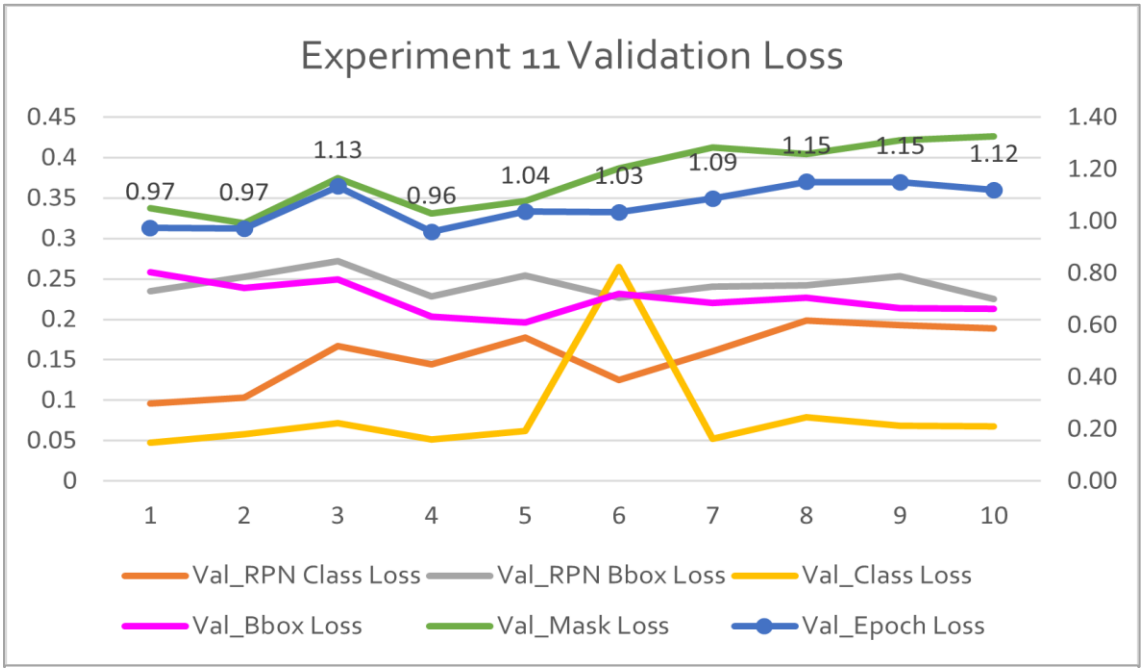

Figure 5.5.3: Validation loss metrics with a learning rate of 0.008, epoch loss plotted on secondary axis. All other loss metrics plotted on primary axis

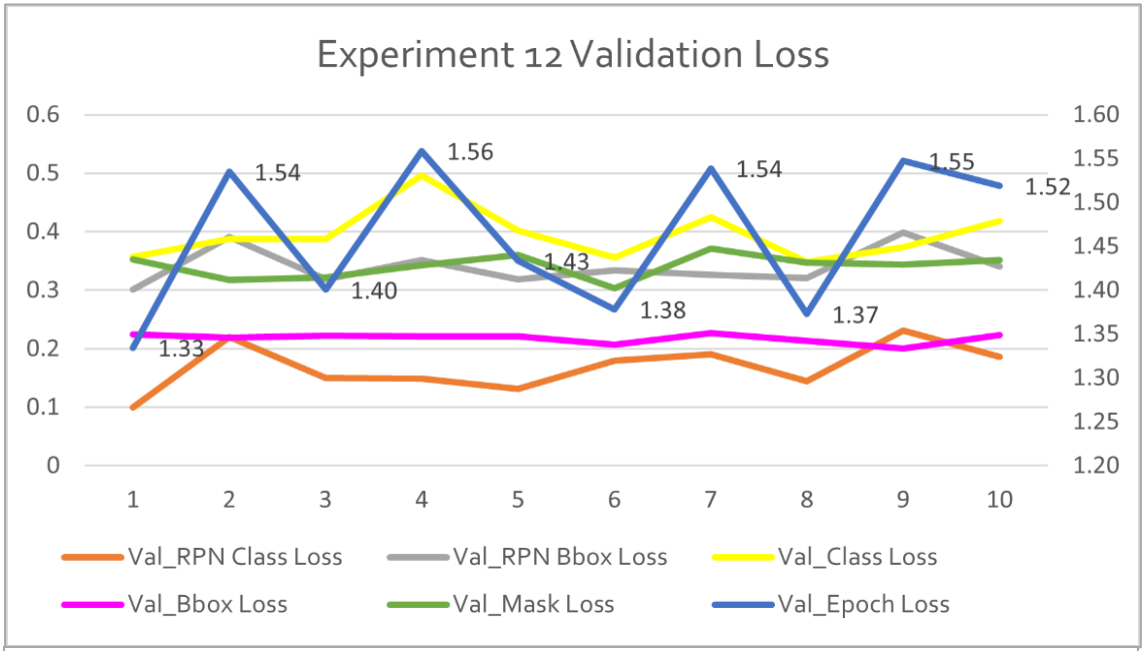

Figure 5.5.4: Validation loss metrics with a learning rate of 0.01, epoch loss plotted on secondary axis. All other loss metrics plotted on primary axis

6 Visualization of results

In this section the weights generated from the trained models are used to visualize object detection using Jupyter Notebook with notebooks that were provided with the source code of Mask RCNN. After adapting these notebooks to perform multi-class object classification, the weights are loaded and run detection is done on images from the validation dataset to evaluate how well the Region Proposal Network stage is performing and the following Regions of Interest are localized and classified through the predictive capability of the model. All images were inspected with run detection and relevant examples were chosen to discuss the results. The evaluation is done separately for moderately dense and very dense forest (Site A and B) and open woodland (Site D) followed by a comparison of the mean Average Precision (mAP) for the study areas.

6.1 Very Dense Forest Landscape (Site B)

Stage 1: Region Proposal Network

The model first generates targets using a grid of anchors that encompass the entire image at varying scales. By computing the Intersection over Union (IoU) of these anchors with ground truth objects, anchors are defined as positive, neutral, or negative. Intersection over Union quantifies the percent overlap between the predicted mask and its target mask, that of ground truth. Positive anchors have an IoU greater than 70%, negative anchors have IoU less than 30% and neutral anchors are any that are between these values. Neutral anchors are then excluded from the training. Thus, we can see that if more ground truth is supplied of an image, more meaningful anchors will be produced allowing the model to generate positive and negative anchors. The first image evaluated is from the very dense forest study area, Site B, for which visual interpretation was performed to identify tree species. Oftentimes many positive anchors are generated for the same objects. These undergo a process of refinement by taking the multiple offsets of these anchors from the ground truth bounding box and by applying a standard deviation function that allows a final refined anchor to be generated

(Fig.6.6.1). After normalization they are converted back to standard x,y coordinates in the image.

Figure 6.1.1: RPN Targets from ground truth

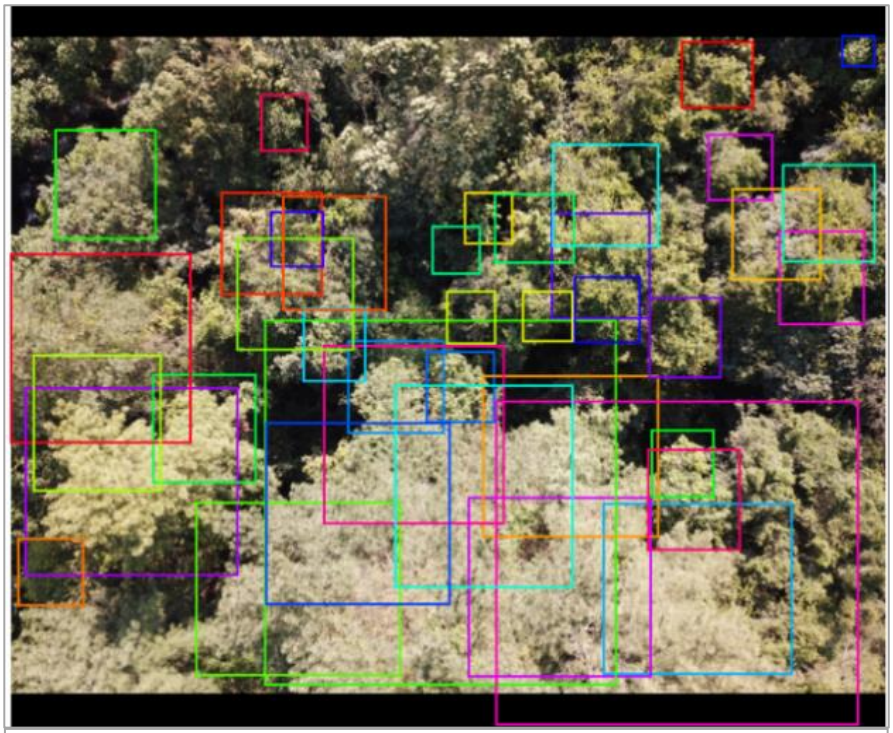

Figure 6.1.2: RPN Predictions after NMS

The model then proceeds to make predictions based on the data it was trained with in the RPN target stage. By setting a limit for these RPNs, the amount of predicted RPNs is controlled as many overlapping RPN anchors are generated, especially in a very dense forest environment with large variance in tree crown area. Then a Non-Max Suppression (NMS) filter is applied which sorts the overlapping anchors over one object by the objectivity score, keeping the proposed region with the highest score.

Stage 2: Proposal Classification

The model now attempts to classify the Regions of Interest (RoIs) by applying model weights and running the classifier heads on the proposals from the first stage to produce class probabilities and bounding box regressions. Using positive RoIs and refined bounding boxes the model outputs the image overlaid with predicted objects, their class probabilities, and bounding boxes. Two more steps refine the results, first a filtering of low confidence detections, with the percentage specified by the user. In this case it was set at 55% because valid objects were identified with low confidence. The final step applies Non-Max Suppression on each class to remove any overlapping anchors that would result in duplicate image detection.

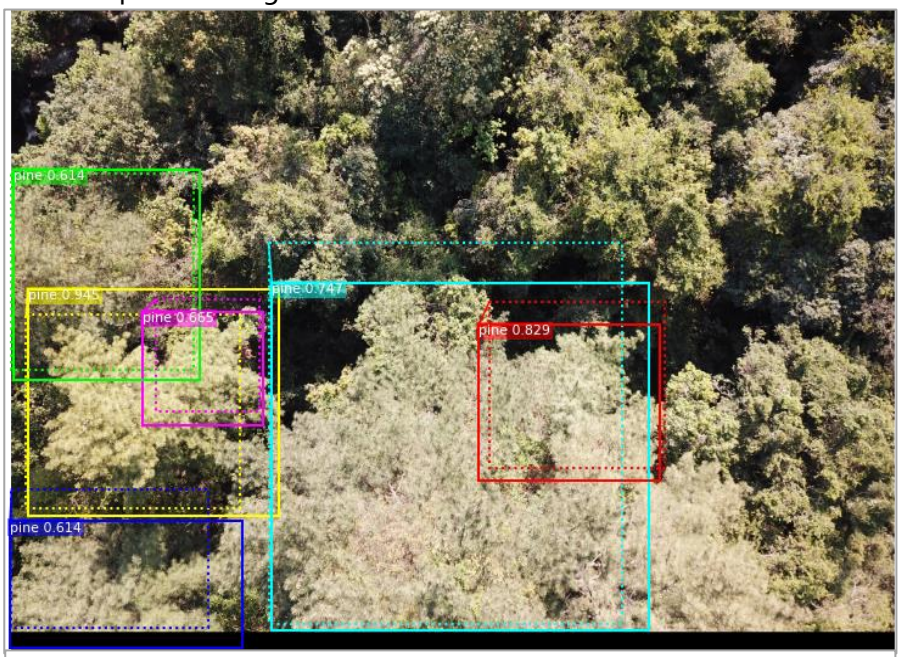

Figure 6.1.3: Final RoIs after per-class NMS

Stage 2: Mask Generation

Mask generation occurs in parallel to the bounding box regression and classification. Using training masks, which are generated from the RPN layer and making predictions of masks, the final masks are generated over the final image RoIs (Figure 6.6.4). Comparing this output with the labelled image provided to the model, classification and bounding boxes generated are mostly for objects with ground truth. Duplicate mask

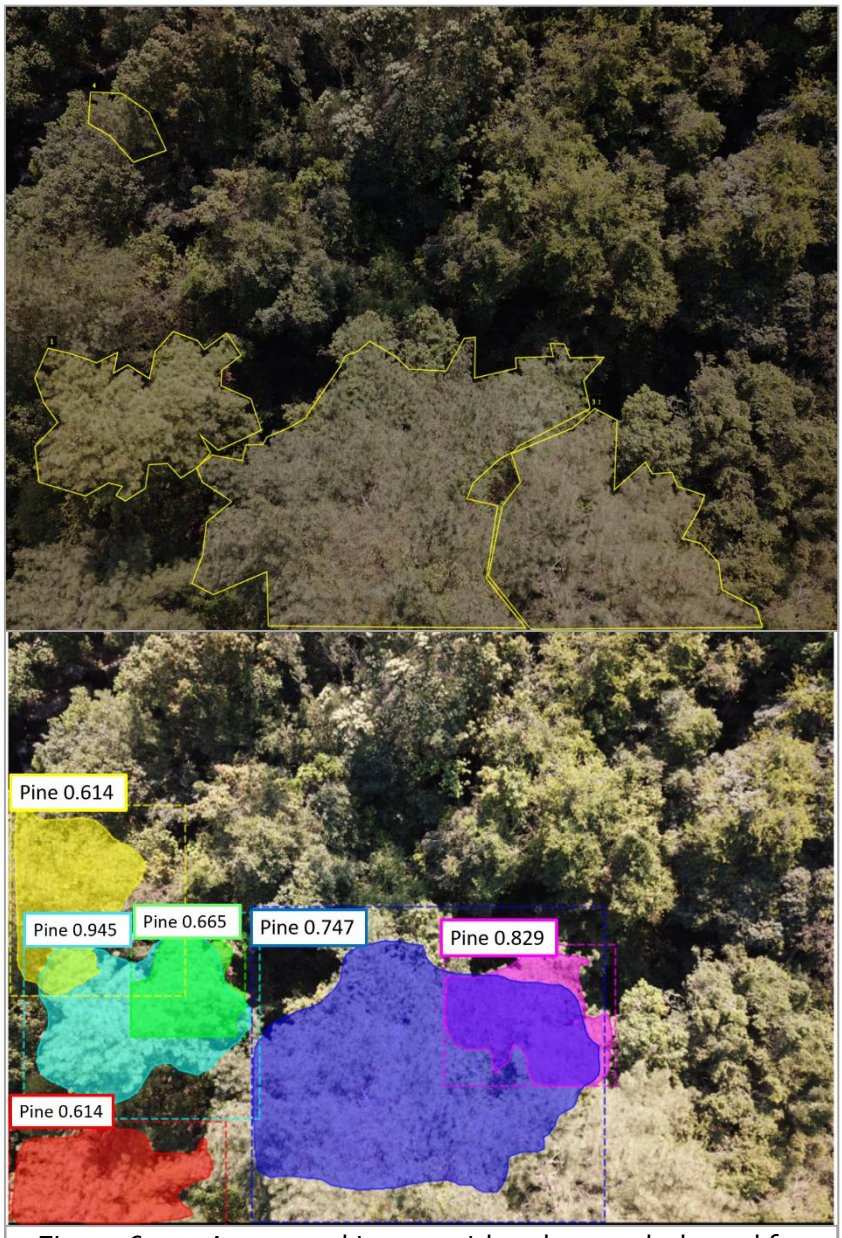

Figure 6.1.4: Annotated image with polygons darkened for contrast(top) and final prediction result (bottom)

generation is occurring between the pink and dark blue polygon and between the teal polygon and the green and yellow polygons. Additionally, a pine tree is detected in the bottom left corner of the image with a low objectivity score of 61.4%.

The pine tree is labelled in other images but was unlabelled in this image because most of the object was occluded. Regarding mask generation, there is a case of duplication following the bounding box error. Using a rectangular bounding box is not intuitive when object overlap is high which is the case for mixed forest canopies. Compared with the ground truth polygons, smoother edges are observed delineating the crowns of the pines in the left of the image, but the dark blue mask is under segmented and the actual tree crown extent is inaccurate. When the final RPN predictions are compared with the final product, we see that the model is struggling to detect trees that do not have supporting ground truth. Accuracies for detecting trees with ground truth verification vary from 61-94%.

6.2 Moderately Dense Forest Landscape (Site A)

Stage 1: Region Proposal Network

Site A has less dense forest cover than Site B and has ground truth. In this image more ground truth was available for the indigenous waterberry tree species. From this the RPN targets were formed and the final RPN predictions were produced. In this

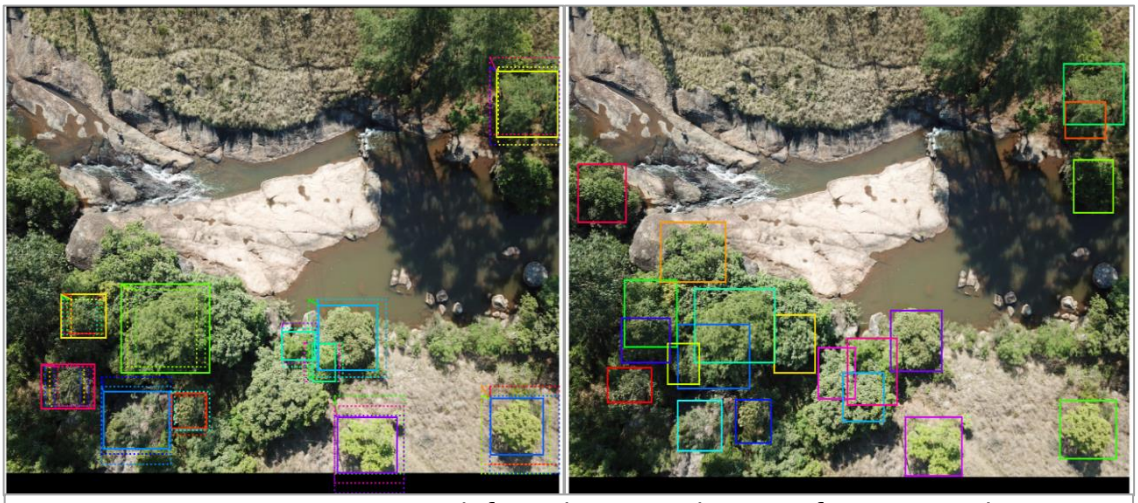

Figure 6.2.1: RPN targets (left) and RPN predictions after NMS (right)

prediction image, instances of duplicate object prediction occur in the dense canopy in the bottom left of the image. However, the bounding boxes are not fitted perfectly over tree crowns compared to the open area in the bottom right of the image. RPN bounding box predictions are also seen over tree crowns in the centre left and amongst the dense canopy in the bottom left where no ground truth is supplied. After Non-Max Suppression the model is making predictions on where objects are in the first stage.

Stage 2: Proposal Classification

Final RoI Predictions are shown in Figure 6.2.2 with tree detections and localization shown next to the ground truth labelled image. Objectivity scores of the species were high: Pine at 92.8%, Gum at 88.6%, Wattle at 72.6% and the average score for waterberry trees was 90.48% with 5 predicted objects. Waterberry trees with tree crowns that were not overlapping other trees had higher objectivity scores whereas with dense canopy, the model has lower confidence predictions. By comparing the labelled images used to train the model with the final predictions it is observed that the model is making predictions on individuals of waterberry that are clustered together. The model failed to detect one individual for pine, gum, and wattle classes. The pine in the top right of the image, although proposed in the RPN layer, was not detected by the second stage classifier after refinement. The wattle was not detected at Stage 2 despite it being proposed in the RPN layer. Total ground truth of wattle in the project was very low, thus detection for this class was greatly impeded.

Stage 2: Mask Generation

In the final output, masks are generated alongside the bounding boxes and classification scores of individual trees. The masks generated accurately delineate almost all the tree crowns in the image save for the gum detected whereby overlap from another tree was included as part of the mask. Additionally, the mask generated for the purple polygon excludes some of the tree crown towards the left and includes partial crown of another waterberry adjacent on the right to it. Masks of the pine were well defined and

waterberry individuals that did not have overlapping canopies exhibited good mask generation.

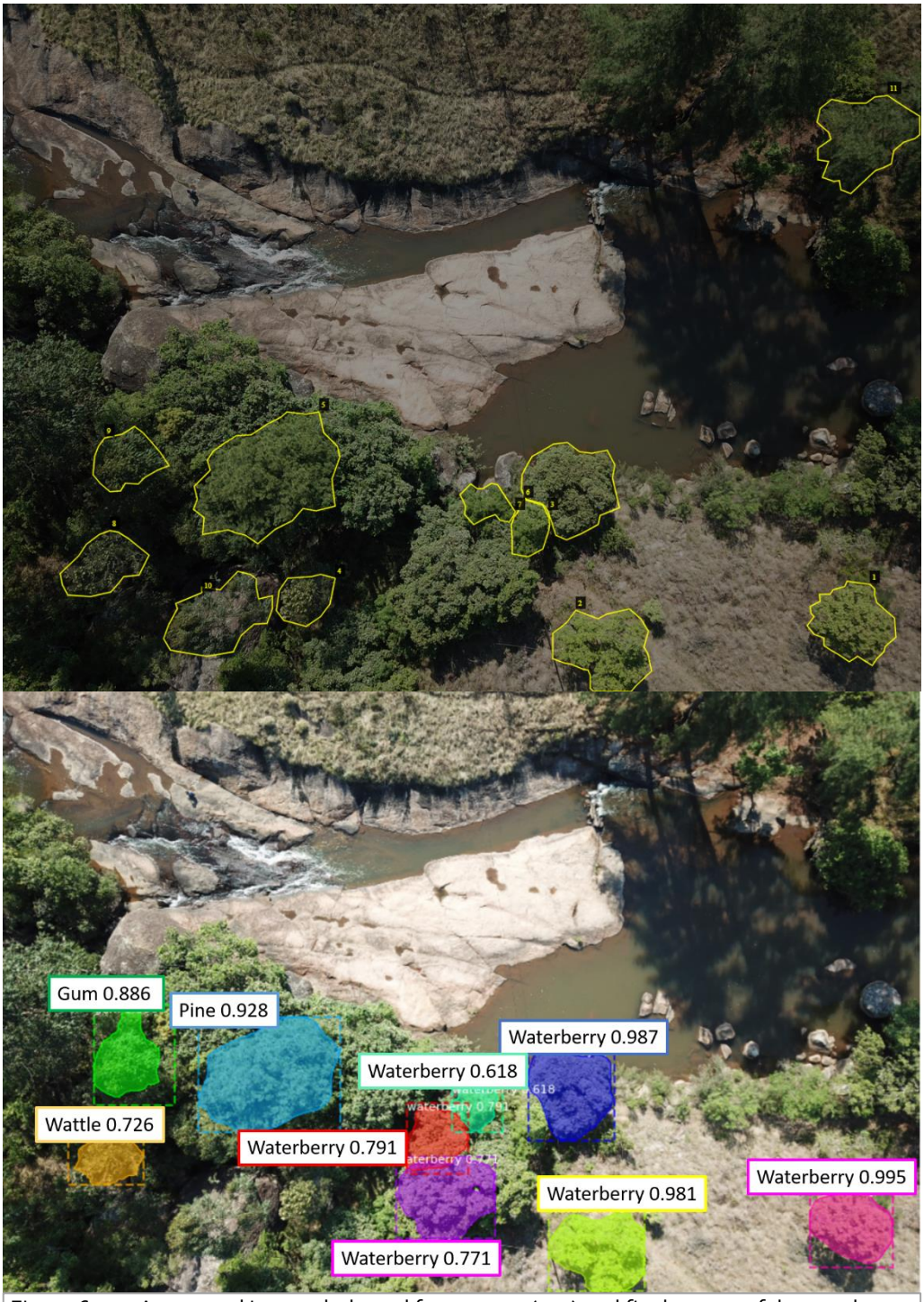

Figure 6.2.2: Annotated image darkened for contrast(top) and final output of detected trees with scores and masks in a moderately dense forest landscape (bottom)

UAV flights provide valuable sweeps of landscapes from above, but flights have varying altitudes. Multi-scale UAV imagery thus has implications on building classifiers for the same species of trees. Since different scenarios require different flight altitudes this could inhibit the model's learning. Additionally, varying heights of trees and the tight interlocked crowns in a very dense forest landscape will result in shadows. These shadows can influence a model's ability to learn from labelled data and make predictions on other images that are captured at different angles.

6.3 Open Forest Landscape (Site D)

Stage 1: Region Proposal Network

Site D differed from other study sites in that the altitude for the drone flight was higher by 6m, the lighting conditions had cloud cover and, the landscape was an open forest with mature trees having distinct crowns that did not overlap as much as the first two study sites. In the RPN target step in the first stage, ground truth instances were available for more trees and the angle as well as crown extent were more distinct than the previous study sites with overlapping canopies. In the RPN prediction stage after NMS we see many more proposed Regions of Interest that encompass the tree crowns of tree individuals. Some errors are also present such as in the bottom left corner where a patch of grass was proposed as an object, as well as some instances of multiple tree

Figure 6.3.1: RPN training targets (left) and RPN predictions after NMS (right)

crowns being conglomerated into one object. In this first stage, the model does well to localize objects proposed using the RPN targets it was trained on.

Stage 2: Proposal Classification

The second stage of the model performs much better for objectivity scores of the objects detected when compared to the scores of the first two study sites. The average scores for pine, gum and waterberry were 93.5%, 96.5% and, 96.75% respectively. A total of 2 pines, 10 gums and 4 waterberry trees were detected in this image. The RoI classification and detection layer again has fewer final detections when compared with the RPN predictions layer that preceded it. Basing the detections on ground truth instances, the model detects species accurately if they are close to the morphology of trees used to train the model, yet here it is evident that overfitting is imparting effects on the model's performance to detect species that are outside of training data. With such a small field sample, the model detects 16 trees from an image with 14 ground truths. Therefore, the model's recall ability is high, but with so many more trees undetected in the image, improvements to further lessen overfitting should be implemented. When compared with the labelled data provided to the model, the RoI predictions manage to predict two trees that are without ground truth instances but fails to detect a waterberry tree numbered "3" in the labelled image. Keeping in mind the grid mission route that the drone flies over the study area, individuals that are labelled in one image may not be annotated in another adjacent image. This may assist the model in making predictions using ground truth from very similar imagery that it also learned on to make predictions.

Stage 2: Mask Generation

Masks generated for this image only had one occurrence of overlap in the fuchsia pink polygon mixing with the yellow polygon. Overall, the masks delineated tree crowns correctly which can be partially attributed to the open woodland form of the landscape as opposed to densely crowded forests with interlocking canopies. Alongside this the

higher altitude allowed for the same ground truth instances to be viewed at various angles resulting in more valuable training data for the model to train on.

Figure 6.3.2: Annotated image (top) and final output of detected trees with scores and masks in an open woodland landscape (bottom)

6.4 Precision and Recall

Evaluation of Object detection models varies with many proposed evaluation approaches. One of the key concepts involved is the Intersection over Union (IoU) value. Many approaches use an IoU threshold at 0.5, where the predicted bounding box overlaps that of the ground truth by at least 50%. This is used to calculate the precision and recall of the model whereby positive values must have an IoU of 50% or higher. Precision in our dataset is a measure of how accurate the predictions are for that class, the percentage of correct predictions. Recall is a measure of how well the model detects all the positives.

True Positives are objects correctly classified in the image and False Positives are objects that have been detected but classified incorrectly. True Negatives and False Negatives in object detection are usually more difficult to quantify as the labelling does not include areas that are not objects in this study. However, the RPN stage defines negative anchors as areas where an IoU < 0.3. By creating a set of negative anchors in the image, the model learns areas that would not be Regions of Interest and then would not consider these regions for object detection. False Negatives are areas labelled as negative anchors but are areas where objects are in fact located in the ground truth. Unlabelled objects in the image would incur this error if not all trees are labelled which was the case due to incomplete ground truth and visual interpretation.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

TP = True positive

TN = True negative

FP = False positive

FN = False negative

An example of the precision and recall for this dataset is the object detection of pine class. The number of pines *correctly* identified in the image by the model is divided by the total number of pines detected. For the recall, the number of correctly detected pines would be divided by the total number of pines that are present in the ground truth dataset for the image. The average precision for each class is calculated separately, and

with these APs, the mean average precision for each class is calculated. Following this the mean Average Precision of the image is calculated by averaging the AP for all the classes within the image.

6.5 mean Average Precision (mAP)

The Mask RCNN and similar RCNN models utilize the evaluation metric of mAP at an Intersection over Union threshold of 0.5. Due to the multi-class nature of our dataset, this was found to be more appropriate as this is defined by the AP for each class averaged for the image. By computing this mAP value for all the validation images, the relevant overall mAP was computed.

$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k$$

$$AP_k = the\ AP\ of\ class\ k$$

$$n = number\ of\ classes$$

Equation 1: Mathematical definition of mean Average Precision

The overall mAP value for the validation dataset was *0.508857*. The disparity between scene views of the study sites however, prompted mAP to be calculated separately for the Study Sites A and B combined, and Study Site D. Site B is a very dense forest with lots of overlapping canopy, Site A is a moderately dense forest landscape, whereas Site D is a simpler landscape with open woodland.

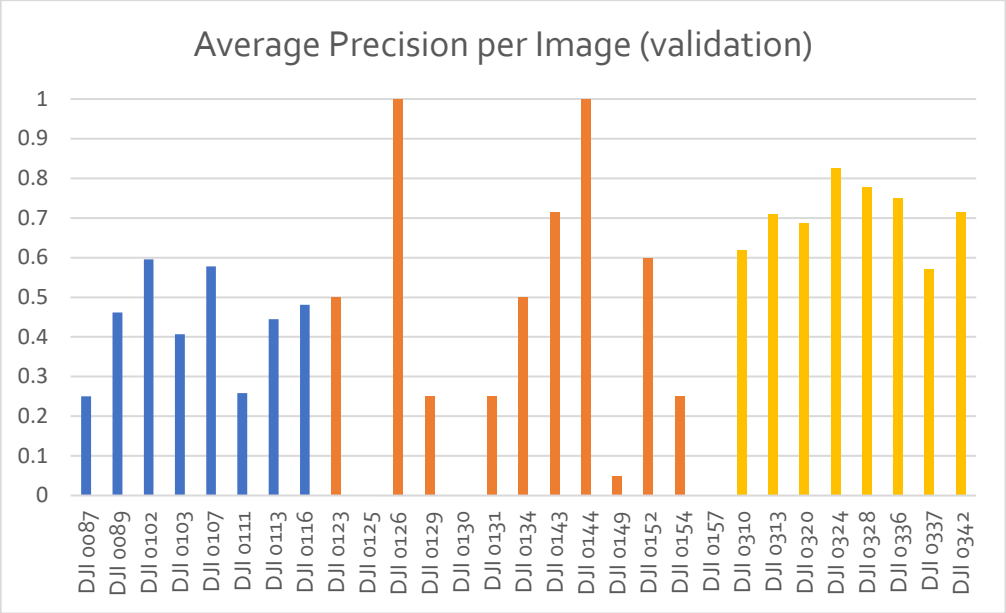

Figure 6.5.1: Average Precision for each image in the validation dataset. Site A(navy blue), Site B (orange) and Site D(yellow)

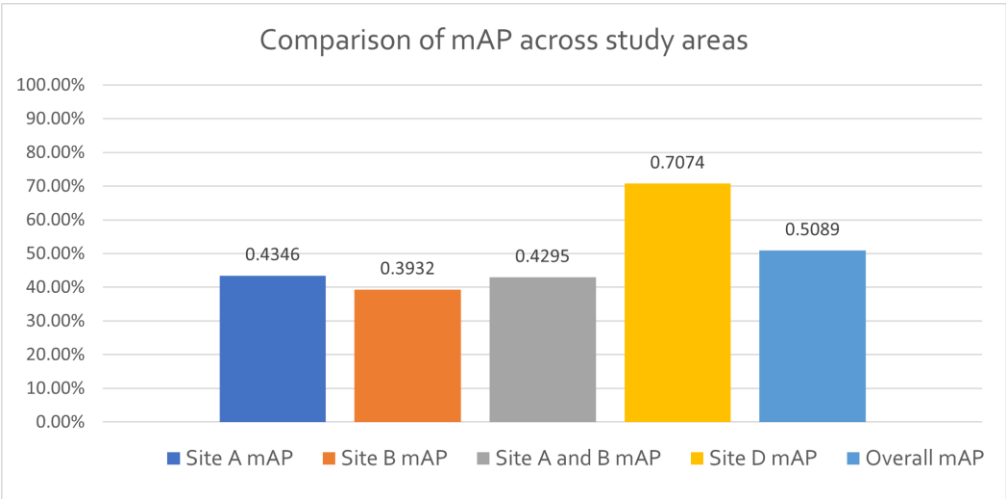

Figure 6.5.2: mean Average Precision across study sites

Findings indicated that in mixed forest landscape, the mean Average Precision was *0.4295* whilst the mAP in Site D was *0.7074*. These findings suggest that the model has higher precision in areas with less overlapping canopies where tree crowns have clear disparity with others.

6.6 Confidence Scores

The confidence scores of the detected trees were averaged for the study sites per class. The overall confidence score for pines, gums, wattles and waterberry were 0.879, 0.896, 0.782 and 0.891, respectively. Noticeably, wattles were only present in the first two study sites and there were no detections of figs by the model which can be attributed to the low sample size provided for the model to learn on.

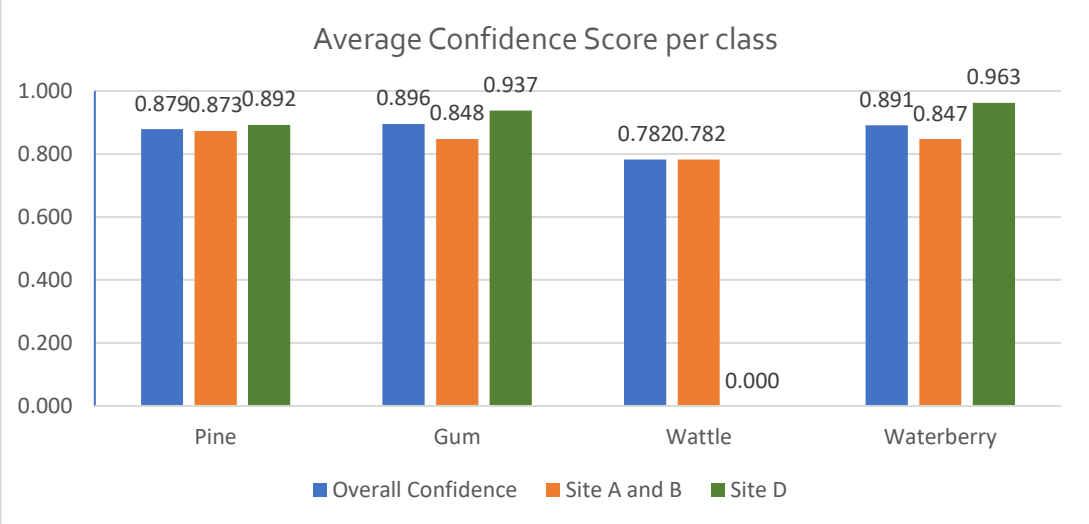

Figure 6.6: Confidence scores per tree species by landscape type

7 Conclusion and final remarks

7.1 Findings

The use of accessible UAV imagery for plant monitoring is being developed on many fronts in the field of remote sensing. This study used readily available UAV capability and coordinated with local plant ecologists to acquire relevant ground truth and imagery without the need for the primary researcher to be present in-situ. This project presents elements for studying tree species distributions and IAPS invasions remotely. Regarding the research question: "Which landscapes are suitable for implementation of the Mask RCNN deep learning model…" it was shown that the Mask RCNN model provided a mean Average Precision of 0.71 in open forest landscapes with less overlapping tree crowns and in cloud cover lighting conditions. However, in environments with dense canopies, the model is shown to have unexceptional results for detection, between 0.44 and 0.39 mean Average Precision in moderately dense forest and very dense forest landscapes.

The results also indicate overfitting with the available dataset despite extensive testing with various hyperparameters, of which a learning rate of 0.008, ResNet50 backbone and pre-trained weights from MS Coco dataset was the best performing. These parameters resulted in an epoch loss of 0.15 after 20 epochs with a validation epoch loss of 1.12 which were found to be the optimal hyper parameters to train this model on for this project answering the first research question of this thesis.

7.2 Limitations

Major limitations for this study were the low amount of ground truth available to train the model and the extensive time periods needed to train the models. Limited ground truth and therefore annotation resulted in overfitting occurring which inhibited the model's ability to detect trees that did not have ground truth available. Images were not tiled with the perspective that in complex landscapes, valuable tree crowns would be occluded since ground truth was lacking. Further, this project annotated individual

UAV imagery that had 70% overlap with the perspective that different angles of the same trees would assist the model to make predictions. Using this approach, default imagery sizes of 4000 x 3000 pixels resulted in extensive training times when input to the Mask RCNN model.

Ideally the final output imagery would be stitched together with the masks generated and these would provide orthomosaics with extracted tree crowns that forestry managers could use to rapidly assess tree diversity and invasive tree presence within landscapes. Due to limitations of the researcher's programming abilities, masks could not be exported from the imagery. Masks are overlaid over the images after resizing and to generate polygons which could be used as layers in GIS software proved to be out of the scope for this project.

7.3 Future Studies and similar works

Combinations of RGB imagery with other data such as hyperspectral, LiDAR or elevation models has been proposed to increase the accuracy of deep learning models with similar plant monitoring objectives.

Multi-temporal collection over the same study areas have been leveraged to obtain images of plants in different seasons which have resulted in high average precision of around 92% by (Santos et al., 2019) albeit with only one tree species. A similar study comparing classification from three acquisition years compared to just one acquisition date with high resolution UAV imagery and ResNet found that classification accuracy increased greatly, from 51% to 80% of two Pine tree species(Natesan et al., 2019). The impact of multi-temporal data for classification improvement identifies one of the future approaches for this study as seasonal differences in tree morphology can be incorporated to provide a better benchmark for tree species object detection libraries.

Object based tree crown segmentation aids in the generation of labelled data to feed CNNs for learning. Whilst Mask RCNN may overcome this by performing both mask generation (segmentation) as well as classification, the advantage of Object based

segmentation lies in producing polygons over large study areas that can be reviewed and labelled, optimizing the workflow of researchers towards annotating data that will enhance the model's performance. This approach has been trialled by (Onishi & Ise, 2021) which classified 7 different tree classes with up to 90% accuracy using an object-based CNN. Further, (Schiefer et al., 2020) have utilized CNNs and a semantic segmentation approach to accurately map tree species in a mixed forest environment with a mean F1 score of 0.73. Employing specialists to check the labels of the trees over the generated segments is an effective way to obtain accurate ground truth. Given that this study had limited ground truth data available to train the model, incorporation of these methods offers avenues for improvement towards faster annotation and generation of objects over diverse landscapes.

Drones can be rapidly deployed in various environments, yet the forest types change for every scenario. This varying forest type was of interest to (Weinstein et al., 2020) who investigated how cross-site learning would affect RGB tree crown detection. By including data from a range of forest types they found that using pre-trained weights generated from various forest types and fine-tuned with hand annotated data from evaluation sites resulted in the same performance as local site models. They determined that a model fit to data from all sites performed better or as well as individual models trained for the local sites. Therefore if further models were to be built from this study, by introducing more sites and then fine tuning with hand-labelled data, more accurate classifiers may be constructed.

A recommendation for future studies with similar objectives is to stitch together individual imagery of the study sites and then to tile these into smaller images. After tiling, manual image annotation can be performed which will enable the researcher to feed smaller images to the deep learning model, but also contain enough relevant data to train it with. The output images could then be restitched together easier allowing for improved transferability.

As drones become more available at lower prices, many flights are being performed by hobbyists and professionals alike. By utilizing UAV-based visual interpretation as an alternative for ground sampling as suggested by (Kattenborn, Lopatin, et al., 2019), plant specialists may approach tree identification studies by identifying trees using annotation software on the giant bank of cloud based drone data. Forestry monitoring could benefit greatly if publicly available drone imagery were labelled by plant specialists for use in further deep learning object detection projects. Perhaps an image library of trees of interest may one day be available for better computer vision techniques involving plant monitoring.

7.4 Conclusion

This thesis presents the use of the Mask RCNN model towards detecting tree species in natural environments. The project leveraged low-cost and accessible drone imagery over areas with presence of well-established invasive tree species. The results indicate that more work needs to be performed if a generalized classifier is to be built using the Mask RCNN architecture. As a simple method with limited resources, Mask RCNN has potential in detecting various classes of tree species at an individual level from UAV imagery but improvements in runtime through image tiling and combinations with other data are necessary.

At times, one species of tree dominates certain landscapes, and in others various IAPS are present at once. Depending on the study area, Mask RCNN is suitable to be implemented as a framework for monitoring multiple IAPS presence within disturbed environments. The advantages that instance segmentation offer over semantic segmentation are only applicable within environments with multiple IAPS presence, in mostly homogeneous IAPS landscapes, simpler CNN's will fulfil objectives of plant cover and distribution mapping sufficiently.

The optimization process of Mask RCNN to multi-class object detection indicates that with the available dataset, less complex backbone architecture using the MS Coco

pretrained weights with a learning rate of 0.008 results in the best performance. Despite this, the model still exhibits overfitting which can be overcome with a variety of suggested methods but at this stage is unable to confidently predict trees out of ground truth sample. The run times necessary to train the model were long and the use of Google Colaboratory greatly assists but the time taken for training is still a huge obstacle that researchers will face using Mask RCNN.

Evaluation of object detection projects is subjective; the mean Average Precision method is proposed in this project. These findings indicate the Mask RCNN model has low performance in dense forest environments and much more encouraging results for open forest landscapes. Confidence scores of the predicted trees indicate high performance but when considering the mean Average Precision, the model's performance in actual forestry monitoring applications has lots of potential for improvement. Natural environment tree monitoring in diverse ecosystems still has the hurdle of dense overlapping canopies to solve, Mask RCNN deep learning models in combination with more complex data (i.e LiDAR or multispectral) could potentially offer solutions.

Bibliographic References

Abdulla, W. (2017). Mask R-CNN for object detection and instance segmentation on Keras and Tensorflow. *GitHub Repository*. https://github.com/matterport/Mask_RCNN

Abutaleb, K., Newete, S. W., Mangwanya, S., Adam, E., & Byrne, M. J. (2020). Mapping eucalypts trees using high resolution multispectral images: A study comparing WorldView 2 vs. SPOT 7. *Egyptian Journal of Remote Sensing and Space Science*, *xxxx*. https://doi.org/10.1016/j.ejrs.2020.09.001

Dlamini, W. M. D. (2020). *National Strategy for the Control & Management of Invasive Alien Plant Species.*

dos Santos, A. A., Marcato Junior, J., Araújo, M. S., Di Martini, D. R., Tetila, E. C., Siqueira, H. L., Aoki, C., Eltner, A., Matsubara, E. T., Pistori, H., Feitosa, R. Q., Liesenberg, V., & Gonçalves, W. N. (2019). Assessment of CNN-based methods for individual tree detection on images captured by RGB cameras attached to UAVS. *Sensors (Switzerland)*, *19*(16), 1–11. https://doi.org/10.3390/s19163595

Essl, F., Lenzner, B., Bacher, S., Bailey, S., Capinha, C., Daehler, C., Dullinger, S., Genovesi, P., Hui, C., Hulme, P. E., Jeschke, J. M., Katsanevakis, S., Kühn, I., Leung, B., Liebhold, A., Liu, C., MacIsaac, H. J., Meyerson, L. A., Nuñez, M. A., … Roura-Pascual, N. (2020). Drivers of future alien species impacts: An expert-based assessment. *Global Change Biology*, *26*(9), 4880–4893. https://doi.org/10.1111/gcb.15199

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-Decem*, 770–778. https://doi.org/10.1109/CVPR.2016.90

Immitzer, M., Vuolo, F., & Atzberger, C. (2016). First experience with Sentinel-2 data

for crop and tree species classifications in central Europe. *Remote Sensing*, *8*(3). https://doi.org/10.3390/rs8030166

IUCN. (2018). *Invasive Species*. https://www.iucn.org/theme/species/our-work/invasive-species

Kattenborn, T., Eichel, J., & Fassnacht, F. E. (2019). Convolutional Neural Networks enable efficient, accurate and fine-grained segmentation of plant species and communities from high-resolution UAV imagery. *Scientific Reports*, *9*(1), 1–9. https://doi.org/10.1038/s41598-019-53797-9

Kattenborn, T., Lopatin, J., Förster, M., Braun, A. C., & Fassnacht, F. E. (2019). UAV data as alternative to field sampling to map woody invasive species based on combined Sentinel-1 and Sentinel-2 data. *Remote Sensing of Environment*, *227*(January), 61–73. https://doi.org/10.1016/j.rse.2019.03.025

Kumar Rai, P., & Singh, J. S. (2020). Invasive alien plant species: Their impact on environment, ecosystem services and human health. *Ecological Indicators*, *111*(January). https://doi.org/10.1016/j.ecolind.2019.106020

Le Maitre, D. C., Blignaut, J. N., Clulow, A., Dzikiti, S., Everson, C. S., Görgens, A. H. M., & Gush, M. B. (2020). Impacts of Plant Invasions on Terrestrial Water Flows in South Africa. *Biological Invasions in South Africa*, 431–457. https://doi.org/10.1007/978-3-030-32394-3_15

Naidoo, L., Cho, M. A., Mathieu, R., & Asner, G. (2012). Classification of savanna tree species, in the Greater Kruger National Park region, by integrating hyperspectral and LiDAR data in a Random Forest data mining environment. *ISPRS Journal of Photogrammetry and Remote Sensing*, *69*, 167–179. https://doi.org/10.1016/j.isprsjprs.2012.03.005

Natesan, S., Armenakis, C., & Vepakomma, U. (2019). Resnet-based tree species classification using uav images. *International Archives of the Photogrammetry,*

Remote Sensing and Spatial Information Sciences - ISPRS Archives, *42*(2/W13), 475–481. https://doi.org/10.5194/isprs-archives-XLII-2-W13-475-2019

Nezami, S., Khoramshahi, E., Nevalainen, O., Pölönen, I., & Honkavaara, E. (2020). Tree species classification of drone hyperspectral and RGB imagery with deep learning convolutional neural networks. *Remote Sensing*, *12*(7). https://doi.org/10.3390/rs12071070

Onishi, M., & Ise, T. (2021). Explainable identification and mapping of trees using UAV RGB image and deep learning. *Scientific Reports*, *11*(1), 1–15. https://doi.org/10.1038/s41598-020-79653-9

Pyšek, P., Jarošík, V., Hulme, P. E., Pergl, J., Hejda, M., Schaffner, U., & Vilà, M. (2012). A global assessment of invasive plant impacts on resident species, communities and ecosystems: The interaction of impact measures, invading species' traits and environment. *Global Change Biology*, *18*(5), 1725–1737. https://doi.org/10.1111/j.1365-2486.2011.02636.x

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

Ricciardi, A., Palmer, M. E., & Yan, N. D. (2011). Should biological invasions be managed as natural disasters? *BioScience*, *61*(4), 312–317. https://doi.org/10.1525/bio.2011.61.4.11

Richardson, D. M., Pyšek, P., Rejmánek, M., Barbour, M. G., Dane Panetta, F., & West, C. J. (2000). Naturalization and invasion of alien plants: Concepts and definitions. *Diversity and Distributions*, *6*(2), 93–107. https://doi.org/10.1046/j.1472-4642.2000.00083.x

Schiefer, F., Kattenborn, T., Frick, A., Frey, J., Schall, P., Koch, B., & Schmidtlein, S.

(2020). Mapping forest tree species in high resolution UAV-based RGB-imagery by means of convolutional neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, *170*(November), 205–215. https://doi.org/10.1016/j.isprsjprs.2020.10.015

Underwood, E., Ustin, S., & DiPietro, D. (2003). Mapping nonnative plants using hyperspectral imagery. *Remote Sensing of Environment*, *86*(2), 150–161. https://doi.org/10.1016/S0034-4257(03)00096-8

Wang, D., Wan, B., Qiu, P., Su, Y., Guo, Q., Wang, R., Sun, F., & Wu, X. (2018). Evaluating the performance of Sentinel-2, Landsat 8 and Pléiades-1 in mapping mangrove extent and species. *Remote Sensing*, *10*(9). https://doi.org/10.3390/rs10091468Abdulla, W. (2017). Mask R-CNN for object detection and instance segmentation on Keras and Tensorflow. *GitHub Repository*. https://github.com/matterport/Mask_RCNN

Abutaleb, K., Newete, S. W., Mangwanya, S., Adam, E., & Byrne, M. J. (2020). Mapping eucalypts trees using high resolution multispectral images: A study comparing WorldView 2 vs. SPOT 7. *Egyptian Journal of Remote Sensing and Space Science*, *xxxx*. https://doi.org/10.1016/j.ejrs.2020.09.001

Dlamini, W. M. D. (2020). *National Strategy for the Control & Management of Invasive Alien Plant Species.*

dos Santos, A. A., Marcato Junior, J., Araújo, M. S., Di Martini, D. R., Tetila, E. C., Siqueira, H. L., Aoki, C., Eltner, A., Matsubara, E. T., Pistori, H., Feitosa, R. Q., Liesenberg, V., & Gonçalves, W. N. (2019). Assessment of CNN-based methods for individual tree detection on images captured by RGB cameras attached to UAVS. *Sensors (Switzerland)*, *19*(16), 1–11. https://doi.org/10.3390/s19163595

Essl, F., Lenzner, B., Bacher, S., Bailey, S., Capinha, C., Daehler, C., Dullinger, S., Genovesi, P., Hui, C., Hulme, P. E., Jeschke, J. M., Katsanevakis, S., Kühn, I.,

Leung, B., Liebhold, A., Liu, C., MacIsaac, H. J., Meyerson, L. A., Nuñez, M. A., … Roura-Pascual, N. (2020). Drivers of future alien species impacts: An expert-based assessment. *Global Change Biology*, *26*(9), 4880–4893. https://doi.org/10.1111/gcb.15199

Forest Survey of India. (2013). *India State of Forest Report 2013*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-Decem*, 770–778. https://doi.org/10.1109/CVPR.2016.90

Immitzer, M., Vuolo, F., & Atzberger, C. (2016). First experience with Sentinel-2 data for crop and tree species classifications in central Europe. *Remote Sensing*, *8*(3). https://doi.org/10.3390/rs8030166

IUCN. (2018). *Invasive Species*. https://www.iucn.org/theme/species/our-work/invasive-species

Kattenborn, T., Eichel, J., & Fassnacht, F. E. (2019). Convolutional Neural Networks enable efficient, accurate and fine-grained segmentation of plant species and communities from high-resolution UAV imagery. *Scientific Reports*, *9*(1), 1–9. https://doi.org/10.1038/s41598-019-53797-9

Kattenborn, T., Lopatin, J., Förster, M., Braun, A. C., & Fassnacht, F. E. (2019). UAV data as alternative to field sampling to map woody invasive species based on combined Sentinel-1 and Sentinel-2 data. *Remote Sensing of Environment*, *227*(January), 61–73. https://doi.org/10.1016/j.rse.2019.03.025

Kumar Rai, P., & Singh, J. S. (2020). Invasive alien plant species: Their impact on environment, ecosystem services and human health. *Ecological Indicators*, *111*(January). https://doi.org/10.1016/j.ecolind.2019.106020

Le Maitre, D. C., Blignaut, J. N., Clulow, A., Dzikiti, S., Everson, C. S., Görgens, A. H. M., & Gush, M. B. (2020). Impacts of Plant Invasions on Terrestrial Water Flows in South Africa. *Biological Invasions in South Africa*, 431–457. https://doi.org/10.1007/978-3-030-32394-3_15

Naidoo, L., Cho, M. A., Mathieu, R., & Asner, G. (2012). Classification of savanna tree species, in the Greater Kruger National Park region, by integrating hyperspectral and LiDAR data in a Random Forest data mining environment. *ISPRS Journal of Photogrammetry and Remote Sensing*, *69*, 167–179. https://doi.org/10.1016/j.isprsjprs.2012.03.005

Natesan, S., Armenakis, C., & Vepakomma, U. (2019). Resnet-based tree species classification using uav images. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, *42*(2/W13), 475–481. https://doi.org/10.5194/isprs-archives-XLII-2-W13-475-2019

Nezami, S., Khoramshahi, E., Nevalainen, O., Pölönen, I., & Honkavaara, E. (2020). Tree species classification of drone hyperspectral and RGB imagery with deep learning convolutional neural networks. *Remote Sensing*, *12*(7). https://doi.org/10.3390/rs12071070

Onishi, M., & Ise, T. (2021). Explainable identification and mapping of trees using UAV RGB image and deep learning. *Scientific Reports*, *11*(1), 1–15. https://doi.org/10.1038/s41598-020-79653-9

Pyšek, P., Jarošík, V., Hulme, P. E., Pergl, J., Hejda, M., Schaffner, U., & Vilà, M. (2012). A global assessment of invasive plant impacts on resident species, communities and ecosystems: The interaction of impact measures, invading species' traits and environment. *Global Change Biology*, *18*(5), 1725–1737. https://doi.org/10.1111/j.1365-2486.2011.02636.x

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards Real-Time Object

Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

Ricciardi, A., Palmer, M. E., & Yan, N. D. (2011). Should biological invasions be managed as natural disasters? *BioScience*, *61*(4), 312–317. https://doi.org/10.1525/bio.2011.61.4.11

Richardson, D. M., Pyšek, P., Rejmánek, M., Barbour, M. G., Dane Panetta, F., & West, C. J. (2000). Naturalization and invasion of alien plants: Concepts and definitions. *Diversity and Distributions*, *6*(2), 93–107. https://doi.org/10.1046/j.1472-4642.2000.00083.x

Schiefer, F., Kattenborn, T., Frick, A., Frey, J., Schall, P., Koch, B., & Schmidtlein, S. (2020). Mapping forest tree species in high resolution UAV-based RGB-imagery by means of convolutional neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, *170*(November), 205–215. https://doi.org/10.1016/j.isprsjprs.2020.10.015

Underwood, E., Ustin, S., & DiPietro, D. (2003). Mapping nonnative plants using hyperspectral imagery. *Remote Sensing of Environment*, *86*(2), 150–161. https://doi.org/10.1016/S0034-4257(03)00096-8

Wang, D., Wan, B., Qiu, P., Su, Y., Guo, Q., Wang, R., Sun, F., & Wu, X. (2018). Evaluating the performance of Sentinel-2, Landsat 8 and Pléiades-1 in mapping mangrove extent and species. *Remote Sensing*, *10*(9). https://doi.org/10.3390/rs10091468

Weinstein, B. G., Marconi, S., Bohlman, S. A., Zare, A., & White, E. P. (2020). Cross-site learning in deep learning RGB tree crown detection. *Ecological Informatics*, *56*(January). https://doi.org/10.1016/j.ecoinf.2020.101061

Annexes

DroneDeploy stitched maps of study areas A,B,C and D

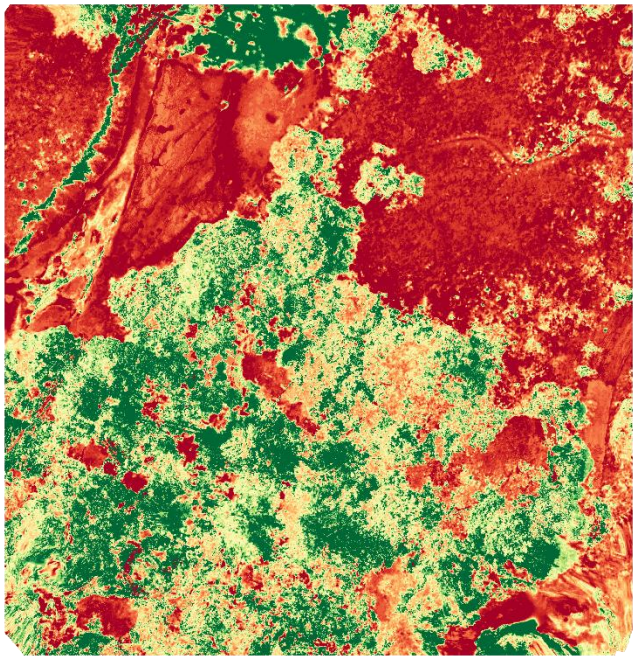

Figure i: NDVI(VARI) map for Site A

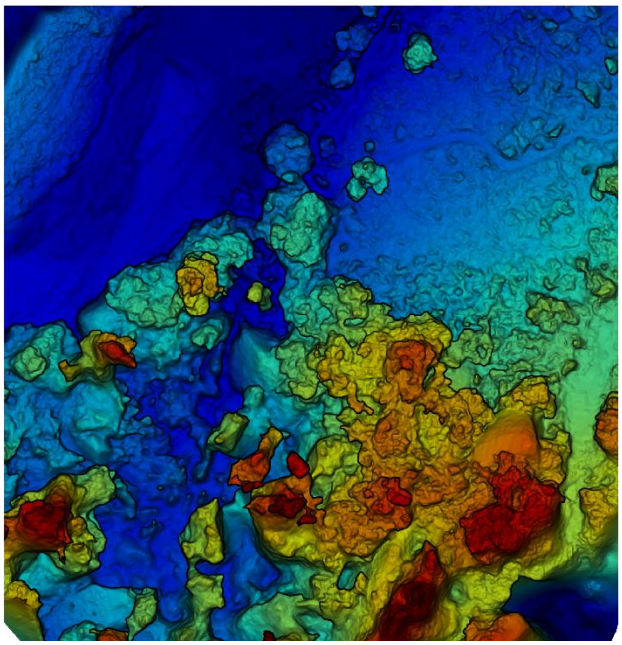

Figure ii: Digital Elevation Model for Site A

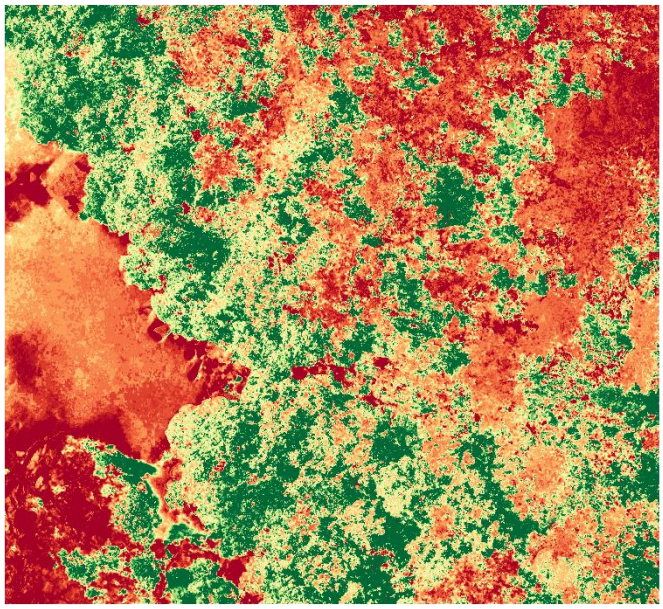

Figure iii: NDVI(VARI) map for Site B

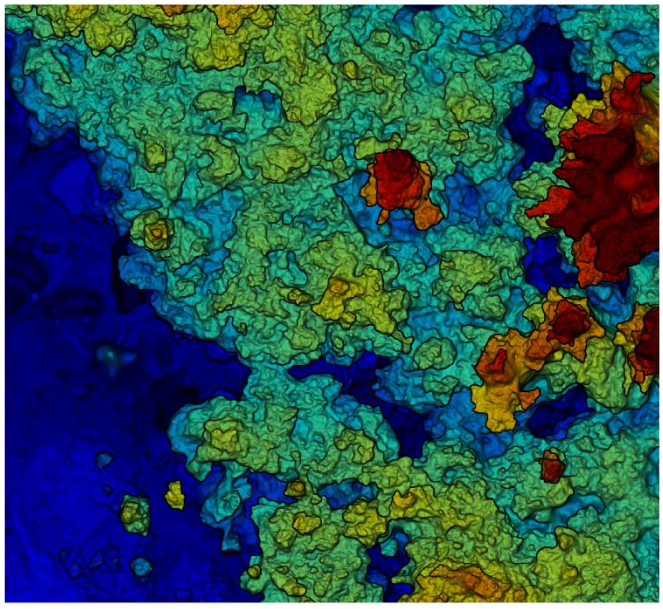

Figure iv: Digital Elevation Model for Site B

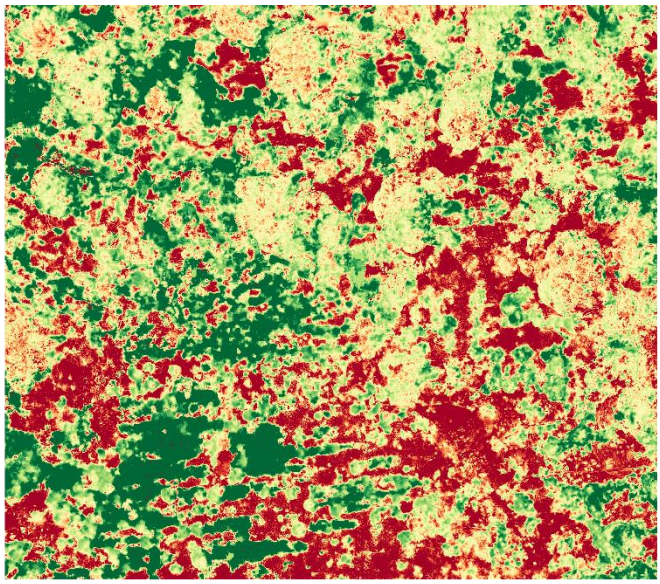

Figure v: NDVI(VARI) map of Site C

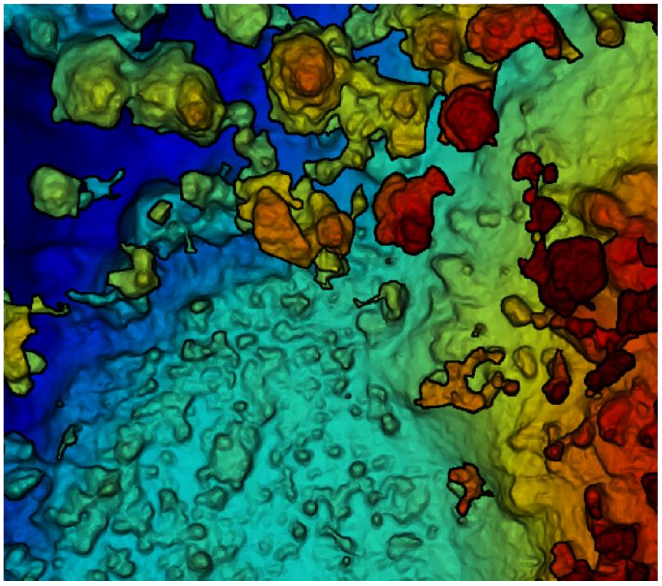

Figure vi: Digital Elevation Map of Site C

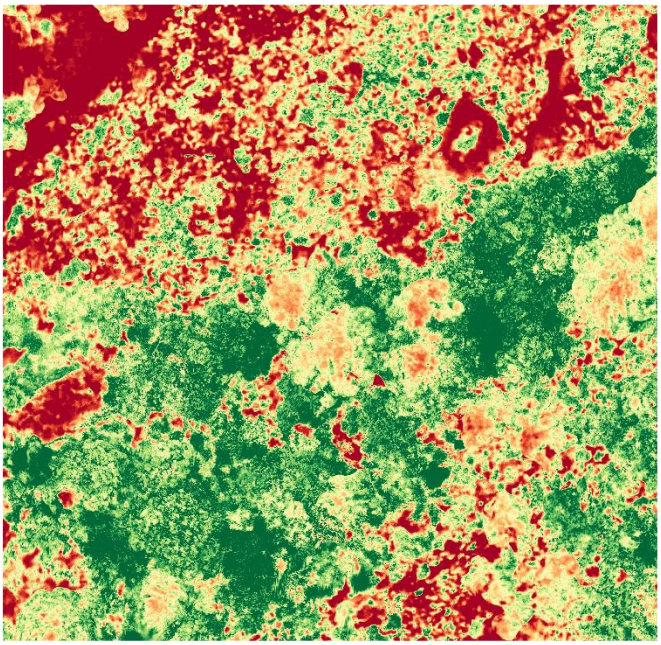

Figure vii: NDVI(VARI) map of Site D

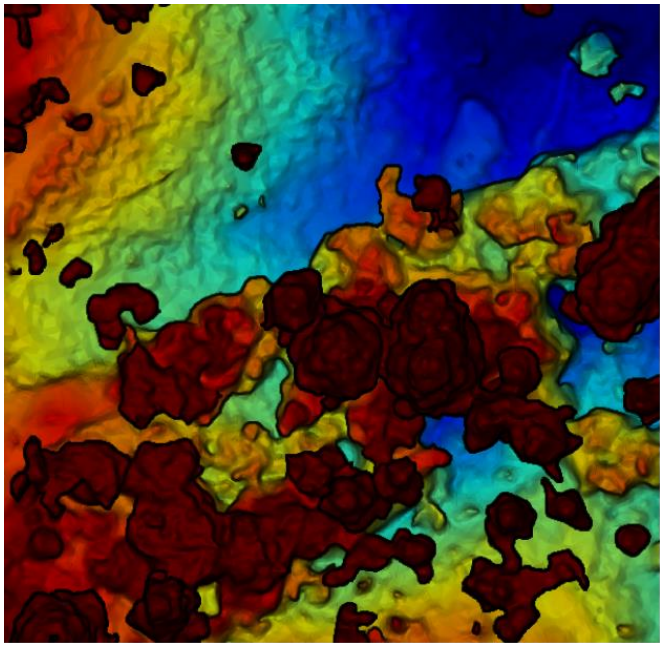

Figure viii: Digital Elevation map of Site D

Loss metrics for Finetuning on pre-trained weights

Figure viiii: RPN Class Loss

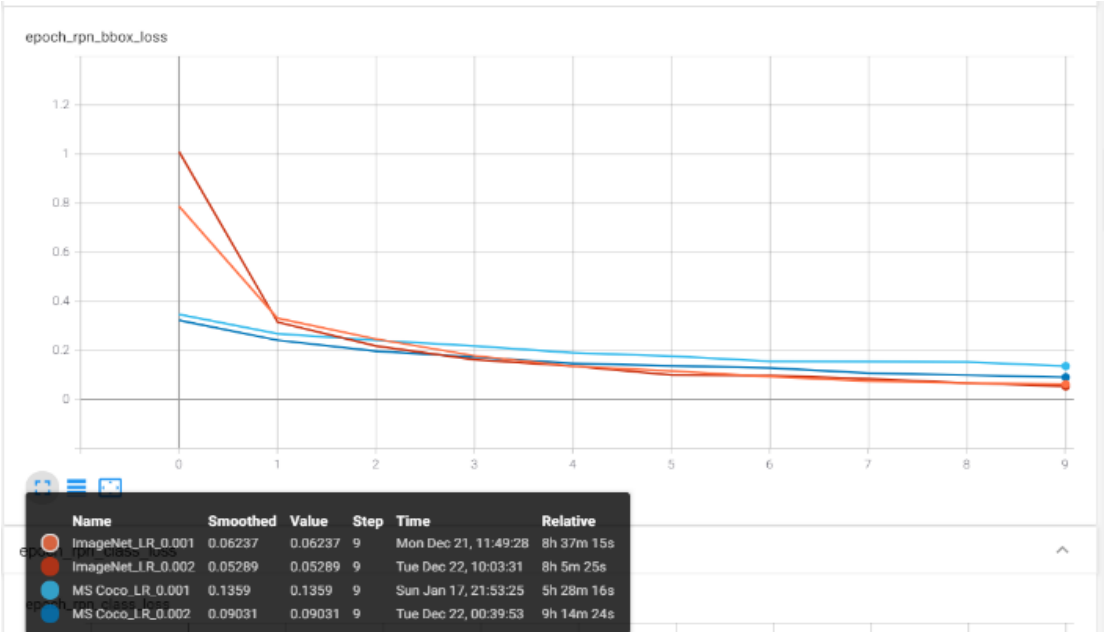

Figure x: RPN Bbox Loss

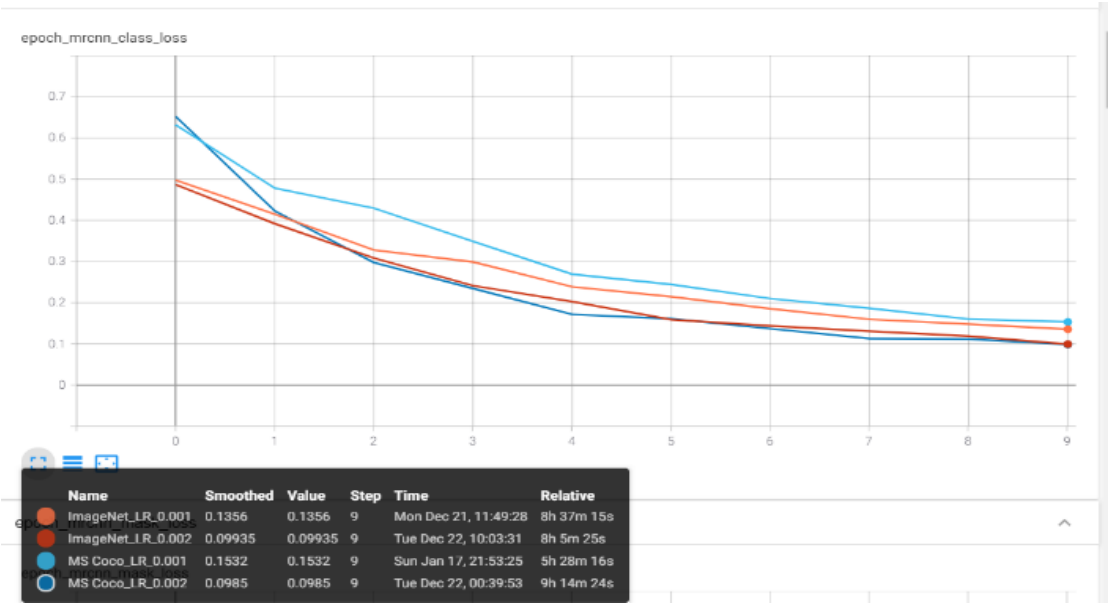

Figure xi: MRCNN Class Loss

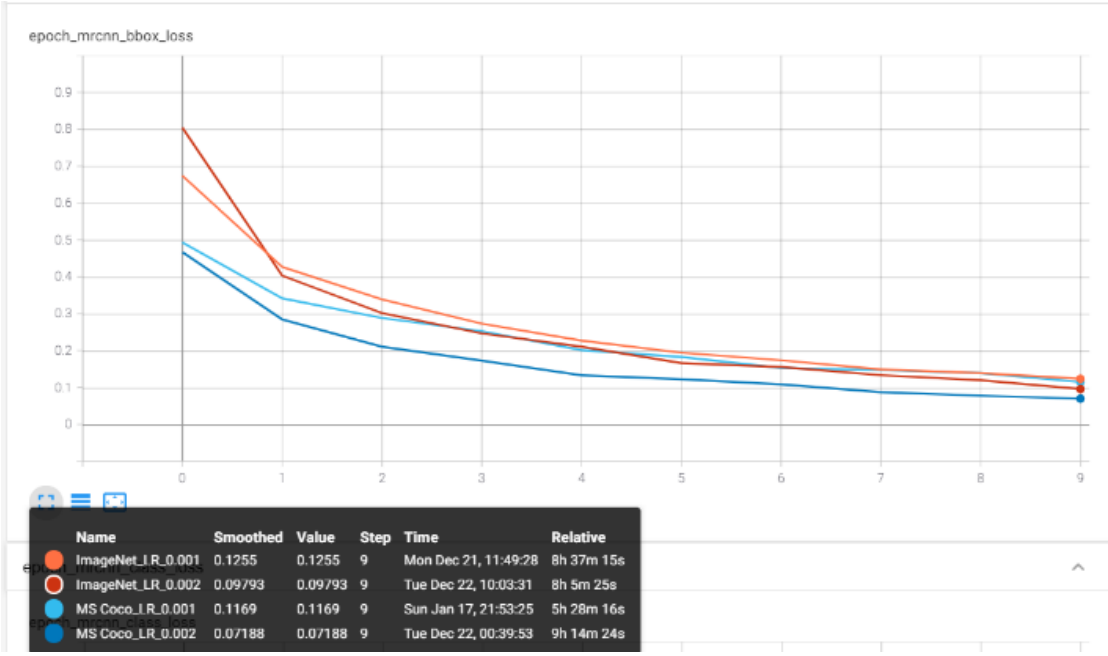

Figure xii: MRCNN Bbox Loss

Figure xiii: MRCNN Mask Loss

Masters Program in **Geospatial Technologies**

Supported by:
Education and Culture
ERASMUS MUNDUS