ForestVISION - Forest Floral Species Classification and Clustering

Versão final após defesa

Gonçalo Gomes Domingos

Dissertação para obtenção do grau de Mestre
Engenharia Informática
(2º ciclo de estudos)

Orientador: Prof. Doutor Luís Filipe Barbosa de Almeida Alexandre
Co-orientador: Dr. António José Marques Abreu

Janeiro de 2023

Declaração de Integridade

Eu, Gonçalo Gomes Domingos, que abaixo assino, estudante com o número de inscrição M11466 do Mestrado em Engenharia Informática da Faculdade de Engenharia, declaro ter desenvolvido o presente trabalho e elaborado o presente texto em total consonância com o Código de Integridades da Universidade da Beira Interior.

Mais concretamente afirmo não ter incorrido em qualquer das variedades de Fraude Académica, e que aqui declaro conhecer, que em particular atendi à exigida referenciação de frases, extratos, imagens e outras formas de trabalho intelectual, e assumindo assim na íntegra as responsabilidades da autoria.

Universidade da Beira Interior, Covilhã 03/01/2024

Agradecimentos

Gostaria de expressar a minha mais profunda gratidão e sinceros agradecimentos a todos os indivíduos envolvidos na realização deste trabalho de dissertação, que representa uma etapa crucial na minha jornada académica.

Em primeiro lugar, gostaria de expressar minha profunda gratidão tanto ao meu orientador quanto ao meu co-orientador, devido à excelente orientação e motivação que forneceram ao longo deste trabalho. As sugestões e críticas construtivas foram fundamentais para o meu desenvolvimento académico e para a conclusão bem-sucedida deste projeto. Sem a devida orientação e encorajamento constante, este trabalho não teria atingido os padrões que alcançou. Agradeço sinceramente por dedicarem tempo e esforço para me guiar e moldar meu percurso académico.

Gostaria de dar também os meus agradecimentos a toda a equipa envolvida neste projeto, bem como ao grupo SOCIALAB, pela generosa disponibilidade de recursos e pelo apoio inestimável que forneceram. O acesso aos recursos e a experiência do grupo SOCIALAB enriqueceu consideravelmente nosso projeto e permitiu que alcançássemos resultados significativos.

Claramente gostaria de expressar meus sinceros agradecimentos a todos os meus colegas que compartilharam esta fase importante da minha vida. As diferentes amizades, colaboração e apoio de cada um foram inestimáveis e tornaram esta jornada académica muito mais significativa. Estou eternamente grato por ter compartilhado este caminho com pessoas tão inspiradoras e dedicadas. Por último agradeço imenso a todos os membros da minha família.

Eles têm sido a base sólida e constante em minha vida, fornecendo o apoio emocional e o incentivo necessário para chegar até este momento. Agradeço do fundo do coração por estar ao meu lado ao longo desta jornada académica.

Resumo

Este trabalho tem como objetivo resolver uma das muitas tarefas relacionadas com a gestão e monitorização de florestas, especificamente em Portugal. A tarefa consiste em classificar diferentes espécies presentes na flora de Portugal. Com a utilização de inteligência artificial, conseguimos distinguir com sucesso diferentes espécies presentes nas imagens fornecidas. Um conjunto de dados também foi criado a partir do zero, uma vez que não havia informações públicas disponíveis sobre as diferentes espécies em Portugal. Foram criados dois conjuntos de dados, embora um deles não tenha atingido os resultados esperados e tenha causado muitos problemas. No entanto, o conjunto de dados de imagens de 8 bandas esperado produziu resultados aceitáveis.
Nesta tese, todos os passos serão abordados em grande detalhe ao longo dos capítulos, desde a criação dos conjuntos de dados até aos testes reais de cada modelo testado. Certificamo-nos de estudar a tecnologia mais recente e, portanto, exploramos os Transformadores que têm chamado a atenção no campo da Visão por Computador devido às suas características. Há também algumas ideias para o desenvolvimento futuro deste trabalho, uma vez que ainda há espaço para exploração adicional para completar este trabalho.

Palavras-chave

Inteligência Artificial; Problema de Classificação; Segmentação Semântica; Pré-Processamento de Data Sets; Aprendizagem Automática

Resumo alargado

A área da Inteligência Artificial tem vindo a aumentar exponencialmente ao longo dos anos. Hoje em dia usamos ou até podemos estar a ser influenciados por mecânismos que usam Inteligência Artificia, como por exemplo o corretor de texto no telemovél ou anuncios presentes em websites. Com o avanço desta área, não é surpreendente que os computadores sejam capazes de compreender o mundo real e tomar decisões com base no que observam. Neste trabalho, apresentamos a utilização da Inteligência Artificial para classificar diferentes espécies encontradas numa floresta, com foco nas espécies predominantemente presentes nas florestas portuguesas. Como um dos objetivos deste trabalho englobava o estudo da tecnologia mais recente na área em questão e aplicá-la, foi realizado um estudo sobre os *Transformers*. Estes modelos desempenham um papel crucial no processamento de linguagem natural e são considerados o estado da arte nessa área. No entanto, recentemente houve um interesse te usar estes modelos no campo da Visão Computacional dado as suas características. Sendo assim o capítulo 2 desta tese apresenta alguns modelos que foram estudados de forma a adquirir um conhecimento prévio da arquitetura dos modelos usados. Para além disso, terá sido feito uma revisão de trabalhos semelhantes, nos quais terão sido utilizados métodos como segmentação semantica e segmentação por instâncias para discernir diferentes copas de árvores. Também dão uso a imagens com diferentes bandas onde algumas espécies conseguem ser mais facilemnte detetadas e diferenciadas. Também foi feito um estudo na possível a deteção de doenças e espécies invasoras num ecosistema, no entanto não será alvo deste trabalho. Todo este estudo poderá ser verificado no capítulo 3. Após uma pesquisa extensa sobre *data sets* públicos que contenham informação sobre características das espécies presentes em florestas portuguesas, ou áreas com maior suscetibilidade de aparecimento de uma espécie, houve a necessidade de criar um *data set* para a respetiva tarefa. Terão sido criados dois *data sets* com diferentes caracteriticas, no entanto apenas um deles nos deu resultados aceitáveis. O capitulo 4 desta tese indica todas as caracteristicas de ambos os *data sets*. Não tendo experiência prévia no uso de modelos de inteligência artificial, foram feitos alguns testes de forma a famirializar-se com o processo de treino e teste de um modelo. O capitulo 5 mostra o uso que um modelo relativamente antigo que no entanto acabou por fazer o seu propósito no âmbito de educar o autor nos aspetos necessários para a composição de um *data set*, bem como treino e teste do modelo no mesmo. Os restantes capítulos acabam com uma compreensão sobre um dos *data sets* não ser apto para a tarefa em questão e daí ser necessário a criação do segundo *dataset* e a sua especificações, conforme descrito no capitulo 6. De seguida temos os resultados onde são comparados diferentes modelos e configurações, bem como uma discussão dos resultados, respetivamente o capitulo 7. Por fim acabamos um breve sumário dos objetivos deste trabalho de tese e algumas ideias para uma melhoria deste trabalho, presente no capitulo 8.

Abstract

This work aims to solve one of the many tasks responsible for management and monitoring of forests, specifically around Portugal. The task is to classify different species present in the Portugal flora. With the use of artificial intelligence, we were able to successfully distinguish different species present within the images provided. A data set was also created from scratch since there was no public available information regarding the different species around Portugal. Two data sets were created although one of them did not meet the expected results and gave us a lot of problems. However the expected 8 band imagery data set gave us acceptable results. In this thesis, all the steps will be covered with great detail along the chapters, from the build up of the data sets to the real testing of each model tested. We made sure to study the most recent technology and therefore explore the Transformers that have been raising awareness in the Computer Vision field for their characteristics. There is also some ideas for a future development of the respective work, since there can still be some exploration to be made to complete this work.

Keywords

Artificial Intelligence; Classification Problem; Semantic Segmentation; Pre-Processing of Data sets; Machine Learning Model;

Contents

1 Introduction **1**

1.1 Thesis Work . . . 1
1.2 Motivation and Objectives . . . 1
1.3 Thesis Organization . . . 1

2 Preliminary Concepts **3**

2.1 Remote Sensing . . . 3
2.2 Spectral Analysis . . . 3
2.3 Image Resolution . . . 3
2.4 Computer Vision . . . 5
2.5 Semantic Segmentation Models . . . 6
2.5.1 Transformers . . . 6
2.5.2 Self-Attention . . . 7
2.5.3 SegFormer [1] . . . 8
2.5.4 Segmenter [2] . . . 9
2.5.5 Vit-Adapter [3] . . . 10
2.5.6 InternImage [4] . . . 11
2.5.7 Lawin [5] . . . 12

3 Related Work **15**

3.1 Introduction . . . 15
3.2 Ludvision [6] . . . 15
3.2.1 Introduction . . . 15
3.2.2 Motivation . . . 15
3.2.3 Methods . . . 15
3.2.4 Utilized Method . . . 15
3.2.5 Results and Discussion . . . 16
3.3 Continuous Monitoring Of Forest Change Dynamics With Satellite Time Series 17
3.3.1 Introduction . . . 17
3.3.2 Material and Methods . . . 17
3.3.3 Study Sites . . . 18
3.3.4 Results and Discussion . . . 19
3.4 Individual tree segmentation and tree species classification in subtropical broad leaf forests . . . 19
3.4.1 Introduction . . . 19
3.4.2 Study Area and Data . . . 20
3.4.3 Methods . . . 20
3.4.4 Results and Discussion . . . 22
3.4.5 Results from feature extraction . . . 23
3.5 Early Detection of Bark Beetle . . . 23

3.5.1 Introduction . . . 23
3.5.2 Reviews on Bark Beetle Damage Detection . . . 24
3.5.3 Bark Beetle-Host Tree Interaction . . . 24
3.5.4 Remote Sensing . . . 24
3.5.5 Machine Learning . . . 26
3.5.6 Classical Methods . . . 26
3.5.7 Deep Learning Methods . . . 26
3.5.8 Metrics . . . 27
3.6 Mapping forest in the Swiss Alps treeline ecotone with explainable deep learning 28
3.6.1 Introduction . . . 28
3.6.2 Data . . . 28
3.6.3 Segmentation Task . . . 28
3.6.4 Explainable deep learning model architecture . . . 29
3.6.5 Forest Mapping . . . 29
3.6.6 Intermediate concepts estimations . . . 30
3.7 Hyperspectral and LiDAR data for the prediction via machine learning of tree species, volume and biomass. . . . 31
3.7.1 Introduction . . . 31
3.7.2 Study Area . . . 32
3.7.3 Spectral feature selection . . . 33
3.7.4 Delineation of tree crowns . . . 33
3.7.5 Classifiers . . . 33
3.7.6 Biomass and volume estimation models . . . 34
3.7.7 Determination of the species . . . 35
3.7.8 Estimation of volume and above ground biomass . . . 36
3.8 Conclusions . . . 37

4 Data sets **39**
4.1 Cartography of North / Center of Portugal . . . 39
4.2 Planet Research and Education Program . . . 39
4.3 Copernicus Land Monitoring . . . 40
4.4 Pre-processing and Methods . . . 40
4.5 Annotations . . . 41

5 Experiments and Ideas **43**
5.1 Data set . . . 43
5.2 Model . . . 43
5.3 Results . . . 44
5.4 Discussion . . . 44
5.4.1 Realization . . . 45
5.5 Relevant Ideas For The Future . . . 46

6 Realization and Annotations **47**
6.1 Realization on the Mentioned Ideas 47
6.2 Annotations 47
6.3 Data set Specifications 49

7 Results **51**
7.1 Introduction 51
7.2 Hardware and Software Specifications 51
7.3 ViT-Adapter 51
7.4 InternImage-H 52
7.5 Lawin 52
7.6 Metrics and Results 52
7.7 Discussion 53

8 Conclusion **55**
8.1 Brief Summary of the Project 55
8.2 Summary of the Research Objectives 55
8.3 Future Work 55

Bibliografia **57**

List of Figures

2.1 Example of Remote Sensing being used in a forest.[7] 4
2.2 Differences between Semantic Segmentation, Instance Segmentation and Panoptic Segmentation. [8] 6
2.3 Example of the weights from an attention mechanism on a sentence. [9] . . . 8
2.4 The proposed SegFormer Framework.[1] 9
2.5 An overview of the Segmenter.[2] 9
2.6 Overall architecture of ViT-Adapter. [3] 11
2.7 Overall architecture of InternImage. ”s2” and ”p1” mean stride = 2 and padding = 1[4] 12
2.8 Highlighting the Differences Between LawinASPP and ASPP.[5] 13
2.9 Overall architecture of Lawin Transformer.[5] 13

3.1 Table of multi spectral specifications.[6] 16
3.2 Altitude, time and number of images collected.[6] 16
3.3 Example of images in each band. *Red band (a), green band(b), blue band (c), red edge band (d), near infra-red band (e), RGB (f)*[6] 17
3.4 Reflectance values (in %) for each band in different image elements.[6] 18
3.5 Architecture of the final model, based on the HRNet.[6] 18
3.6 Performance comparison between HRNet and our model.[6] 19
3.7 Reference phenology of the study site, Genteguela (Ivory Coast).[10] 20
3.8 Sample size and mapped portion for each site and stratum.[10] 21
3.9 Area weighted accuracy’s (OA = Overall Accuracy; UA = User Accuracy; PA = Producer Accuracy).[10] 22
3.10 Examples of the segmentation results.[11] 22
3.11 Accuracy assessment results of individual tree segmentation of WST-Ncut algorithm.[11] 23
3.12 The mean and ± standard deviation of the reflectance of 18 tree species. Each specie has it’s own value represented on the graph.[11] 23
3.13 Box plots of the 6 most important LiDAR features.[11] 24
3.14 Box plots of the 6 most important textural features.[11] 25
3.15 The classification results of 18 tree species using seven feature sets.[11] 25
3.16 Detection of different stages of a bark beetle attack.[12] 26
3.17 An overview of the 3 key aspects, along with all the aspects involved behind the key aspects.[12] 27
3.18 Examples of different bark beetle species and host trees (left) and symptoms of attacked trees (right).[12] 27
3.19 Comparison of Remote Sensing systems to detect bark beetle attacks (only for satellite imagery).[12] 28

3.20 Most effective SVIs for early detection of bark beetle attacks. The multi spectral and hyper spectral analyses are denoted by MS and HS, and RX denotes reflectance at wavelength X nm.[12] 29
3.21 Classical machine learning methods for detecting bark beetle attacks.[12] . . . 30
3.22 Study area, displaying the general appearance with Sentinel-2 imagery (a) and the partition into training, validation, and test sets (b).[13] 31
3.23 NFI plots and SwissImage tiles spatial arrangement (easternmost part of the test set).[13] 31
3.24 Aerial image and associated targets extracted from SwissTLM3D annotations. p/a: presence/absence. [13] 32
3.25 Flowchart of the forest mapping methods.[13] 32
3.26 Rules enforced by the rule module.[13] 32
3.27 Overall accuracies for both methods.[13] 33
3.28 Overall accuracies for both methods in forest type and the forest presence/absence.[13] 33
3.29 Visual extracts of the forest mapping results.[13] 33
3.30 Visual extracts of the tree height predictions (in m) on the test set (SB)[13] . . 34
3.31 Segmentation results by applying rules on intermediate concept targets. Rules: map obtained by applying the rules to the intermediate targets. Rules-TLM: Rules map with TLM disambiguation.[13] 34
3.32 Study area and the inventories of the private forests.[14] 35
3.33 List of species and number for samples used for the training and test phases.[14] 35
3.34 Hyperspectral classification comparison.[14] 36
3.35 Accuracy metrics from test data sets.[14] 36
3.36 Biomass and volume results from observed ground truth values (Ob) and estimated predicted values (Pr) for volume (V) and above ground biomass (AGB).[14] 37

4.1 Example of the carthography using when using the QGES software. [15] 39
4.2 Example of the Planet.com plaftorm [16] 40
4.3 Example of the information from the Copernicus Land Monitoring Service. [17] 41

5.1 Example of the class masks from an annotation. 44
5.2 Overall architecture of MaskR-CNN.[18] 44
5.3 Example of an output from the MaskR-CNN model. 45
5.4 Hand-made annotation of the supposed right masks of the output from figure5.3. *(The 1 present in the green patches is to differentiate from the other green patches with no number in it, since it is from another class.)* 46

6.1 RGB Scene from Cartography of North / Center of Portugal. 48
6.2 Scene from the Planet data set. 48
6.3 Scene from figure 6.2 but with rearranged bands. 49

List of Acronyms

CNN	Convolutional Neural Networks
CFIUTE	Centro de Formação Interação UBI Tecido Empresarial
LiDAR	Laser Imaging, Detection, and Ranging
LSTM	Long Short-Term Memory
NLP	Natural Language Processing
GPU	Graphical Processing Unit
NN	Neural Networks
MLP	Multi Layer Perceptron
MSA	Multi-headed Self-Attention
TCP	*Transmission Control Protocol*
UBI	Universidade da Beira Interior
FCN	Fully Convolutional Network
ViT	Vision Transformer

Chapter 1

Introduction

1.1 Thesis Work

This thesis project aims to take advantage of the latest computer vision techniques for the accurate classification and clustering of diverse plant species found in the Portuguese flora, particularly within the Cova da Beira region. Given the extensive terrain to be covered, this project will be designed to use satellite imagery while taking into account the resolutions and spectral bands necessary for detecting to detect and classify the different plant species in the respective area. This research will focus on a select number of species due to some difficulties that were faced later on in the research.

1.2 Motivation and Objectives

In the context of a forest ecosystem, correctly identifying which species are present is really important. This can be used either for inventory management, monitoring or planning. Remote sensing has become a prevalent tool in forestry applications since it can detect enormous areas of landscape with the use of various sensors that capture information at different wavelengths. Each specie tend to give away different reflectance values in across bands [19] thus using different bands is essential to help in the classification of different species. While RGB images can be helpful for humans to make annotations and have a perception of the scene, when it comes to the machine the use of different bands is the right way to classify different species [19] [6].
This project explores most of the aspects that are generally associated with training a machine learning model for a specific task. From gaining the insights into the required data, to the model training and testing to obtain the final results. Each step will be documented along the documented and explained in clear detail.

1.3 Thesis Organization

To help the reading experience, the following list shows how this article is organized:

1. **Introduction** - It expresses the motivation and the objectives for the respective project. It also includes how the thesis is organized.
2. **Preliminary Concepts** - Gives an introduction to the reader about some fundamental concepts to this thesis.
3. **Related Work** - Lays out some works done by other authors, along with a brief summary, that were used as base to this project.

4. **Data set** - Explains all the data set properties and some pre-processing methods used on the data.
5. **Experiments and Ideas** - Presents us some experimentation done on a similar task and some thought to take in from it.
6. **Realization and Annotations** - Gives us an idea on how the annotations were made and the characteristics of the data set.
7. **Results** - This chapter has the results of all the experimentation with different models and a comparison between them.
8. **Conclusion** - Ending chapter with a brief summary of the project, all objectives achieved and future work that can be useful to the completion of this work.

Chapter 2

Preliminary Concepts

2.1 Remote Sensing

Remote Sensing is the act of acquiring information about an object, area or phenomenon without making physical contact with it. Its applications can be seen in many different fields such as geography, hydrology, ecology and military operations. Typically, remote sensing is associated with the use of a satellite or aircraft that can hold sensors for capturing data on the surface of the planet.
There are two types of Remote Sensing: Active and Passive. Passive Remote Sensing can gather radiation emitted or reflected by an object or surrounding areas. Active Remote Sensing emits energy upon the target area and measures the reflected radiation from the objects in the area. From the time delay of the emission returned it is possible to detect the characteristics of an object, such as speed, location and direction.
Therefore, the use of this technology can be considered an adequate approach for tree classification, since it can provide large views of a landscape and detect the flora present in it. Different species react in different manners upon different wavelengths [19], so finding the right type of spectral data is a necessary task to distinguish different species. The same species also change according to the season and thus changing their characteristics, like their leafs or even their blooming.

2.2 Spectral Analysis

When Remote Sensing is performed on a specific area, the imagery taken usually consists of different bands that are specific to their respective frequency and wavelength. Normally, every object has their unique value of light reflectance thus helping in the classification, and for trees/plants it can be their chlorophyll fluorescence emitted from their leafs. Using many bands is the strategy to identify distinct species that can look the same relative to the human eye, although it can be quite expensive asking for every band possible in some range. Having this in mind, there needs to be a study behind the species that need to be classified in order to find the best sensor and spectral data to detect each specie.

2.3 Image Resolution

There are three types of image resolution: Spatial, Temporal and Spectral. When taking an image, either with a regular camera or a satellite sensor, the term resolution always comes along. This type of image resolution is the spatial resolution and can be considered as the

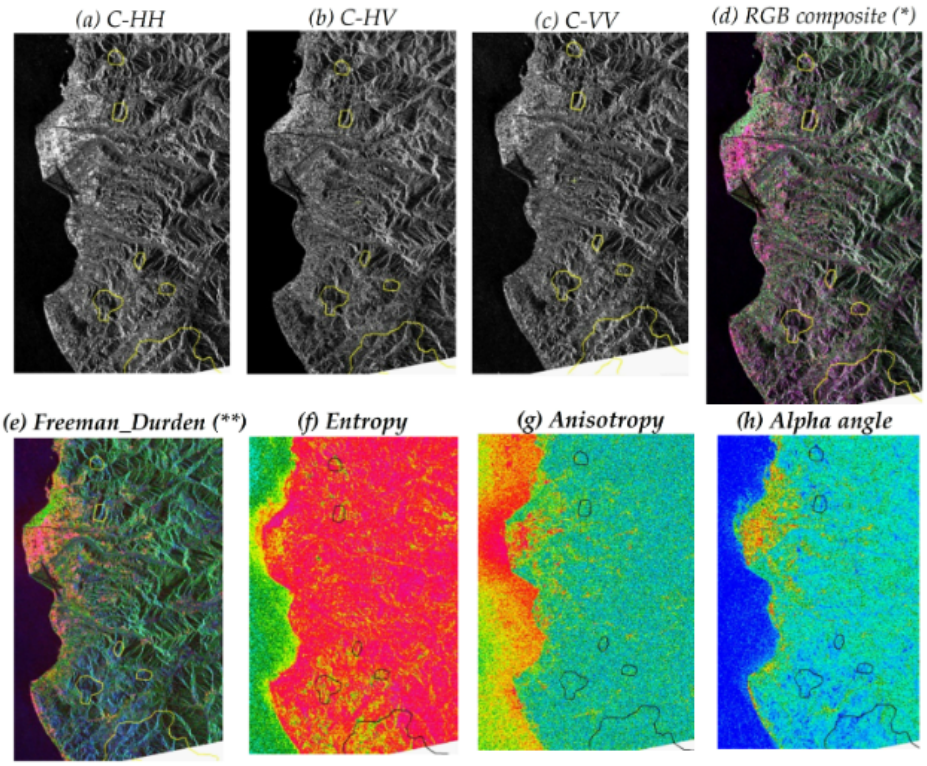


Figure 2.1: Example of Remote Sensing being used in a forest.[7]

detail that an image holds. Higher resolution meaning more detail since there are more pixels. When it comes to both aerial imagery and satellite imagery, the spatial resolution of the image often comes at play. If it is too low, it can not give us much information since there is more real characteristics agglomerated in one pixel, and thus giving bad results for certain tasks, for instance, individual tree classification [19].

Most public accessible data from these satellites, such as the Landsat[20], often range between 20 to 50 meters. Some platforms often have a subscription plan where it is possible to obtain images with higher resolutions. Most commercial data often ranges between 2 and 5 meters, although it is possible to achieve better resolutions such as 30 centimeters. However most projects tend to use the standard commercial data and perfect their methods or work with those resolutions.

Spectral Resolution is the ability to hold different bands in one image, and as it was mentioned before, it is a must to complete the objective of this work.

Temporal Resolution is how much frequency a particular area is captured and revisited. As you will see later on in chapter 4, the images from the data sets were obtained from a single day and therefore the concept of Temporal resolution will not be necessary for this work. Although some species can change with their bloom, none of the studied species change along the year and therefore there is no need for a temporal resolution

2.4 Computer Vision

This area of Artificial Intelligence focuses on enabling computer systems to interpret information from images and videos, comprehending the visual world. Numerous algorithms and systems have been developed to analyze both images and videos, extracting pertinent features from them. Additionally, it enables these systems to take actions based on the extracted information, as further discussed in this section.
To undertake such tasks, recent approaches require training the machine with extensive data. Computer Vision applications can be observed across many fields such as autonomous driving, video content analysis, pose tracking, gesture recognition, and more.
The use of deep learning methods, propelled the computer vision field along with the use of CNNs (Convolutional Neural Network). Once enough data is given to these types of models, they will be able to differentiate between visual inputs. CNNs extract features from the input no matter where they are in the input itself. A feature extractor with a defined size slides through the input and outputs a value, based on it's size and values. Along with the convolution term often comes striding, pooling, and padding. Regular NNs (Neural Network) on larger inputs such as images tend to be less efficient due to the high computational power needed for the weights of each neurons. In order to be efficient in those tasks CNNs tend to reduce the size of the input after each convolutional layer. That task is either done by pooling or striding, where pooling consists of dividing the input area into various non-overlapping region while taking a rule-based value from itself. The output from pooling will be smaller than the input since we took a singular value that will represent the divided region from the input. Striding is a value that is used to skip some pixels each time the matrix slides across the input, for example, skipping 2 or 3 pixels, thus resulting again in a smaller output. Padding comes into play when we do not want to reduce the size of the image to extract low level features. For instance, applying 0 values to the outside of the image is called padding. These transformations help to process an image by reducing the amount of information from the input, while maintaining some of the characteristics and thus making NNs capable of classifying the images given.
This leads to the application of CNNs in various areas for example: Image Classification; Object Recognition; Audio Visual Matching; Object Reconstruction; Semantic Segmentation; Speech Recognition.
Although the last item is not a visual task, CNNs can also be used in this area. Image classification only gives us the labels of the different objects on images. Object Recognition identifies and mark the objects on an image, often it's useful to have these in real-time. Audio Visual Matching is used by streaming platforms to improve their search algorithms and the user experience, since streaming platforms is where users mainly have specific requests and there's the need to satisfy such requests. Object Reconstruction is used to model real life objects. Nowadays CNNs models can create 3D face models based on one image, or even create a representation of a product that can be used in manufacturing.
Semantic Segmentation is the task that's most important for this project, since it can give a label to each pixel. This area can have other fields like Instance Segmentation where it segments each object in the image, even when they are from the same class, and Panoptic

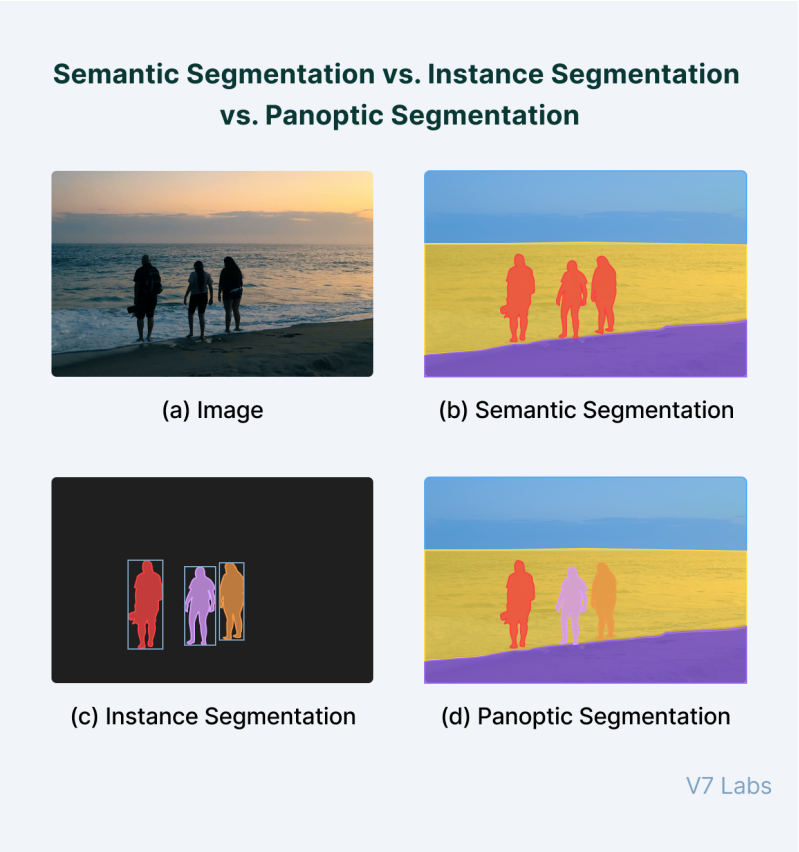


Figure 2.2: Differences between Semantic Segmentation, Instance Segmentation and Panoptic Segmentation. [8]

Segmentation that can be considered as a mix between Semantic and Instance Segmentation. The figure 2.2 shows us an example of these three.

2.5 Semantic Segmentation Models

One of the base models for Semantic Segmentation task was the FCN (Fully Convolutional Network) along with the U-Net. Although FCNs and CNNs are both used in Computer Vision, they are quite different. FCNs perform labeling for each pixel while traditional CNNs just label the whole image. The U-Net is a CNN that is modified to produce segmentation, in other words, it is able to label each pixel in the image. Having those types of model as the foundation, researchers focused on improving these types of networks, by processing the contextual information, enlarging the receptive field of each pixel, using the so called dilated convolutions, boundary information, designing various attention modules, etc. In this section we'll talk about some of the most interesting models for the Semantic Segmentation task.

2.5.1 Transformers

This type of architecture is becoming more common in the vision task area. It can be considered as a general MLP whereas CNNs and LSTM are biased models, since connections on transformers are computed on the fly, while in CNNs or MLP everything is connected and

weights remain there to be changed. Therefore Transformers can grasp the relations of different parts of an input on the first layer of the model. Given enough data it can surpass the results obtained by CNNs on the same tasks. The attention mechanism present in this type of architecture is what makes them different from other architectures. Upon receiving an input, this mechanism relates each and every part of the input, and gives a certain weight to other parts of the input. Figure 2.3 shows a clear example of how this mechanism works. It takes a phrase as input and each word is considered a token. Words that end up relating with each other tend to have different weights than those that do not relate to them. From the example on figure 2.3, we can see that the word "*it*" heavily relates to "*The*" and "*Animal*". The overall architecture is composed of 2 parts, an encoder and a decoder. The encoder is constituted by a stack of many blocks of encoders, all performing the same task. Each encoder is divided in two layers, the Self-Attention layer and a Feed Forward Neural Network, and only takes numbers as input, so in other words, the original input needs to be transformed into a vector. This characteristic makes it more interesting since we can train any transformer with any type of data, whether it is a pixel value or a word doesn't affect the understanding of the transformer. The first layer is the key component of this architecture, where it takes a look at every single part of the input and helps it encode the single point of data that is observing. Section 2.5.2 will go into more detail on this topic, since it's the sole characteristic that makes this architecture so appealing.
The decoder it's the same as the encoder, with just as many stacked blocks of decoders. The only difference is that the decoder has one more layer in between the 2 previously mentioned. It's an Encoder-Decoder Attention layer that takes the input of the $\boldsymbol{N}$ encoder block so that it takes in consideration the important parts of the input.
On other cases, such as vision tasks, we can take advantage of this mechanism on the pixels that make an image. So in other terms, with the attention mechanism we can see which pixels relate to the specific pixel we take in consideration.

2.5.2 Self-Attention

This part of the architecture is the most important part, so it obviously needs a deeper explanation. By taking the input, either from the original input transformed or from another encoder block, three vectors are created from each of the input vectors. Let's take the sentence from the example in figure 2.3, each word from that sentence will create three vectors associated with the word vector. These vectors will be called: Query vector; Key vector; Value vector. These vectors are made by multiplying the word vector with a respective matrix, that was built on the training process. With these vectors for each vector input, or following the example each word, the next step it's to obtain the score by multiplying the query vector of the considered word with all the key vectors from all the words in the input. That gives us the importance of each word of the input when encoding the word considered. The following step is to take the score and divide by the square root of the dimension of the key vectors, and then applying a softmax operation. This softmax value actually correlates to how much each word is relevant to the word taken into account. The next step is multiplying this softmax value with the Value vector. This step makes it so that the Value vector of each word isn't

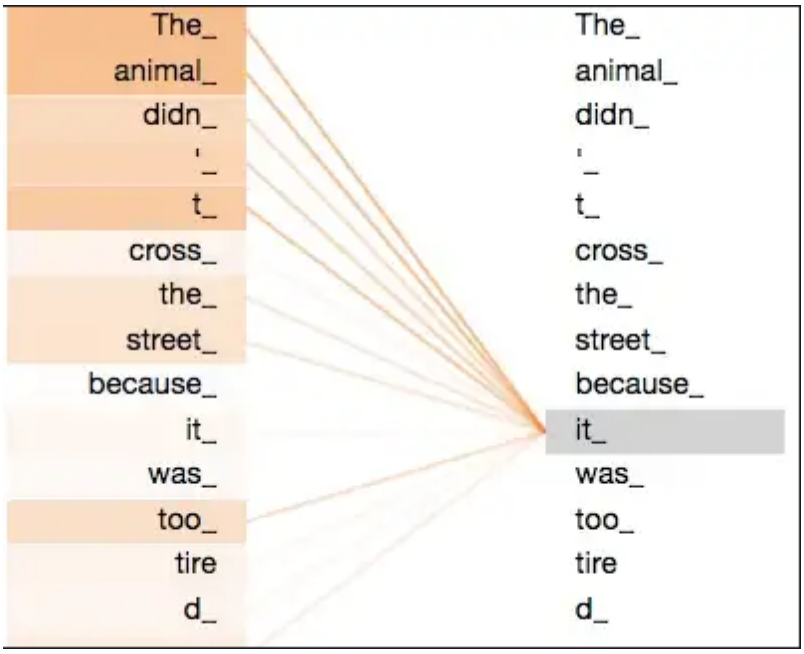


Figure 2.3: Example of the weights from an attention mechanism on a sentence. [9]

changed but the words that are the most meaningless to the respective word, will have lower softmax-Values vector values, since that is one of the characteristics of the softmax operation. The final step of calculations is to sum all of these softmax-Values vectors and then feed them into the next layer, the Feed Forward Neural Network layer. Usually these calculations are made with matrices since it is faster and efficient.

2.5.3 SegFormer [1]

This framework unifies Transformers with lightweight MLP decoders. After the great success that transformers had in NLP, the introduction of Transformers to the visual tasks was of great interest. The first proposed Vision Transformer (ViT) for image classification was done by Dosovitskiy et al.[21] achieving great results on ImageNet.
The key novelties of this model are: A new positional-encoding-free and hierarchical Transformer encoder; A lightweight All-MLP decoder design that yields a powerful representation without complex and computationally demanding modules; it produces state-of-the-art efficiency, accuracy and robustness in three Semantic Segmentation data sets[1].
One of the features of SegFormer is that the model performs inference on a different resolution used for training. The positional-encoding-free means that when testing in different resolutions the encoder can adapt to arbitrary resolutions without impacting the performance since it does not interpolate positional codes on different types of resolutions from the training one. In addition, the hierarchical parts enable the encoder to generate high-resolution fine features and low-resolution coarse features. The key idea of using an MLP decoder is to take advantage of the Transformer features where the attentions of the lower layers stay local, where the higher layers are highly non-local. The MLP decoder combines both local and global attention by joining the information from different layers. The figure 2.4 shows the respectively framework.

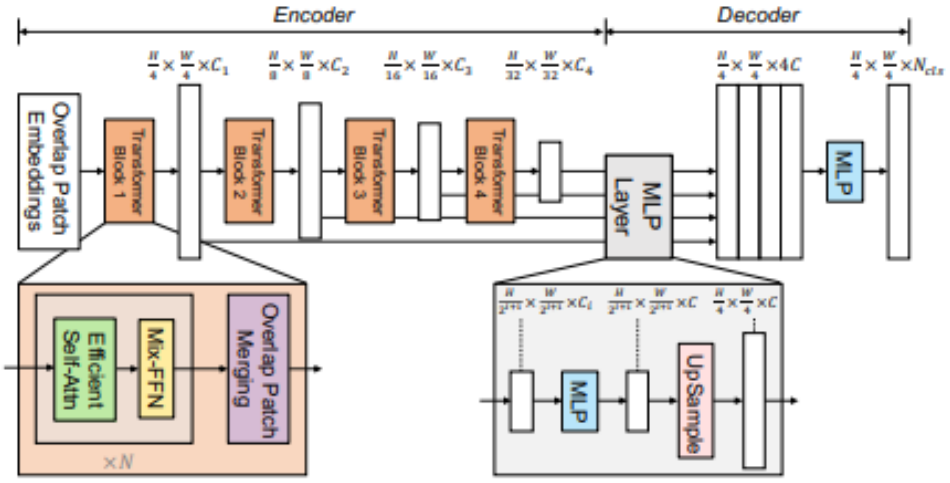


Figure 2.4: The proposed SegFormer Framework.[1]

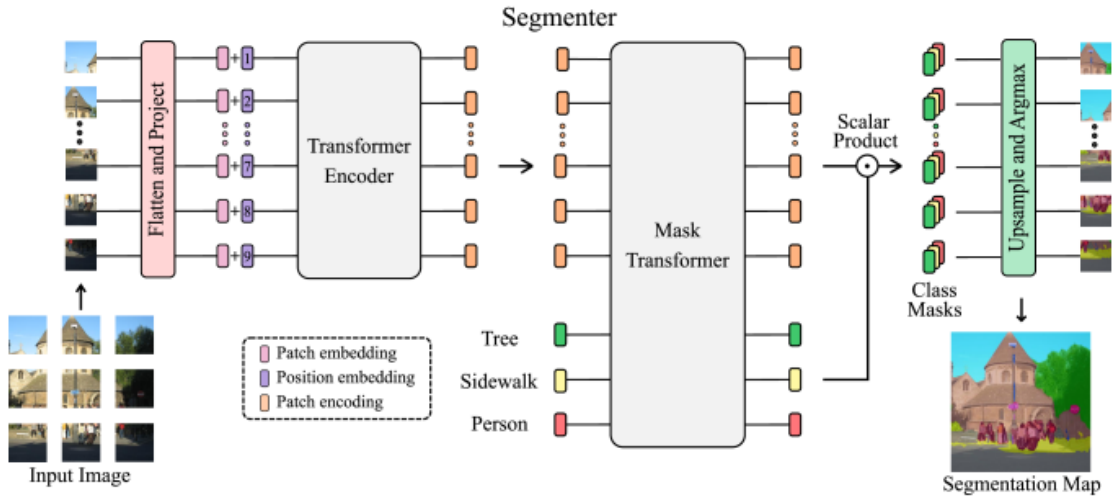


Figure 2.5: An overview of the Segmenter.[2]

2.5.4 Segmenter [2]

This model is also derived from ViT [21] and extended to Semantic Segmentation. By design, transformers can capture global interaction between elements of a scene and have no built-in inductive biases like CNNs can have, such as spatial or spectral inductive biases. With those biases, for instance, one can assume that a certain type of spatial structure is present in the data or take in account lower frequencies first in the learning process. The appealing features of Segmenter is that it does not use any convolutions, captures global image context by design and achieves competitive performance on standard image segmentation benchmarks.
In contrast to the method above in Segformer2.5.3, it uses patch embedding. They also propose a transformer-based decoder that generates class masks and can be extended to more general image segmentation tasks. The authors also created a family of models by changing its parameters size. It can be used in different tasks since it is a trade off between precision and time. Figure 2.5 shows an overview of the model.
It starts by dividing the input image in a sequence of patches and embedding them with a

position, based on the input image. These embedded patches go through L encoder layers where basically each layer is a MSA block followed by a point-wise MLP block of two layers with normalization layer before every block, with also a residual connection added after every block. The encoding patches are then processed by the decoder and turned into a segmentation map. The decoder learns how to map patch-level encodings to patch-level class scores which are then upsampled to pixel-level scores by bilinear interpolation. The authors proposal is the Mask Transformer in the decoder part where they introduce a set of learnable classes embeddings K, where K is the number of classes. Each class is randomly assigned to a single semantic class. They state that the decoder is a transformer encoder composed of multiple layers. The mask transformer generates K masks by calculating the scalar product between a normalized patch embeddings and class embeddings. Each mask sequence is reshaped to 2D and upsampled to the original size. Softmax is also applied to ensure the class scores forming the final segmentation map. The key idea is that by creating the mask along with the decoder the filters used are changed by the input since it's one of the features of the base transformers.

2.5.5 Vit-Adapter [3]

This fairly recent model with great results in most of the Semantic Segmentation data sets, state-of-the-art on Cityspaces and COCO-Stuff, ranked two on Pascal context and ranked seven on ADE20K, has caught our interest. It immediately changed the course of the work, since it is a more recent adaptation of the original ViT [21]. This model specifically takes advantage of inductive biases and introduces it to the original ViT and also takes in account local spatial context with the spatial prior module. It also adds that additional information from the spatial prior module into the ViT resulting in a Transformer with good global and local attention.

The Spatial Prior Module consists of three convolutions and a max-pooling layer, afterwards 3x3 convolutions of stride two are stacked in order to double the amount of channels and reduce the feature maps size. This type of module in the respective work [3], makes use of the convolutions to aid the transformers in apprehending local spatial contexts of images, therefore obtaining better results.

They also use the so called Spatial Feature Injector and Multi-Scale Feature Extractor, two feature interaction modules that bridge the module described before and the ViT. The Injector module adds the spatial context information from the Spatial Prior Module into the ViT. The Extractor merely extracts hierarchical features from the output of the block. After many interactions on the N blocks, different resolutions of features such as 1/8, 1/16 and 1/32 are obtained.

The ViT-Adapter has various sizes, depending on the parameter numbers such as the number of attention heads or depth, and lacks precision in the smaller ones. However, the large module can achieve state-of-the-art results and great results overall on the various benchmark data sets. Figure 2.6 shows the overall architecture of the ViT-Adapter.

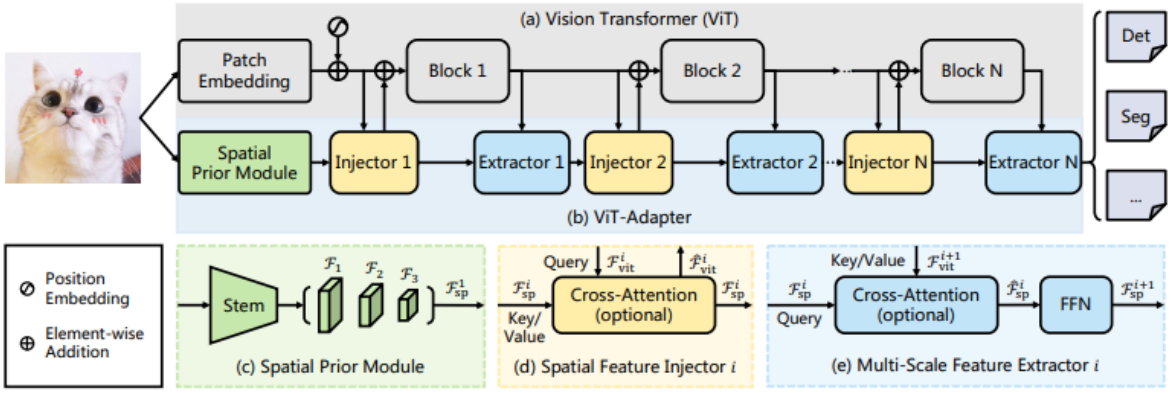


Figure 2.6: Overall architecture of ViT-Adapter. [3]

2.5.6 InternImage [4]

InternImage is a CNN-based foundation model that adapts some of the properties from Vision Transformers into the CNN architecture. It achieves good results on the most used benchmarking data sets and can be considered as state-of-the-art in most data sets. The fundamental part about this model is the convolutions utilized. The authors make use of deformable convolutions in order to adapt the kernel to gather data from larger training data and have a similar approach on one of the characteristics of a Vision Transformer, the ability to obtain dependencies on the whole data at very early stages. However, the authors also had in consideration the efficiency of this adaptation in terms of memory. Therefore it is a CNN model with less cost / efficiency, long-range dependence with flexible spatial pooling. The authors main purpose is to give into account that modified CNN models that can work in larger training data and increased parameters can also achieve the same results or even better than some modified Vision Transformers.

Figure 2.7 shows us the whole idea of the architecture from this model. The idea is to make the whole basic block similar to a Vision Transformer block. The input is always downsampled before each interaction in order to reduce its size by 4 and therefore create various sized featured maps. Each block is made of a normalization layer, followed by a a feed forward network and residual connections. The core operator is the deformable convolution and in this model, as you can see from the respective figure, a third version of a modified deformable convolution series is used. It basically makes use of a 3x3 kernel to be able of capturing long range dependencies. It is more efficient than a multi head attention layer commonly used in all types of transformers.

A model variant is defined with only 4 hyper parameters, (C_1, C', L_1, L_3), C_i means the number of channels in the i-th stage and L_i meaning the number of blocks in the i-th stage. The authors used the rules visible on the right bottom figure 2.7 to obtain the best variants of models of for the respective problem. Since the search space is too large, by using these rules it can be reduced heavily.

It has various sizes and configurations as we will see later in chapter 7, since this is one of the models that we used on this project.

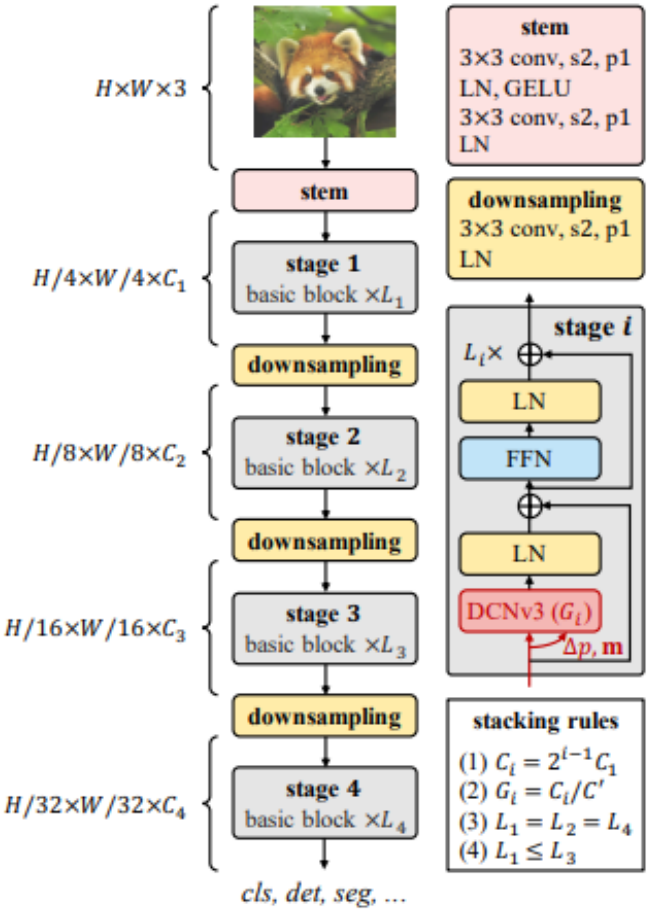


Figure 2.7: Overall architecture of InternImage. ”s2” and ”p1” mean stride = 2 and padding = 1[4]

2.5.7 Lawin [5]

The last model that we will talk about in this section is the Lawin. This model is also a modification of the original ViT, where it successfully brings a multi-scale representation onto the ViT. These multi-scale representations are very important in tasks such as semantic segmentation. The approach chosen by the authors to enhance the original ViT was to make use of larger attention windows. These larger attention windows allow the model to consider a larger area of context with little computation overhead. Figure 2.8 show how the large attention windows works in comparison with atrous convolution. We can see that the context area is different with different spatial sizes, respective to the different colors. The area is always relative to the size of the query, for example, the side of the orange square is 8 times the query size. However, if the area is too large, it can get too computational complex and therefore be too hard to calculate. To solve this issue the authors downsample the area to a specific size named P.

The whole structure of this model can be seen in the figure 2.9. It consists of an encoder and decoder, very general in this type of transformer architecture. The encoder is made of four blocks that extracts various feature maps with different levels of detail from an image. The last three of these maps are then send the into the decoder part. However, the feature map from the first block of the encoder is only used in the last part of the decoder, where it en-

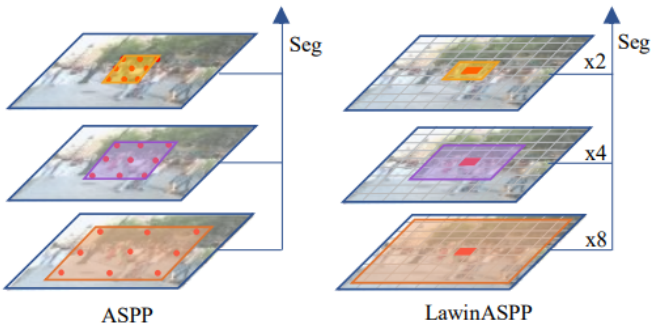


Figure 2.8: Highlighting the Differences Between LawinASPP and ASPP.[5]

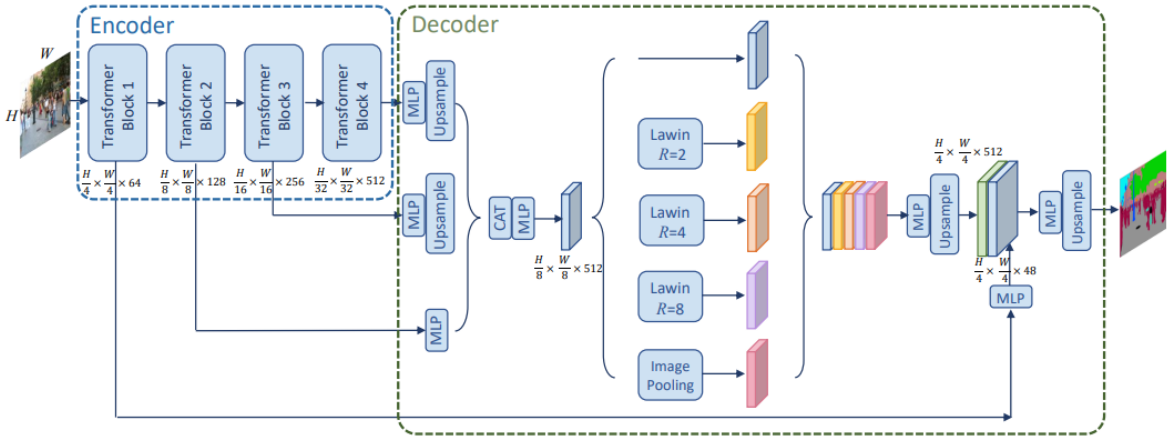


Figure 2.9: Overall architecture of Lawin Transformer.[5]

hances the resulted feature from the decoder. In order to capture multi-scale representations the authors used the LawinASPP module, located in the middle of the decoder. This module incorporates the spatial pyramid pooling architecture to complement the large window attention mechanism and therefore composed of five parallel branches: a shortcut connection, three large window attentions with different ratios, R, and an image pooling branch. The final segmentation feature map is achieved by concatenating all the resulting features and subsequently reducing their dimensionality.

The Lawin model has many different configurations from tiny to large, and the larger the configuration the better results it achieves. The authors also show that it can obtain the same results as other implementations with far less parameters and therefore making it more sustainable for some equipment. For this project we will only work with one type of configuration as we will see later on.

Chapter 3

Related Work

3.1 Introduction

To be able to understand and grasp the state-of-the-art methods used in this type of work many articles were read. This section contains some summaries of read articles that were related to the respective work.

3.2 Ludvision [6]

3.2.1 Introduction

This article makes use of technologies like Remote Sensing to detect the presence of an invasive specie in an ecosystem.

3.2.2 Motivation

Invasive species present in an ecosystem may cause catastrophic problems in different fields such as agriculture, fishing and navigation in terms of human activities or to the species involved in the ecosystem. It can also cause problems to the surrounding ecosystems. Having those problems in mind, finding a way to detect invasive species from Remote Sensing information was the motivation for this article.

3.2.3 Methods

The region selected for this study was *Reservatório da Barragem de Toulica (Zebreira, Portugal)*. The specie highlighted in this study is the Ludwigia peploides. It is from South America and it can invade rivers and lakes, although it also can be present in rice fields. It can grow fully or partially submerged under water. The figures 3.1 and 3.2 give us the details of the aircraft used to make a collection of data in the region and the metadata of the images, respectively.
Each instance of data acquired by the aircraft is composed of five different images from five different bands and one RGB image just for scene visualization. Figure 3.3 shows the 6 images taken at each instance of data collection.

3.2.4 Utilized Method

The authors assessed the spectral radiance on the acquired data, since the data would allow them to perform photo physiological measurements. By executing that, it was noticed a clear distinction between elements that could be used to classify the specie from the surroundings.

Aircraft	
Hover Accuracy Range	RTK enabled and functioning properly: Vertical: ± 0.1 m; Horizontal: ± 0.1 m RTK disabled: Vertical: ± 0.1 m (with vision positioning); ± 0.5 m (with GNSS positioning) Horizontal: ± 0.3 m (with vision positioning); ± 1.5 m (with GNSS positioning)
Camera	
Sensors	Six 1/2.9" CMOS, including one RGB sensor for visible light imaging and five monochrome sensors for multispectral imaging. Each Sensor: Effective pixels 2.08 MP (2.12 MP in total)
Filters	Blue (B): 450 nm ± 16 nm; Green (G): 560 nm ± 16 nm; Red (R): 650 nm ± 16 nm; Red edge (RE): 730 nm ± 16 nm; Near-infrared (NIR): 840 nm ± 26 nm
Max Image Size	1600×1300 (4:3.25)

Figure 3.1: Table of multi spectral specifications.[6]

Altitude	Time	Number of images
10 m	11h - 12:45h 15:30h - 17h	435
15 m	11h - 12h	365
40 m	10h - 12:45h	135
70 m	11h - 12h	27

Figure 3.2: Altitude, time and number of images collected.[6]

Completing this step, the authors classified this as a Semantic Segmentation problem and adapt the methods to this area. After a study on the state-of-the-art of Remote Sensing and semantic segmentation, was then decided to take an hybrid approach. The base model used was an an High Resolution Network (HRNet) in its HRNet + Object Context Recognition (OCR) implementation, but later on changes were made in order to fine-tune to obtain better results on their own data. One of them was the input, since the base model only takes RGB images and the data acquired in this project consists of five different bands. Other changes were made while testing for better results like increasing the *stride* value on the fusion models between stage 1 with the remaining stages and the use of dilated convolutions to enlarge the perceived FoV of the images, making them similar to satellite imagery. Figure 3.5 shows the architecture of the final model.

3.2.5 Results and Discussion

The metrics used to evaluate the final model were: Producer's accuracy, User's accuracy and IoU. The figure 3.6 shows the results obtained from the tests performed. Although the results are still high, they were a bit lower than the base model in lower altitudes, but since the authors intention is to upscale the model to satellite imagery it can be ignored.

Since the final model was a tweaked version from an outstanding model, it is noted that the time to develop the final model was reduced. The training times of this final model were cut in half when compared to the base model. It also outperforms the base model although it showing already demonstrated excellent results in correctly classifying the invasive specie.

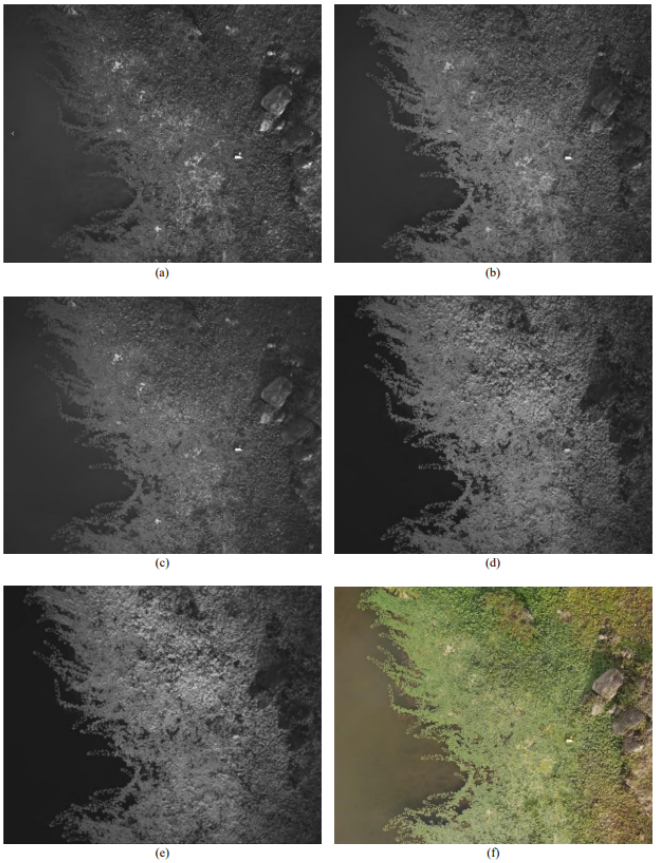


Figure 3.3: Example of images in each band. *Red band (a), green band(b), blue band (c), red edge band (d), near infra-red band (e), RGB (f)*[6]

3.3 Continuous Monitoring Of Forest Change Dynamics With Satellite Time Series

3.3.1 Introduction

There are innumerous change detection algorithms to track and quantify deforestation based on dense time series satellite data, but only a few capture regrowth on various types of forests. This article presents a new change detection algorithm that uses the flexibility of kernel density estimation.

3.3.2 Material and Methods

NDMI was used in this project (Normalized Difference Moisture Index) to create the reference phenology by the means of a kernel density estimation along the years. The figure 3.7 shows the reference phenology of one of the study sites as an example.
To better interpret this figure, we can look at the red area around 2012. Some of the points around the 260 days up to 345 days are most likely an anomaly, since it has a really low probability to be around those values. That was considered as a potential disturbance by the authors, but only if the values are higher or equal than 0.95. There are other arguments like the how many consecutive data points with the characteristics explained before. Looking at

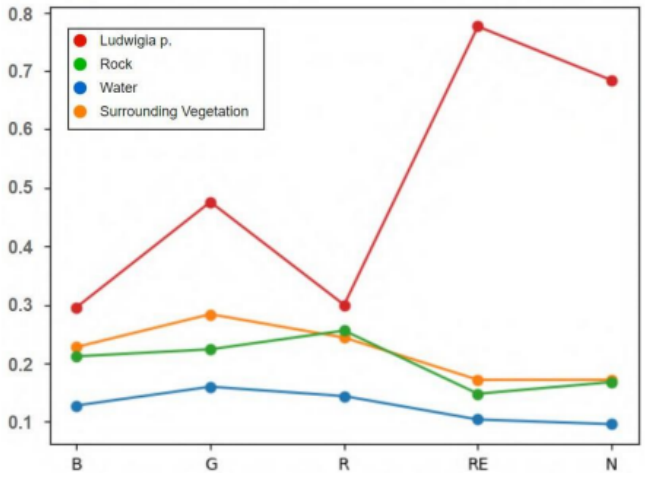


Figure 3.4: Reflectance values (in %) for each band in different image elements.[6]

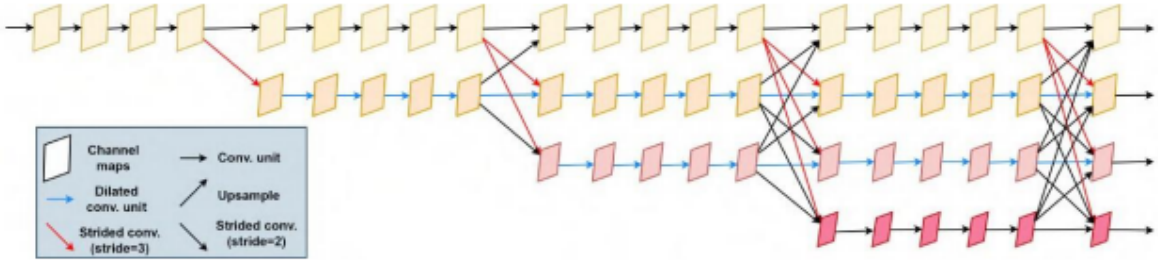


Figure 3.5: Architecture of the final model, based on the HRNet.[6]

the green area now, around 2015, we can see that around the same time of the year, those points are closer to the most frequent values. If this happens, the authors consider this as a forest regrowth and, once again, there is the argument of how many data points consecutively are needed to be detected as regrowth.

3.3.3 Study Sites

Three study sites were chosen for this study. Figure 3.8 shows the location of each site along with all the information of the area. The authors used a buffer and separate the disturbances prior to the year 2000 and after that. The buffer of the intact forest is represented as an area of 500 meters around villages and roads where there could be human activity leading to different results than those expected. The buffer for the regrowth is located in rivers since those areas are more dynamic than the rest. Although they made this distinction, the results in 3.3.4 have all the disturbances accounted for. The distinction between disturbances means that the authors, when they applied the algorithm, assumed that there was always a forest there, and since the data availability before the year 2000 was quite low, the results weren't accurate so it was made a clear distinction.

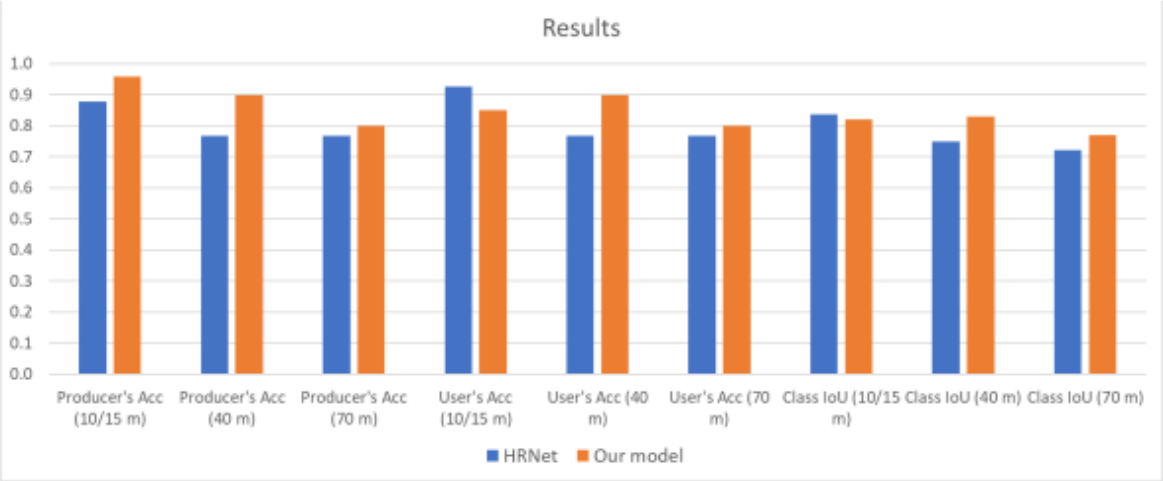


Figure 3.6: Performance comparison between HRNet and our model.[6]

3.3.4 Results and Discussion

The authors start this section by talking about their overall accuracy being over 90% in most of the sites, although there is an exception the Ivory Coast (33%). The explanation behind it is that in the 1990s or before the most of the area was disturbed, and then most of the recent disturbances were due to small scale selective logging. The other two areas were considered more stable since there is expected change due to agriculture practices in areas around rivers, although they had their fair share of events like mining in the site of Peru or the presence of a national Park in Tanzania. The high accuracy's in the site of Peru most likely are due to the low seasonal variability and the high amount of data availability from that site. Ivory Coast has a stronger seasonality and it's more heterogeneous in land cover than the study site of Peru which might have affected the producer's accuracy (PA) of the disturbance detection. The site in Tanzania has lowers disturbance and regrowth accuracies compared to the other sites due to its strong seasonality. Most of the disturbance errors were because of a late detection due to the lack of data in the site.

3.4 Individual tree segmentation and tree species classification in subtropical broad leaf forests

3.4.1 Introduction

Correct and precise classification of tree species is a must for many tasks such as inventorying, managing, and protecting forest resources. Usually, tree segmentation algorithms tend to lean into over segmentation, thus leading to incorrect classification, especially in more exotic ecosystems where same species have multiple crown peaks. This study, proposes a watershed-spectral-texture-controlled normalized cut (WST-Ncut) algorithm to better segment individual trees and furthermore correctly classify them.

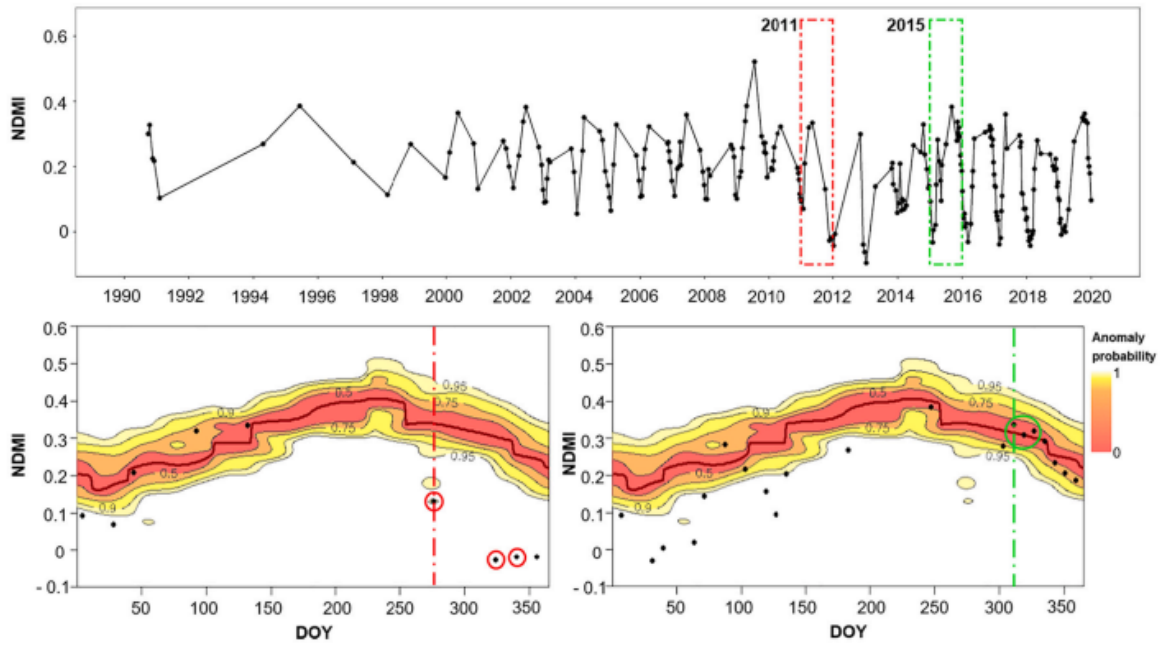


Figure 3.7: Reference phenology of the study site, Genteguela (Ivory Coast).[10]

3.4.2 Study Area and Data

The chosen region for this study was on Jalongshan Park, situated in Shenzhen City of southern China. This subtropical broad leaf's structure is single-layered with few small trees. Average temperature of this area is around 22.4 ◦C and a total annual precipitation of 1933.3 mm. The area covered is around 15.23 ha with the altitude varying from 47 m to 105.6 m. Consists of 18 different tree species: Terminalia neotaliala, Elaeocarpus grandifloras, Cinnamomum camphora, Delonix regia, Litchi chinensis, Bischofia javanica, Trachycarpus fortunei, Lagerstroemia speciosa, Melia azedarach, Acacia mangium, Acacia auriculiformis, Schima superba, Ficus concinna, Bauhinia purpurea, Mangifera indica, Artocarpus communis, Leucaena leucocephala, and Falcataria moluccana. They used three different types of data, LiDAR data, Hyper spectral data and Ultrahigh-resolution RGB imagery.

3.4.3 Methods

On recent works, CHM-based watershed algorithms have been successfully used for this type of work, in forests which tree species have unique crown picks. Based on that, the authors took this approach and improved by proposing this new type of watershed-spectral-texture-controlled normalized cut (WST-Ncut) algorithm. First, they took the patches that the CHM-based watershed segmentation created and made a graph out of it. The patches correspond to the vertices of the graph and the edges were the spectral and textural distances between the patches. If different patches have similar features that only means that it probably is the same tree specie therefore a cut criteria is determined. The next equations show us the way the authors calculated the distances between patches, leading to the normalized cut.

Site	Stratum	Sample size	Mapped area proportion
Peru	Disturbance <2000	118	0.314
Peru	Disturbance ≥2000	105	0.229
Peru	Intact forest (inside buffer)	100	0.038
Peru	Intact forest (outside buffer)	191	0.418
Peru	Regrowth	100	0.244
Peru	No regrowth (inside buffer)	109	0.537
Peru	No regrowth (outside buffer)	100	0.220
Ivory Coast	Disturbance <2000	327	0.637
Ivory Coast	Disturbance ≥2000	287	0.330
Ivory Coast	Intact forest (inside buffer)	100	0.003
Ivory Coast	Intact forest (outside buffer)	100	0.030
Ivory Coast	Regrowth	100	0.003
Ivory Coast	No regrowth (inside buffer)	180	0.475
Ivory Coast	No regrowth (outside buffer)	198	0.522
Tanzania	Disturbance <2000	100	0.085
Tanzania	Disturbance ≥2000	100	0.058
Tanzania	Intact forest (inside buffer)	100	0.073
Tanzania	Intact forest (outside buffer)	207	0.784
Tanzania	Regrowth	100	0.198
Tanzania	No regrowth (inside buffer)	102	0.548
Tanzania	No regrowth (outside buffer)	100	0.254

Figure 3.8: Sample size and mapped portion for each site and stratum.[10]

$$w_{i,j} = \frac{e^{-|S(i)-S(j)|^2}}{a} * \frac{e^{-|T(i)-T(j)|^2}}{b} \quad (3.1)$$

$$S(i) = \frac{\sum_{p=1}^{n} SI_p^i}{n} \quad (3.2)$$

$$T(i) = \frac{\sum_{p=1}^{n} TI_p^i}{m} \quad (3.3)$$

The first equation gives us the total weight of the edge between patch i and j. We can see that similar patches have a weight close to 0. Second and third equation gives us the composite spectral values and the composite textural values, respectively.

The next equation gives us the normalized cut criteria made by the authors. The detailed explanation of the mathematics can be found in [22].

Disturbance strata	OA (±CI)	UA (±CI)	PA (±CI)	Sample based area estimate (ha) (SE)	Map area (ha)	Δ Sample based area estimate and map area (ha)
Peru	95.1 (2.3)					
Disturbance <2000		93.2 (4.6)	99.4 (1.1)	60,626 (1540)	48,575	12,051
Disturbance ≥2000		92.4 (5.1)	98.5 (2.8)	33,211 (1042)	35,424	−2213
Intact forest		99.1 (1.7)	88.9 (5.0)	60,755 (1802)	70,592	−9837
Ivory Coast	93.2 (2.0)					
Disturbance <2000		92.7 (2.8)	99.4 (0.7)	6018 (95)	6453	−435
Disturbance ≥2000		93.7 (2.8)	99.8 (0.3)	3136 (48)	3339	−203
Intact forest		97.2 (3.0)	33.8 (7.0)	979 (103)	341	638
Tanzania	94.0 (2.2)					
Disturbance <2000		82.0 (7.6)	79.4 (18.1)	22,710 (2154)	21,978	732
Disturbance ≥2000		77.0 (8.3)	73.9 (13.6)	15,769 (2063)	15,126	643
Intact forest		96.4 (2.4)	97.0 (0.9)	221,097 (2927)	222,473	−1376
Regrowth strata	**OA (±CI)**	**UA (±CI)**	**PA (±CI)**	**Sample based area estimate (ha) (SE)**	**Map area (ha)**	**Δ Sample based area estimate and map area (ha)**
Peru	98.1 (1.8)					
No regrowth		98.1 (2.1)	99.4 (0.8)	39,287 (459)	48,575	12,051
Regrowth		98.0 (2.8)	93.6 (6.7)	11,730 (459)	12,425	−2213
Ivory Coast	96.0 (2.0)					
No regrowth		92.7 (2.8)	99.4 (0.7)	6018 (95)	6453	−435
Regrowth		93.7 (2.8)	99.8 (0.3)	3136 (48)	3339	−203
Tanzania	91.2 (3.1)					
No regrowth		82.0 (7.6)	79.4 (18.1)	22,710 (2154)	21,978	732
Regrowth		77.0 (8.3)	73.9 (13.6)	15,769 (2063)	15,126	643

Figure 3.9: Area weighted accuracy's (OA = Overall Accuracy; UA = User Accuracy; PA = Producer Accuracy).[10]

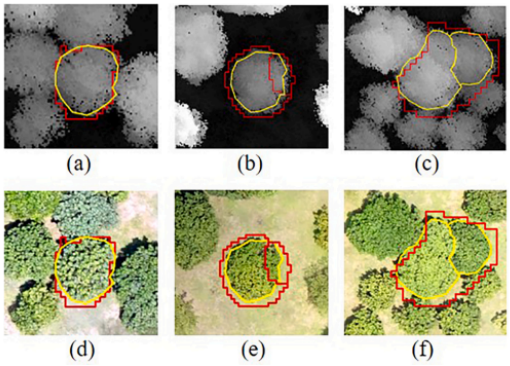


Figure 3.10: Examples of the segmentation results.[11]

$$Ncut(A,B) = \frac{cut(A,B)}{assoc(A,V)} + \frac{cut(A,B)}{assoc(B,V)} \quad (3.4)$$

3.4.4 Results and Discussion

The examples of the segmentation results can be shown in the figure 3.10. (a) is a True positive on CHM, where a real tree is segmented correctly, (b) is a False positive, where a real tree is segmented in small partitions, making it more trees than it should be, and lastly (c) is a False negative, where a real tree is not segmented right.

Figure 3.11 shows the amount of trees in the site and all the counted examples of each result. We can give more value on the True positives, since they are the most important result in this whole experience. It also shows different metrics like Recall, Precision, F-Score and their values are really high, all above 84%.

Site	Tree	True positive	False positive	False negative	Recall	Precision	F-score
Site 1	350	336	39	14	0.96	0.90	0.93
Site 2	257	245	43	12	0.95	0.85	0.90
Site 3	369	351	66	18	0.95	0.84	0.89
All sites	976	932	148	44	0.95	0.86	0.91

Figure 3.11: Accuracy assessment results of individual tree segmentation of WST-Ncut algorithm.[11]

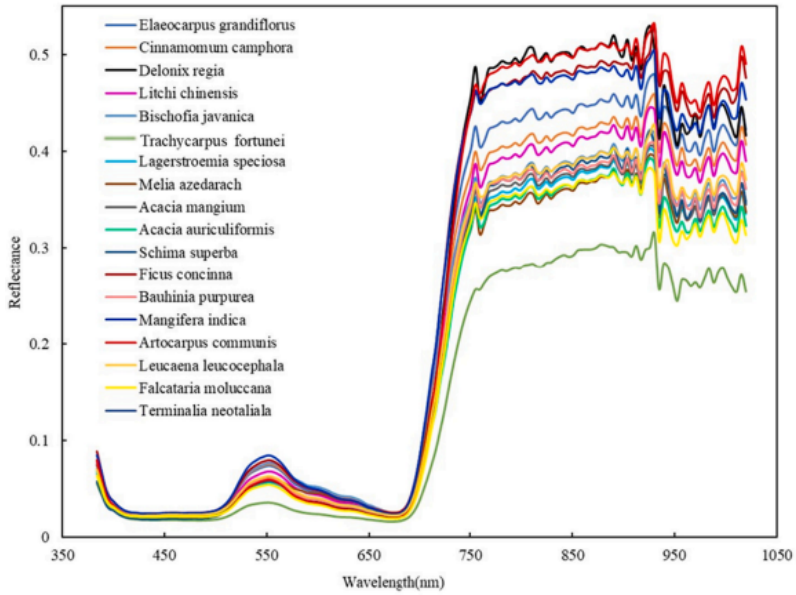


Figure 3.12: The mean and ± standard deviation of the reflectance of 18 tree species. Each specie has it's own value represented on the graph.[11]

3.4.5 Results from feature extraction

Having those patches segmented, the last step for the authors was to extract the features from them. Figure 3.12 shows the different reflectance values from each specie. From there we can clearly see that some of them are distinct from others, thus able to be classified. These next 2 figures, 3.13 and 3.14 present us the 6 most important features from the LiDAR data and textural features, respectively. The box plots from each specie is relatively different from the others, thus with the help from the three types of data it is easier to identify with higher accuracy the correct tree specie. Figure 3.15 consolidates that with the results of the three data types joined being higher than all of the ones before.

3.5 Early Detection of Bark Beetle

3.5.1 Introduction

Early and accurate detection of a bark beetle infestation is crucial to mitigate further damage on the attacked specie or species, develop proactive forest management activities and minimize economic losses, since it can result in a devastating impact on a forest ecosystem, biodiversity, forest structure and function. This paper provides a comprehensive review of past and current early detection of bark beetle attacks. Figure 3.16 gives an idea of the different stages of a bark beetle attack.

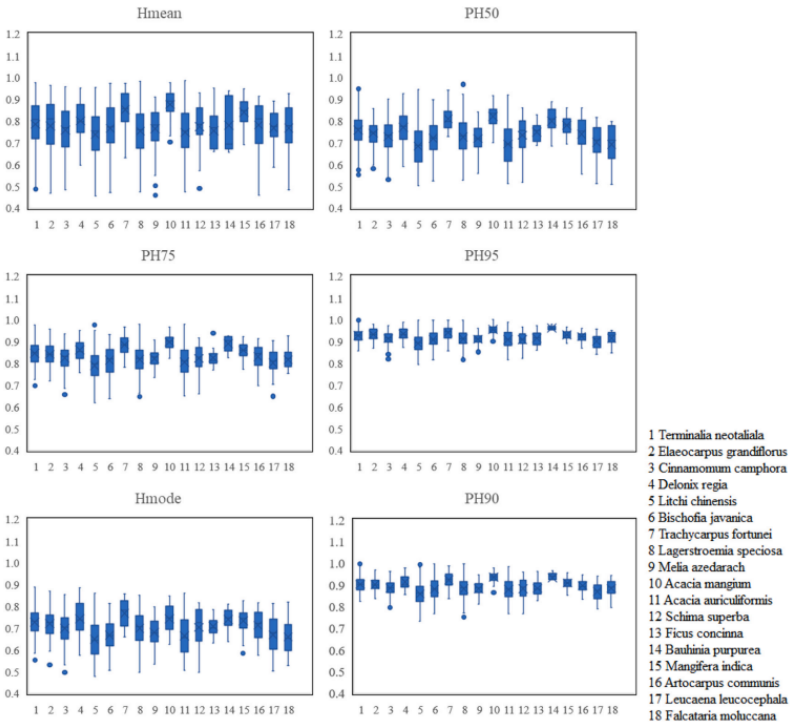


Figure 3.13: Box plots of the 6 most important LiDAR features.[11]

3.5.2 Reviews on Bark Beetle Damage Detection

A systematic study was done by the authors. Exploring every aspect involving the detection of a beetle attack, from the bark beetle species, host trees, study regions, to the data collection dates, different types of images and sensors and machine learning concepts. Figure 3.17 shows the aspects that were reviewed along the study.

3.5.3 Bark Beetle-Host Tree Interaction

Bark beetles are a group of insects that have received particular attention due to their ability to kill healthy trees over vast areas in many forest ecosystems. There is also the case where more than one specie of bark beetles are attacking the same tree. Figure 3.18 shows some of the species and hosts along with the symptoms of analyzed trees.
The beetles contributions to an ecosystem may vary. They can cycle nutrients or even make alterations to the soil and its chemistry. After mating successfully, female beetles excavate oviposition galleries and lay eggs along the walls. From there emerging larvae excavate larval galleries and feed on the nutritious phloem, the living tissue that transports the organic compounds made during photosynthesis. From shutting down both nutrients and water flow between tree canopy and roots, tree foliage changes and other symptoms occur.

3.5.4 Remote Sensing

Remote Sensing is used to gather information about a target with the use of one or more sensors that record capture electromagnetic radiation reflected or back-scattered from the

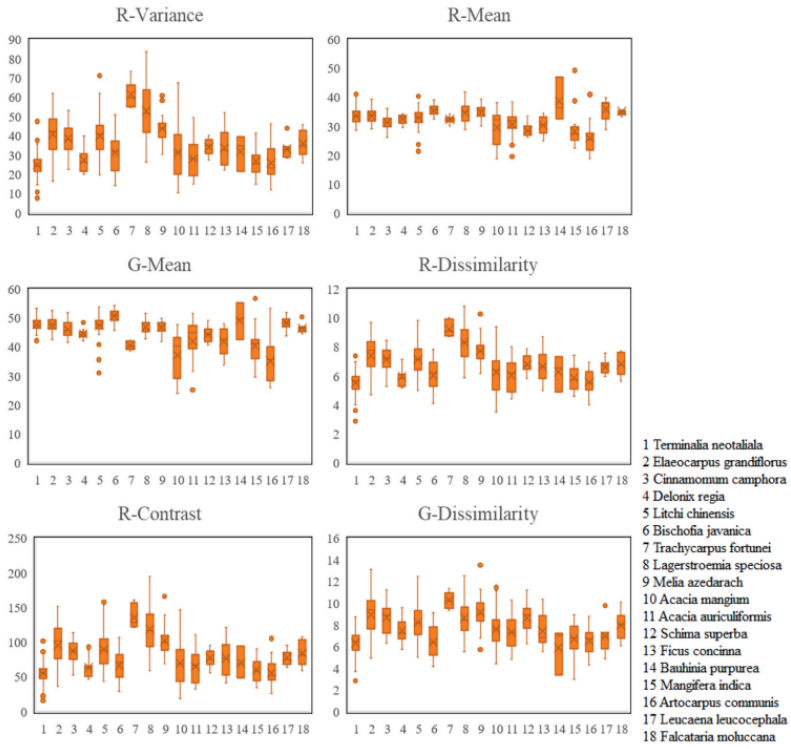


Figure 3.14: Box plots of the 6 most important textural features.[11]

Data	Feature	Overall Accuracy	Kappa
Hyperspectral	Spectral	81.6%	0.797
Ultrahigh-resolution RGB	Textural	72.8%	0.711
LiDAR	Structural	78.2%	0.730
Hyperspectral+LiDAR	Spectral+Structural	88.5%	0.865
Hyperspectral+Ultrahigh-resolution RGB	Spectral+Textural	86.1%	0.846
Ultrahigh-resolution RGB + LiDAR	Textural+Structural	82.6%	0.811
Hyperspectral+LiDAR+Ultrahigh-resolution RGB	Spectral+Structural+Textural	91.8%	0.910

Figure 3.15: The classification results of 18 tree species using seven feature sets.[11]

specific target. These recordings can eventually be used to identify a target due to they biophysical properties like its temperature.

To successful detect bark beetle-induced tree mortality across vast areas, it is required the use of Remote Sensing systems with suitable spatial, spectral and temporal resolutions. Since all data collection is subject to errors depending on the platforms, sensors and imaging conditions, there is the need to do a pre-processed treatment of the data, including enhancing vegetation signals like the presence of soil background effects, since it is one of the characteristics of the role played by these beetles. Figure 3.19 shows the Remote Sensing platforms using along with the meta data of the images, the attacking phase, the specie of beetle and the year that took place only for the satellite imagery.

A deep study was performed and analyses from the imagery of each source to achieve the best results. Since each satellite has their own specifications, the imagery taken are quite different. Due to that, they took most of their time explaining what bands and satellite were better for each task, like detecting the best source for the green, yellow, red and gray crowns, water stress signals. Some thermal data offered the plants' physiological and biochemical properties and can be used to detect diseases before it becomes apparent to the human eye.

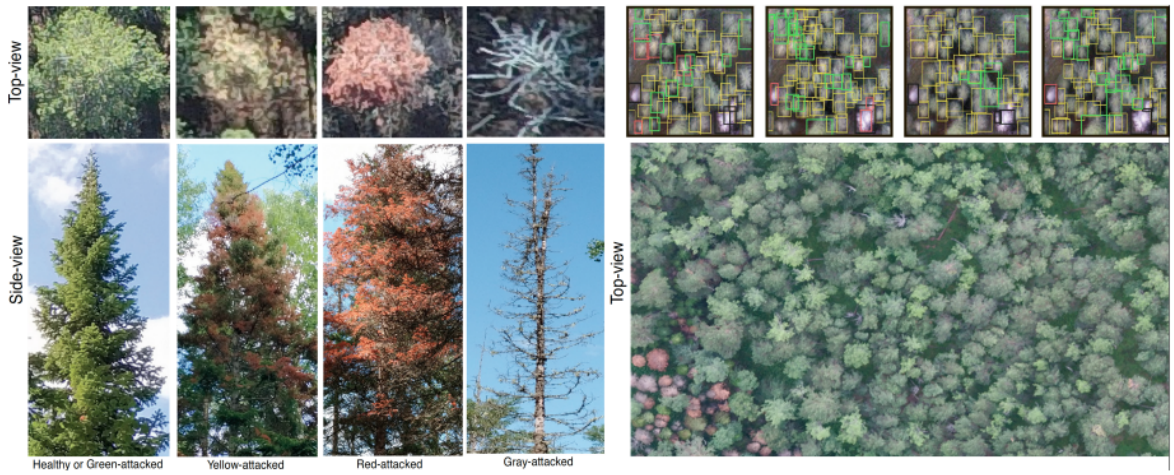


Figure 3.16: Detection of different stages of a bark beetle attack.[12]

Many aspects were observed, from temperature, photosynthesis activity, humidity levels, leaf water content, leaf pigment changes, etc. Figure 3.20 shows the most important spectral vegetation indices from all the data acquired to identify the bark beetle attacks.

3.5.5 Machine Learning

Machine learning is the primary component of artificial intelligent system to learn from data automatically. It also can be divided into 2 categories: Classical and Deep Learning based methods. Classical relies on hand-engineered features, eventually leading to a more time-consuming process than when Deep Learning methods are used since they have the ability to learn the relevant features based on the data.

3.5.6 Classical Methods

Different classical algorithms were used to detect these attacks. The corresponding figure 3.21 present us the different classical methods used along the years. Each method has its advantages and disadvantages.
From here we can observe that Random Forest (RF) is the most frequently used algorithm for this purpose. And as explained by the authors, might be because it is fast, scalable, robust to noise, avoids over fitting and it is simple to visualize and explain the input. However, as the numbers of trees increases, it can become a bit slower to execute. The pixels from belonging to non-forest areas are discarded and only forest stand species that were prone to bark beetle attacks were classified. Having other factors like wind fallen trees or small clear-cuts often led to incorrect classified forest stands. Other methods were used but since the results were not accurate, irrelevant, ambiguous or not usable for an actual solution of the problem it is not worth mentioning them.

3.5.7 Deep Learning Methods

Although Deep Learning approaches have been used and extensively developed to solve many computer vision problems, there is a lack of usage of DL methods on this type of detection.

Figure 3.17: An overview of the 3 key aspects, along with all the aspects involved behind the key aspects.[12]

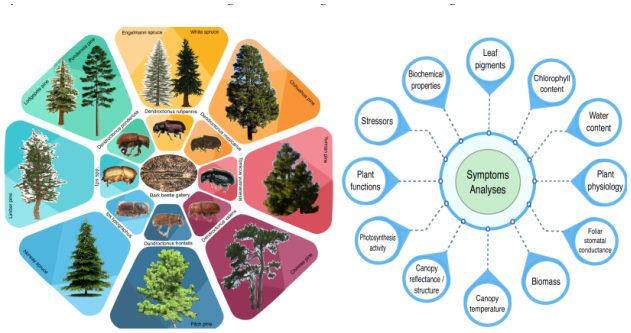


Figure 3.18: Examples of different bark beetle species and host trees (left) and symptoms of attacked trees (right).[12]

Three convolutional neural networks (CNN) were evaluated for their potential to classify infested trees from MS images. Two of them were created from scratch and one was a pretrained DenseNet-169 network applying transfer learning. Overall the 2 handcrafted networks had better results than the DenseNet, but all three were more accurate than the Random Forest classifiers that used raw spectral bands.

3.5.8 Metrics

The most common evaluation metrics used on the methods for detecting bark beetle induced tree mortality are: average precision, root mean square error, the confusion matrix, coefficient of determination and intersection-over-union.

	Year	Method	Attack Phase	Bark Beetle	Platform, Camera	Imaging Technique	Spectral Bands (ranges)	Spatial Resolution	Pre-Processing Software
Satellite / Focus of Our Review	2013	[81]	GAtt	ESBB	RapidEye, TerraSAR-X	MS	REye: RGB, Red-edge, NIR; TSAR: X-band	5/2m	GAMMA
	2014	[45]	GAtt, RAtt, GRAtt	ESBB, SBB	WorldView-2	MS	RGB, NIR1, NIR2, Red-edge, Coastal, Yellow, PAN	0.5/2cm	ENVI, ERDAS Imagine
	2017	[26]	GAtt	SBB	Landsat TM	MS	RGB, NIR, SWIR1, SWIR2	30m	ENVI, ViewSpec Pro
	2018	[75]	GAtt	MPB	WorldView-2	MS	RGB, NIR1, NIR2, Red-edge, Coastal, Yellow, PAN	1.85/0.5m	ENVI, ERDSA Imagine, ArcMap
	2019	[2]	GAtt, GRAtt	ESBB	Landsat-8 (OLI & TIRS)	MS	9 bands from OLI, thermal band from TIRS	100/30m	FLAASH
	2019	[3]	GAtt	ESBB	Sentinel-2, Landsat-8	MS	L8OLI: RGB, NIR, SWIR1-2; Sen2: RGB, Red-edge 1-3, NIR, NIR(a), SWIR 1-2	10/30m	MODTRAN4, ENVI, SEN2COR, ArcMap
	2020	[24]	GAtt, RAtt, GRAtt	ESBB	Sentinel-2	MS	RGB, Coastal aerosol, Red-edge 1-3, NIR, Narrow NIR, Water vapour, Cirrus, SWIR1-2	10cm	-
	2021	[7]	GAtt, RAtt	IPSEB	Sentinel-2	MS	RGB, B5-B7, B8a, B11-B12	20m	Google Earth Engine
	2021	[44]	GAtt	ESBB	Sentinel-1, Sentinel-2, Pleiades	MS	Sen1: C-band; Sen2: RGB, Coastal aerosol, B5-7, B8, B8a, B11-12	10/20/60m	SNAP
	2003	[27]	RAtt	MPB	Landsat TM	MS	B1-B5, B7	30m	-
	2005	[98]	RAtt	MPB	IKONOS	MS	RGB, NIR, PAN	4m	ImageStation Automatic Triangulation
	2006	[100]	RAtt	MPB	Landsat-7 ETM+	MS	B1-5, B7	30m	SPSS
	2006	[11]	RAtt	MPB	Landsat-5 TM, Landsat-7 ETM+	MS	B1-B7	30m	-
	2009	[40]	RAtt	MPB	QuickBird	MS	RGB, NIR	2.4m	ENVI
	2010	[13]	RAtt, GRAtt	MPB	GeoEye-1	MS	RGB, NIR, PAN	0.5m	ENVI
	2011	[70]	RAtt, GRAtt	MPB	QuickBird	MS	RGB, NIR	2.4-24m	METRO
	2013	[71]	RAtt	Unspecified	Landsat-5 TM, Landsat-7 ETM+	MS	RGB, B4-B5, B6, B7, PAN	30m	Exelis Visual Information Solutions
	2013	[29]	GRAtt	Unspecified	QuickBird, WorldView-2	MS	WV2 & QB: RGB, NIR, PAN	0.6m	GENIE
Late Detection	2014	[57]	RAtt, GRAtt	ESBB	Landsat-5 TM, SPOT-2, Landsat-7 ETM+, SPOT-4	MS	Landsats: B1-B5, B7; SPOTs: red, green, NIR	30/20m	-
	2014	[56]	RAtt, GRAtt	ESBB	Landsat-5 TM, SPOT-2, Landsat-7 ETM+, SPOT-4	MS	Landsats: B1-B5, B7; SPOTs: B1-B3; PAN	30/20/15m	Quantum GIS, ENVI
	2014	[62]	GRAtt	MPB	Landsat-5	MS	B4, B7	30m	-
	2015	[37]	GRAtt	SBB	Landsat TM	MS	B1-B7	30/120m	Ecosystem Disturbance Adaptive Processing
	2016	[67]	RAtt, GRAtt	MPB	Landsat-8	MS	B1-B7, B10-B11	30/100m	-
	2018	[91]	RAtt	DSBB	WorldView-2, WorldView-3, Landsat-8 (OLI)	MS	Landsat OLI: B1-B7, B9	1.5m	ENVI
	2018	[58]	RAtt, GRAtt	ESBB	RapidEye, MODIS	MS	REye: RGB, NIR, Red-edge; MODIS: NIR, red	5/250m	ATCOR-3
	2019	[90]	RAtt, GRAtt	ESBB	WorldView-2, Landsat-8 (OLI)	MS	WV2: RGB, Coastal, Yellow, Red-edge, NIR1-2; L8OLI: RGB, Coastal, NIR, SWIR 1-2	0.46/1.84/15/30m	ENVI
	2020	[106]	RAtt, GRAtt	RTB	Sentinel-2, Gaofen-2	MS	RGB, NIR, VEG1-VEG4, SWIR1-2, PAN	1/4/10/20/30/60m	Agisoft, eCognition
	2021	[104]	RAtt, GRAtt	MPB, SBB	Landsat (ARD Tier 1)	MS	Red, NIR, SWIR1 & SWIR2	30m	Google Earth Engine
	2022	[93]	GRAtt	ESBB	Landsat (time-series)	N/A	N/A	30m	QGIS
	2022	[72]	RAtt	SPB	Landsat-5 TM, Landsat-8 (OLI)	MS	L5TM: RGB, NIR, SWIR, TIR, MIR; L8OLI: RGB, Coastal, NIR, B6-7, PAN, Cirrus	15/30/120m	Google Earth Engine, Google Earth Pro
	2021	[65]	RAtt, GRAtt	ESBB	WorldView-3	MS	RGB, NIR1, NIR2, Red-edge, Yellow, Coastal, PAN	2m/0.5m	eCognition, ENVI, PCI Geomatics
	2021	[51]	RAtt, GRAtt	Unspecified	WorldView-2, Worldview-3	MS	RGB, PAN	0.3m	-

Figure 3.19: Comparison of Remote Sensing systems to detect bark beetle attacks (only for satellite imagery).[12]

3.6 Mapping forest in the Swiss Alps treeline ecotone with explainable deep learning

3.6.1 Introduction

Forest mapping leans to important tasks such as monitoring, inventory and get information about different tree species in the area. This work proposes an explainable deep learning method for forest mapping, that uses prior knowledge of others forests to explain its decisions.

3.6.2 Data

The study region chosen is a composed area of the Valais canton and a subregion of the Vaud canton around Southern Switzerland. They use SwissImage-10 aerial images from 2017, consisting of red, blue and green bands. They divided the study area in square tiles of 1 km x 1 km, and only took the squares around the altitude range of 1500-2500m to use as data. Figure 3.22 shows the way the training validation and test sets were divided.
They also used the *NFI Vegetation Height Model* as an additional target data source to further expand the explainable model. Similar to the square tiles of the study area mentioned above, they also divided the data with a similar idea. Figure 3.23 illustrates the process.

3.6.3 Segmentation Task

They approach this as a Semantic Segmentation problem. Every pixel is mapped to one forest type, Open Forest (OP), Closed Forest (CF), Shrub Forest (SF) or Non-Forest (NF). They

Analysis	Spectral Vegetation Indices	Formula	Used in
MS	Disease Water Stress Index (DWSI)	$\frac{NIR+Green}{SWIR1+Red}$	[3, 44]
	Normalized Difference Red-edge 2 (NDRE-2)	$\frac{NIR-RedEdge1}{NIR+RedEdge1}$	[3]
	Normalized Difference Red-edge 3 (NDRE-3)	$\frac{NIR-RedEdge2}{NIR+RedEdge2}$	[3]
	Simple Ratio/Short Wave Infrared (SR-SWIR)	$\frac{SWIR1}{SWIR2}$	[3]
	Normalized Difference Water Index (NDWI or NDVI 0.819/1.649)	$\frac{NIR-SWIR}{NIR+SWIR}$	[3, 7, 44]
	Leaf Water Content Index (LWCI)	$\frac{\log(1-(NIR-SWIR))}{-\log(1-(NIR-SWIR))}$	[3]
	Ratio Drought Index (RDI)	$\frac{SWIR2}{NIR}$	[3, 44]
	Enhanced Normalized Difference Vegetation Index (ENDVI)	$\frac{(NIR+Green)-(2\times Blue)}{(NIR+Green)+(2\times Blue)}$	[73]
	Tasseled Cap - Wetness (TCW)	$0.1509 \times Blue + 0.1973 \times Green + 0.3279 \times Red + 0.3406 \times NIR - 0.7112 \times SWIR1 - 0.4572 \times SWIR2$	[7]
	Normalized Distance Red SWIR (NDRS)	$\frac{DRS-DRS^{r}_{min}}{DRS^{r}_{max}+DRS^{r}_{min}}$	[44]
	Distance Red SWIR (DRS) †	$\sqrt{(Red)^2+(SWIR)^2}$	[44]
HS	Moisture Stress Index (MSI)	$\frac{R1.599}{R0.819}$	[59]
	Carotinoid Reflectance Index 1 (CRI-1)	$\frac{(1/R0.51)}{(1/R0.55)}$	[59]
	Anthocyanin Reflectance Index 2 (ARI-2)	$R0.8 \times \left[\frac{2}{R0.55}-\frac{1}{R0.7}\right]$	[59]
	Green Normalized Difference Vegetation Index (GNDVI)	$\frac{(R0.78-R0.55)}{(R0.78+R0.55)}$	[59]
	Normalized Difference Soil Moisture Index (NSMI)	$\frac{(R1.8-R2.119)}{(R1.8+R2.119)}$	[59]
	Normalized Water Index 2 (NWI-2)	$\frac{(R0.97-R0.85)}{(R0.97+R0.85)}$	[59]
	Green Optimized Soil Adjusted Vegetation Index (GOSAVI)	$\frac{(1+0.16)(R0.8-R0.67)}{(R0.8+R0.67+0.16)}$	[59]
	Normalized Pigment Chlorophyll Ratio Index (NPCI)	$\frac{(R0.68-R0.43)}{(R0.68+R0.43)}$	[59]
	Transformed Chlorophyll Absorption in Reflectance Index (TCARI)	$3\times[(R0.7-R0.67)-0.2\times(R0.7-R0.55)\frac{R0.7}{R0.67}]$	[59]
	Difference Index 1 (DI-1)	$R0.8-R0.55$	[59]
	Red-Green Index (RGI)	$\frac{Red}{Green}$	[26]
	Water Index (WI)	$\frac{R0.9}{R0.97}$	[26]
	Normalized Difference Photochemical Reflectance Index (PRI)	$\frac{R0.531-R0.570}{R0.531+R0.570}$	[8]
	Laboratory Index 3 (LI-3)	$1.5\times[(R0.724-R0.716)-(R0.716-R0.709)]-2\times[(R0.549-R0.541)-(R0.49-R0.483)]+0.5\times[(R0.541-R0.534)-(R0.52-R0.512)]$	[39]
	HySpex Index 1 (HI-1)	$\frac{R0.75252-R0.73067}{2}$	[39]
	Red-edge inflection point (REIP)	$0.75+0.035\left(\frac{\frac{R0.665+R0.783}{2}-R0.705}{R0.74+R0.705}\right)$	[8]
	ANCB index CR (0.65–0.72) ‡	$\frac{1}{2}\sum_{i=1}^{n-1}(\lambda_{i+1}-\lambda_i)(R_{CR(\lambda_{i+1})}+R_{CR(\lambda_i)})\frac{1}{BD_{0.68}}$	[8]
MS & HS	Normalized difference Vegetation Index (NDVI or NDVI - NIR or NDVI 0.8/0.65)	$\frac{NIR-Red}{NIR+Red}$	[7, 26, 52, 73]
	Greenness Index (GI)	$\frac{R0.554}{R0.677}$	[8, 52]

†† ranges of the DRS values for all spruce pixels in the image

‡ the area under the continuum removed (CR) reflectance (0.65-0.72 μm), normalized by the CR band depth (0.68 μm) ($RCR(\lambda_{i+1})$ and RCR (λ_i) are values of CR reflectance at the i and $i+1$ bands, λ_i and λ_{i+1} are wavelengths of the j and $j+1$ bands, n is number of bands)

2021	[65]	RAtt, GRAtt	ESBB	WorldView-3	MS	RGB, NIR1, NIR2, Red-edge, Yellow, Coastal, PAN	2m/0.5m	eCognition, ENVI, PCI Geomatics
2021	[51]	RAtt, GRAtt	Unspecified	WorldView-2, Worldview-3	MS	RGB, PAN	0.3m	-

Figure 3.20: Most effective SVIs for early detection of bark beetle attacks. The multi spectral and hyper spectral analyses are denoted by MS and HS, and RX denotes reflectance at wavelength X nm.[12]

also divided this task into a sub-task, as if there is a Forest (F) or NF. Figure 3.24 shows an example of the mentioned classes.

The authors' baseline model would be a *black-box CNN* that outputs a forest segmentation map. This map has size $\boldsymbol{C \times W \times H}$, for $\boldsymbol{C}$ classes and images of width $\boldsymbol{W}$ and height $\boldsymbol{H}$.

3.6.4 Explainable deep learning model architecture

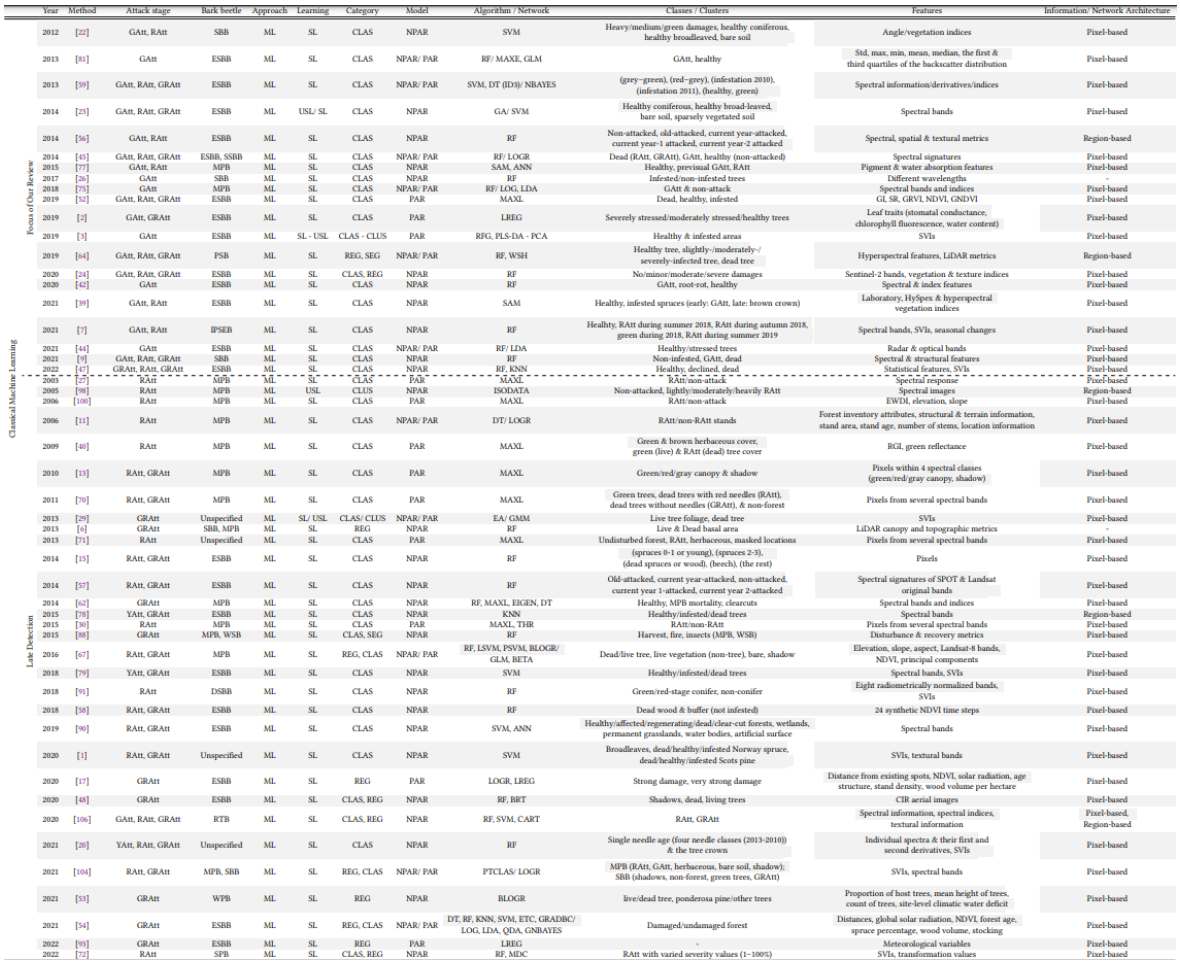
To build the explainable model they used the semantic bottleneck approach, with some modifications. First they added a *concepts extractor* that predicts intermediate concepts that are relevant to forest mapping and second a *rule module* that translates these intermediate predictions into class probabilities using simple heuristics. This basically creates a basis for the prediction of the classes via intermediate products and combinatory rules. They also used a *correction module* to further emphasize the model to learn and adjust the rule-based outputs by feeding it more features from the base feature extractor. Figure 3.25 is a flowchart of the model.

The *intermediate concept predictions* quantifies concepts that are relevant to the task. Tree height and Tree Canopy Density were chosen as intermediate concepts.

The rule model combines the intermediate concepts estimated in the semantic bottleneck to obtain forest class probabilities. They hard code probabilities to enforce the rules from the model. Table 3.26 shows one type of rule hard-coded.

3.6.5 Forest Mapping

The overall accuracies are shown in the figures 3.27 and 3.28. We can see that very good results were obtained. All of them being over 90% and consistent from training to test on figure 3.27. Figure 3.28 show us that the binary classification results are similar in all models, however there is a noticeable degradation of accuracy in the forest type classification. The



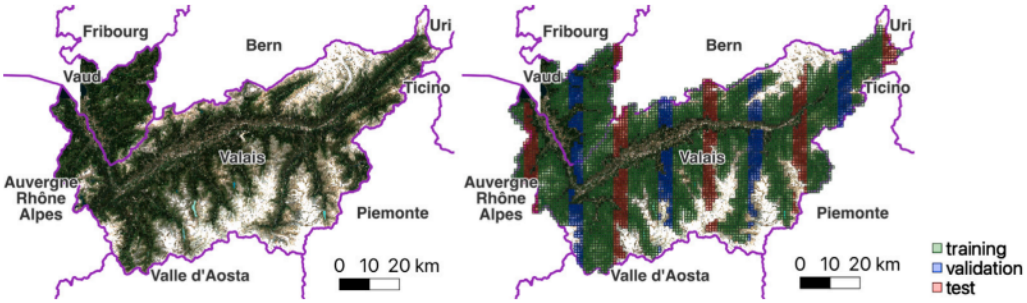
	Year	Method	Attack stage	Bark beetle	Approach	Learning	Category	Model	Algorithm / Network	Classes / Clusters	Features	Information/ Network Architecture
Focus of Our Review	2012	[22]	GAtt, RAtt	SBB	ML	SL	CLAS	NPAR	SVM	Heavy/medium/green damages, healthy coniferous, healthy broadleaved, bare soil	Angle/vegetation indices	Pixel-based
	2013	[81]	GAtt	ESBB	ML	SL	CLAS	NPAR/ PAR	RF/ MAXE, GLM	GAtt, healthy	Std, max, min, mean, median, the first & third quartiles of the backscatter distribution	Pixel-based
	2013	[59]	GAtt, RAtt, GRAtt	ESBB	ML	SL	CLAS	NPAR/ PAR	SVM, DT (ID3)/ NBAYES	(grey–green), (red–grey), (infestation 2010), (infestation 2011), (healthy, green)	Spectral information/derivatives/indices	Pixel-based
	2014	[23]	GAtt, RAtt, GRAtt	ESBB	ML	USL/ SL	CLAS	NPAR	GA/ SVM	Healthy coniferous, healthy broad-leaved, bare soil, sparsely vegetated soil	Spectral bands	Pixel-based
	2014	[56]	GAtt, RAtt	ESBB	ML	SL	CLAS	NPAR	RF	Non-attacked, old-attacked, current year-attacked, current year-1 attacked, current year-2 attacked	Spectral, spatial & textural metrics	Region-based
	2014	[45]	GAtt, RAtt, GRAtt	ESBB, SSBB	ML	SL	CLAS	NPAR/ PAR	RF/ LOGR	Dead (RAtt, GRAtt), GAtt, healthy (non-attacked)	Spectral signatures	Pixel-based
	2015	[77]	GAtt, RAtt	MPB	ML	SL	CLAS	NPAR	SAM, ANN	Healthy, previsual GAtt, RAtt	Pigment & water absorption features	Pixel-based
	2017	[26]	GAtt	SBB	ML	SL	CLAS	NPAR	RF	Infested/non-infested trees	Different wavelengths	-
	2018	[75]	GAtt	MPB	ML	SL	CLAS	NPAR/ PAR	RF/ LOG, LDA	GAtt & non-attack	Spectral bands and indices	Pixel-based
	2019	[52]	GAtt, RAtt, GRAtt	ESBB	ML	SL	CLAS	PAR	MAXL	Dead, healthy, infested	GI, SR, GRVI, NDVI, GNDVI	Pixel-based
	2019	[2]	GAtt, GRAtt	ESBB	ML	SL	CLAS	PAR	LREG	Severely stressed/moderately stressed/healthy trees	Leaf traits (stomatal conductance, chlorophyll fluorescence, water content)	Pixel-based
	2019	[3]	GAtt	ESBB	ML	SL - USL	CLAS - CLUS	PAR	RFG, PLS-DA - PCA	Healthy & infested areas	SVIs	Pixel-based
	2019	[64]	GAtt, RAtt, GRAtt	PSB	ML	SL	REG, SEG	NPAR/ PAR	RF, WSH	Healthy tree, slightly-/moderately-/severely-infected tree, dead tree	Hyperspectral features, LiDAR metrics	Region-based
	2020	[24]	GAtt, RAtt, GRAtt	ESBB	ML	SL	CLAS, REG	NPAR	RF	No/minor/moderate/severe damages	Sentinel-2 bands, vegetation & texture indices	Pixel-based
	2020	[42]	GAtt	ESBB	ML	SL	CLAS	NPAR	RF	GAtt, root-rot, healthy	Spectral & index features	Pixel-based
	2021	[39]	GAtt, RAtt	ESBB	ML	SL	CLAS	NPAR	SAM	Healthy, infested spruces (early: GAtt, late: brown crown)	Laboratory, HySpex & hyperspectral vegetation indices	Pixel-based
	2021	[7]	GAtt, RAtt	IPSEB	ML	SL	CLAS	NPAR	RF	Healthy, RAtt during summer 2018, RAtt during autumn 2018, green during 2018, RAtt during summer 2019	Spectral bands, SVIs, seasonal changes	Pixel-based
Classical Machine Learning	2021	[44]	GAtt	ESBB	ML	SL	CLAS	NPAR/ PAR	RF/ LDA	Healthy/stressed trees	Radar & optical bands	Pixel-based
	2021	[9]	GAtt, RAtt, GRAtt	SBB	ML	SL	CLAS	NPAR	RF	Non-infested, GAtt, dead	Spectral & structural features	Pixel-based
	2022	[47]	GRAtt, RAtt, GRAtt	ESBB	ML	SL	CLAS	NPAR	RF, KNN	Healthy, declined, dead	Statistical features, SVIs	Pixel-based
	2003	[27]	RAtt	MPB	ML	SL	CLAS	PAR	MAXL	RAtt/non-attack	Spectral response	Pixel-based
	2005	[98]	RAtt	MPB	ML	USL	CLUS	NPAR	ISODATA	Non-attacked, lightly/moderately/heavily RAtt	Spectral images	Region-based
	2006	[100]	RAtt	MPB	ML	SL	CLAS	PAR	MAXL	RAtt/non-attack	EWDI, elevation, slope	Pixel-based
	2006	[11]	RAtt	MPB	ML	SL	CLAS	NPAR/ PAR	DT/ LOGR	RAtt/non-RAtt stands	Forest inventory attributes, structural & terrain information, stand area, stand age, number of stems, location information	Pixel-based
	2009	[40]	RAtt	MPB	ML	SL	CLAS	PAR	MAXL	Green & brown herbaceous cover, green (live) & RAtt (dead) tree cover	RGI, green reflectance	Pixel-based
	2010	[13]	RAtt, GRAtt	MPB	ML	SL	CLAS	PAR	MAXL	Green/red/gray canopy & shadow	Pixels within 4 spectral classes (green/red/gray canopy, shadow)	Pixel-based
	2011	[70]	RAtt, GRAtt	MPB	ML	SL	CLAS	PAR	MAXL	Green trees, dead trees with red needles (RAtt), dead trees without needles (GRAtt), & non-forest	Pixels from several spectral bands	Pixel-based
	2013	[29]	GRAtt	Unspecified	ML	SL/ USL	CLAS/ CLUS	NPAR/ PAR	EA/ GMM	Live tree foliage, dead tree	SVIs	Pixel-based
	2013	[6]	GRAtt	SBB, MPB	ML	SL	REG	NPAR	RF	Live & Dead basal area	LiDAR canopy and topographic metrics	-
	2013	[71]	RAtt	Unspecified	ML	SL	CLAS	PAR	MAXL	Undisturbed forest, RAtt, herbaceous, masked locations	Pixels from several spectral bands	Pixel-based
	2014	[15]	RAtt, GRAtt	ESBB	ML	SL	CLAS	NPAR	RF	(spruces 0-1 or young), (spruces 2-3), (dead spruces or wood), (beech), (the rest)	Pixels	Pixel-based
	2014	[57]	RAtt, GRAtt	ESBB	ML	SL	CLAS	NPAR	RF	Old-attacked, current year-attacked, non-attacked, current year 1-attacked, current year 2-attacked	Spectral signatures of SPOT & Landsat original bands	Pixel-based
	2014	[62]	GRAtt	MPB	ML	SL	CLAS	NPAR	RF, MAXL, EIGEN, DT	Healthy, MPB mortality, clearcuts	Spectral bands and indices	Pixel-based
	2015	[78]	YAtt, GRAtt	ESBB	ML	SL	CLAS	NPAR	KNN	Healthy/infested/dead trees	Spectral bands	Region-based
	2015	[30]	RAtt	MPB	ML	SL	CLAS	PAR	MAXL, THR	RAtt/non-RAtt	Pixels from several spectral bands	Pixel-based
	2015	[88]	GRAtt	MPB, WSB	ML	SL	CLAS, SEG	NPAR	RF	Harvest, fire, insects (MPB, WSB)	Disturbance & recovery metrics	Pixel-based
Late Detection	2016	[67]	RAtt, GRAtt	MPB	ML	SL	REG, CLAS	NPAR/ PAR	RF, LSVM, PSVM, BLOGR/ GLM, BETA	Dead/live tree, live vegetation (non-tree), bare, shadow	Elevation, slope, aspect, Landsat-8 bands, NDVI, principal components	Pixel-based
	2018	[79]	YAtt, GRAtt	ESBB	ML	SL	CLAS	NPAR	SVM	Healthy/infested/dead trees	Spectral bands, SVIs	Pixel-based
	2018	[91]	RAtt	DSBB	ML	SL	CLAS	NPAR	RF	Green/red-stage conifer, non-conifer	Eight radiometrically normalized bands, SVIs	Pixel-based
	2018	[58]	RAtt, GRAtt	ESBB	ML	SL	CLAS	NPAR	RF	Dead wood & buffer (not infested)	24 synthetic NDVI time steps	Pixel-based
	2019	[90]	RAtt, GRAtt	ESBB	ML	SL	CLAS	NPAR	SVM, ANN	Healthy/affected/regenerating/dead/clear-cut forests, wetlands, permanent grasslands, water bodies, artificial surface	Spectral bands	Pixel-based
	2020	[1]	RAtt, GRAtt	Unspecified	ML	SL	CLAS	NPAR	SVM	Broadleaves, dead/healthy/infested Norway spruce, dead/healthy/infested Scots pine	SVIs, textural bands	Pixel-based
	2020	[17]	GRAtt	ESBB	ML	SL	REG	PAR	LOGR, LREG	Strong damage, very strong damage	Distance from existing spots, NDVI, solar radiation, age structure, stand density, wood volume per hectare	Pixel-based
	2020	[48]	GRAtt	ESBB	ML	SL	CLAS, REG	NPAR	RF, BRT	Shadows, dead, living trees	CIR aerial images	Pixel-based
	2020	[106]	GAtt, RAtt, GRAtt	RTB	ML	SL	CLAS, REG	NPAR	RF, SVM, CART	RAtt, GRAtt	Spectral information, spectral indices, textural information	Pixel-based, Region-based
	2021	[20]	YAtt, RAtt, GRAtt	Unspecified	ML	SL	CLAS	NPAR	RF	Single needle age (four needle classes (2013-2010)) & the tree crown	Individual spectra & their first and second derivatives, SVIs	Pixel-based
	2021	[104]	RAtt, GRAtt	MPB, SBB	ML	SL	REG, CLAS	NPAR/ PAR	PTCLAS/ LOGR	MPB (RAtt, GAtt, herbaceous, bare soil, shadow); SBB (shadows, non-forest, green trees, GRAtt)	SVIs, spectral bands	Pixel-based
	2021	[53]	GRAtt	WPB	ML	SL	REG	NPAR	BLOGR	live/dead tree, ponderosa pine/other trees	Proportion of host trees, mean height of trees, count of trees, site-level climatic water deficit	Pixel-based
	2021	[54]	GRAtt	ESBB	ML	SL	REG, CLAS	NPAR/ PAR	DT, RF, KNN, SVM, ETC, GRADBC/ LOG, LDA, QDA, GNBAYES	Damaged/undamaged forest	Distances, global solar radiation, NDVI, forest age, spruce percentage, wood volume, stocking	Pixel-based
	2022	[93]	GRAtt	ESBB	ML	SL	REG	PAR	LREG	-	Meteorological variables	Pixel-based
	2022	[72]	RAtt	SPB	ML	SL	CLAS, REG	NPAR	RF, MDC	RAtt with varied severity values (1–100%)	SVIs, transformation values	Pixel-based

Figure 3.21: Classical machine learning methods for detecting bark beetle attacks.[12]

authors explain this degradation by every class contributing equal values to the metrics and the minority of cases actually overcoming the influence of the others.
Figure 3.29 shows some visual extracts from different areas. Both models are quite good in the sharp forest boundaries (Extract A), but we can clearly see a difference between models on the most diffuse forest expansion. They explain it further by stating that the lack of large shadows can make it difficult to predict the type of forest. Nevertheless, they found the results on extract A and B on the BB model to be satisfying, and stated that the SB model reflect better the definitions underlying the target classes.

3.6.6 Intermediate concepts estimations

Due to the wide range of heights, the scarcity of intermediate height values, as well as the use of monocular input images rather than stereoscopic images, tree height prediction is expected to be a challenging task. Figure 3.30 shows a visual example of the quantitative results of the tree height prediction task. As the authors noticed, it underestimates the height of the tree, where more than half is predicted lower than the true value. The metrics and values used for tree height, respectively, were R^2 and $RMSE$, with 0.75 and 4.2.
With tree canopy density prediction, distribution of values shows a similar trend, with low values and high values being more frequent inside and outside the forest, respectively, and



(a) Sentinel-2 imagery over the study area (b) Image tiles repartition

Figure 3.22: Study area, displaying the general appearance with Sentinel-2 imagery (a) and the partition into training, validation, and test sets (b).[13]

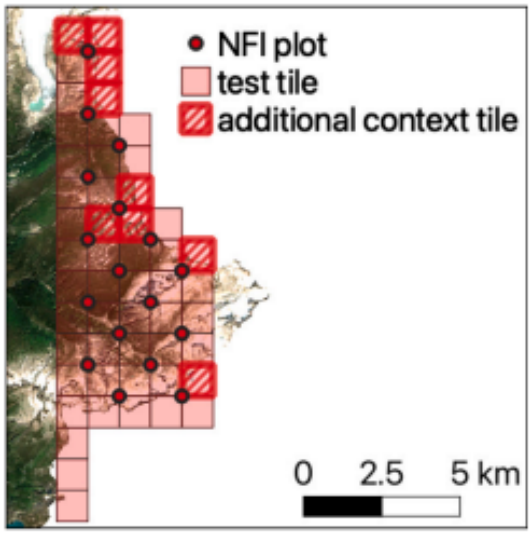


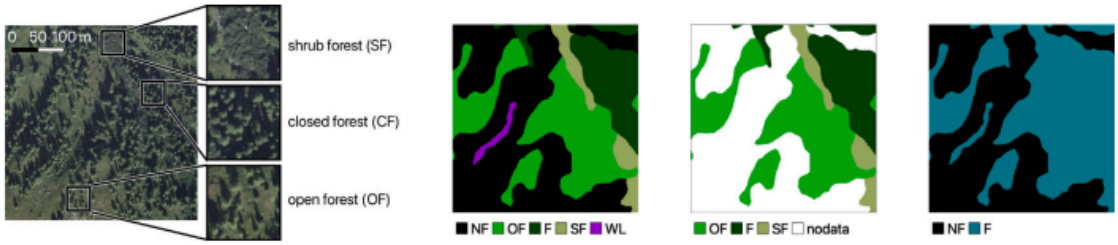
Figure 3.23: NFI plots and SwissImage tiles spatial arrangement (easternmost part of the test set).[13]

the complete opposite on the forest borders. The prediction scores from the metrics used in the tree height task were 0.89 and 12.1, respectively, indicating good estimation of the tree canopy density. Figure 3.31 shows the segmentation results by applying rules on tree canopy density.

3.7 Hyperspectral and LiDAR data for the prediction via machine learning of tree species, volume and biomass.

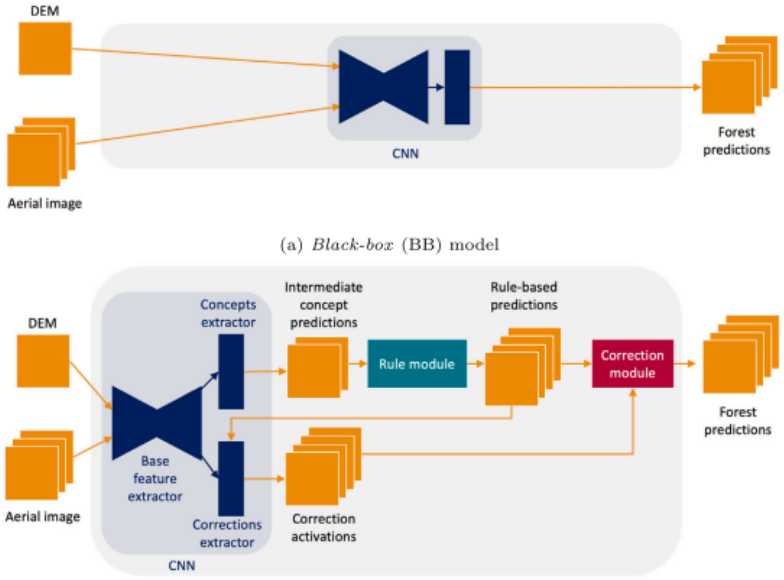
3.7.1 Introduction

Using available LiDAR and Hyperspectral data from the Autonomous Province of Trento, this works intends to lay the foundations for correctly identifying the forest types and the representation of single trees withing the specific private forest. These studies are quite important in the context of managing, inventorying and monitoring scenarios.



(a) Aerial image (b) SwissTLM3D annotations (c) *Forest type* target (d) *Forest p/a* target

Figure 3.24: Aerial image and associated targets extracted from SwissTLM3D annotations. p/a: presence/absence. [13]



(a) *Black-box* (BB) model

(b) Semantic bottleneck (SB) model

Figure 3.25: Flowchart of the forest mapping methods.[13]

3.7.2 Study Area

The study areas covered are located in the southern part of PAT in Italy around Calliano, Flogaria, Rovereto and Volano. Figure 3.32 shows the study area along with the geographic coordinates of the locations.

The data set used to train and test was composed of approximately 900 samples. The train and test proportions were 65/35. Figure 3.33 shows a table of the samples used on each specie.

TH (m)	TCD (%)		
	[0, 20)	[20, 60)	≥ 60
[0, 1)	NF	NF	NF
[1, 3)	NF	NF	NF/SF
≥ 3	NF	OF	CF/SF

Figure 3.26: Rules enforced by the rule module.[13]

	Training	Validation	Test
BB	**0.93**	**0.93**	**0.94**
SB	**0.93**	0.92	0.93

Figure 3.27: Overall accuracies for both methods.[13]

	Forest type	Forest presence/absence
BB	**0.89**	**0.93**
SB	0.80	**0.93**
SBrules^{-}	0.83	**0.93**
SBcorr^{+}	0.81	**0.93**

Figure 3.28: Overall accuracies for both methods in forest type and the forest presence/absence.[13]

3.7.3 Spectral feature selection

122 were used, bands distributed over all the visible and near infrared spectrum. However a set of suboptimal bands were selected by the selection operation based on the Sequential Forward Floating Selection (SFFS) along with the Jeffries-Matusita distance.

3.7.4 Delineation of tree crowns

The authors from this work used the Canopy Height Model (CHM) to identify the forest units, similar to what the authors in section 3.4 did. They opted to take the *itcLiDAR* approach inside the *itcSegment* library withing the R environment.

3.7.5 Classifiers

Two non-parametric supervised machine learning classifiers were used, K-Nearest Neighborhood and a Support Vector Machine (SVM). The first one defines the average values of each class defined by the input features n-dimensional space. The pixel values are attributed to the respective class which the Euclidean distance is minimum from the original value. The second classifier is used having in mind that the starting features can be transformed into a higher-dimensional space, which the classes are linearly separable.

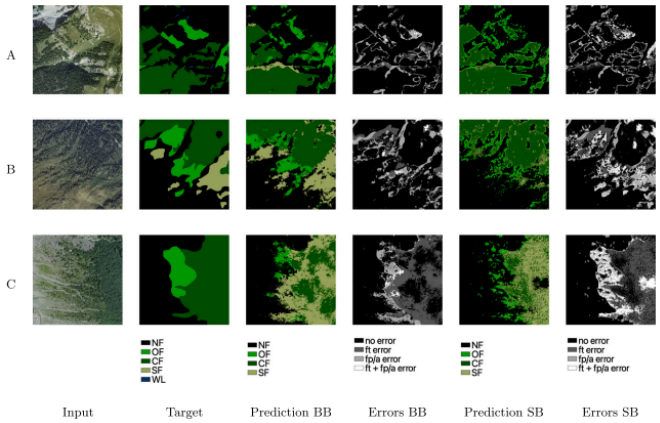


Figure 3.29: Visual extracts of the forest mapping results.[13]

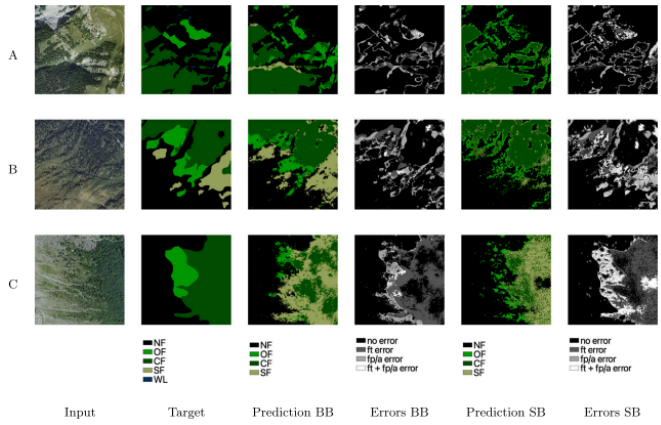


Figure 3.30: Visual extracts of the tree height predictions (in m) on the test set (SB)[13]

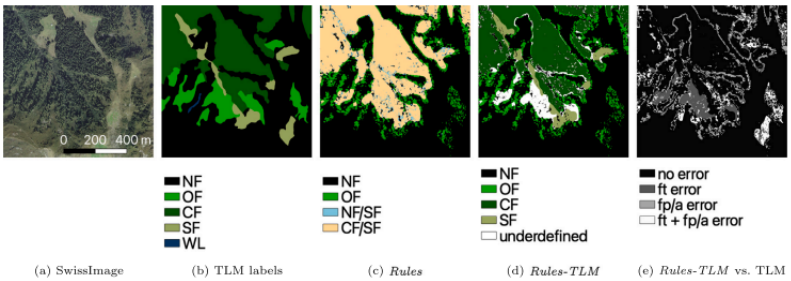


Figure 3.31: Segmentation results by applying rules on intermediate concept targets. Rules: map obtained by applying the rules to the intermediate targets. Rules-TLM: Rules map with TLM disambiguation.[13]

3.7.6 Biomass and volume estimation models

The biomass estimate was calculated from the work done in section 3.7.4. Two equations were used to identify Above Ground Biomass (AGB) and volume, and these equations are based on the diameter of the tree to estimate the volume of the tree. The authors point the fact that trees tend to grow taller to reach the maximum amount of light as possible, but when they can't reach higher they develop their diameter. From this simple fact, different individuals from the same specie might have different characteristics. Although this is a fact, the relation between these 2 characteristics remains constant over time as the stem of the tree must continue to grow to maintain the stability and water supply. The equations 3.5 and 3.6 are the most suitable equations identified for the respective characteristics, considered by the authors, taking account the different relationships between angiosperms, gymnosperms and the Scrinzi tariff volume equation.

$$AGB_{predicted} = (0.016 + a_G) * (H * CD)^{(2.013+B_G)} * exp(\frac{0.204^2}{2}) \quad (3.5)$$

$$V = b_0 + b_1 G + b_2 GP_s + b_3 GP_s I_t + b_4 GP_s B_d \quad (3.6)$$

In 3.5, α_G and β_G are functional group-dependent parameters. ***H*** and ***CD*** are Height and

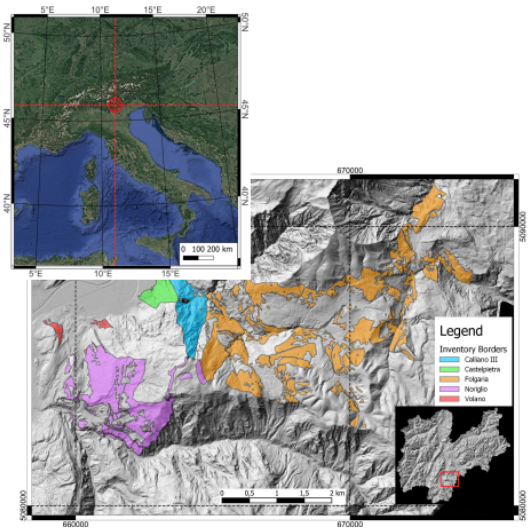


Figure 3.32: Study area and the inventories of the private forests.[14]

Species name	Training	Test
Norway spruce (*Picea abies* Karst.)	115	61
Silver fir (*Abies alba* Mill.)	53	24
Larch (*Larix decidua* Mill.)	78	43
Scots pine (*Pinus sylvestris L.*)	60	30
Black pine (*Pinus nigra* Arn.)	31	15
Beech (*Fagus sylvatica* L.)	71	39
Downy oak (*Quercus pubescens* Willd.)	17	8
Hop-hornbeam (*Ostrya carpinifolia* Scop.)	28	18
Manna ash (*Fraxinus ornus* L.)	14	9
European ash (*Fraxinus excelsior* L.)	12	7
Sycamore (*Acer pseudoplatanus* L.)	16	12
Birch (*Betula pendula* Roth)	10	10
Turkey oak (*Quercus cerris* L.)	3	2
Other conifers	12	5
Other broadleaves	60	28

Figure 3.33: List of species and number for samples used for the training and test phases.[14]

Crown Diameter, respectfully. In 3.6, $\boldsymbol{G}$ is the basal area per hectare, b_0, b_1, b_2, b_3, b_4 are the regression coefficients, P_s is the stereo metric potential index of the species, I_t is the tariff index and B_d is the barycentric dimensional index.

3.7.7 Determination of the species

Using the two classifiers mentioned in section 3.7.5 and a third one, SVM with aggregation of species that basically is a modification of the previous SVM mentioned, new images were obtained, figure 3.34 displays them. The A panel represents the classification comparison of the K-Nearest neighborhood, panel B represents the SVM mentioned in 3.7.5 and panel C represents the modified aggregated species SVM.

Figure 3.35 show us the table of accuracy and precision from all models on predicting the respective specie. As stated, the best results were the ones related to the SVM all species. Since the most important for managing is the accuracy of the prediction, we can see that 96% for black spine and 73% for beech were good results.

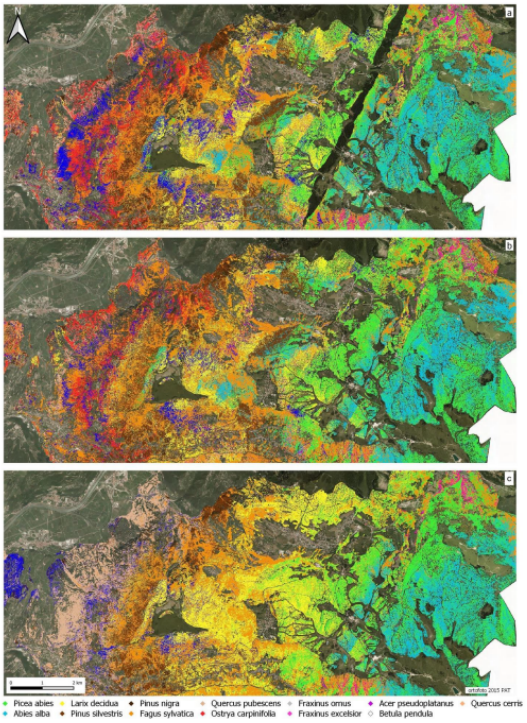


Figure 3.34: Hyperspectral classification comparison.[14]

3.7.8 Estimation of volume and above ground biomass

Since volume and above ground biomass are different for every specie, and even among the same species, each individual tree crown that was delineated in the initial steps was labeled with a specie using the classification step results. 11 areas were selected, distant form one another with different compositions. The results are shown in figure 3.36. We can see some of them being pretty close to reality while others being very far, for example areas 6,7,8,10,11 where the volume predicted isn't even close to the real values.

Species	K-means			SVM all species			SVM subset of species		
	Acc.	Prec.	F	Acc.	Prec.	F	Acc.	Prec.	F
P. abies	78%	64%	56%	85%	75%	70%	69%	53%	52%
A. alba	85%	44%	55%	90%	55%	62%	80%	39%	41%
L. decidua	76%	44%	50%	85%	59%	65%	65%	33%	42%
P. silvestris	78%	37%	42%	82%	42%	49%	73%	32%	32%
P. nigra	94%	63%	69%	96%	91%	71%	91%	60%	60%
F. sylvatica	74%	64%	49%	73%	51%	48%	65%	54%	47%
Q. pubescens	92%	17%	12%	95%	100%	44%	77%	0%	-
O. carpinifolia	85%	63%	45%	88%	73%	44%	70%	36%	32%
F. ornus	92%	-	-	93%	100%	12%	76%	7%	9%
F. excelsior	92%	0%	-	93%	25%	13%	89%	17%	12%
A. pseudoplatanus	90%	83%	34%	92%	83%	36%	88%	67%	31%
B. pendula	94%	-	-	94%	100%	33%	82%	15%	17%
Q. cerris	97%	40%	44%	99%	100%	67%	86%	11%	16%
Other conifers	99%	100%	67%	99%	100%	75%	97%	100%	33%
Other broadleaves	83%	68%	51%	84%	70%	55%	78%	61%	49%

Figure 3.35: Accuracy metrics from test data sets.[14]

	Area	1	2	3	4	5	6	7	8	9	10	11
V (m^3)	Ob	3.00	3.33	57.87	21.32	15.41	12.22	14.53	25.09	18.42	6.74	20.51
	Pr	1.02	1.32	63.92	13.14	12.22	3.19	5.67	12.40	11.57	0.34	0.23
AGB (Mg)	Ob	1.84	1.99	26.95	11.82	7.49	6.10	7.24	12.76	9.21	4.04	12.25
	Pr	1.07	1.45	37.01	8.99	9.19	2.68	5.39	9.53	7.43	5.48	3.68

Figure 3.36: Biomass and volume results from observed ground truth values (Ob) and estimated predicted values (Pr) for volume (V) and above ground biomass (AGB).[14]

3.8 Conclusions

Overall, the methods used in these papers are relatively old as some of them do manual extraction of features like 3.4, although 3.5 refers to the use of newer technologies like convolutional neural networks. Nonetheless it's a great base to give an idea of the technologies used such as sensors, bands and think about recent models and methods that can be implemented in this thesis.

Chapter 4

Data sets

Since the beginning of this work, one of the main concerns was the data necessary to correctly identify the different species involved. It's an easy task to differentiate dense forests from not so dense forests, since the trees are grouped together, the pixel values don't differ much from each other, and the same species tend to be around the same area so there's not a problem in those types of forests.
The problem is when we have a really low resolution image and the trees are dispersed from each other, making the surroundings of the tree actually being taken into account. These surroundings can take the form of bushes, grass, sand, rocks, and other environment participants. Having this in mind, there was the need to study what results each type of resolution would give us and the expectation for the work at hand.

4.1 Cartography of North / Center of Portugal

This data set is a complete record of satellite imagery of north and center of Portugal from 2021[15]. The resolution of these images is 25cm and they only have 3 bands: Red; Blue; Green. Although this data set only has these 3 bands, its a good data set because of its high resolution. Some images were taken from this source in order to clearly annotate the area and have knowledge about the respective area.

4.2 Planet Research and Education Program

Planet[16] is a platform with the objective to image all landscape making it easily available for business, researchers, government and other parties. With their special program, we

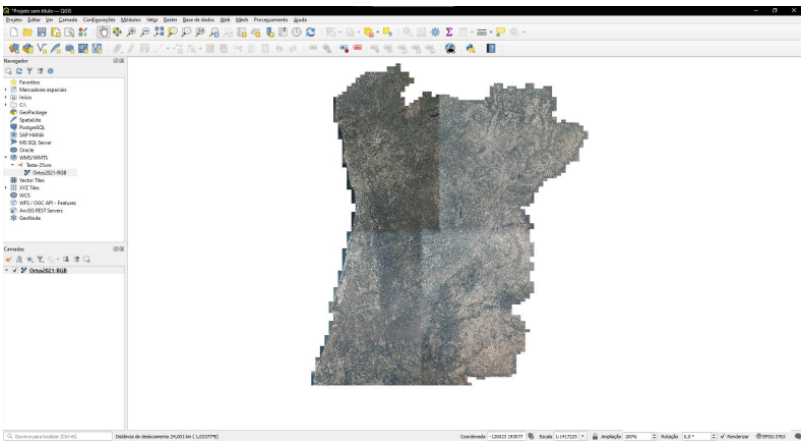


Figure 4.1: Example of the carthography using when using the QGES software. [15]

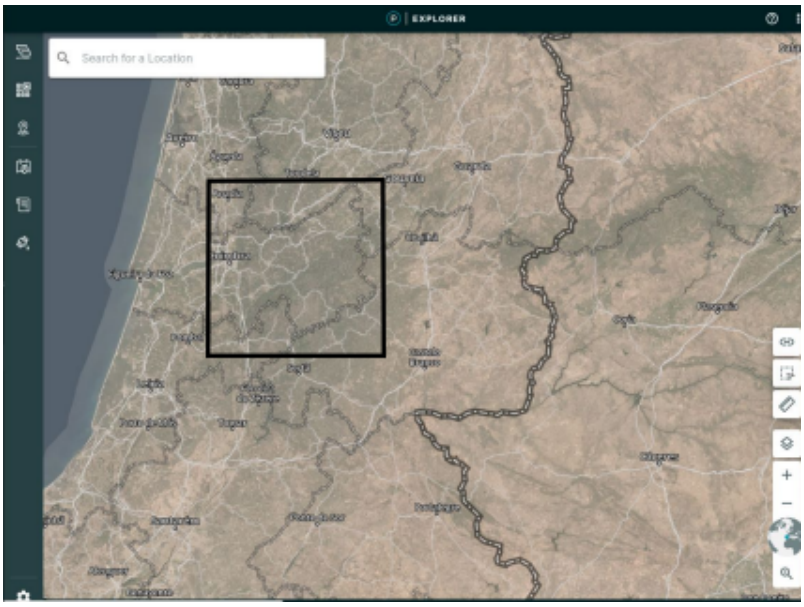


Figure 4.2: Example of the Planet.com plaftorm [16]

could obtain access data up to five thousand squared kilometers, with 3m resolution and eight different bands: Red, Green, Blue, Coastal Blue, Green II, Yellow, Red-Edged and Near Infra-Red. This data set becomes really useful since it has many different bands. As seen before, bands can help distinguish between different species since the reflectance values may differ between species. Most images obtained in this website are around June 2021 since it was around the same time as the previous dataset was obtained and therefore we can compare the differences between the datasets.

4.3 Copernicus Land Monitoring

The Copernicus platform[17] provides different types of data relevant to this work, such as forest types, tree cover density or dominant leaf type, all in ten meter resolution. Since it contains information about different types of forests, the data from this platform was mainly used to help in annotations, and although the resolutions differ from the other data sets and the data is not that recent, it was still helpful.

4.4 Pre-processing and Methods

Since the most efficient way to download this type of imagery is to download them as a really big image like 9000 x 9000 pixels, there is the need to treat this data so that the model can handle it without crashing. One of the ways is to partition the image into smaller samples. Since most of these data sets have a reasonable resolution themselves, from 25cm to 3m per pixel, we can get a partition image with a considerable size. Transformers already do the slicing of the image into tiny parts, since it is the most efficient way for them to work due to their attention mechanism, but giving an image with 7000x5000 or more pixels can be too much for them to handle.

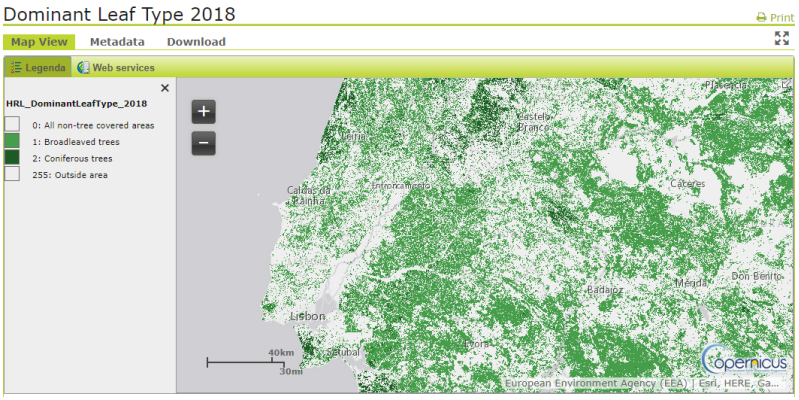


Figure 4.3: Example of the information from the Copernicus Land Monitoring Service. [17]

Having this in mind, a python script was built in order to partition the images into smaller images. There is no size that needs to be respected, since the python script was built to slice an image into the number of parts we chose, but the images never go above the 1024x1024 or less than 240x240. This size was chosen because most of the models that were tested on the benchmark data sets never train on images smaller or larger than those sizes, respectively. It's also easier to annotate these smaller images than to annotate a big image in one go.

4.5 Annotations

The annotations on the imagery were done with the help of Hasty.ai [23], a platform that allows efficient data building with many different tools for annotating data. Since there is no prior area of expertise on the species needed to be identified, one of the guides for the annotations is the information from the Copernicos Land Monitoring data 4.3. It has a 10 meter resolution but since most of the areas for annotation contain a lot of trees of the same specie grouped in the same area, we can take this 10 meter resolution imagery and use it to our advantage.

Figure 4.3 shows an example of the dominant leaf type information around the center of Portugal. It shows where the broadleaved trees and the coniferous trees are present on the map. With this information it's safe to assume that taking imagery from an area where there is only one type of tree, the annotations wont be wrong thus not misleading the model in the classification process or the person annotation the data. Since we want certain species and for starters the most predominant ones, knowing the type of leaf it has can lead to certain areas of observation for acquiring the data. One of the species in mind is the *Eucalyptus* and therefore we must search areas with the respective type of leaf.

Chapter 5

Experiments and Ideas

Having read the articles presented in section 3, the next step was to perform some experiments and gather some thoughts about the data that was needed to be acquired and the type of segmentation we would want to perform. This was a very early experiment and at the time the idea of using satellite imagery was not feasible, therefore some data from another project mentioned in section 3.2 was provided to us and some experiments were made in order to set course for the main project, the classification of different plant species. After performing some research on different types of Segmentation, the main idea was to perform Instance Segmentation on the limited data that was given to us, since the spacial resolution was good enough for the models to detect each and every single tree on an image and apply a mask to it and therefore identify the respective species and cluster them in the image. To test this idea, there was the need to acquire some data and find a model that was able to perform this type of Segmentation.

5.1 Data set

The data acquired belonged to a former student from the same institution, that also worked on a similar project [6]. It must be noticed that at the time we did not have the data sets stated in section 4.
The specifications of the images are similar to the ones in the section 3.2, where they were captured with a drone and had 5 different bands, although for this test only RGB imagery was used. This data set contains around 98 RGB images and was partitioned in 80/10/10 for training, validation and test respectively.
The annotations were also done by hand with the help of a tool described earlier in 4.5, the Hasty.ai platform [23]. The annotation of the trees were distributed into three classes, the normal tree that could be identified with a darker green color, the wide tree that had the same green color but in lighter tone and the dead tree that was completely white. Some trees appear more than others, for instance, the dead trees rarely appears in the images but it is still relevant to annotate them. An example of the annotations can be found in figure 5.1.

5.2 Model

The model used was an implementationof the MaskR-CNN in TensorFlow 2.0[18] model from 2017[24]. This model has been used in other projects that also use satellite and aerial imagery achieving good results[24]. Although there are recent models that can also perform Instance Segmentation better than this model, this model was enough to acquire knowledge

Figure 5.1: Example of the class masks from an annotation.

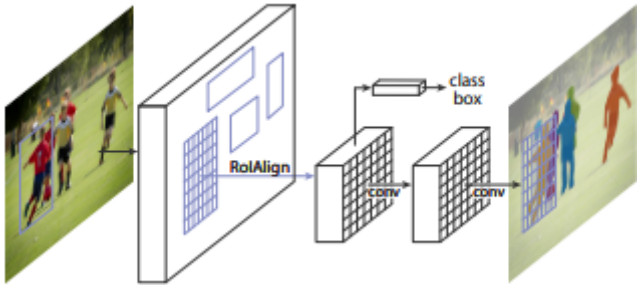


Figure 5.2: Overall architecture of MaskR-CNN.[18]

on the data specifications and Segmentation for the future work. Figure 5.2 shows an overall architecture of the model.

5.3 Results

Having done the train and testing, some inference was done on some individual images. No metrics or values were noted down for this test since the results from the few inferences done were pretty clear. Unfortunately the results were not satisfactory at all and an example can be seen in figure 5.3.
As we can see in that same figure, even when the mask around the predicted tree by the model is a perfect coverage of the tree area, it predicts the wrong class. For instance the red mask should be a dead tree and in that example it corresponds to a normal tree. Every class present in the example is always the same class, the normal tree. There is the overlapping of boxes that should not happen since there is only one tree in the bounding box, for instance the orange and the purple masks. There is also some incomplete detection and by taking a look at the hand-made mask in figure 5.4 it is clear that the results were not satisfactory at all and can not be accepted.

5.4 Discussion

After some thoughts about these results, one main point needs to be approached. Since there is no prior experience or knowledge on the area of the data, the annotations made could be wrong since anyone could pick up one image from the data set and annotate it differently from anyone else just by guiding themselves on what the person defines what could be a tree.

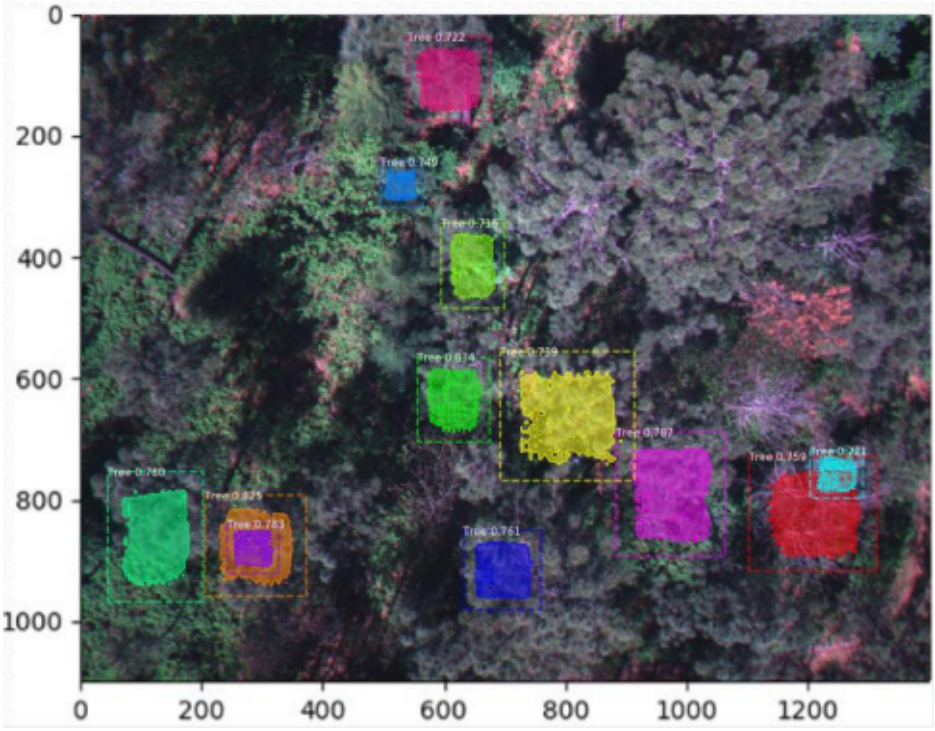


Figure 5.3: Example of an output from the MaskR-CNN model.

Differentiating large bushes from smaller trees is also a difficult task and most importantly, defining the boundaries of a singular tree crown can also be quite hard.

Other reason is the band usage for this test. As stated in section 5.1, we only used RGB imagery and perhaps with the use other bands, the results could have been better. However it goes back to the point stated before, without the exact information from the area that could help the annotator, it could lead us to the same quality of results.

Overall the blame should be put on the annotator for not being able to distinguish the different trees or bushes and not on the model itself, therefore a reminder for the future annotations on the data sets from section 4.

5.4.1 Realization

As explained in section 4.5, the same specie of trees usually are grouped together, hence, the idea of performing Instance Segmentation was dropped and we started to focus on Semantic Segmentation.

The annotations are easier to perform in this case, since it is not needed to delineate singular tree crown peaks and guess what is a single tree, but rather just annotate the area that the certain specie is present in. It also gives us the same end goal of the classification of the different species and cluster them just like Instance Segmentation would do except there is only one instance and that is the class of the specie.

The usage of different bands also needs to be looked upon, since it could give us a clear distinction between different species in the new data acquired from section 4 and therefore an helpful bonus information to feed into the model.

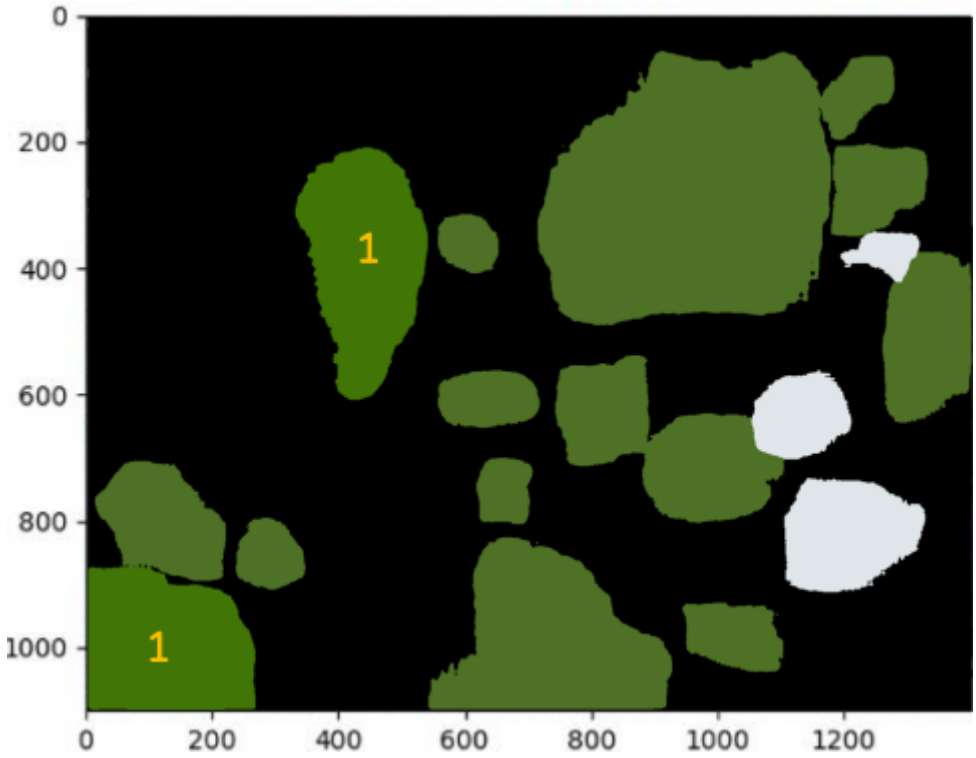


Figure 5.4: Hand-made annotation of the supposed right masks of the output from figure5.3. *(The 1 present in the green patches is to differentiate from the other green patches with no number in it, since it is from another class.)*

5.5 Relevant Ideas For The Future

One of the ideas that came up during this acquisition of new data, section 4, is to use the data set from Planet, that has eight bands, to help create bands for the twenty-five centimeter resolution data set. With the use of all eight bands from the Planet data set, we can make a classifier to take the red, blue, green bands and predict the other bands values. With that classifier trained we can take the three bands from the Cartography of North / Center of Portugal and predict the values of the other bands that were in the Planet data set.
Although the images from the data sets are in different resolutions, since the values taken are from the same type of band, it can give us an opportunity to use the Cartography data set with an eight band prediction, leading to more data available to be used.
The other idea was to take the multi-band low resolution imagery from Planet[16], and use a model to detect the zones that have trees, then these areas are sent to another model. This next model then works on the zones flagged by the first model but uses the higher resolution RGB imagery from the Cartography[15] to perform a finer segmentation on the respective zone and detect the species present on those areas.

Chapter 6

Realization and Annotations

6.1 Realization on the Mentioned Ideas

One of the ideas was to enhance the segmentation process by using both data sets presented in chapter 4. However, due to the contrasting characteristics of the sources and some lack of experience on working with QGIS [25], manually delineating the exact area of extraction turned out rather difficult and thus dropped. The other alternative was to increase the RGB data set 4.1 by adding new bands to it. The whole process would consist in training a classifier using the data from the Planet.com[16]. However, the data values from the cartography ranged between 0 and 65535 while the Planet data set ranged from 0 to 255. Although some normalization could be performed, the values were not evenly distributed along the interval with the majority residing within the first 4000 values and therefore, considering the risk of poor normalization and potential for negative outcomes it was decided to forget this approach in order to avoid losing time.

6.2 Annotations

Image annotations is a time-consuming task that becomes slightly harder the more classes it has. Even a couple of small mistakes could bias a module to give inaccurate results. Images need to be clear for the annotator so that he is able to know what he is annotating along the way. However, identifying tree species solely by using Red, Green and Blue bands it is almost impossible. Therefore an expert had to intervene in this process. This person was assigned to this work, through a company collaboration, due to their many years experience in the field. The expert provided insights on how to identify different trees based on their unique characteristics and provided areas where those species tend to be present. With this guidance, the RGB data set4.1 was completed.
The Planet data set was different. Since the images are in lower resolution, it is impossible to differentiate between tree species using Red, Green and Blue bands. However, since this data set has five other bands, by rearranging them in the QGIS software[25], it was possible to annotate the images. Utilizing bands such as Red, Near Infra-Red and Red-edged combined with having a prior knowledge on the geographic location and predominant tree in the area, it is possible to annotate a certain specie. A set of images can be seen in figures 6.1,6.2 and 6.3.
Figure 6.1 by itself offers a challenge in distinguishing different species due to the similarity of colors, those ranging from darker green to lighter green. However, when taking in information from figure 6.3 this distinction becomes easier, since differentiating light blue from black is effortless. To confirm this distinction, we can use the information from the previous

Figure 6.1: RGB Scene from Cartography of North / Center of Portugal.

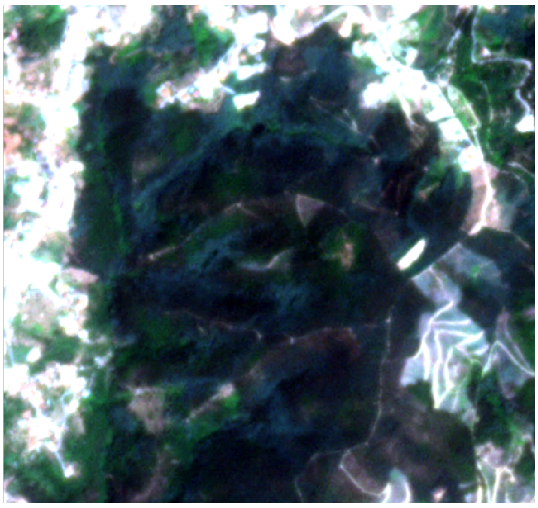


Figure 6.2: Scene from the Planet data set.

mentioned figure 6.1 to aid the annotator in case of doubt. The difficulty on annotating these images is when there are two or more colors on the same patch. To solve that problem, since we want to mainly know if a certain specie is in the mix, if most of the area is colored the same type as the known specie, it is annotated as if the whole area contains the specie. In simple terms, in the figures mentioned previously, 6.1,6.2 and 6.3, the specie is identified by a dark color, if an area is dark blue, it means that there is at least one more specie besides the one that we want to identify, and consequently it is considered as dark, since the specie is dark.

Most of mixed species cases are a result of human intervention, where multiple species are intentionally mixed in their terrain in order to create a forest mosaic. Some companies adopt this approach in order to preserve the ecosystem. It can also happen naturally but human influence is the predominant cause.

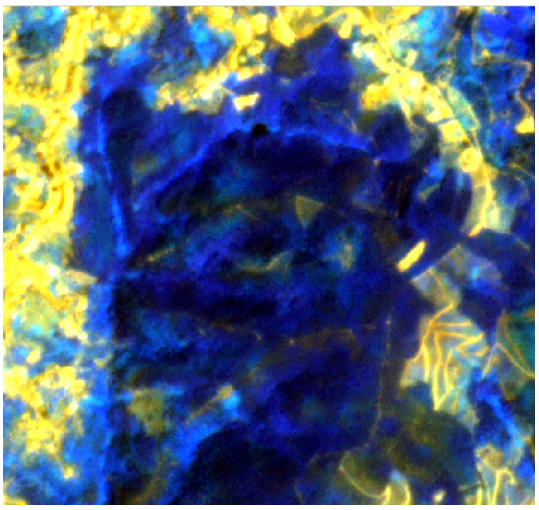


Figure 6.3: Scene from figure 6.2 but with rearranged bands.

6.3 Data set Specifications

As mentioned before in 6.2, two distinct data sets were made. Both cover the same area but possess unique characteristics from each other.
The Cartography data set 4.1, consists of 182 images with 3 bands: Red, Green and Blue. Each image measures 233x285 pixels and none of them contain overlapping information from each other. This specific size was used in order to better align with the dimensions required by the first model that was trained, the ViT-Adapter2.6. Since we use a smaller version of the models as we will see later on, the training data sets that were used to obtain the pre-trained weights for the backbone of this model were also data sets that had around the same size of images. Therefore it was decided to crop the images around that size.
The second data set created using images from Planet.com4.2 consists of 48 images with eight bands as mentioned previously in section 4.2. Although it has far less images, it contains a greater amount of information per image due to the number of bands. These images were obtained by slicing them side by side from a larger image so there is no overlapping of data and have size of 400x400 pixels. It has a larger size compared to the previous data set but still rather small and made to align with the dimensions required the configuration of the small configuration of the models.

Chapter 7

Results

7.1 Introduction

Having finished creating the data sets it was time to work with the models. For this work several models were used, such as the ViT-Adapter, Lawin [5] and InternImage-H[4]. These models have their differences in terms of architecture and procedures, however all of them work with the roughly the same libraries generally used in semantic segmentation, making it easier for the user to work with.

7.2 Hardware and Software Specifications

The research was conducted using the hardware resources available in the SociaLab university laboratory. All these results were obtained using a single machine equipped with a single NVIDIA GeForce GTX 1080 Ti, with 12GB of dedicated RAM. This GPU was capable of performing the training and testing on the smaller configuration of the models, since it could only hold up to 12GB of memory. However it still achieved some productive and acceptable results as we will see next. The specific versions of PyTorch[26] and its dependencies can be refered to the authors' GitHub repositories, as each model comes with its unique specifications. PyTorch was preferred over TensorFlow[27] due to the author's familiarity with it.

7.3 ViT-Adapter

This was the first model used and therefore the most studied. It has many different configurations depending on the size of the model itself. It ranges from the large configuration, ViT-Adapter-L, to the smallest, ViT-Adapter-T. In our test studys, we used the small configuration, ViT-Adapter-S, since we only had a GPU available that only could handle configurations of that size. This model managed to fit within the 10GB margin, leaving some working space on the model. Some changes were made such as the number of heads of the transformer, the learning rate and number of iterations of the training phase. Other changes were made such as the type of some layers and the width of the input bands of the model, since we are working with imagery with 8 channels, and therefore some other calculations and normalization needed to be tweaked in order to work. The best results out of various tests can be observed in the table 7.1.

7.4 InternImage-H

InternImage-H was utilized solely to observe differences in results between models, however, just like the next model in the list, no major changes were made, since most of the time was taken by the ViT-Adapter to be able to perform on the data sets. The small configuration was used, InternImage-S, due to the same reasons stated in section 7.3. The learning rate was the only parameter that was tweaked in this model.

7.5 Lawin

This model is slightly older than the two models previously mentioned, however it still is a notable model that can perform semantic segmentation. The same reason from the model above applies to this one, where it is only used for results comparison. Nonetheless some changes on the learning rate were made in order to achieve better results.

7.6 Metrics and Results

To evaluate each model on the data sets, metrics such as mean F-Score and mean Intersection over Union were utilized. Since we are working with multi-class semantic segmentation, we can take the Intersection over Union to measure the accuracy of segmentation models and its effectiveness. The F-Score provides us a quantitative measure on how the model performed. It takes in account Recall and Precision in its formula and therefore a good quantitative representation of how good or bad a classification was. As mentioned previously, we are working with multi-class problem, so we need to calculate the mean of each F-Score of each class to represent the model.

Table 7.1: Model Evaluation Metrics - Each training had 40k iterations of the data set, with a validation step every 4k iterations. (Mean Intersection over Union - mIoU, Mean F-Score - mFscore)

Model Configuration	**mFscore**	**mIoU**
ViT-Adapter-S | Base Configuration	0.757	0.614
ViT-Adapter-S | Configuration A | Dilated Convolutions	0.806	0.678
ViT-Adapter-S | Configuration A | Normal Convolutions	0.808	0.679
Lawin B2 | Base Configuration | Changed LR	0.827	0.697
InternImage-H | Base Configuration | Changed LR	0.835	0.705

Table 7.1 provides us all the tests made only using the 8 channel imagery. As we can see the base configuration of the ViT-Adapter-S obtained a 61% mIoU, however with some changes on the hyper parameters it rises up to around 68%. A reminder that using one of the smallest size of the ViT-Adapter brings us a to almost 68% mIoU, there is plenty of room to improve this metric by using larger configurations. One of the surprises from this tests is the results of Lawin, a model that is older than the other two models used but nonetheless still going pair to pair with them, obtaining almost 70% mIoU with one of the smallest configurations. To be precise, the B2 configuration from the default Ade20k on the authors GitHub was used. The best results overall come from the InternImage-H model, with a 70% mIoU.

7.7 Discussion

At the start of this project, we only had access to the 3 channel imagery and therefore, only worked with those images. For months, these images would only give us bad results and weird metric values on some of the steps of training, such as the evaluation step. Various modifications were performed in order to get to the root of the problem although they were all ineffective. As last resource, in a desperate attempt,a re-annotation of a subset from the data set was made and not even that fixed the bad results. Having tried everything at our disposal, one could only assume that using only RGB imagery is not fit to solve this classification problem.

After this realization, we were able to obtain the 8 band imagery from the Planet.com Research Program and use it to compare the results from the previous failed RGB tests. After making the necessary changes to each model, it was obvious that the necessary imagery for this classification problem needed to contain various bands that would give us more information. Therefore, the 3 band imagery was discarded from this work and we only focused our attention on the 8 band imagery.

As we can see from the tests performed in the table 7.1, the best results reach about 70% mIoU. These results are quite acceptable, since the number of images on the data set is minimal, rounding about 50 images. Every model performs data augmentation but the base number of images in the data set is still small. Especially when we are talking about Transformers or adaptations of this type of architecture that usually require large amounts of data to have good performance. Also, these results were obtained only using the smallest or the second smallest configuration available on the respective author's GitHub, since we did not have much memory to work with as mentioned previously.

Chapter 8

Conclusion

8.1 Brief Summary of the Project

The whole idea of this project was to successfully distinguish different species of trees in a forested environment by making use of satellite imagery and machine learning models. In this project, the concept of semantic segmentation was utilized in order to solve this problem, therefore the models that were studied needed to be able to perform this type of segmentation. Different types of data set were tested such as 4 band imagery and 8 band imagery, however only good results were achieved on the 8 band imagery data set. With the proper tools such as recent models that can perform semantic segmentation on images we were able to obtain a 70% mIoU when classifying the different classes on the images, however since there was a single GPU provided for this project, there was not much memory to work with and therefore only small configurations of the mentioned models were used for this task.

8.2 Summary of the Research Objectives

This project was composed of different objectives. Since this is a thesis work, the author needed to learn all the concepts of remote sensing, spectral analysis and computer vision necessary for the completion of the project. Having mastered them the author also studied all the necessary parts of each model utilized on this project in order to grasp the idea of how a model would take the data and process it in order to achieve the given results. Since most of the recent models are adaptations of older models, it was also necessary to study the original architecture that everyone took reference of, the ViT. Since there were no public available data sets with information about the different characteristics of the Portugal flora, it was decided to created a data set from scratch. Choosing what data to use and how to annotate and pre-process it was also one of the objectives, although it was harder than it seemed to be due to the characteristics of the portuguese flora. Finding images where one specie is completely isolated was almost impossible and therefore, the help of a specialist was needed to know what areas should be looked into in order to obtain the best possible imagery for the wanted specie, and some tips on how to distinguish one specie from another with a naked eye.

8.3 Future Work

There are many things that we can add do this project. Although we were successful in identifying one specie, this type of work can clearly handle more than one specie. Acquiring imagery that meets the requirements for a specific specie can be quite hard, since most of the

forests in Portugal are usually a combination of various species and therefore hard to acquire training data for the specific wanted specie. However with some time and patience it can be done as it was proven by this project. One adaptation that can be made to this project is to correctly distinguish certain species in different times of the year. Since species have their bloom at different times of the year and some blooms might even change the species characteristics completely, it would be quite helpful to correctly identify the specie even with their bloom. One of the ideas of this project was to check if RGB imagery would have enough information to correctly identify different species in a forested environment. However we were not able to get absolute results about this topic, therefore it should be a priority as future work to solve this problem and compare the results with the 8 band imagery.

Bibliography

[1] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers," *arXiv e-prints*, p. arXiv:2105.15203, May 2021. xiii, xvii, 8, 9

[2] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for Semantic Segmentation," *arXiv e-prints*, p. arXiv:2105.05633, May 2021. xiii, xvii, 9

[3] Z. Chen, Y. Duan, W. Wang, J. He, T. Lu, J. Dai, and Y. Qiao, "Vision Transformer Adapter for Dense Predictions," *arXiv e-prints*, p. arXiv:2205.08534, May 2022. xiii, xvii, 10, 11

[4] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, X. Hu, T. Lu, L. Lu, H. Li, X. Wang, and Y. Qiao, "Internimage: Exploring large-scale vision foundation models with deformable convolutions," 2023. xiii, xvii, 11, 12, 51

[5] H. Yan, C. Zhang, and M. Wu, "Lawin transformer: Improving semantic segmentation transformer with multi-scale representations via large window attention," 2022. xiii, xvii, 12, 13, 51

[6] A. J. Abreu, L. A. Alexandre, J. A. Santos, and F. Basso, "LudVision – Remote Detection of Exotic Invasive Aquatic Floral Species using Drone-Mounted Multispectral Data - Submitted," *arXiv e-prints*, p. arXiv:2207.05620, Jul. 2022. xiii, xvii, 1, 15, 16, 17, 18, 19, 43

[7] "Use of remote sensing in wildfire," https://www.intechopen.com/chapters/38093. xvii, 4

[8] Nilesh-Barla, "Panoptic segmentation: Definition, datasets & tutorial [2023]," https://www.v7labs.com/blog/panoptic-segmentation-guide, Jan. 2023. xvii, 6

[9] J. Alammar, "The illustrated transformer – jay alammar – visualizing machine learning one concept at a time." https://jalammar.github.io/illustrated-transformer/, Jul. 2018. xvii, 8

[10] M. Decuyper, R. O. Chávez, M. Lohbeck, J. A. Lastra, N. Tsendbazar, J. Hackländer, M. Herold, and T.-G. Vågen, "Continuous monitoring of forest change dynamics with satellite time series," *Remote Sensing of Environment*, vol. 269, p. 112829, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425721005496 xvii, 20, 21, 22

[11] H. Qin, W. Zhou, Y. Yao, and W. Wang, "Individual tree segmentation and tree species classification in subtropical broadleaf forests using uav-based lidar, hyperspectral, and ultrahigh-resolution rgb data," *Remote Sensing of Environment*, vol. 280, p. 113143, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425722002577 xvii, 22, 23, 24, 25

[12] S. Mojtaba Marvasti-Zadeh, D. Goodsman, N. Ray, and N. Erbilgin, "Early Detection of Bark Beetle Attack Using Remote Sensing and Machine Learning: A Review," *arXiv e-prints*, p. arXiv:2210.03829, Oct. 2022. xvii, xviii, 26, 27, 28, 29, 30

[13] T.-A. Nguyen, B. Kellenberger, and D. Tuia, "Mapping forest in the swiss alps treeline ecotone with explainable deep learning," *Remote Sensing of Environment*, vol. 281, p. 113217, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425722003248 xviii, 31, 32, 33, 34

[14] D. Michelini, M. Dalponte, A. Carriero, E. Kutchartt, S. E. Pappalardo, M. De Marchi, and F. Pirotti, "Hyperspectral and lidar data for the prediction via machine learning of tree species, volume and biomass: A contribution for updating forest management plans," in *Geomatics for Green and Digital Transition*, E. Borgogno-Mondino and P. Zamperlin, Eds. Cham: Springer International Publishing, 2022, pp. 235–250. xviii, 35, 36, 37

[15] "Cartography of north / center of portugal," https://cartografia.dgterritorio.gov.pt/ortos2021/service?service=WMTS&REQUEST=GetCapabilities&VERSION=1.3.0. xviii, 39, 46

[16] "Planet homepage," https://www.planet.com/. xviii, 39, 40, 46, 47

[17] "Copernicus land monotoring service homepage," https://land.copernicus.eu/. xviii, 40, 41

[18] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *arXiv e-prints*, p. arXiv:1703.06870, Mar. 2017. xviii, 43, 44

[19] H. Qin, W. Zhou, Y. Yao, and W. Wang, "Individual tree segmentation and tree species classification in subtropical broadleaf forests using uav-based lidar, hyperspectral, and ultrahigh-resolution rgb data, *Remote Sensing of Environment*,280:113143," 2022. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0034425722002577 1, 3, 4

[20] "Landsat science website - this joint nasa/usgs program provides the longest continuous space-based record of earth's land in existence." [Online]. Available: https://landsat.gsfc.nasa.gov/ 4

[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2020. [Online]. Available: https://arxiv.org/abs/2010.11929 8, 9, 10

[22] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000. 21

[23] "Hasty.ai homepage," https://hasty.ai/. 41, 43

[24] W. Abdulla, “Mask r-cnn for object detection and instance segmentation on keras and tensorflow,” https://github.com/matterport/Mask_RCNN, 2017. 43

[25] “Qgis software,” https://qgis.org/en/site/. 47

[26] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “Pytorch: An imperative style, high-performance deep learning library,” 2019. 51

[27] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, “Tensorflow: A system for large-scale machine learning,” 2016. 51