Technische Hochschule für angewandte Wissenschaften

Würzburg-Schweinfurt

Masters Thesis

Weakly Supervised Semantic Segmentation of Multispectral Satellite Images for Land Cover Mapping

Submitted in Partial Fulfillment of the Requirements for the Degree Programme Masters of Science (M.Sc) in Artificial Intelligence

Author:	**Esther IfeOluwa Ademola**
Matriculation number:	**5122738**
1. Examiner:	**Prof. Dr. Magda Gregorova**
2. Examiner:	**Prof. Dr. Frank Deinzer**
Submission Date:	**30.8.2024**

Abstract

This thesis addresses the challenge of weakly supervised semantic segmentation in the domain of multispectral satellite imagery, particularly emphasising its application in the context of deforestation monitoring. The research proposes a novel methodology leveraging high-resolution multispectral data from the Sentinel-2 satellite over a forest area in Bavaria, Germany. The primary objective is to develop a high-performance algorithm capable of semantic segmentation, emphasising the unique nature of multispectral data compared to conventional RGB images. The study explores custom model development to adapt existing architectures to the distinctive characteristics of the dataset, addressing challenges related to weak annotation. Evaluation strategies involve benchmarking against datasets to assess the model's effectiveness under varying conditions. The significance of this work lies in advancing weakly supervised semantic segmentation techniques for multispectral images, with potential applications in scenarios with limited labelled data, contributing to the broader field of artificial intelligence.

Acknowledgement

This master's thesis was written during my studies in the Master of Artificial Intelligence program at the University of Applied Sciences Würzburg-Schweinfurt.

I would like to express my deepest gratitude to my supervisor, Prof. Dr. Magda Gregorova, for her invaluable guidance and support throughout this thesis.

A heartfelt thank you to Gunther and Denise at Green Spin for their generous assistance and collaboration, which greatly enriched this research.

To my parents, thank you for your unwavering support and belief in me. Your encouragement has been my driving force.

Lastly, I thank God for providing me with the strength and perseverance needed to complete this journey, e don teh.

Würzburg, on 30.8.2024

Esther IfeOluwa Ademola

Contents

List of Figures

List of Tables

Abbreviations

ESA ..... **E**uropean **S**pace **A**gency

FN ..... **F**alse **N**egative

FP ..... **F**alse **P**ositive

IoU ..... **I**ntersection **o**ver **U**nion

NIR ..... **N**ear **I**nfra**R**ed

SGD ..... **S**tochastic **G**radient **D**escent

SWIR ..... **S**hort-**W**avelength **I**nfra**R**ed

TN ..... **T**rue **N**egative

TP ..... **T**rue **P**ositive

1. Introduction

1.1. Background and Motivation

Forests play a crucial role in maintaining ecological balance, supporting biodiversity, and mitigating climate change by acting as significant carbon sinks. However, these vital ecosystems face various threats, one of which is infestation by parasites such as the beetle bug, which has become increasingly prevalent in forested areas. These beetles infect trees, leading to widespread tree mortality and consequent ecological disruption. To prevent the spread of these infestations and to manage forest health, affected trees are typically identified and removed.

Traditionally, identifying and mapping deforested areas due to such infestations involve manual surveys where forest personnel physically walk through the forests to detect and record regions of tree loss. This method, while effective, is inherently time-consuming, labor-intensive, and costly. Moreover, manual surveys are prone to human error and may not always capture the extent of deforestation accurately and in a timely manner, which is critical for effective forest management and disease control.

With the advent of remote sensing technologies, it has become feasible to monitor large forested areas more efficiently. Satellite imagery provides a valuable tool for observing changes in forest cover over time. Specifically, the Sentinel-2 satellites operated by the European Space Agency (ESA) offer high-resolution multispectral images that cover a wide range of wavelengths. These images provide detailed information about vegetation health,

land cover, and land use changes, making them ideal for monitoring forest conditions.

The manual approach to detecting deforested areas in response to beetle infestations poses several challenges. Firstly, the extensive human resources and time required for such surveys make it impractical for large-scale monitoring, especially in remote or inaccessible regions. Secondly, the dynamic nature of forest ecosystems demands more frequent monitoring to promptly identify new infestations and mitigate further spread. Given these limitations, there is a pressing need for an automated and scalable solution that can accurately detect and map deforested regions.

Advancements in deep learning, particularly in the domain of semantic segmentation, offer promising avenues for addressing this challenge. Semantic segmentation is a computer vision task that involves classifying each pixel in an image into a predefined category, allowing for detailed mapping of various land cover types, including deforested areas. When combined with multispectral satellite imagery, semantic segmentation techniques can potentially automate the detection of deforested regions, providing a more efficient, cost-effective, and accurate alternative to manual surveys.

However, traditional deep learning models typically require large amounts of labeled data for training, which is often unavailable or expensive to obtain in the context of satellite imagery. This constraint necessitates the exploration of weakly supervised learning methods, where models are trained with limited or imprecise labels. This approach reduces the dependency on extensive labeled datasets while still enabling effective model training. The availability of open-source satellite imagery provides a valuable resource for this purpose. These images, freely accessible, cover vast geographic areas and offer sufficient resolution to detect changes in land cover. Leveraging these data, the goal is to develop a weakly supervised deep learning model capable of identifying and mapping deforested areas efficiently.

The primary motivation behind this research is to harness the information in multispectral satellite imagery and advanced machine learning techniques to create a robust, scalable

tool for monitoring forest health. By automating the detection of deforested regions, we aim to provide a more effective means of managing forest resources and combating the spread of tree infestations, thereby contributing to sustainable forest management and environmental conservation.

1.2. Problem Statement

The manual identification and mapping of deforested areas in forest ecosystems, particularly those affected by beetle infestations, are time-consuming, labour-intensive, and costly. These traditional methods are not scalable for large and remote forested regions, leading to delays in detection and mitigation efforts. To address this issue, there is a critical need for an automated solution that leverages multispectral satellite imagery and advanced deep learning techniques to efficiently and accurately detect and map deforested regions.

1.3. Research Objective

This research aims to develop a weakly supervised semantic segmentation model capable of analysing satellite images to identify deforested areas, thereby providing a more effective and scalable approach to forest monitoring and management.

1.4. Research Questions and Hypothesis

This thesis seeks to address the following research questions:

1. **How can weakly supervised or self-supervised semantic segmentation be effectively applied to multispectral satellite imagery with limited or weak annotations?**

Given the challenges of acquiring comprehensive labelled datasets for multispectral satellite images, this question explores methods to leverage incomplete annotations for training effective semantic segmentation models. The goal is to determine how weakly supervised or self-supervised learning techniques can be adapted to handle the complex and diverse data provided by multispectral imagery.

2. **Do existing segmentation methodologies generalise effectively for multispectral data, considering its unique characteristics and variance from the data used to develop existing model architecture?**

   This question investigates whether current semantic segmentation techniques, typically developed using standard Red, Green, Blue (RGB) images, can be effectively applied to multispectral satellite imagery. It seeks to evaluate the generalisability of these methodologies to handle the additional spectral information and the differences in data characteristics, such as spatial resolution and spectral variability, inherent to multispectral datasets.

3. **How does the spectral band variability of multispectral satellite imagery influence the performance and generalisation capabilities of segmentation models?**

   This research aims to investigate the impact of spectral variability in multispectral satellite imagery on the effectiveness of segmentation models. By examining how different combinations of spectral bands—including traditional RGB bands and additional non-visible bands—affect model performance, this study seeks to determine whether including the full range of spectral information enhances model accuracy and generalisation across diverse land cover types and environmental conditions. Additionally, it will explore whether the performance of models trained on various band combinations aligns with established domain knowledge in remote sensing, providing insights into the necessity and efficiency of using additional spectral bands

beyond RGB for high-performance segmentation.

1.5. Overview of Research Methodology

This research employs a systematic methodology to acquire, preprocess, and analyse multispectral satellite imagery for the purpose of land cover mapping using weakly supervised semantic segmentation techniques. The initial step involves the acquisition of data from the Sentinel-2 satellite, provided by the ESA. These images, which include multiple spectral bands such as visible (RGB) and non-visible wavelengths, are specifically selected for their high resolution and relevance to the task of identifying deforested regions affected by beetle infestations.

The raw satellite imagery undergoes several preprocessing steps to prepare it for model training and analysis. These steps include the selection of pertinent spectral bands that offer crucial information for vegetation and deforestation analysis, and the standardisation of image resolution to ensure uniformity, focusing on the 10m and 20m bands. The images are then normalised to mitigate the effects of lighting variations, thereby enhancing the model's generalisation capabilities. Additionally, data augmentation techniques, such as rotation, flipping, and scaling, are applied to increase the diversity of the training dataset and to prevent overfitting.

Furthermore, the methodology presents the approach to finding answers to the research questions. Initially, despite the weak annotation of the dataset, model training begins with a supervised learning approach. Off-the-shelf models such as UNet are employed for their proven effectiveness in semantic segmentation tasks, which involve capturing both spatial and contextual information. The models are trained on the weakly annotated dataset as though it is fully supervised. The performance of the model is then evaluated on a validation set, with metrics such as Intersection over Union (IoU), calculated to assess the segmentation quality.

Following the initial supervised learning phase, a weakly supervised learning approach is adopted to better exploit the limited and inaccurate annotations available. This involves selecting a suitable weakly supervised algorithm, such as self-training, pseudo-labelling, or multiple instance learning, and adapting supervised learning model to incorporate the weak supervision signals.

Additionally, the methodology delves into the significance of multispectral satellite imagery, taking into consideration the spectral bands and their respective unique characteristics in analysing land covers.

Finally, the model's performance is evaluated using a separate test set to ensure it can accurately identify deforested areas. The results from the weakly supervised learning approach are compared with those from the initial supervised learning phase to demonstrate the improvements achieved through weak supervision.

This methodology provides a comprehensive framework for developing and assessing a weakly supervised semantic segmentation model for multispectral satellite imagery, aiming to leverage weakly annotated data to improve the accuracy and efficiency of deforestation mapping.

1.6. Structure of Thesis

This section outlines the organisation of the thesis, which comprises six main chapters:

Chapter 1 (Introduction) establishes the problem statement, objectives, aims, research methodology overview, and provides an outline of the thesis structure.

Chapter 2 (Literature Review) presents a comprehensive review of remote sensing, semantic segmentation, focusing on weakly supervised techniques, and discusses state-of-the-art methodologies in these areas.

Chapter 3 (Data Description) outlines the attributes of the Sentinel-2 satellite imagery, focusing specifically on the dataset used in this research.

Chapter 4 (Methodology) explores the methodology employed in the research, including data processing, model selection, training strategies, and evaluation metrics.
Chapter 5 (Experiments) details the experimental setup, evaluates the experiments, analyses the results, and interprets the findings.
Chapter 6 (Conclusion) summarises the key findings of the study, highlights contributions and the significance of the research, and suggests directions for future research in the field of weakly supervised semantic segmentation of multispectral satellite images for land cover mapping.

2. Literature Review

Computer vision and what not then a good segue to segmentation

2.1. Remote Sensing Techniques for Land Cover Mapping

2.2. Semantic Segmentation

SS, State-of-the-Art Semantic Segmentation Architectures. Chat briefly about semantic segmentation, supervised and unsupervised with literature (brief history till now). Introduce weakly as well and state that it would be talked about in detail in the subsection or something

Also, evaluation theory, IoU, TP; FP; TN; FN

2.2.1. Weakly Supervised Semantic Segmentation

2.2.2. Models such as Unet

2.3. Evaluation Metrics

In the context of a semantic segmentation task, where the goal is to classify each pixel in an image into a predefined set of categories, the concepts of true positives (TP), false positives

(FP), true negatives (TN), and false negatives (FN) are fundamental for evaluating model performance. These metrics are defined as follows:

- **TP**: Pixels that are correctly classified as belonging to the positive class (i.e., the target class of interest).
- **FP**: Pixels that are incorrectly classified as belonging to the positive class when they actually belong to the negative class (i.e., background or other non-target classes).
- **TN**: Pixels that are correctly classified as belonging to the negative class.
- **FN**: Pixels that are incorrectly classified as belonging to the negative class when they actually belong to the positive class.

These classifications can be visualised in the form of a confusion matrix specific to each pixel in the image, providing the basis for various evaluation metrics.

Pixel Accuracy

A commonly used metric in segmentation, pixel accuracy is the ratio of the correctly classified pixels to the total number of pixels in an image. Although pixel accuracy is straightforward, easy to calculate and computationally efficient, it has setbacks. It can be misleading in cases of *class imbalance*. If the majority of the pixels belong to a dominant class, high pixel accuracy can be achieved by simply predicting the dominant class for most pixels, ignoring minority classes. In addition, it does not provide detailed information about the performance of individual classes, making it difficult to diagnose specific issues with the segmentation model. Pixel accuracy is calculated by:

$$Pixel\ accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

F1 Score

Intersection Over Union

The Intersection Over Union (IoU), also known as Jaccard index, is one the most common metrics used widely to evaluate semantic segmentation. (reason why Iou is commonly used)

As expressed by the formula 2.2 and illustrated in Fig. 2.1, the IoU is the area of overlap between predicted segmentation and the ground truth divided by the area of union between the predicted segmentation and the ground truth. This metric ranges from 0 to 1 (or 0 - 100%), where 0 indicates no overlap and 1 indicates a perfect overlap.

$$IoU = \frac{|A \cap B|}{|A \cup B|} \tag{2.2}$$

Using binary bitmaps, the IoU can be rephrased in terms of TP, FP and FN as follows:

$$IoU = \frac{TP}{TP + FP + FN} \tag{2.3}$$

For multi-class segmentation, the mean IoU (mIoU) of an image is calculated by taking the IoU of each class and averaging them. If there are N classes, the mIoU is computed as follows:

$$\text{mIoU} = \frac{1}{N} \sum_{i=1}^{N} \text{IoU}_i \tag{2.4}$$

where IoU_i is the IoU for class i. This approach ensures that the performance across all classes is considered, providing a more comprehensive evaluation of the segmentation model.

To compare pixel accuracy with Intersection over Union (IoU), consider a scenario with class imbalance. Suppose we have an image with a dominant background class and a minority object class. If the model correctly classifies 950,000 out of 1,000,000 total pixels, pixel accuracy is 95%. However, for the object class, if the model fails to detect

it, the IoU is 0. For the background class, with 950,000 correctly classified pixels out of 950,000 actual background pixels, the IoU is 1.0. The mean IoU, averaging across classes, might be significantly lower, reflecting poor performance on the object class despite high pixel accuracy. This shows that IoU provides a more nuanced evaluation by highlighting issues in class-specific performance, making it a better indicator of overall segmentation quality in the presence of class imbalance.

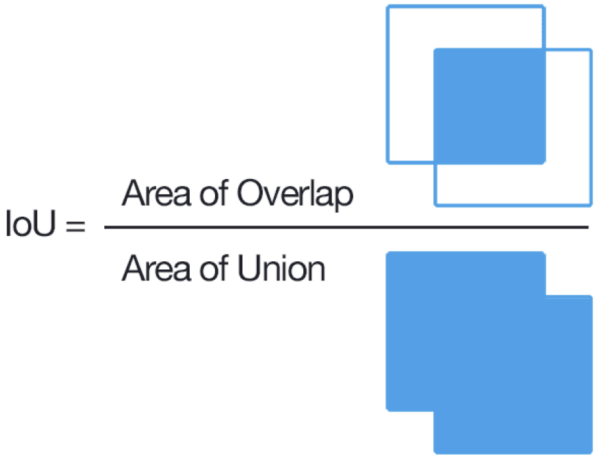

Figure 2.1.: Evaluation metric, Intersection over Union (IoU)

2.4. Augmentation

2.5. Previous Studies in Multispectral Image Segmentation

3. Data Description

This chapter provides a comprehensive overview of the data used in this research, focusing on multispectral images from the Sentinel-2 satellite operated by the European Space Agency (ESA). The Sentinel-2 satellite provides high-resolution imagery across multiple spectral bands, enabling detailed land cover analysis and vegetation monitoring.

3.1. Description of Sentinel-2 satellite data

The Sentinel-2 satellite, part of the ESA's Copernicus Programme, is designed to provide comprehensive Earth observation data. Equipped with a multispectral instrument, Sentinel-2 captures imagery across a wide range of spectral bands, from visible to shortwave infrared wavelengths. The satellite provides data that is crucial for environmental monitoring and land cover mapping, particularly in applications such as deforestation monitoring.

The images from Sentinel-2 encompass 13 spectral bands, each varying in spatial resolution to cater to different observational needs. The visible bands, consisting of blue, green, and red, have a resolution of 10 metres and are the standard for many visual tasks and general remote sensing applications. In addition to these, the near-infrared (NIR) band, also with a resolution of 10 metres, is particularly useful for vegetation monitoring due to its sensitivity to chlorophyll. This band captures the reflectance of vegetation, aiding in the assessment of plant health and biomass.

Sentinel-2 also includes four red-edge bands with a resolution of 20 metres. These bands are critical for vegetation analysis as they cover the spectral region where the reflectance of vegetation changes rapidly. This is particularly useful for identifying different types of vegetation and assessing their health. Additionally, the shortwave infrared (SWIR) bands, also with a resolution of 20 metres, provide valuable information on moisture content and are useful for distinguishing between various land cover types, including bare soil and vegetation.

Furthermore, three bands have a spatial resolution of 60 meters: the coastal aerosol band, the water vapour band, and one of the SWIR bands. The coastal aerosol band is primarily used for atmospheric corrections, particularly in coastal and inland water studies. The water vapour band assists in atmospheric water vapour correction, and the SWIR band aids in cloud detection and snow/ice monitoring.

For this research, only the bands with 10-metre and 20-metre resolutions are utilised as shown in table 3.1. The 60-metre resolution bands are excluded as they are are primarily used for atmospheric corrections and cloud detection which is not relevant for fine-scale vegetation and land cover analysis. The selected bands offer a balance between spatial resolution and spectral coverage, making them ideal for identifying and analysing deforested areas, which is the focus of this thesis.

Band	**Spatial Resolution (m)**	**Central Wavelength (nm)**
B2 - Blue	10	490
B3 - Green	10	560
B4 - Red	10	665
B05 - Red edge 1	20	705
B6 - Red edge 2	20	740
B7 - Red edge 3	20	783
B8 - NIR	10	842
B8A - Red edge 4	20	865
B11 - SWIR 1	20	1610
B12 - SWIR 2	20	2190

Table 3.1.: Spatial resolution and central wavelength of the Sentinel-2 bands utilised in this research

To enhance the reliability of the data, it is important to understand the technical aspects of how Sentinel-2 imagery is processed. The imagery is initially recorded with a bit depth of 12 bits per pixel, which allows for the representation of 4096 different levels of intensity for each spectral band. This high level of detail is crucial for accurately analysing various land cover types, vegetation health, and other environmental features.

For storage and processing purposes, the imagery is typically stored in a 16-bit format. This practice offers several advantages that are crucial for effective data analysis. First, the 16-bit format allows for enhanced compatibility with various data processing tools, making it easier for researchers to manipulate and analyse the data across different software platforms. More importantly, the use of a 16-bit format provides sufficient precision for subsequent analyses. This increased precision is vital because it ensures that any calculations or transformations applied to the data—such as atmospheric corrections or spectral indices—are based on the highest quality input, thereby maintaining the integrity of the original observations.

The increased bit depth significantly impacts the quality of the data analysis. By allowing for a greater range of values, the 16-bit format helps to reduce the risk of data loss during processing, which can occur if information is rounded or truncated in lower bit depths. This precision is particularly important when distinguishing subtle differences in reflectance values, such as those found in vegetation health assessments or land cover classifications. In essence, the 16-bit format enables more sophisticated analysis techniques, allowing researchers to detect fine variations in the data that might otherwise go unnoticed. This capability is essential for accurate environmental monitoring, ultimately leading to more informed decision-making in resource management and conservation efforts.

In terms of data processing, the imagery utilised in this research follows level 2A processing, which is essential for obtaining high-quality data. Level 2A processing involves the derivation of atmospherically corrected surface reflectance values, which range from 0 to

1, from the original 12-bit imagery. This correction is critical as it removes atmospheric effects—such as scattering and absorption caused by air molecules and aerosols—that can distort the true reflectance values of the Earth's surface.

The importance of Level 2A processing cannot be overstated. By providing atmospherically corrected data, this processing level ensures that the reflectance values accurately represent the surface features being analyzed. This level of precision is particularly vital for applications such as vegetation monitoring and land use classification, where subtle variations in reflectance can indicate significant changes in plant health or land cover. As a result, Level 2A products deliver accurate and reliable data, facilitating a wide range of scientific analyses and applications.

In summary, the Sentinel-2 satellite provides invaluable data for environmental monitoring and land cover analysis, particularly through its diverse range of spectral bands and high spatial resolution. The careful processing of this imagery, including the transition to a 16-bit format and the application of Level 2A atmospheric corrections, enhances the reliability and accuracy of the data. This ensures that we can confidently analyse vegetation health and monitor land use changes with precision. In the next section, we will delve deeper into the specifics of the dataset used in this research, exploring its characteristics and the state in which it was obtained, further contextualising its significance for our analysis.

3.2. Characteristics of the dataset

The dataset utilised in this research is specifically designed for the task of semantic segmentation in forested regions, focusing on identifying and mapping deforested areas. The data encompasses two distinct geographical locations in Bavaria, Germany: one in the western region (Bunsdorfer Forst) and another in the eastern region (Forst Schwarzenbach am Wald) (see Figure 3.1. Each location includes a variety of land cover types, such as

farmland, human settlements, and forest areas. However, for the purposes of this research, only the forest areas are considered.

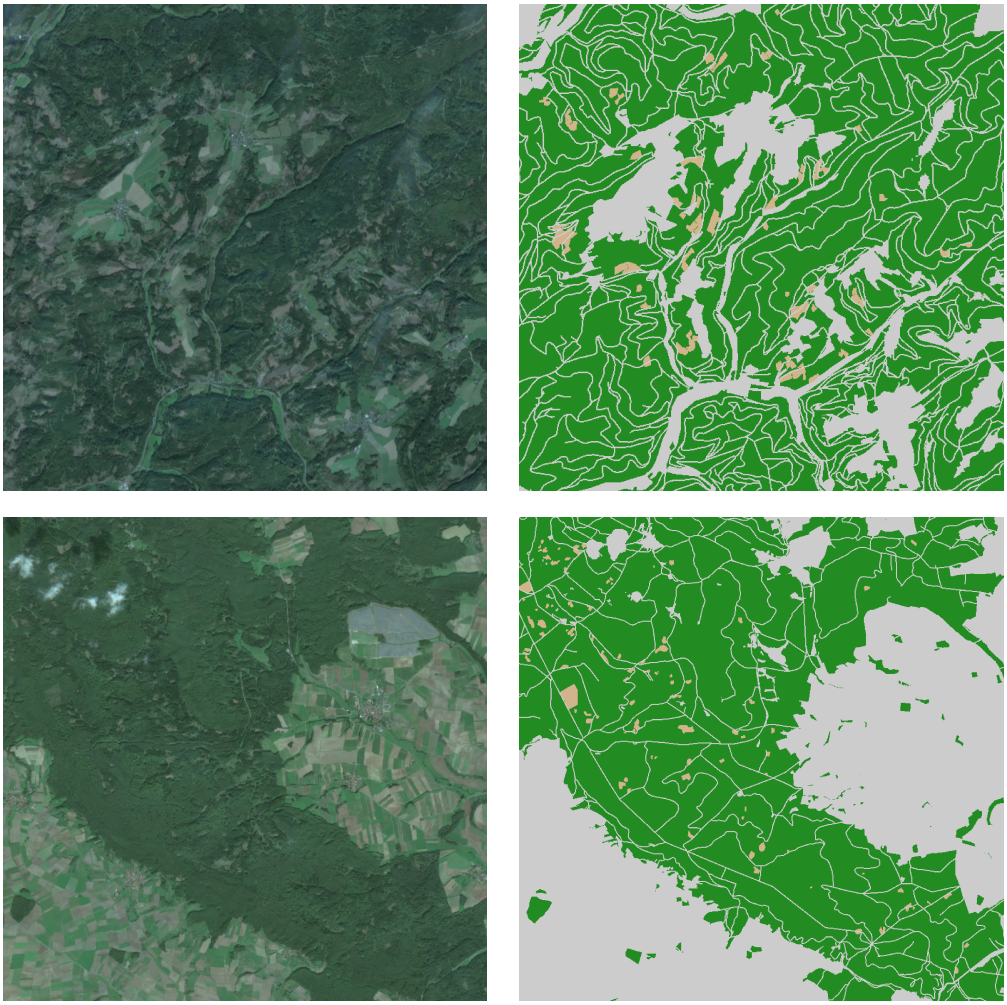

Figure 3.1.: Visible bands (RGB) images of the Bunsdorfer Forst and the Forst Schwarzenbach am Wald, along with their respective segmentation mask. In the mask, green indicates forested areas, brown indicates deforested areas within the forest, and gray indicates human-made areas such as farmlands and roads.

Geographical Scope and Temporal Coverage

The dataset contains Sentinel-2 satellite images for each of the two regions, captured at three different time points: May 2023, September 2023, and February 2024. This temporal coverage allows for the observation of seasonal changes and the detection of deforestation

events over time. The images cover 10 spectral bands, including all the available bands from Sentinel-2 except for those with a 60-metre resolution, which were excluded due to their lower spatial detail that is not suited for detailed vegetation analysis. This selection ensures high spatial resolution and comprehensive spectral information, crucial for distinguishing between different land cover types and detecting changes within the forest areas.

Additionally, each image contains an 11th band that serves as a manually created mask. This mask identifies valid and invalid values, where a pixel value of 1 indicates valid forest area, and a value of 0 represents invalid areas such as urban regions, farmlands, streets, and bodies of water. All these invalid areas fall under the umbrella term "anthropogenic areas", because they mostly originate from human activity. This mask is crucial for isolating the forest areas from other types of land cover, ensuring that the analysis focuses solely on the relevant regions.

Segmentation Labels

For each land area, the dataset also includes two additional arrays that serve as labels for the segmentation task. These arrays provide detailed information on deforested areas within the forest. They contain four distinct channels, each providing critical data for the segmentation process as shown in table 3.2 (see Figure 3.2).

The dataset is meticulously curated for the task of segmenting forest areas, distinguishing between forested regions, newly deforested areas, and older deforested areas. The combination of multispectral images with high spatial and spectral resolution, along with detailed segmentation labels, provides a robust foundation for developing and validating semantic segmentation models. The dataset's structure allows for the comprehensive analysis of land cover changes and the effective identification of deforested regions, supporting the objectives of this research.

Attribute	Description
ID	Assigns a unique ID to each deforested area within the forest, enabling the identification and tracking of specific clearings over time.
Clearing type	Indicates if the clearing was already established or newly added: 1: Existing clearing 2: Newly added clearing
Vegetation cover	Classifies the vegetation cover in cleared/deforested areas: 1: Soil 2: Low grass 3: High grass 4: Sparse trees
Timestamp	Date of recognition of the clearing

Table 3.2.: Description of segmentation task label attributes.

Overall, the dataset's characteristics, including its geographical scope, temporal coverage, and detailed segmentation masks, make it a good resource localising deforestation.

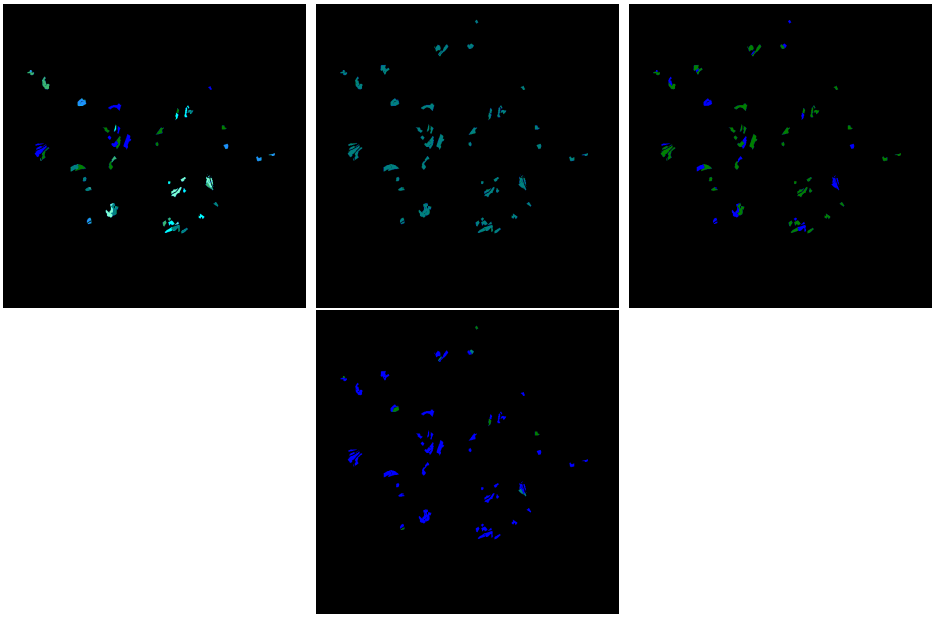

Figure 3.2.: Labels indicating the unique ID of deforested areas pixels, their clearing type, the vegetataion cover and the timestamp

4. Methodology

This chapter outlines a detailed step-wise methodology employed for training a semantic segmentation task, with our particular focus on identifying deforested regions. I leverage both simple and extended segmentation masks to capture the nuances of forest segmentation, utilising various combinations of multispectral bands as training inputs.
Additionally, I will discuss the evaluation methods used to assess the model's effectiveness, ensuring a comprehensive understanding of its capabilities and limitations. Each aspect of the training and evaluation process will be examined in detail in the subsequent sections, providing a thorough insight into the techniques and strategies that underpin this research work. Sit back and read on as we explore the intricacies of my approach.

4.1. Annotation and Labelling Process

The annotation and labelling process for the dataset used in this research is a meticulous and multi-step procedure designed to ensure that the data is accurately prepared for the semantic segmentation task. This process involves creating detailed masks and labels that represent deforested regions within the forested areas. By following a systematic approach, we ensure that the dataset is well-structured and suitable for analysing deforestation.
Firstly, I extracted region-specific labels for each timestamp. Each geographical region and its corresponding timestamp are associated with a mask that differentiates valid forest areas from anthropogenic areas — areas influenced by human activities, such as farmlands,

urban regions, human settlements, and streets. This step is crucial for isolating the forest areas that are the primary focus of our research.

Afterwards, I created a mask to identify deforested regions within the forests. This deforested area mask is generated by combining the region mask with attributes from the deforested areas data, specifically the ID and timestamp channels. A pixel is considered part of a deforested area if it is marked as valid forest in the region mask and has a positive ID value in the deforested areas data. The timestamp is used to ensure that only clearings that occurred at or before the specific timestamp are marked as deforested. This temporal consideration is vital for tracking the progression of deforestation over time. The result of this is a new channel of 3 classes representing the anthropogenic regions, deforested regions and the forested regions

Building on this three-class segmentation mask, I introduce a new channel that extends it by representing different vegetation cover types within the deforested areas. This vegetation cover mask categories the deforested pixels into four types: soil, low grass, high grass, and sparse trees, based on the vegetation type cover channel.

Seasonal variations are accounted for by creating a season mask and adding it to the data array. This mask assigns specific values based on the timestamp of the data. Incorporating this seasonal information allows for a more nuanced analysis of deforestation patterns across different seasonal times of the year.

After all these steps, the resulting array has 13 channels: 10 spectral bands, a seasonal mask representing the dataset features, and two types of segmentation masks serving as the dataset labels. One segmentation mask categorises data into three segments: anthropogenic, deforested, and forest. The other segmentation mask categorises data into six segments: anthropogenic, soil, low grass, high grass, sparse trees, and forest. Either of these segmentation masks is used as the labels of the dataset in a semantic segmentation task.

4.2. Data Processing

In the data processing stage, the primary goal is to prepare the dataset for effective training and validation of the semantic segmentation models. This involves splitting the data into sizeable patches and applying various augmentation techniques to enhance the size, diversity, and robustness of the training data.

The initial step in the processing pipeline is to split the large arrays into smaller patches. Given the size of the satellite images and the need to focus on specific regions, the arrays are divided into patches of size 32×32 pixels with a stride of 8 pixels. This stride ensures that there is a significant overlap between adjacent patches, which helps in capturing finer details and contextual information within the forested areas. Additionally, this patching approach is necessary to manage the high computational demands of processing large images, ensuring that the data can be handled within our available resources.

As mentioned in section 3.1, the values for the spectral bands were processed from 16 bits to surface reflectance values, which range from 0 to 1. However, there were some areas in the images that were affected by artefacts, specifically clouds, which cased their corresponding pixel values to exceed 1. To address this issue, I made the informed decision to exclude such areas from the training process. Each patch was then scaled to a range of -1 and 1 to introduce symmetry and create more consistent gradient descent behaviour. This scaling inherently leads to faster convergence and enhances the stability of the neural networks driving the models.

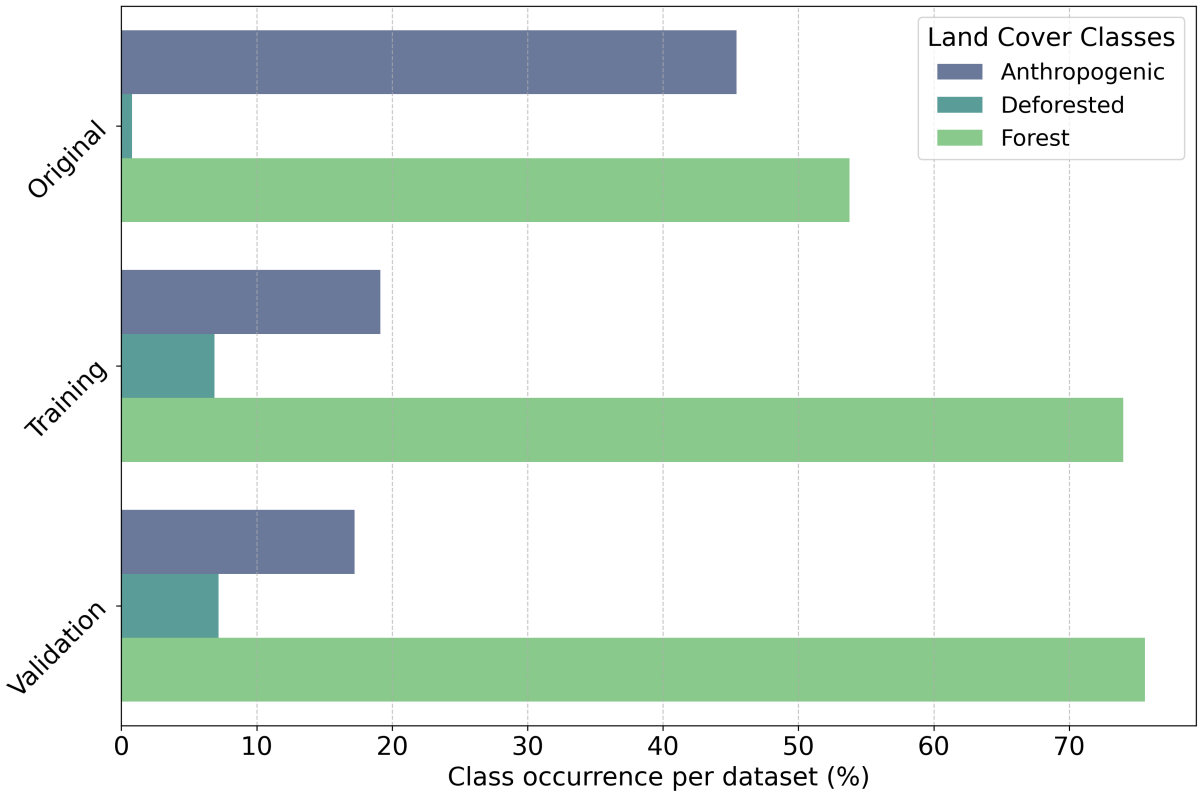

Figure 4.1.: The distribution for the different land cover classes for the original, training and validation datasets for the simple segmentation mask.

As shown in Fig. 4.1 and 4.2, the original dataset is unevenly distributed: deforested regions are sparse, while anthropogenic areas are abundant. To address this imbalance, it is crucial to select patches that specifically contain deforested pixels. By focusing on these patches, we ensure that the training data is rich in the features that are most relevant to our research objectives. This targeted selection helps in creating a more balanced dataset that can effectively train the segmentation model to identify deforested areas.

The dataset is divided into training and validation sets, with an 80:20 split, resulting in 5,869 patches for training and 1,471 for validation. To enhance the diversity and robustness of the training set, each patch undergoes augmentation five times, producing multiple variations of the original patches. This process expands the training dataset to approximately 35,214 image patches.

To ensure the integrity of the features relevant to deforestation detection, only geometric

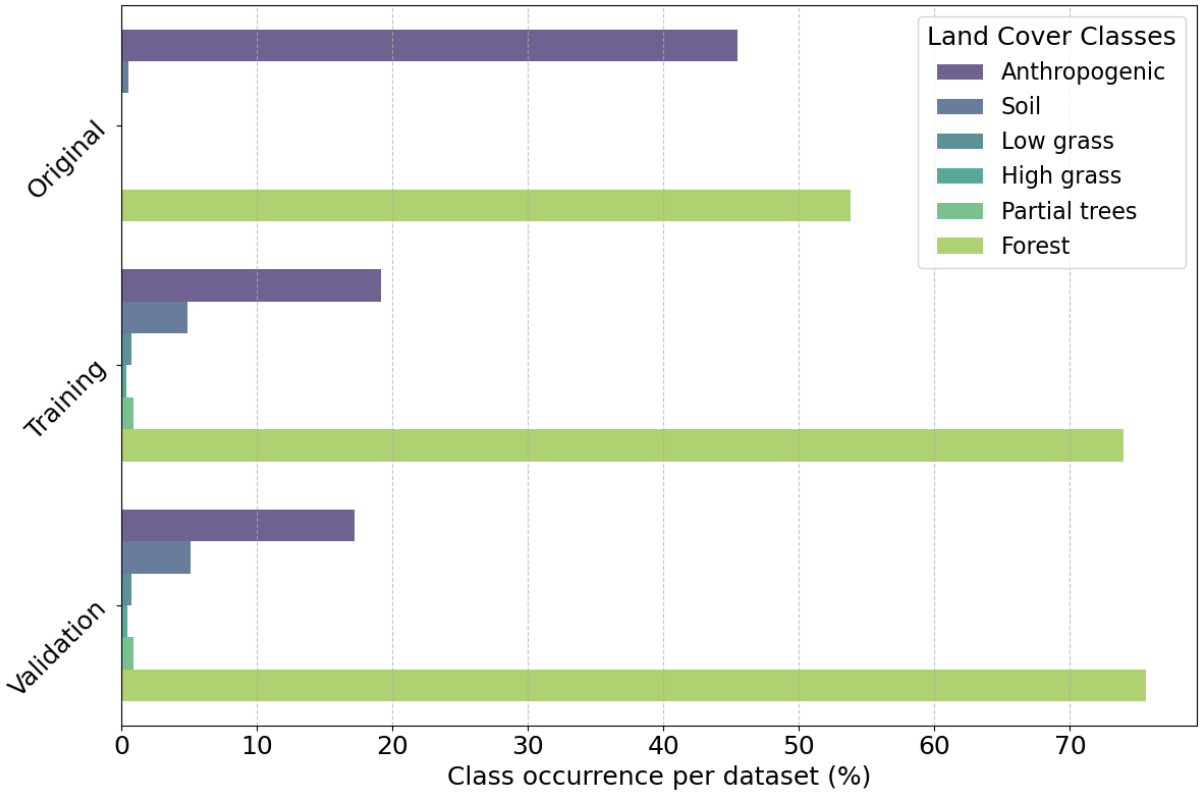

Figure 4.2.: The distribution for the different land cover classes for the original, training and validation datasets based for the extended segmentation mask.

transformations—such as flipping, rotation, and shift-scale-rotate—are applied (see Figure 4.3). These transformations introduce substantial variability while maintaining the original spectral information of the patches. This approach helps avoid any distortion or noise that could arise from altering colour properties or other spectral characteristics, which are critical for accurate vegetation and biomass analysis. The corresponding segmentation masks are subjected to the same geometric transformations, ensuring that the augmented data remains consistent and accurately represents the original features.

It's important to note that while augmentation increases the size of the training dataset, it does not alter the balance between forest and deforested areas. This is because the patches have been filtered to include only those containing deforested pixels, ensuring that the ratio of forest to deforested areas remains constant throughout the augmented dataset. I tackle this data imbalance problem and it's solution in section 4.3.1.

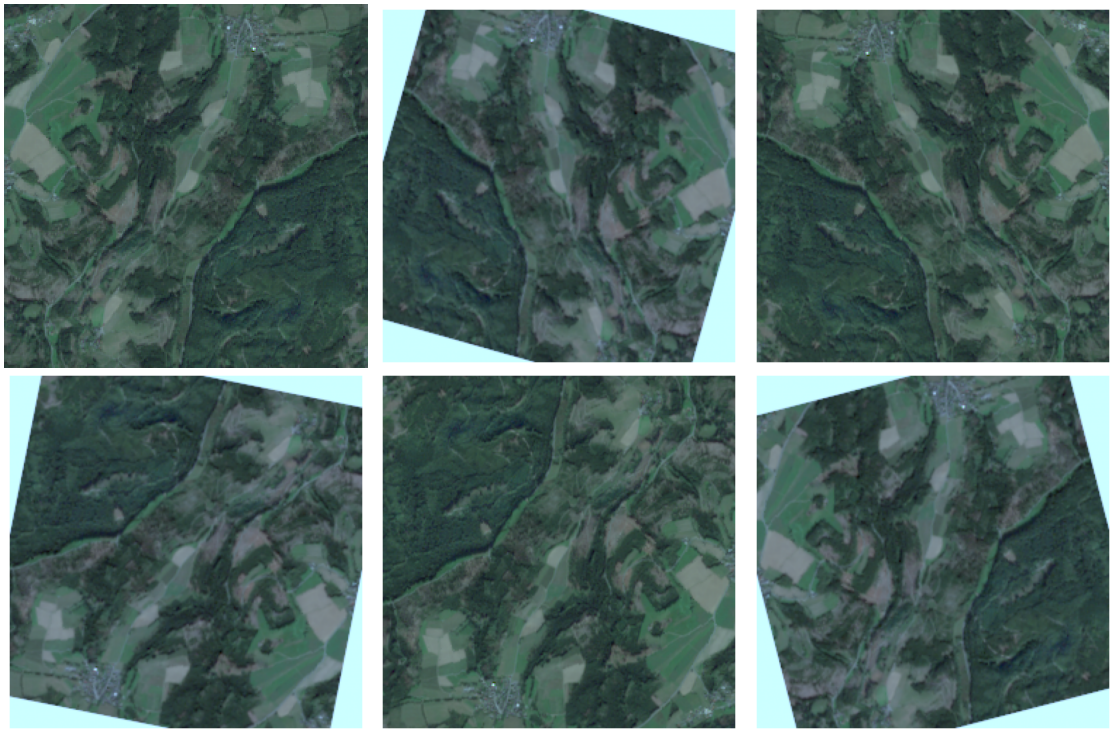

Figure 4.3.: Input image and five subsequent augmentations using only geometric transformations so as to preserve spectral information of the bands

4.3. Training Strategies

The training process in this thesis was designed with dual objectives: optimising the segmentation performance for identifying deforested regions and rigorously addressing the unique challenges posed by multispectral satellite imagery and limited annotations.

The first objective addresses the first research question of working with weak annotations, which is incomplete or inaccurate in our case. Initially, the problem was approached as a supervised learning task (considering the weak labels as groundtruth) to establish a baseline understanding of the data and model performance. This step was crucial before introducing weakly supervised techniques, allowing for a clearer analysis of how these methods could be effectively applied. The rationale behind this approach will be discussed in detail in a later section.

The second objective centres on assessing the value and necessity of the additional spectral information provided by multispectral data in comparison to conventional RGB inputs. A fundamental question arises, the basis for our second and third research question: can typical architectures designed for RGB images effectively adapt to the unique characteristics of multispectral imagery? To explore this, the training process involved varying the combinations of spectral bands used as input. This methodology aimed to identify which, if any, of the extra bands significantly enhance segmentation performance.

Furthermore, it is crucial to investigate whether the models' performance corroborates the established knowledge in the field of remote sensing regarding the unique characteristics of different spectral bands. In remote sensing, each spectral band corresponds to specific wavelengths that capture distinct reflectance characteristics of various land covers, including vegetation, soil, and water. For example, the Near-Infrared (NIR) band is particularly sensitive to vegetation health and biomass, facilitating the differentiation between healthy and stressed vegetation. This insight into the relevance of multispectral data for this task underscores its potential advantages over traditional RGB imagery.

In the subsequent sections, I will delve into the methodologies used, starting with the supervised approach, followed by the weakly supervised approach, and finally, examining the impact of spectral band variation. These discussions will provide a comprehensive analysis aimed at answering the research questions posed.

4.3.1. Supervised Approach

To address the challenge of segmenting multispectral satellite imagery, we leveraged existing, well-established segmentation architectures: U-Net and DeepLabv3. Both models were selected after extensive research and careful consideration due to their proven effectiveness in various segmentation tasks, including those related to remote sensing and land cover classification.

U-Net is a CNN architecture that excels in image segmentation tasks, where the goal is to classify each pixel of an image into a specific category [1]. The architecture is structured around a symmetric encoder-decoder framework as shown in Figure 4.4. The encoder, or contracting path, gradually reduces the spatial dimensions of the input image while increasing the depth of the feature maps. This process allows the network to capture essential contextual information by applying a series of convolutional layers followed by max-pooling, which distils the input into increasingly abstract features.

At the core of U-Net lies the bottleneck, which represents the most compressed version of the input, containing the key features needed for accurate reconstruction. The decoder, or expansive path, then reverses this process by upsampling the feature maps. What sets U-Net apart is the use of skip connections, which link corresponding layers in the encoder and decoder. These connections help preserve spatial details that might otherwise be lost during downsampling, allowing the network to combine both high-level and fine-grained information for precise segmentation.

During training, U-Net optimises a loss function that compares its predictions to the

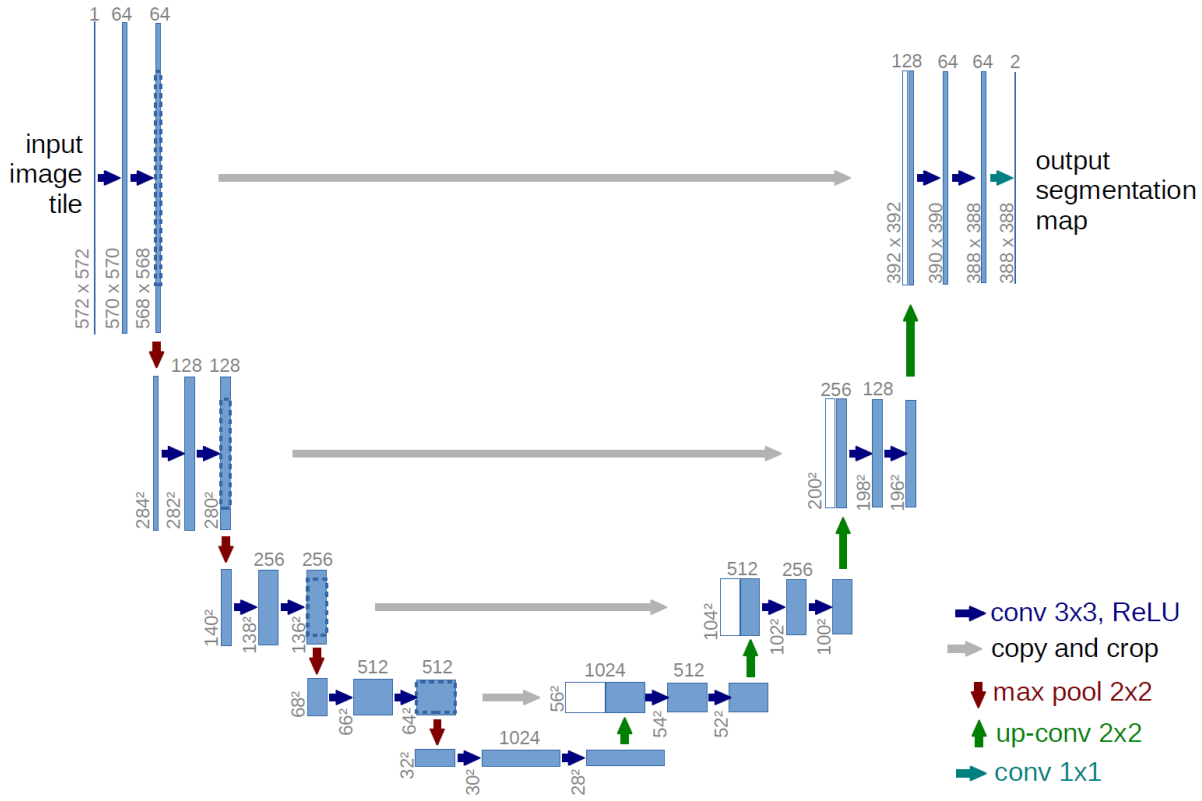

Figure 4.4.: A standard U-Net architecture

ground truth, gradually refining its ability to classify pixels correctly. This learning process, powered by backpropagation, adjusts the network's weights to minimise errors.

While originally developed for biomedical image segmentation, U-Net's ability to maintain spatial accuracy has made it particularly useful in remote sensing, where it can effectively differentiate between various land cover types in high-resolution satellite images.

Similarly, DeepLabv3 is another advanced CNN architecture designed for semantic segmentation, with a focus on capturing detailed contextual information while maintaining spatial resolution [2]. Its distinguishing feature is the use of atrous (or dilated) convolutions, which allow the network to expand the receptive field without reducing the resolution of the feature maps. This capability enables DeepLabv3 to analyze multi-scale information, making it adept at handling complex scenes where objects vary in size and context is crucial.

A key component of DeepLabv3 is the Atrous Spatial Pyramid Pooling (ASPP) module, which applies atrous convolutions with varying dilation rates. This technique captures information at multiple scales simultaneously, enriching the network's understanding of the scene. The ASPP module allows DeepLabv3 to effectively balance detail and context, making it highly versatile for different segmentation tasks.

The network also employs an encoder-decoder structure, though its emphasis is more on the encoder's ability to process and pool information across scales. The final output layer generates a segmentation map by assigning each pixel a probability of belonging to a specific class, which the network refines through a loss function during training.

DeepLabv3's architecture makes it particularly suitable for remote sensing applications, where it can efficiently segment diverse land cover types even in complex and varied landscapes. Its ability to maintain detail while understanding broader spatial contexts makes it a powerful tool for analysing satellite imagery.

Adapting these architectures to handle multispectral data involved several modifications. For U-Net, the primary adaptation was altering the input layer to accommodate the

increased number of spectral bands. While the original U-Net is designed for three-channel RGB images, our multispectral imagery includes additional bands, necessitating a modification to the input channels to fully utilise the available spectral information. This adjustment was crucial for ensuring that the model could learn from all relevant data rather than being constrained to RGB inputs alone.

Similarly, DeepLabv3 was adapted to handle the multispectral data by modifying its input layer. Additionally, for both architectures, the number of output classes was tailored to match the specific categories within our dataset. This step ensured that the models were aligned with the unique characteristics of our data, allowing for more accurate segmentation.

To evaluate the effectiveness of these adaptations, I conducted experiments comparing the performance of the modified U-Net and DeepLabv3 on both RGB and multispectral datasets. I tested the models' ability to accurately segment deforested regions using only the RGB bands and then assessed whether the inclusion of additional spectral bands improved performance. This approach provided insights into the relevance of multispectral data and helped identify the most significant spectral bands for this specific task.

Loss Function

In supervised learning, the choice of loss function is crucial as it defines the objective that the model strives to optimise. The cross-entropy loss is commonly used in classification tasks due to its effectiveness in measuring the difference between predicted and true class probabilities. For a single instance, the cross-entropy loss can be expressed as:

$$\text{Cross Entropy Loss} = -\log(p_t) \tag{4.1}$$

where p_t is the predicted probability of the true class t.

Initially, I selected cross entropy loss as the criterion, however, I came to realise that it does not take into consideration an imbalanced dataset which leads to suboptimal

performance. As stated earlier, the dataset used in this research exhibits substantial imbalance with forested areas vastly outnumbering deforested ones (0.915 : 0.085). This imbalance prompted the adoption of the focal loss function, a modified version of cross-entropy loss designed to address such issues.

Focal loss introduces a scaling factor to the cross-entropy loss, focusing more on hard-to-classify examples [3]. The focal loss is given by:

$$\text{Focal Loss} = -\alpha_t(1 - p_t)^\gamma \log(p_t) \tag{4.2}$$

Here, α_t is a weighting factor that adjusts the loss contribution of each class, and γ is a focusing parameter that controls the rate at which easy examples are down-weighted. The term $(1 - p_t)^\gamma$ reduces the loss for well-classified examples (where p_t is close to 1) and puts more emphasis on hard-to-classify examples (where p_t is closer to 0).

This adjustment is particularly beneficial in our case, where the deforested class is under-represented. By emphasising the learning of these difficult, minority class examples, focal loss improves the model's accuracy in detecting deforested areas, which is critical for this task.

Furthermore, during training, it is necessary to handle invalid pixels, such as those belonging to the anthropogenic class, which could introduce noise into the learning process. These pixels are masked out with a weight of zero in the loss calculations, ensuring that irrelevant areas did not influence the model's learning. This focused approach helps to improve the segmentation quality of the relevant classes.

To further address class imbalance, specific class weights are assigned according to the dataset distribution, significantly increasing the weight for the underrepresented deforested class. This ensures that the model penalises misclassifications of deforested areas more heavily, thereby encouraging better performance in predicting these regions. Overall, the combination of focal loss and class weighting significantly enhances model's performance and solves the problem of data under-representation.

4.3.2. Weakly Supervised and Self-Supervised Learning Approaches

In the realm of semantic segmentation, the availability of accurate labels is often a significant challenge. In this research, the dataset includes two types of weakly supervised annotations: inaccurate and incomplete labels. Traditional supervised learning methods rely heavily on fully annotated datasets where each pixel in the image is correctly labelled as seen in section 4.3.1. However, in many real-world scenarios, like the one addressed in this research, labels can be incomplete, inaccurate, or noisy. This issue is particularly prevalent in large-scale remote sensing applications, where manual pixel-wise annotation is time-consuming and prone to error. To address these challenges, weakly supervised learning (WSL) techniques are employed to effectively leverage the imperfect data for both inaccuracies and incompleteness.

In this study, the primary focus is on localising deforested areas within the forest, and here I would address the issue of inaccurate weak annotation in the dataset. Typically, deforested pixels are in clusters, meaning that accurately identifying every pixel in these regions is not as critical as ensuring proper localisation. The goal is not to match real-world features pixel by pixel but to identify the general extent of deforestation. However, this aspect directly impacts the loss calculations during model training. To mitigate the effects of inaccuracies in the labelling, a reweighting strategy was implemented in the loss function. Pixels confidently classified as deforested—those within the core of the deforested regions—are assigned a weight of 1, indicating high confidence. Conversely, outer or boundary pixels, carry a higher degree of uncertainty, are assigned a weight of 0.5. This reweighting approach allows the model to focus more on the high-confidence pixels, thus improving the overall learning process and ensuring that the localisation of deforested areas is both effective and robust.

Incomplete strategy...

4.3.3. Spectral Bands Variation

Given the unique capabilities of each spectral band in multispectral satellite imagery, I explore how different combinations of these bands affect the performance and generalisation of segmentation models. The third research question centred around understanding whether the inclusion of additional spectral bands, beyond the conventional RGB, improves the accuracy in distinguishing between deforested and forested regions.

From a remote sensing perspective, each spectral band offers distinct and complementary information crucial for land cover classification. For example, the RGB bands provide a basic visual representation of the Earth's surface, capturing general colour and texture. In contrast, NIR band is known for its sensitivity to chlorophyll, making it particularly effective for vegetation analysis. Other bands, such as SWIR and Red Edge, offer insights into soil moisture, vegetation health, and other surface properties that are not visible in the RGB spectrum.

To assess the impact of these bands on segmentation performance, we designed a series of experiments where models were trained using different combinations of spectral bands. The selected combinations includes:

- **RGB Bands**: The baseline input using only the Red, Green, and Blue bands, which capture the general visual appearance of the land surface.
- **RGB + NIR Bands**: Incorporating the NIR band to enhance the model's ability to detect and differentiate vegetation, leveraging the band's sensitivity to chlorophyll content.
- **RGB + SWIR Bands**: Adding the SWIR band, which is particularly useful for distinguishing between different soil types and moisture levels, and for identifying water bodies.
- **RGB + Red Edge Bands**: Including the Red Edge band, which is effective for monitoring vegetation stress and health, as it captures the transition between visible

red light and NIR.

- **All Invisible Bands**: Utilising only the NIR, SWIR, and Red Edge bands, omitting the RGB channels, to investigate how these non-visible wavelengths contribute to land cover classification.
- **All spectral bands**: Using the full spectrum of available bands from the Sentinel-2 imagery to maximise the information provided to the model.

The underlying hypothesis was that the inclusion of additional spectral bands would lead to more accurate segmentation results, given that different types of land cover have distinct spectral signatures. For example, dense forests reflect significantly more in the NIR and absorb more in the red bands, while water bodies absorb more in both the NIR and SWIR bands, appearing dark in these wavelengths. Similarly, soils typically have higher reflectance in the red and SWIR bands, which can help distinguish them from vegetation and water as shown in Figure 4.5.

By training and evaluating models across these different band combinations, the study aims to determine whether using the full range of available spectral data would yield superior segmentation results or if a more targeted selection of bands could provide a balance between accuracy and computational efficiency. This analysis also provids insights into which spectral bands are most critical for effective land cover segmentation in this context, potentially informing future remote sensing applications and model design.

4.3.4. General Training Setup

The training process employs the conventional parameters used in training a neural network. For optimisation, I utilise the Adam optimiser, which is renowned for its effectiveness in managing large datasets and its rapid convergence properties [4]. Initially, stochastic gradient descent (SGD) was tested; however, after further research, the Adam optimiser was adopted, leading to significant performance improvements in the models.

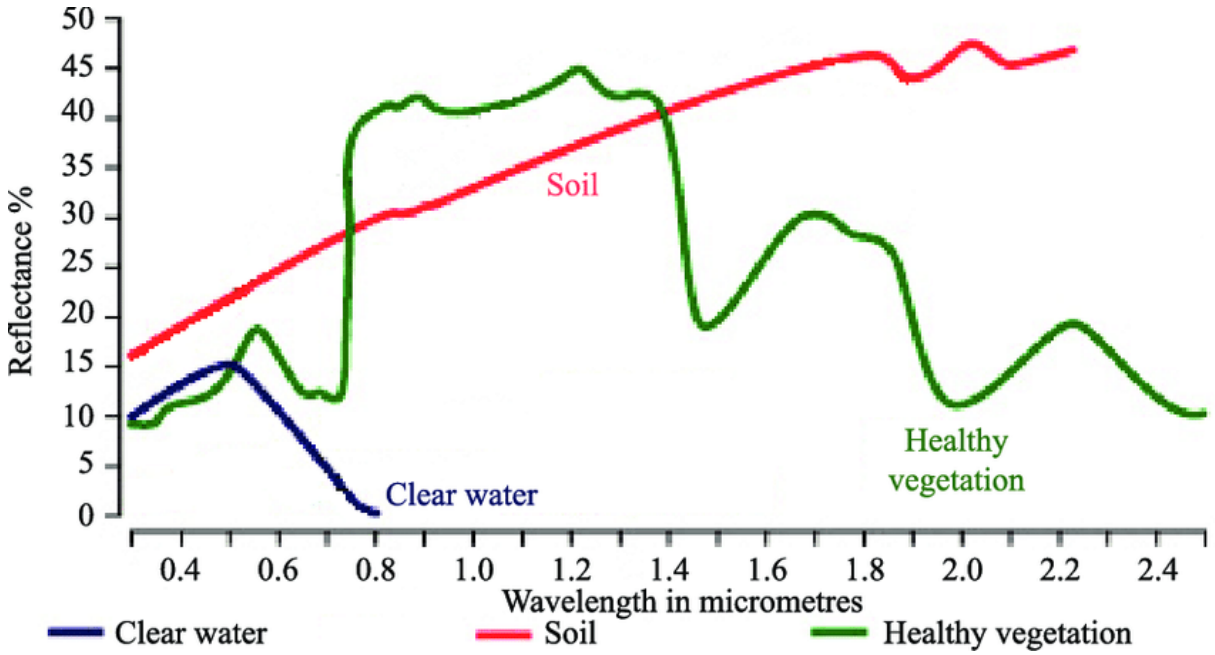

Figure 4.5.: Spectral reflectance of Earth surfaces: vegetation, soil, water

Adam combines the benefits of various extensions of SGD by adaptively adjusting the learning rate for each parameter, which is especially advantageous when handling sparse gradients.

For the training setup, an initial learning rate was selected to achieve a balance between rapid convergence and stability, minimising the risk of overshooting optimal weights during the early training phases. Additionally, a learning rate scheduler was implemented, gradually decreasing the learning rate over epochs. This approach allows for aggressive learning at the beginning of training while enabling finer adjustments in later stages. The scheduler plays a crucial role in preventing the model from becoming trapped in local minima, thereby enhancing convergence towards a global minimum. Such strategies are particularly beneficial in segmentation tasks, where precise pixel-wise predictions are essential. The model was trained with a specific batch size and for a predetermined number of epochs to ensure a comprehensive exploration of the parameter space and to achieve robust performance.

4.3.5. Conclusion

This research presents a comprehensive approach to detecting deforested areas. The focus is on localising deforested regions, recognising that precise pixel-level accuracy is less critical than accurately identifying clusters of deforestation. The methodology begins with the processing of satellite imagery, employing data augmentation and preprocessing steps to enhance dataset quality. It details the supervised and weakly supervised learning strategies implemented to address the problem, followed by a description of the training setup and parameters used for neural network training. As we proceed, the next chapter will outline the experiments conducted, present the results obtained, and discuss their implications, thereby highlighting the effectiveness of the proposed methodologies for advancing deforestation detection research.

5. Experiments

5.1. Experimental Setup

The parameters used in the experiments, summarised in Figure 5.1, were determined through a process of trial and error combined with careful consideration of the model's performance. These models were trained on a GPU with the following specifications:

Parameter	Value
Optimiser	Adam
Criterion	Focal loss
Learning rate	0.001
Maximum epochs	600
Batch size	512

Table 5.1.: Training Parameters

5.2. Experimental Results and Discussion

In the previous chapter, we mentioned that two segmentation masks were generated: a simple one with three classes and a more complex one with six classes. We performed experiments using both masks and also varying the spectral bands. It is to find the answer to the third research question of whether all the Sentinel-2 spectral bands are really required in training to produce an efficient model.

In this research, we are trying to differentiate regions of trees from regions with little to no trees. Based on domain knowledge from the field of remote sensing, it is understood that

different spectral bands provide unique and complementary information that can significantly enhance land cover classification. For instance, the visible bands (Red, Green, and Blue) capture the general appearance of the land surface, while the Near Infrared (NIR) band is particularly effective for vegetation analysis due to its sensitivity to chlorophyll content. The Shortwave Infrared (SWIR) bands, though not used in our study due to their lower resolution, are known for their utility in assessing soil and moisture content.

Supervised

The results presented in Table 5.2 offer insight into the impact of different spectral band combinations on the performance of the U-Net model for the binary semantic segmentation of forest and deforested areas.

Spectral Bands	**IoU (%)**		
	Deforested	**Forest**	**Overall**
RGB	95.3	99.5	97.4
RGB + S	94.4	99.4	96.9
RGB + NIR	96.7	99.6	98.2
RGB + NIR + S	94.3	99.4	96.9
RGB + Red Edge	96.6	99.7	98.1
RGB + Red Edge + S	96.5	99.7	98.0
NIR + Red Edge + SWIR	96.6	99.7	98.2
NIR + Red Edge + SWIR + S	95.0	99.5	97.3
All Bands	96.7	99.7	98.2
All Bands + S	96.9	99.7	98.3

Table 5.2.: IoU validation scores for various spectral bands combination trained on a binary segmentation dataset using U-Net architecture. RGB indicates red, green, blue; NIR, near infrared; SWIR, shortwave infrared; S the seasonal Band.

The baseline model using only the conventional RGB bands achieved a high overall IoU of 97.4%, with class-specific IoUs of 95.3% for deforested and 99.5% for the forest classes. When additional spectral bands were incorporated, some combinations resulted in marginal improvements, while others did not significantly enhance or even slightly

reduced the model's performance.

For example, the inclusion of the NIR band with RGB led to an increase in overall IoU to 98.2%, suggesting that the NIR band provides valuable information that enhances the model's ability to differentiate between forested and deforested areas. However, adding the seasonal band (S) to this combination did not result in further improvement. This indicates that while NIR is beneficial, the seasonal band might introduce variability that does not always translate into better segmentation performance.

Contrary to the hypothesis that incorporating seasonal data would improve the model's performance, the results show that the seasonal band did not consistently enhance segmentation accuracy. In most cases, the addition of the seasonal band had a negligible effect. This outcome suggests that the seasonal variation, rather than providing additional discriminative features, may introduce noise or redundancy that the model cannot effectively leverage. Therefore, the temporal aspect of the data does not seem to offer a significant advantage in this context, contrary to the initial hypothesis.

Overall, while the inclusion of certain spectral bands, particularly NIR and Red Edge, does enhance the model's performance, the improvements are modest. The seasonal data, expected to be beneficial, does not contribute significantly to model accuracy and in some cases may slightly detract from it. These findings indicate that for the task of forest and deforestation segmentation using Sentinel-2 imagery, conventional RGB data, could be enough in getting a robust performance.

For the multiclass segmentation, the results are shown in Table 5.3

Weakly Supervised

Spectral Bands	IoU (%)					
	Soil	**Low grass**	**High grass**	**Sparse trees**	**Forest**	**Overall**
RGB				–	–	–
RGB + S				–	–	–
R + NIR				–	–	–
R + NIR + S				–	–	–
NIR + Red Edge				–	–	–
NIR + Red Edge + S				–	–	–
NIR + SWIR				–	–	–
NIR + SWIR + S				–	–	–
All Bands				–	–	–
All Bands + S				–	–	–

Table 5.3.: IoU validation scores for various spectral bands combinations trained on a multiclass segmentation dataset using U-Net architecture. RGB indicates Red, Green, Blue; NIR, Near infrared; SWIR, shortwave infrared; S, the seasonal Band.

6. Conclusion

6.1. Limitations

6.2. Conclusion

6.3. Future Research Direction

7. Summary and Outlook

A. Bibliography

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[2] L. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *CoRR*, vol. abs/1706.05587, 2017.

[3] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *CoRR*, vol. abs/1708.02002, 2017.

[4] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.