DEFORESTATION DETECTION IN SAR IMAGES USING DEEP NEURAL NETWORKS

IGOR BISPO DE MORAES COELHO CORREIA

DISSERTAÇÃO DE MESTRADO
EM ENGENHARIA ELÉTRICA

DEPARTAMENTO DE ENGENHARIA ELÉTRICA

FACULDADE DE TECNOLOGIA

UNIVERSIDADE DE BRASÍLIA

Universidade de Brasília

Faculdade de Tecnologia

Departamento de Engenharia Elétrica

Deforestation Detection in SAR Images Using Deep Neural Networks

Detecção de Desmatamento em Imagens SAR usando Redes Neurais Profundas.

Igor Bispo de Moraes Coelho Correia

DISSERTAÇÃO DE MESTRADO SUBMETIDA AO PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA DA UNIVERSIDADE DE BRASÍLIA COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE.

APROVADA POR:

Mylène Q. C. Farias, Dra. (Universidade de Brasília)
(Orientadora)

Daniel Guerreiro e Silva, Dr. (Universidade de Brasília)
(Examinador Interno)

Eduardo Peixoto, Dr. (Universidade de Brasília)
(Examinador Interno)

Hélcio Vieira Junior, Dr.
(Instituto Tecnológico de Aeronáutica)

Brasília/DF, Maio de 2024.

FICHA CATALOGRÁFICA

B. M. C. CORREIA, IGOR

Deforestation Detection in SAR Images Using Deep Neural Networks. [Brasília/DF] 2024.

xix, 70p., 210 x 297 mm (ENE/FT/UnB, Mestre, Dissertação de Mestrado, 2024).

Universidade de Brasília, Faculdade de Tecnologia, Departamento de Engenharia Elétrica.

Departamento de Engenharia Elétrica

1. Deforestation detection	2. SAR images
3. Data fusion	4. Deep learning
5. Sentinel-1	6. Sentinel-2
7. UNet	8. Environmental monitoring
I. ENE/FT/UnB	II. Title (series)

REFERÊNCIA BIBLIOGRÁFICA

B. M. C. CORREIA, IGOR (2024). Deforestation Detection in SAR Images Using Deep Neural Networks. Dissertação de Mestrado, Publicação PPGEE.DM-816/2024, Departamento de Engenharia Elétrica, Universidade de Brasília, Brasília, DF, 70p.

CESSÃO DE DIREITOS

AUTOR: Igor B. M. C. Correia

TÍTULO: Deforestation Detection in SAR Images Using Deep Neural Networks

GRAU: Mestre ANO: 2024

Igor B. M. C. Correia
Universidade de Brasília (UnB)
Campus Darcy Ribeiro
Faculdade de Tecnologia - FT
Departamento de Engenharia Elétrica(ENE)
Brasília - DF CEP 70919-970

I dedicate this work to everyone who helped me complete another important milestone in my life. From classmates, teachers, advisors to people outside the academy like my friends and family.

ACKNOWLEDGEMENTS

I dedicate this work to everyone who helped me complete another important milestone in my life. From classmates, teachers, advisors to people outside the academy like my friends and family.

ABSTRACT

Deforestation has broad and significant impacts, becoming a major global environmental threat. This problem endangers biodiversity, contributes to climate change, and compromises the sustainability of natural ecosystems. Therefore, monitoring and detecting deforested areas is a critical issue. Although deforestation affects many regions, the Amazon rainforest is one of the most prominent and frequently discussed cases.

Optical satellites are extremely powerful and important tools in remote sensing, prevention, and mitigation of deforestation. However, this type of sensor is not robust to climatic variations and is sensitive to cloud occlusion. On the other hand, synthetic aperture radars (SARs) stand out for their resilience to adverse weather conditions, as they are active sensors that operate in microwave bands which penetrate water particles. Nonetheless, accurately recognizing deforested areas in SAR images is challenging due to the amount of speckle noise and the variability in the appearance of objects between different captures.

In this study, we conducted an online experiment with voluntary participants who identified deforested areas in SAR images. For this purpose, we developed software that allows participants to annotate SAR images, delineating deforested areas. With the results of this experiment, it was possible to analyze the relationship between the participants' self-reported experience level and the accuracy in detecting deforested areas.

We also compared human performance with the performance obtained from an automatic model based on the UNet architecture. The results show that greater knowledge in remote sensing or SAR does not guarantee quality annotations. Moreover, the UNet's performance surpasses human performance in the task.

To explore SAR image segmentation in greater depth, a second experiment was conducted using state-of-the-art models for segmentation in fused SAR and optical data. This second part of the study showed that the most modern models, despite having a smaller number of trainable parameters, can outperform a heavier model. The study reinforces the potential of deep learning in deforestation detection, emphasizing the need for continuous improvements in architectures and specialist training.

Keywords: SAR, optical, data, fusion, convolutional, neural, network, transformer, unet, remote, sensing

RESUMO

O desmatamento tem impactos amplos e significativos, tornando-se uma grande ameaça ambiental global. Esse problema coloca em risco a biodiversidade, contribui para as mudanças climáticas e compromete a sustentabilidade dos ecossistemas naturais. Portanto, monitorar e detectar áreas desmatadas é uma questão crítica. Embora o desmatamento afete muitas regiões, a floresta amazônica é um dos casos mais destacados e frequentemente discutidos.

Os satélite ópticos são ferramentas extremamente poderosas e importantes no sensoriamento remoto, prevenção e mitigação de desmatamento. Entretanto, esse tipo de sensor não é robusto a variações climáticas e é sensível a oclusão por nuvens, em contrapartida, os radares de abertura sintética (SARs) destacam-se por sua resistência a condições climáticas adversas visto que são sensores ativos que operam em faixas de microondas as quais atravessam particulas de água. No entanto, o reconhecimento preciso de áreas desmatadas em imagens SAR é desafiador devido à quantidade de ruído speckle e à variabilidade de aparência dos objetos entre capturas diferentes.

Neste trabalho, realizamos um experimento online com participantes voluntários que identificaram áreas desmatadas em imagens SAR. Para isso, desenvolvemos um software que permite aos participantes anotarem as imagens SAR, delimitando áreas desmatadas. Com os resultados desse experimento, foi possível analisar a relação entre o nível de experiência autodeclarado dos participantes e a precisão na detecção de áreas desmatadas.

Também comparamos o desempenho humano com o desempenho obtido com um modelo automático baseado na arquitetura UNet. Os resultados mostram que maior conhecimento em sensoriamento remoto ou SAR não garante qualidade nas anotações. Além disso, o desempenho do UNet supera o desempenho humano na tarefa.

Para explorar a segmentação de imagens SAR com mais profundidade, um segundo experimento foi realizado utilizando modelos de última geração para segmentação em dados SAR e ópticos fundidos. Essa segunda parte do estudo mostrou que os modelos mais modernos, apesar de disporem de uma menor quantidade de parâmetros treináveis, pode sobrepujar um modelo mais pesado. O estudo reforça o potencial do aprendizado profundo na detecção de desmatamento, enfatizando a necessidade de melhorias contínuas nas arquiteturas e no treinamento de especialistas.

Palavras-chave: SAR, óptico, fusão, rede neural convolucional, transformer, unet, sensoriamento remoto

CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

ANN	Artificial Neural Network
DNN	Deep Neural Network
SAR	Synthetic Aperture Radar
INPE	Instituto Nacional de Pesquisas Espaciais
PRODES	Projeto de Monitoramento do Desmatamento na Amazônia Legal por Satélite
VV	Vertical Transmit-Vertical Receive Polarisation
VH	Vertical Transmit-Horizontal Receive Polarisation
IoU	Intersect over Union
PPCDAm	Plano de Ação para Prevenção e Controle do Desmatamento na Amazônia Legal (PPCDAm)
NASA	National Aeronautics and Space Administration
CBERS	China-Brazil Earth-Resources Satellite
CENSIPAM	Centro Gestor e Operacional do Sistema de Proteção da Amazônia
GRD	Ground Range Detected
LSTM	Long short-term memory
ViTs	Vision Transformers
NLP	Natural Language Processing
GCPs	Ground Control Points
RGB	Red Green Blue

CHAPTER 1

INTRODUCTION

Deforestation has multiple and comprehensive impacts, becoming an increasingly worrying environmental issue throughout the world. In addition to posing a threat to biodiversity, deforestation contributes significantly to climate change (NOBRE *et al.*, 2007) and compromises the sustainability of natural ecosystems(VIEIRA *et al.*, 2014). Therefore, monitoring and detecting deforested areas is one of the most urgent and relevant problems of today. Although many regions of the world face deforestation issues (CLEMENT *et al.*, 2015; FERNANDES *et al.*, 2023), the Amazon forest is one of the most debated and cited in this context.

Founded in 1961, the National Institute for Space Research (INPE)[1] was created to tackle these challenges within Brazil, amidst a worldwide growth in space research and the country's pursuit to enhance its own expertise in this field. INPE was established to achieve positive impacts on ecology objectives, including the development of technology in the space sector, conducting scientific research, monitoring the environment, forming specialized human resources, and fostering international cooperation.

In addition to that, INPE plays a crucial role in collecting data that support the formulation of public policies, especially on environmental and land use issues. Its contributions are vital for understanding and addressing challenges such as deforestation, climate change, and natural disasters. The institute's advanced technologies and methodologies provide essential insight and data for policymakers and researchers.

INPE stands out in Latin America for its technical capacity and innovations in satellites and other observation technologies that allow detailed monitoring of vast areas such as the Amazon. The institute's efforts not only contribute to the global scientific community, but also enhance Brazil's capabilities in managing and protecting its natural resources, ensuring sustainable development and national sovereignty.

[1] Official INPE website: <https://www.gov.br/inpe/pt-br>, visited in May 2024

One of the main INPE initiatives in the context of remote sensing in the Amazon is the Amazon Satellite Monitoring Program (PRODES)[2], established in 1988. PRODES was specifically developed to monitor the annual deforestation rate by clear-cutting in the Legal Amazon using satellite images to detect changes in forest coverage in areas larger than 6.25 hectares.

Clear-cutting refers to a method of deforestation in which all trees in a designated area are uniformly cut down, leaving the land completely devoid of forest cover. This practice is often used for agriculture, logging, or other land development purposes, and results in significant environmental impacts, such as loss of biodiversity, disrupting ecosystems, and changes in local climate patterns (MARQUES *et al.*, 2019).

The PRODES program is fundamental to the Brazilian government's position in discussions on climate change under the United Nations Framework Convention on Climate Change. The numbers of annual deforestation, according to INPE, are shown in Figure 1.1.

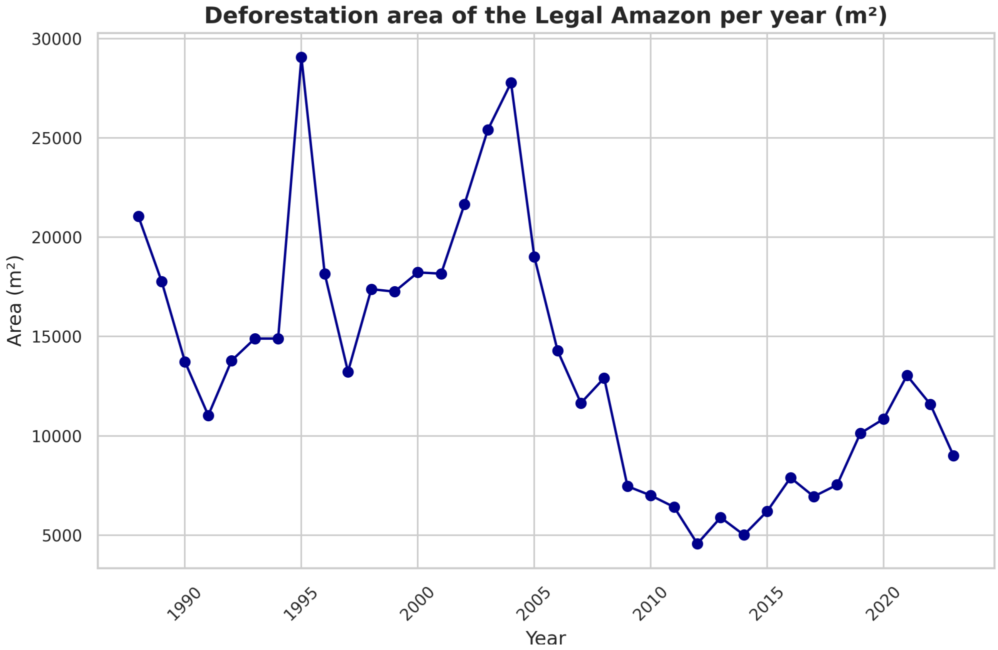


Figure 1.1. Annual deforestation area of the Legal Amazon as measured in square meters. Data obtained from PRODES (Brazilian Amazon Forest Satellite Monitoring Program), a program developed by the Brazilian National Institute for Space Research (INPE) to track deforestation using satellite images.

PRODES is part of the Action Plan for the Prevention and Control of Deforestation in the Legal Amazon (PPCDAM), coordinated by the Ministry of Environment and the Civil House of the Presidency of the Republic. This plan includes a series of policies and strategic actions

[2] Official PRODES website: <http://www.obt.inpe.br/OBT/assuntos/programas/amazonia/prodes>

aimed at reducing deforestation and promoting the sustainable use of natural resources in the region.

In addition to producing annual deforestation rates, PRODES also provides images, vector maps, and tables that detail these forest losses exclusively by clear-cutting, significantly contributing to the actions of environmental monitoring and control, and serving as a valuable tool for academic research and public awareness.

It should be noted that the discussion on the deforestation of the Amazon rain forest is a highly politically charged topic. Recent governments, from 1999 to 2020, have had mixed positions on this issue (FONSECA *et al.*, 2022). There was a strengthening of regulatory institutional capacities between 1999 and 2012, followed by active dismantling starting in 2013, which dramatically intensified from 2019 onward. This latter period witnessed a significant reduction in the density of forest conservation policies and an increase in deforestation rates.

From a technical point of view, PRODES uses a combination of several satellites to map the Amazon rain forest. The primary satellite used is Landsat (NASA, 2023), operated by NASA and the United States Geological Survey, which captures images of the Amazon every 16 days with a resolution of 30 meters. In addition to Landsat, PRODES also relies on images from the CBERS satellites [3] (a partnership between Brazil and China) and Sentinel (from the European Space Agency) to supplement information in areas where there is cloud cover that Landsat cannot penetrate.

Figure 1.2 shows an image acquired with synthetic aperture radars (SARs) and an image acquired with optical satellites. Optical satellites record spectral bands within the visible spectrum and also have some capabilities beyond this range. Images that incorporate data from three or more different spectral bands are known as multispectral images. The images obtained by optical satellites are heavily affected by weather conditions, such as heavy rain, cloud cover, and fog, which affect the visibility of areas in the acquired images. Unlike optical satellites, SARs have active sensors that emit microwave pulses at a typical operating frequency that varies from hundreds of megahertz (MHz) to several gigahertz (GHz) (GEUDTNER *et al.*, 2014). These microwave pulses are directed towards the Earth's surface, and then the SAR records the return of these pulses, reflected by objects on the surface. Based on the time difference

[3] <http://www.cbers.inpe.br/>

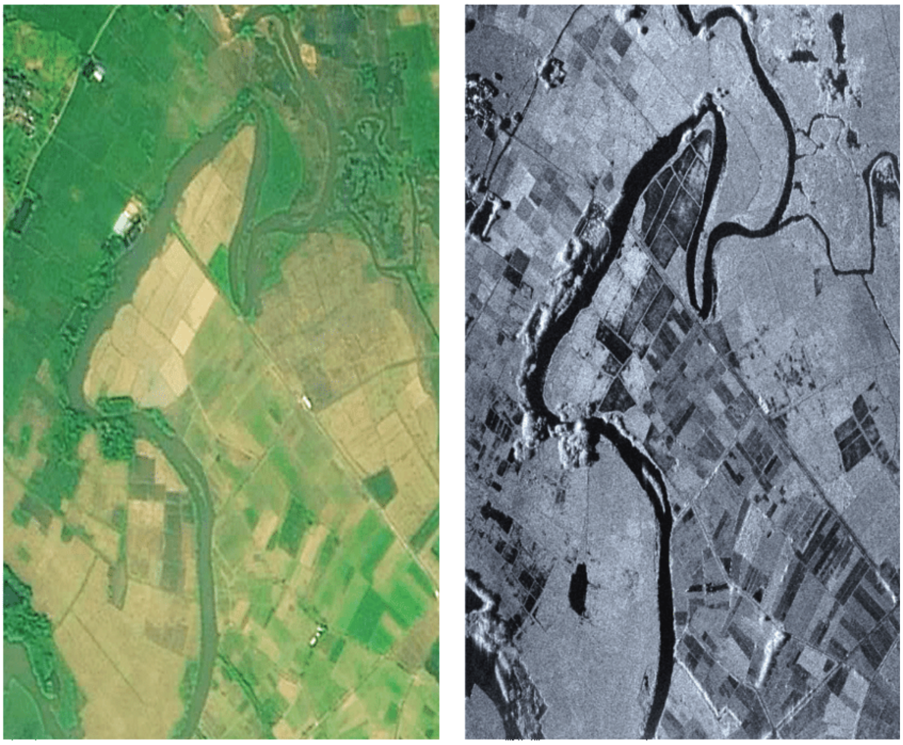


Figure 1.2. Comparison between a SAR image (on the right) and an optical image (on the left) of the same area. Source: *Very-High-Resolution SAR Imaging with DGPS-Supported Airborne X-Band Data*(ZHOU *et al.*, 2020)

between when the pulses are emitted and when the reflected pulses return to the radar, SAR creates images, using the synthetic aperture radar interferometry (InSAR) technique (YAGÜE-MARTÍNEZ *et al.*, 2016) to achieve high spatial resolution. The InSAR technique combines and processes the various microwave pulses received at different radar positions along the satellite's movement. The waves emitted have the ability to penetrate most cloud covers, allowing the radar to gather information even under adverse atmospheric conditions.

Therefore, SAR images are valuable for detecting deforested areas in tropical regions, such as the Amazon, where the presence of clouds and fog is frequent. Despite the advantage of weather independence, the accurate detection of deforested areas in SAR images is a challenging task due to the high amount of speckle noise in the captured images and the variability in the appearance of objects. The speckle noise is caused by the random interference of coherent returns from multiple scatterers present on the Earth's surface at the scale of the radar's wavelength (SINGH; SHREE, 2016). The speckle manifests as a grainy salt-and-pepper pattern in SAR images, complicating accurate image interpretation.

Figure 1.3 shows a close-up of a deforested region seen in an optical image, Sentinel-2, and

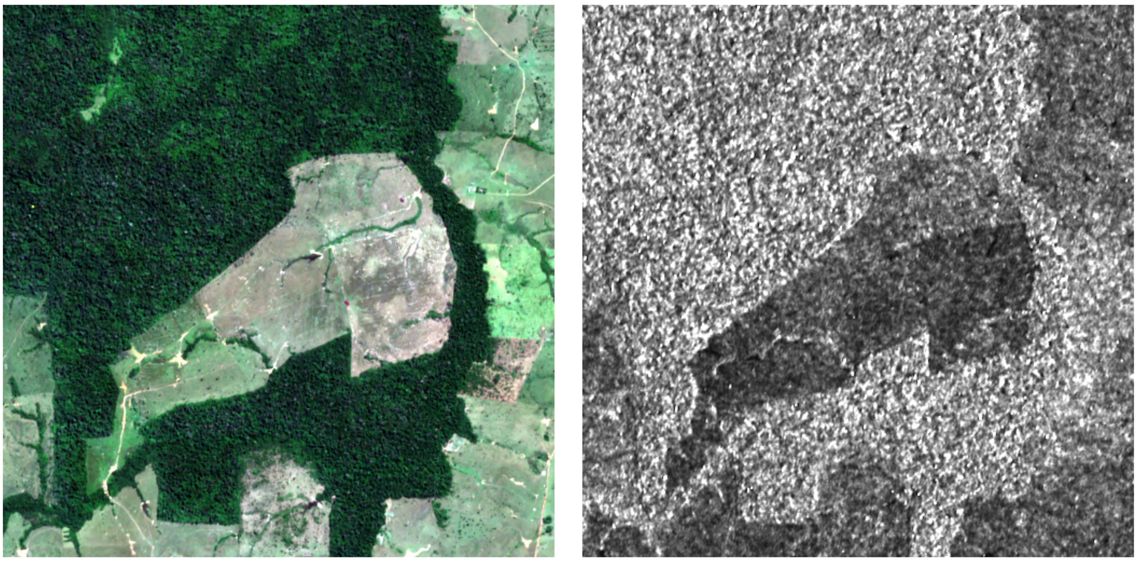


Figure 1.3. The same deforested region seen by an optical and a SAR sensor.

a SAR image, Sentinel-1. In the SAR image, it is possible to see a large amount of noise represented as a granulation of very bright and dark pixels randomly distributed throughout the image.

The manual process of identifying deforestation in satellite images by professionals from INPE and CENSIPAM is laborious and intricate. The INPE PRODES project relies heavily on the manual analysis of numerous Landsat TM images. Analysts face significant challenges during this process, including variable scales of different scenes and the complexity of closing polygons on interpretation maps due to intricate patterns of deforestation. These difficulties require a considerable amount of manual effort to convert the original Landsat bands into vegetation, soil, and shade fraction images, which are then segmented and classified manually to generate the final deforestation maps (SHIMABUKURO *et al.*, 2000).

Semi-automatic and automatic methodologies have been developed to assist in this arduous task. For example, Shimabukuro *et al.* (SHIMABUKURO *et al.*, 2000) proposed an approach to map and monitor deforested areas in the Amazon, automating PRODES's manual interpretation tasks and building a GIS database. This methodology combines digital analysis and manual editing, which still requires significant manual intervention to ensure accuracy. The combination of automated segmentation algorithms and manual adjustments demonstrates the persistent need for human expertise to interpret complex deforestation patterns.

In addition, techniques such as fuzzy C-Means (FCM) have been used to segment forest land cover, highlighting the role of fuzzy soft computing techniques in distinguishing between forested and deforested areas. However, these methods also require an extensive manual effort to overcome the limitations of automatic classification, such as reducing computational time and refining the results of the segmentation (PERUMAL *et al.*, 2021).

Despite advances in technology, the manual interpretation of satellite imagery remains a cornerstone of deforestation monitoring. (ADARME *et al.*, 2020) evaluate deep learning-based strategies for automatic deforestation detection but acknowledge that many approaches still require some level of human intervention or are dependent on manually selected thresholds. This highlights the continued reliance on labor-intensive processes to ensure the accuracy and reliability of deforestation data.

In general, while automated methods are gradually being integrated into deforestation monitoring, the laborious manual processes performed by INPE and CENSIPAM professionals are crucial for managing and verifying the vast and complex data derived from satellite imagery. The expertise and meticulous work of these analysts is indispensable in accurately mapping and monitoring deforestation, ensuring that the data used for environmental policies and conservation efforts are precise and reliable.

More recently, there are some proposals for automatic systems for the detection of deforested areas. For example, Pimenta *et al.*(PIMENTA *et al.*, 2022) developed a deforestation detection system for tropical forests based on the neuroevolution technique (NEAT). This method demonstrated significant efficacy in identifying recently deforested areas, outperforming traditional monitoring techniques. Zhu *et al.*(ZHU *et al.*, 2018) and Zheng *et al.*(ZHENG *et al.*, 2019) proposed methods based on convolutional neural networks (CNNs) for target detection in SAR images. Zhu *et al.* used transfer learning to deal with data scarcity, optimizing the network for the target detection task, resulting in a faster detection speed and a lower number of false positives. On the other hand, Zheng *et al.* introduced a method that uses both features learned by a CNN and features manually extracted. These features are processed in parallel subnetworks and later merged for the final classification, resulting in improved detection performance. These approaches represent advances in target detection in SAR images, but applying these methods to deforestation detection in SAR images, considering speckle noise and object variability, remains a challenge.

To study and develop techniques that can help specialists identify deforestation, as well as raising the capabilities of specialists and ordinary people in the task of labeling in SAR images, we propose three objectives. The first objective of this work is to estimate the capacity of humans to identify deforested areas in SAR images, taking into account each individual's experience in the field. To this end, we conducted a subjective experiment in which volunteer participants labeled deforested areas in a set of 50 SAR images of the Amazon region captured with the Sentinel-1A satellite(ESA, 2023b). The experiment was carried out online using image labeling software developed specifically for this purpose. Chapter 3 is dedicated to explaining in more depth the setup of the experiment and the results obtained.

The second objective of our work is to evaluate the performance of an automatic model to detect deforested areas, comparing this performance with that of humans. The model considered is based on the UNet architecture, as described in Section 2.7, proposed in 2015 by Olaf *et al.* (RONNEBERGER *et al.*, 2015). Although UNet was originally developed for medical image segmentation, this architecture performs well in a wide range of segmentation tasks, from detecting cracks (ALI *et al.*, 2022) and defects in fabrics (JING *et al.*, 2022), to detecting deforestation in satellite images (JOHN; ZHANG, 2022).

The final objective of this work is to develop a robust deforestation segmentation technique in satellite captures of the Amazon forest that uses the advantages of optical satellites and synthetic aperture radars, to achieve this, we implement a data fusion strategy, described in Chapter 4, using single-look complex SAR Sentinel-1 and optical Sentinel-2 data from the same region, co-registered. In this last task, the models EfficientFormerV2 (LI *et al.*, 2022b), SegFormerB0 (XIE *et al.*, 2021), PPLiteSeg (PENG *et al.*, 2022) and UNet (RONNEBERGER *et al.*, 2015) were used, with modifications in the architecture to support five input channels (R, G, B, VV and VH) instead of three (R, G, B).

These architectures were chosen for this work because of their ease of implementation and adaptation, as they are available in the PaddleSeg framework (LIU *et al.*, 2021), good performance demonstrated in general segmentation tasks (LI *et al.*, 2022b; XIE *et al.*, 2021; PENG *et al.*, 2022) and, specifically, in satellite images (THAI *et al.*, 2022; LI *et al.*, 2023; LI *et al.*, 2022a), and low training and inference time. This last point is especially important because the computational resources available for this work did not allow the training, refinement, and experiment cycles of models with a very high number of parameters in a timely manner.

CHAPTER 2

THEORETICAL FRAMEWORK

In this chapter, the relevant concepts for the more advanced chapters of this work will be presented. In Section 2.1, a theoretical presentation on remote sensing, the operation of SARs and optical satellites, and some examples of captures from various satellites will be given. Then, in Section 2.3.1, a basic theoretical introduction to Convolutional Neural Networks and Transformer Networks will be presented. Finally, the structure of the database used in this study to train the models used in Chapters 3 and 4 will be described in Section 2.2

2.1 REMOTE SENSING

Remote sensing is a technique used to observe and measure the characteristics of an area or object from a distance, commonly through the use of satellite or airborne sensor technologies. One of the foundational works in remote sensing is the development and planning of airborne photogrammetry missions. Modern platforms and sensors, such as rotary and fixed-wing aircraft, gliders, and unmanned aerial vehicles (UAVs), are essential in ensuring successful data acquisition. (PEPE *et al.*, 2018) provide a comprehensive overview of mission planning techniques using passive optical sensors, discussing methods, procedures, and tools for various airborne missions. Remote sensing satellites and airborne sensors play a crucial role in automated satellite image understanding systems. Various satellites such as Landsat, SPOT, IRS, and Worldview, and airborne systems such as AVIRIS and DAIS 7915, provide essential data for environmental monitoring and analysis. (UNSALAN; BOYER, 2011) discuss the properties, historical development and applications of these remote sensing systems.

Recent advances in remote sensing technology have also significantly affected wildfire detection and monitoring. (ALLISON *et al.*, 2016) review the state-of-the-art in fire detection using hyperspectral cameras, thermal cameras, and unmanned aircraft, highlighting the operational constraints and opportunities provided by these sensor systems. These studies demonstrate significant progress and diverse applications of remote sensing technologies. Using modern platforms, sensor systems, and integrated data solutions, remote sensing continues to be a vital tool for environmental monitoring, disaster management, and resource assessment. Subsequent sections will provide further insights into the satellite technologies employed in this study.

2.1.1 Optical Satellites

Optical satellites are a type of Earth observation satellite that utilize optical sensors to capture images of the Earth's surface. These satellites play a crucial role in various applications, including environmental monitoring, urban planning, agriculture, and disaster management (LEYVA-MAYORGA *et al.*, 2022; ZHANG; LIU, 2010). The use of optical satellites allows for detailed observation and analysis of surface conditions, making them invaluable tools for scientific research and practical applications.

Optical satellites function by capturing reflected sunlight from the Earth's surface using different types of sensors, including multispectral, hyperspectral, and panchromatic sensors. These sensors can detect and record data across various wavelengths of the electromagnetic spectrum, enabling the extraction of detailed information about the observed area.

One of the main advantages of optical satellites is their ability to provide images that are easier and more interpretable than SAR images and, using RGB bands, a natural coloring true to reality, essential for detailed analysis and monitoring. The resolution of these images can vary, but some modern optical satellites are capable of capturing images with a spatial resolution of less than one meter, allowing highly detailed observations (NIU *et al.*, 2023).

2.1.1.1 Sentinel-2

The Sentinel-2 mission, part of the European Space Agency's Copernicus program, involves a constellation of two identical satellites, Sentinel-2A and Sentinel-2B, which were launched in June 2015 and March 2017 respectively. These satellites operate in sun-synchronous orbit and are crucial for monitoring variations in land surface conditions(ESA, 2023c). Figure 2.1 shows

Figure 2.1. Graphical representation of the Sentinel-2 satellite. Source: Sentinel-2 Overview (ESA, 2023c).

a representation of this satellite made by ESA.

Equipped with the MultiSpectral Instrument (MSI), Sentinel-2 satellites capture images across 13 spectral bands, as shown in Table 2.1 enabling detailed observations of vegetation, soil, inland and coastal waters, and urban areas. With a swath width of 290 km and spatial resolutions ranging from 10m to 60m, Sentinel-2 offers a five-day re-visit time on the equator, providing frequent and reliable data for various applications, including agriculture, forest monitoring, and disaster management(ESA, 2023c).

The MSI operates by capturing sunlight reflected off the Earth, with a shutter mechanism that prevents direct sunlight from contaminating the images and also serves as a calibration device. This makes Sentinel-2 an invaluable tool for continuous and detailed observation of the planet. Sentinel-2 data products are available at different processing levels. Level-1C delivers top-of-atmosphere reflectances in cartographic geometry, while Level-2A provides surface reflectances corrected for atmospheric conditions, making them immediately useful for land applications. For this study, the B2, B3, and B4 (B, G, R, respectively) bands of the signal were used, as shown in Figure 2.2.

Table 2.1. Sentinel-2 Multispectral Instrument (MSI) bands and their applications. These bands are designed to provide data for various earth observation purposes such as vegetation, water bodies, and soil monitoring.

Band	Wavelength (nm)	Resolution (m)	Application
B1	443	60	Coastal and aerosol monitoring
B2	490	10	Visible (blue)
B3	560	10	Visible (green)
B4	665	10	Visible (red)
B5	705	20	Vegetation red edge
B6	740	20	Vegetation red edge
B7	783	20	Vegetation red edge
B8	842	10	NIR for vegetation monitoring
B8a	865	20	Narrow NIR
B9	945	60	Water vapor
B10	1375	60	Cirrus cloud detection
B11	1610	20	SWIR for moisture content
B12	2190	20	SWIR for geological and vegetation applications

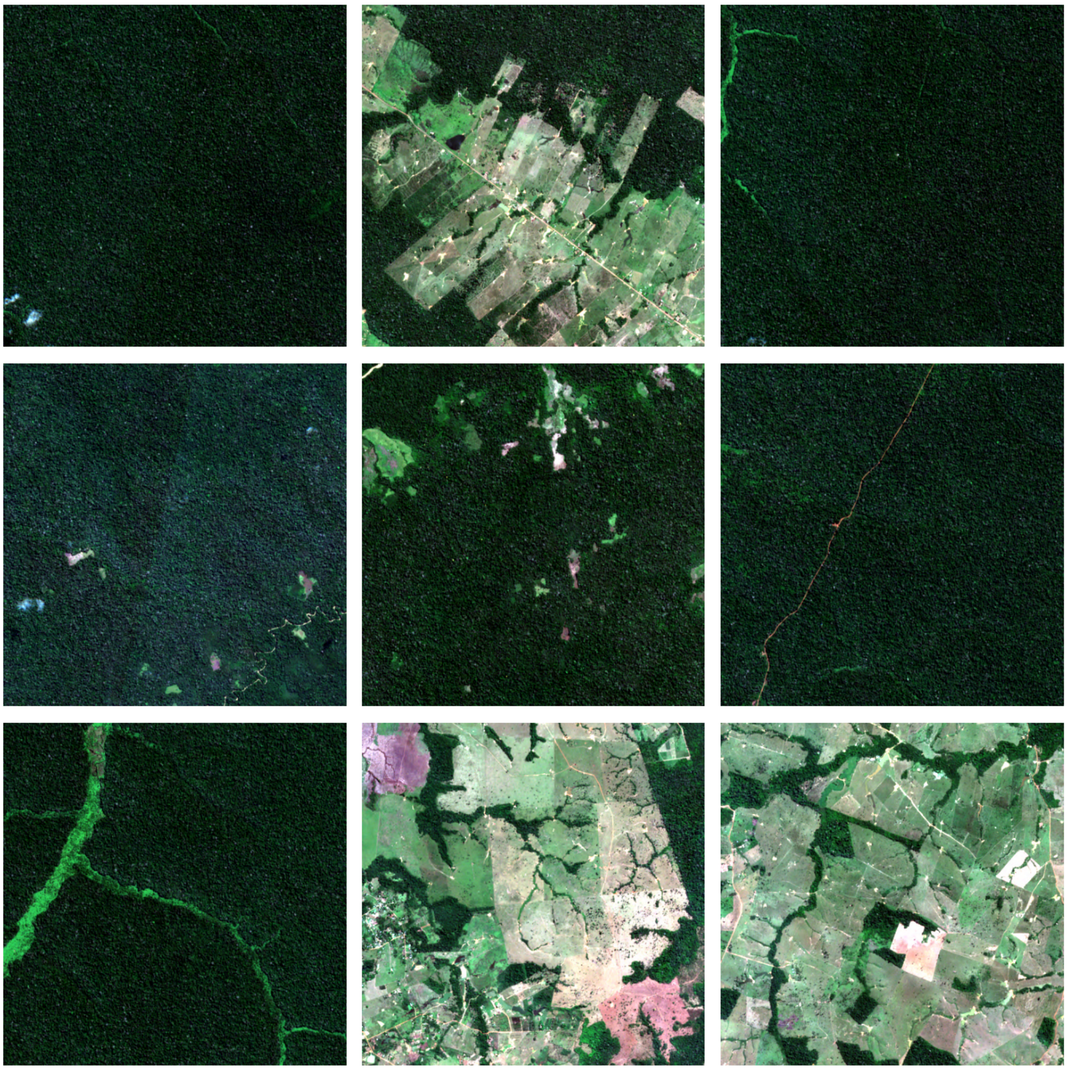


Figure 2.2. Examples of crops extracted from a Sentinel 2A capture in 07/27/2019 from the Amazon rainforest. The B4, B3 and B2 bands were stacked to create a 3 channel RGB image.

2.1.2 Synthetic Aperture Radar Satellites

Synthetic aperture radars (SAR) satellites have active sensors that emit waves in the microwave spectrum and receive the wave reflected from the Earth's surface back. In addition to that, SAR technology utilizes the coherent processing of radar signals to generate high-resolution images of the Earth's surface, which are invaluable for various applications including environmental monitoring and management. Figure 2.3 shows an illustration of the principles of SAR technology that allows one to capture an image of a surface. The *azimuth* refers to the radar trajectory on the mobile platform, the *slant range* is the line of sight of the SAR, the *nadir track* is the path directly below the radar, projected onto the surface of the Earth.

The coherent nature of SAR allows for the synthesis of a large antenna aperture electronically, effectively enhancing spatial resolution beyond what the physical antenna size would normally allow (MCCOY; TANENHAUS, 1992). This is achieved by combining signals received at different times as the radar platform moves, allowing for a fine resolution that is independent of the altitude of the sensor.

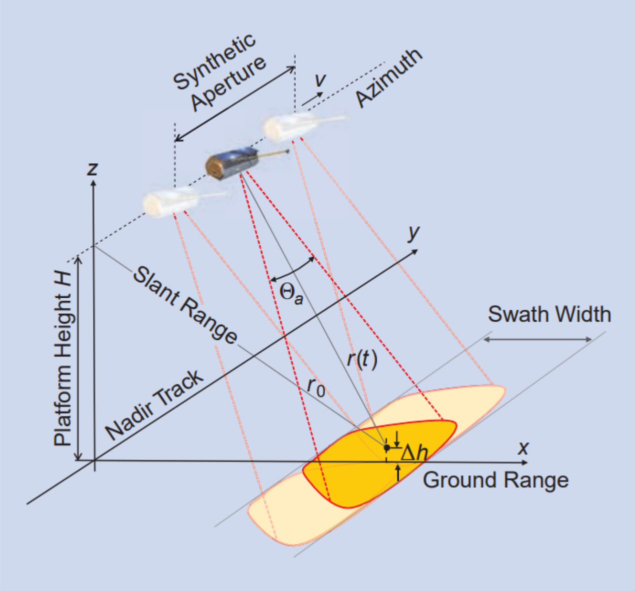


Figure 2.3. Illustration of SAR image geometry. The parameter r_0 refers to the closest approach distance; θ_a is the beamwidth, and v is the sensor speed. Source: *A tutorial on synthetic aperture radar*(MOREIRA *et al.*, 2013), adapted.

SARs operate in the radio wave spectrum and different radars can operate in different frequency bands, as described in Table 2.2. The L-Band, due to its longer wavelengths, is exceptionally effective in penetrating through the canopy cover, making it ideal for the esti-

Table 2.2. Most commonly used SAR bands (MOREIRA *et al.*, 2013). The L band is typically used for biomass estimation due to its penetration into foliage; C, S, and X bands are used for monitoring oceans and ice; X and K bands are used for snow monitoring.

Band	**Ka**	**Ku**	**X**	**C**	**S**	**L**
Frequency (GHz)	40-25	17.6-12	12-17.5	7.5-3.75	3.75-2	2-1
Wavelength (cm)	0.75-1.2	1.7-2.5	2.5-4	4-8	8-15	15-30

mation of biomass and the analysis of soil moisture content (MACEDO *et al.*, 2021). It is particularly beneficial in forest applications where it is necessary to assess vegetation and tree density, even in dense forest regions. On the other hand, the C-Band is commonly used to monitor environmental changes such as flood inundation beneath the forest canopy and wetland dynamics (TOWNSEND, 2002). It is a preferred choice for agricultural monitoring due to its moderate penetration capabilities and sensitivity to surface roughness, making it suitable for crop condition and type monitoring.

As described by Marzano *et al.* (MARZANO *et al.*, 2009), the X-Band is known for its high resolution and high effectiveness in monitoring ice and snow, including tracking changes in glacier dynamics and snow cover. It is also beneficial in urban planning and infrastructure monitoring due to its ability to detect small-scale features. The Ka-band offers the highest resolution among the SAR bands and is particularly useful in detailed surface characterization and target identification. This band is also utilized in high-precision topographic mapping and complex urban area analysis because of its sensitivity to finer details.

Further advancements in SAR technology have introduced techniques like polarimetry and interferometry, which provide additional data dimensions to improve the analysis of surface characteristics and temporal changes. Polarimetric SAR (PolSAR) involves the use of various polarization states of microwave signals to extract detailed information about surface textures and features (BOERNER, 2000). By analyzing how the polarized wave is scattered upon hitting the surface, PolSAR can determine the geometric and dielectric properties of the targets. This technique allows for the discrimination between different types of surface materials and can be crucial in applications such as vegetation mapping, soil moisture estimation, and even the assessment of urban infrastructure. The enhanced textural information obtained through polarimetric measurements considerably improves the interpretation of SAR images over amplitude-only radar systems (MOREIRA *et al.*, 2013).

Interferometric SAR (InSAR) utilizes the phase differences between successive radar pulses to create maps of surface topography and its temporal changes with high precision (CLOUDE; PAPATHANASSIOU, 1998). By comparing the phase of the waves returned from successive passes over the same area, InSAR can measure minute changes in distance, making it invaluable for applications like earthquake and volcano monitoring, where shifts in the Earth's surface need to be tracked over time. The integration of these phase measurements can reveal sub-centimeter changes in elevation, providing critical data for geological and environmental studies.

The ability of SAR to operate independently of lighting conditions and in all weather conditions, including through cloud cover, makes it an indispensable tool in the remote sensing arsenal, especially for monitoring regions like the Amazon rainforest, where cloud cover and precipitation can obstruct optical sensors. (LEITE-FILHO *et al.*, 2021).

2.1.2.1 Sentinel-1

The images used in the experiment were collected by the Sentinel-1A radar(ESA, 2023b), Sentinel-1 is the first constellation of SARs from the Copernicus program (ESA, 2023a), conducted by the European Space Agency starting in 2014. The constellation consists of Sentinel-1A and Sentinel-1B radars, both operating on the C-band with frequency specifications shown in Table 2.2. The primary goals of the Sentinel-1 mission include the surveillance of forests, farmlands, seas, and the observation of glaciers to keep track of climatic variations. These satellites have a recharge time of 12 days and operate at an altitude of 400 km. The figure 2.4 shows a graphical representation of this radar.

Sentinel-1 satellites provide products across three processing levels. Level 0 consists of raw data. Level 1 includes Single Look Complex (SLC) and Ground Range Detected (GRD) products. Lastly, Level 2 offers oceanic data. These levels cater to different applications, with Level 1 providing more processed and interpretable data than Level 0, and Level 2 focusing specifically on marine environments.

An SLC product is an image that represents a single radar capture at a specific time. SLC images have pixels with complex values that represent the phase and amplitude values of the signal in the region. These products are typically composed of two images with different

Figure 2.4. Graphical representation of the Sentinel-1 radar. Source: Sentinel-1 Overview (ESA, 2023b).

polarizations, one vertical and the other horizontal.

GRD type products are obtained from multiple capture samples projected at ground level from an ellipsoidal Earth model and considering the terrain altitude. In this type of product, the image pixels represent only the magnitude of the signal while the phase is disregarded. In Sentinel-1 radars, GRDs can have three different resolutions: full resolution, high resolution, and medium resolution. For this project, preprocessed SLC products from CENSIPAM were used.

2.2 GROUNDTRUTH DEFORESTATION DATA

The labels used as ground truth for the data were provided by CENSIPAM in the context of a collaborative research project between UnB and the Ministry of Defense, a project in which the author of this work participated. The deforestation labels were generated by CENSIPAM specialists using the technique described by Paulo Tavares *et al.*(TAVARES *et al.*, 2019) which uses both the SAR data from Sentinel 1 and the optical data from Sentinel-2. The images are organized in a coorbital arrangement, with revisits made at a maximum interval of 3 days. These images were selected in the period preceding winter in the Amazon, with cloudiness rates below 2%.

In this technique, the Sentinel 1 and Sentinel 2 images are colocated and have a terrain sampling of about 10m, being referenced in the WGS84 geographic coordinate system. To generate the ground truth, between 5,000 to 10,000 vertices were collected, using the Random Forest algorithm with the following parameters: N-try equal to 1,000, M-tree equal to 7, 5,000 random samples divided into 30% for validation and 70% for training (TAVARES *et al.*, 2019). According to the methodology of Inpe's Prodes (ALMEIDA *et al.*, 2021), four labeling classes were defined: forest, deforestation, water and nonforest.

The method involves a sophisticated process to ensure an accurate detection of deforested areas. Initially, SAR and optical images were collected from Sentinel-1 and Sentinel-2, respectively. Sentinel-1 SAR data, which is capable of penetrating clouds and capturing data regardless of weather conditions, is crucial in tropical regions such as the Amazon where cloud cover is prevalent. Optical data from Sentinel-2 provides high-resolution imagery that is beneficial for identifying vegetation types and changes.

The preprocessing steps for Sentinel-2 images involve atmospheric correction using the Sen2Cor algorithm to derive surface reflectance values (JOSHI *et al.*, 2016a; ZHANG *et al.*, 2018; CHATZIANTONIOU *et al.*, 2017). All spectral bands of Sentinel-2 are resampled to a uniform spatial resolution of 10 meters using bilinear upsampling. For Sentinel-1, the preprocessing includes applying orbit files, radiometric calibration, thermal noise removal, and debursting. A slice assembly technique is used to combine multiple scenes and a range Doppler terrain correction is applied using the UTM WGS84 projection and 30-meter SRTM data, resampled to 10 meters to match the Sentinel-2 resolution.

For Sentinel-1, texture characteristics are derived using the Grey-Level Co-occurrence Matrix (GLCM) with a 5x5 sliding window to calculate the mean, variance, and correlation for both VV and VH polarizations. This results in six texture products that enhance the classification process. For Sentinel-2, radiometric indices like the Normalized Difference Vegetation Index (NDVI), Normalized Difference Water Index (NDWI), and Soil-Adjusted Vegetation Index (SAVI) are computed to enhance the identification of various land cover types.

Table 2.2 shows the equations to calculate this indexes, where NIR is the near infrared band, 842 nm for S-2, for NDVI, NDWI, and SAVI; Red is 665 nm for S-2 for NDVI and SAVI; MIR (Medium Infrared) is 2190 nm for S-2, for NDWI; P(i,j) is a normalized gray-tone spatial

Table 2.3. S-2 Indexes and S-1 GLCM Textural Measures used to generate ground-truth, as described in (TAVARES *et al.*, 2019)

S-2 Indexes

Index Applied	Equation
NDVI	$\frac{\text{NIR}-\text{Red}}{\text{NIR}+\text{Red}}$
NDWI	$\frac{\text{NIR}-\text{MIR}}{\text{NIR}+\text{MIR}}$
SAVI	$L \times \frac{(\text{NIR}-\text{Red})}{\text{NIR}+\text{Red}+0.5}$

S-1 GLCM Textural Measures

Measure	Equation
Mean	$\sum_{i,j=0}^{N-1} iP_{i,j}$
Variance	$\sum_{i,j=0}^{N-1} iP_{i,j}(i-\mu)^2$
Correlation	$\frac{\sum_{i,j=0}^{N-1} iP_{i,j} - \mu_x\mu_y}{\sigma_x\sigma_y}$

dependence matrix such that SUM(i,j = 0, N - 1) (P(i,j)) = 1; i and j represent the rows and columns, respectively, for the measures of Mean, Variance and Correlation; μ is the mean, for the Variance textural measure; and N is the number of distinct grey levels in the quantized image; μ_x, μ_y, σ_x, and σ_y are the means and standard deviations of p_x and p_y, respectively, for the correlation textural measure.

The processed images from Sentinel-1 and Sentinel-2 are then stacked using nearest neighbor resampling, with Sentinel-1 as the master image. This integrated dataset undergoes a segmentation process to aggregate pixels with similar values, employing a mutual best fitting region merging criteria (BAATZ; SCHäPE, 2000; LASSALLE *et al.*, 2015), resulting in approximately 82,246 segments.

The segmented data are classified using the Random Forest algorithm implemented in ArcGIS 10.4. This algorithm requires specifying the number of trees (N) and the maximum depth of each tree (n). The parameters are optimized on the basis of tests to ensure the best accuracy. The classification process uses various attributes of the segment, including color, mean, standard deviation, count, compactness, and rectangularity.

The accuracy of the classification is assessed using metrics such as overall accuracy (OA) and the Kappa coefficient. The overall accuracy is given by:

$$OA = \frac{\sum TP_i}{N}, \tag{2.1}$$

where $\sum TP_i$ is the sum of true positives for all classes and N is the total number of pixels.

Analogously, the Kappa coefficient, a statistical measure that evaluates the agreement be-

Data Combination	Overall Accuracy (%)	Kappa Coefficient
S-1 with S-2	91.07	0.8709
S-2 Only	89.53	0.8487
S-2 with Indexes	89.45	0.8476
All	87.09	0.8132
S-1 with Textures	61.61	0.4870
S-1 Only	56.01	0.4194

Table 2.4. Performance Table of Different Data Combinations

tween the observed classification and the reference data, taking into account the agreement that could occur by chance. The equation for the Kappa coefficient is given by:

$$\kappa = \frac{OA - P_e}{1 - P_e}, \tag{2.2}$$

where P_e is the expected agreement by chance, which is calculated as follows

$$P_e = \sum_{i=1}^{k} \left(\frac{(R_i \times C_i)}{N^2} \right),$$

and R_i is the total number of pixels in row i of the confusion matrix, C_i is the total number of pixels in column i of the confusion matrix, N is the total number of pixels, and k is the number of classes.

The results of the automatic method were manually checked by CENSIPAM experts and interns to assess the accuracy metrics and generate the results shown in Table 2.4. These metrics are derived from cross-validation with high-resolution Planet[1]imagery and statistical approaches such as Jeffries-Matusita and Transformed Divergence (SEN *et al.*, 2019) for evaluating the separability of classes.

Using this automatic method, the CENSIPAM team generated labels for six Sentinel-1 image scenes. This classification was reviewed by 5 experts and 10 interns and scholars. Figure 2.5 represents two examples of images and their respective ground truth masks. Note that even though the masks have been manually reviewed by the CENSIPAM team, there is still a significant amount of noise and sparse pixels with deforestation. This amount of noise will be a problem in training the model described in Chapter 3. This problem was addressed in the second stage of this work, Chapter 4, using morphological operations.

[1] Planet is a company that operates a fleet of Earth-imaging satellites, providing high-resolution imagery and data for various applications such as environmental monitoring, agriculture, and urban planning. More information available at: <https://www.planet.com/>

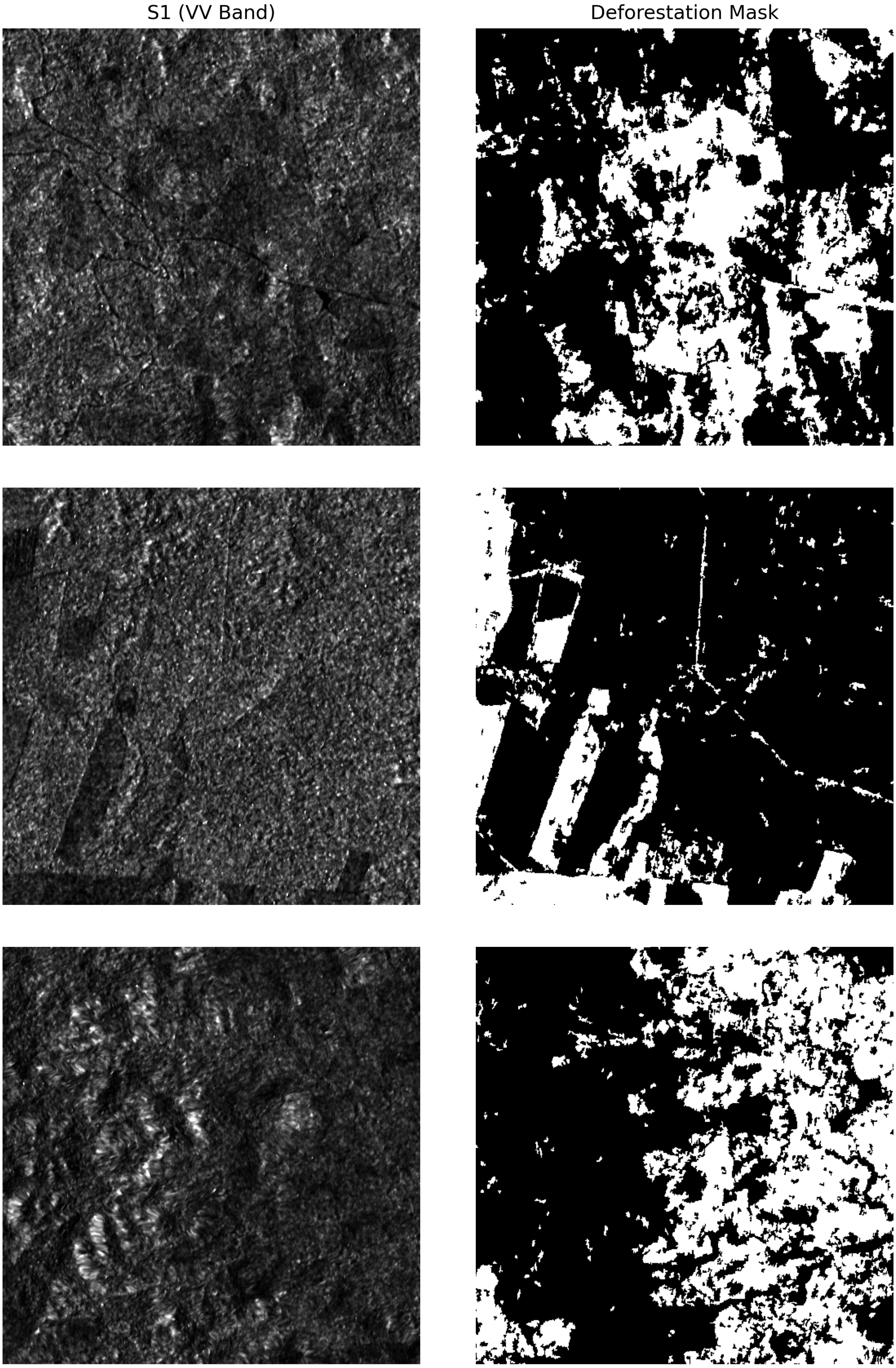


Figure 2.5. Crops from captures of the processed VV band Sentinel 1 next to the corresponding mask obtained by the described process.

2.3 SEMANTIC SEGMENTATION WITH ARTIFICIAL NEURAL NETWORKS

Semantic segmentation is a task in computer vision that involves partitioning an image into distinct regions and assigning a class label to each pixel (ZHENG *et al.*, 2021). This process allows for the detailed understanding of the image content by distinguishing different objects and their boundaries within the scene. Unlike traditional image classification, which only assigns a single label to an entire image, semantic segmentation provides a pixel-level classification, making it essential for applications that require precise localization and identification of various elements.

The goal of semantic segmentation is to generate a mask that categorizes each pixel of the input image into one of several predefined classes, such as roads, buildings, people, and trees. This pixel-wise annotation enables detailed analysis and understanding of complex scenes, facilitating tasks such as autonomous driving, medical imaging, and aerial imagery analysis (WANG *et al.*, 2022; AZIMI *et al.*, 2019; YANG *et al.*, 2020). In autonomous driving, for example, semantic segmentation helps the vehicle understand its surroundings by identifying and locating objects like pedestrians, vehicles, and traffic signs, ensuring safe navigation.

In this section, we briefly describe the artificial intelligence methods used in this work to segment deforested areas. More specifically, we describe the basic concepts of artificial neural networks, the architectures used in this work, and the performance metrics considered.

2.3.1 Fundamentals of Neural Networks

Artificial neural networks (ANN) are computational models inspired by the structure and function of the human brain (GOODFELLOW *et al.*, 2016). They consist of interconnected layers of nodes or "neurons," each performing a simple computation. These neurons are organized into an input layer, one or more hidden layers, and an output layer. Each connection between neurons has an associated weight, which is adjusted during training to minimize the error in the network's predictions using optimization techniques and backpropagation for gradient calculation. ANNs have become the state-of-the-art in many tasks within computer vision (WANG *et al.*, 2023c). Their ability to automatically learn features from raw data without the need for manual feature extraction has significantly advanced the field. For example, convolutional neural networks (CNNs) (GOODFELLOW *et al.*, 2016), a type of ANN particularly effective for image-related tasks, have been instrumental in achieving high performance in image classification, object detection, and segmentation.

CNNs operate by applying convolution operations to input images, which involve sliding filters (or kernels) across the image to detect various features. These features are then combined in subsequent layers to form more complex representations (BROWNLEE, 2018). The main components that distinguish convolutional neural networks from other types of ANNs are the convolutional layers and the pooling layers. Convolutional layers apply convolution operations to extract features from the input image. Each layer learns multiple filters that capture different aspects of the data, such as edges, textures, and patterns (PAJANKAR; JOSHI, 2022; TEAM, 2020).

As a result of the convolution operations, a feature map is generated. Feature maps are essentially grids of numbers representing the presence of detected features across the spatial dimensions of the input (TEAM, 2020). Each point in the feature map corresponds to the application of a filter at a specific location on the input image, indicating where certain features are found and their intensity. These maps maintain spatial hierarchies and allow the network to understand the positional context of features (BROWNLEE, 2018). As the network deepens, subsequent layers use these feature maps to detect higher-level patterns and objects, leading to a more refined and comprehensive understanding of the image content. This ability to automatically learn and adapt filters through training makes CNNs highly effective for visual tasks, allowing them to detect and analyze various elements within an image. Figure 2.6 illustrates how a convolution operation works in a neural network architecture. The input image (left) is processed by a filter (center) to produce an output array (right). The example shows the computation of a single output value (16) by performing element-wise multiplication and summation of the overlapping regions of the input image and the filter.

The pooling layers play a crucial role in CNNs by performing down-sampling operations, which reduce the spatial dimensions of the feature maps. This reduction helps to reduce computational load and minimizes the risk of overfitting (DIAMANTIS; IAKOVIDIS, 2020). Common types of pooling, such as max pooling and average pooling, ensure that the most significant

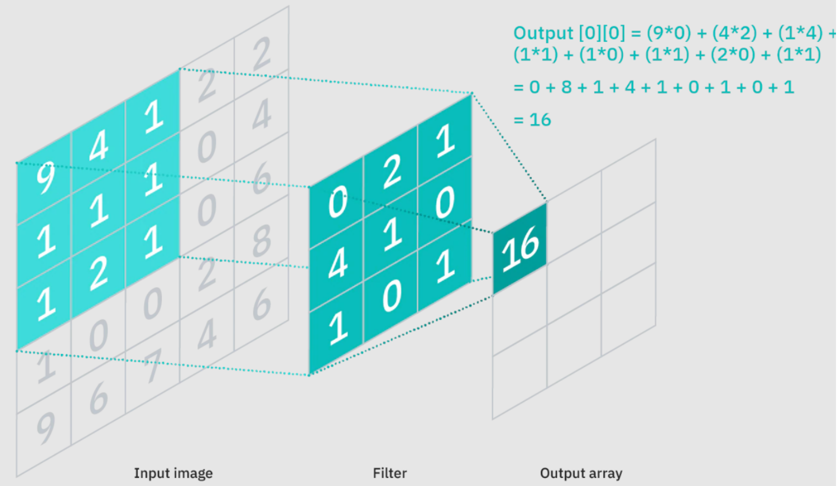


Figure 2.6. Illustration of a convolution operation in a neural network.

features are retained while the size of the data is reduced, maintaining essential information for further processing (BROWNLEE, 2019). Following the convolutional and pooling layers, the output is usually flattened and passed through one or more fully connected layers. These layers are responsible for high-level reasoning and classification based on the features extracted by the convolutional layers (MADHUGIRI, 2020). They combine the features in a way that allows the network to make final predictions, categorizing the input images into predefined classes with high accuracy. To enhance the learning capability of CNNs, non-linear activation functions are applied after each convolutional and fully connected layer. Functions like ReLU (Rectified Linear Unit)[2] introduce non-linearity into the model, enabling it to learn complex patterns and representations. This non-linearity is essential for the network to capture intricate relationships within the data, making neural networks versatile and powerful for various computer vision tasks (MCCULLUM, 2019).

More recently, transformer networks have revolutionized the field of natural language processing (NLP) and are increasingly making significant impacts on computer vision. Transformers are designed to handle sequential data, making them highly effective for tasks involving text, such as translation, summarization, and sentiment analysis (VASWANI *et al.*, 2017). Unlike traditional recurrent neural networks (RNNs) and long short-term memory networks (LSTMs), transformers rely on self-attention mechanisms to process entire sequences of data simultaneously, which allows for more efficient training and the capture of long-range dependencies. The core innovation of Transformers lies in the self-attention mechanism, which enables the model

[2] The ReLU funciton is defined by $f(x) = \max(0, x)$.

to weigh the importance of different words in a sentence when making predictions. This mechanism computes attention scores that highlight which parts of the input sequence are most relevant for generating each part of the output sequence. This approach not only improves the model's ability to understand context, but also allows for parallelization, significantly speeding up training times compared to RNNs and LSTMs.

Transformers have been adapted for computer vision tasks with notable success. Vision Transformers (ViTs) apply the same self-attention principles to image patches, treating them as sequences. This adaptation allows Transformers to excel in tasks such as image classification, object detection, and segmentation. Studies have shown that ViTs can achieve state-of-the-art performance on various benchmarks. For example, (DOSOVITSKIY *et al.*, 2020) demonstrated that ViTs could outperform traditional convolutional neural networks (CNNs) on image classification tasks when pretrained on large datasets.

For the purpose of semantic segmentation of deforested regions, four benchmark segmentation architectures were evaluated. The initial model implemented was a modified UNet architecture (RONNEBERGER *et al.*, 2015), configured to process five input channels (R, G, B, VV, VH) rather than the usual three (R, G, B). Next, three advanced segmentation techniques from the PaddleSeg toolkit (LIU *et al.*, 2021) were examined: EfficientFormerV2 (LI *et al.*, 2022b), SegFormerB0 (XIE *et al.*, 2021), and PPLiteSeg (PENG *et al.*, 2022). The following subsections provide a brief examination of each architecture employed in this study.

2.3.2 UNet

The UNet architecture, proposed by (RONNEBERGER *et al.*, 2015), as illustrated in Figure 2.7, employs a symmetrical encoder-decoder structure designed for precise segmentation tasks. The encoder path, composed of repeated convolutional blocks, progressively reduces the spatial dimensions while increasing the feature depth. Each convolutional block consists of sequences of convolutional layers (Conv 3x3), followed by ReLU activations and batch normalization to enhance feature extraction and normalization. Downsampling is achieved through max-pooling layers, reducing the spatial resolution and capturing contextual information at various scales.

The decoder path mirrors the encoder structure but focuses on upsampling the feature maps to the original image resolution. This is achieved through bilinear upsampling, followed by convolutional layers that refine the feature maps. A crucial aspect of the UNet architecture is the inclusion of skip connections, which concatenate the feature maps from the encoder to the corresponding decoder layers. These connections, indicated by the green arrows in Figure 2.7, allow the network to retain fine-grained spatial information lost during downsampling.

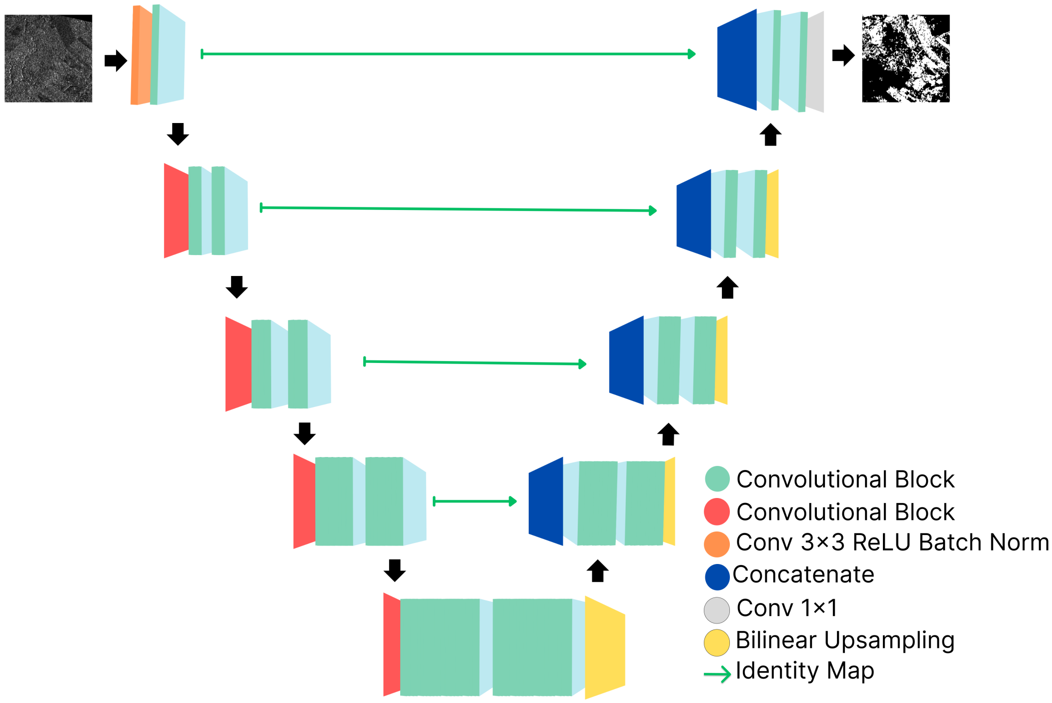


Figure 2.7. Representation of the UNet architecture. The network input is the processed 512x512 cut and the output is a binary mask containing the deforestation marking pixel by pixel.

In addition, 1x1 convolutional layers are employed at various stages to adjust the dimension of the feature map and facilitate smooth information flow between different scales. The concatenation operations (blue blocks) integrate high-resolution features from the encoder with the upsampled features in the decoder, enhancing the network's ability to perform precise and context-aware segmentation. In general, the UNet architecture combines deep feature extraction with efficient spatial information preservation, making it highly effective for medical image segmentation and other dense prediction tasks.

Although other more modern models are also used for image segmentation, such as the InternImage framework proposed by Wang *et al.*(WANG *et al.*, 2023a) which is based on transformer networks, we chose UNet due to its simplicity of implementation, rapid training, and demonstrated efficacy on satellite images(MCGLINCHY *et al.*, 2019). The choice is justified, in particular, by the model's ability to operate efficiently on limited hardware resources and its proven suitability for images produced by SARs. This model will be used as a baseline in comparison with human performance in the labeling task in Chapter 3 and will also be expanded to five channels in Chapter 4 for use in the segmentation task using fused data.

2.3.3 PPLite-Seg

PPLiteSeg (PENG *et al.*, 2022), shown in Figure 2.8, is a lightweight and efficient architecture, specifically designed for devices with limited computational capabilities. This is achieved through a simple yet powerful network design that employs advanced convolutional strategies such as depthwise separable convolutional layers and a pyramidal structure for efficient image processing. In the context of fusing optical and radar data for deforestation segmentation, PPLiteSeg is particularly appealing due to its efficiency in processing multiple input channels without significantly increasing computational complexity.

Depthwise separable convolutions are a type of convolutional operation used in deep learning to reduce computational cost and model size. They decompose the standard convolution into two separate layers: depthwise convolution, which applies a single convolutional filter per input channel, and pointwise convolution, which uses a 1x1 convolution to combine the outputs of the depthwise layer. This separation significantly decreases the number of parameters and calculations required, making them efficient for mobile and embedded applications.

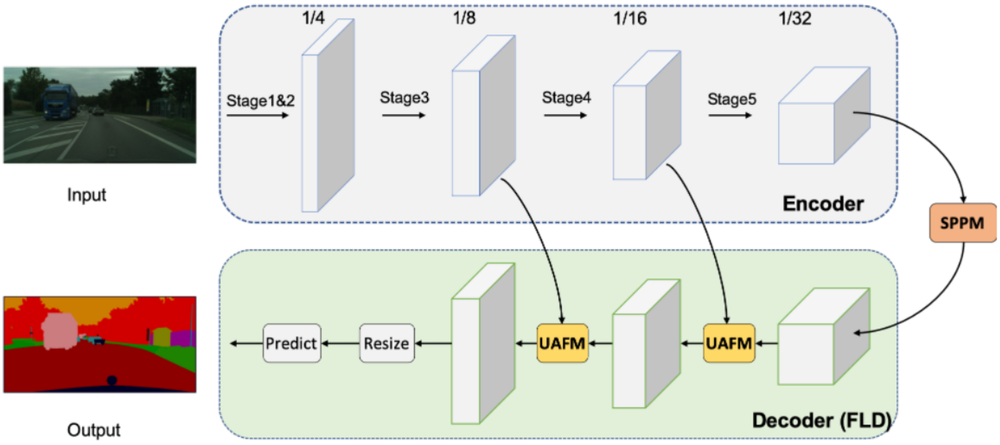


Figure 2.8. Schematic representation of the encoder-decoder architecture for semantic segmentation, demonstrating the flow from input through various stages to the output. Source: PP-LiteSeg: A Superior Real-Time Semantic Segmentation Model (PENG *et al.*, 2022)

The PP-LiteSeg, which architecture is shown in Figure 2.8, adopts an encoder-decoder architecture optimized for real-time processing. The encoder uses lightweight convolution layers to efficiently extract multi-scale features from the input image. The decoder, known as the Flexible and Lightweight Decoder (FLD), is crucial for up-sampling and combining these features into a final segmentation map.

The encoder, represented in the upper block of Figure 2.8, consists of several stages that progressively downsample the input image, extracting hierarchical features with increasing depth. This process balances computational efficiency by reducing spatial resolution while increasing the number of feature channels at each stage. The Simple Pyramid Pooling Module, SPPM, is strategically positioned between the encoder and decoder to aggregate global context information with minimal computational overhead. It achieves this by applying pyramid pooling with different bin sizes, followed by convolution and upsampling operations to refine the pooled features.

The decoder in PP-LiteSeg, represented in the lower block of Figure 2.8, employs the FLD, which gradually upsamples the features back to the original resolution. The FLD incorporates UAFM to enhance feature representation by applying both spatial and channel attention mechanisms. These attention mechanisms generate weights that emphasize the most relevant features, improving the fusion of multi-level features from the encoder. The UAFM ensures that the combined features retain critical spatial and semantic information, leading to more accurate segmentation results. The detailed working mechanism of UAFM and SPPM is described in the following subsections.

2.3.3.1 Unified Attention Fusion Module (UAFM)

The UAFM (PENG *et al.*, 2022), represented in Figure 2.9, is a key component of the decoder that improves feature maps by applying attention mechanisms. It integrates both spatial and channel attention to dynamically prioritize relevant features dynamically throughout the network. This module first computes attention weights, which are then used to scale the feature maps, thus ensuring that salient features are enhanced while less important ones are suppressed.

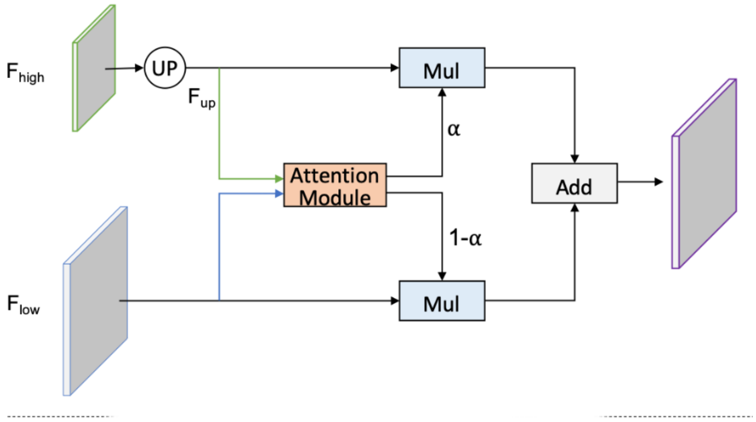


Figure 2.9. Illustration of the Unified Attention Fusion Module. Source: PP-LiteSeg: A Superior Real-Time Semantic Segmentation Model. (PENG *et al.*, 2022)

Then, high-level features F_{high} are upsampled to F_{up} which, along with low-level features F_{low}, are fed into the attention module. The output is modulated by a scaling factor α before the addition operation that integrates the processed high and low-level features to produce the final enhanced feature map. The attention module, represented by the orange box in Figure 2.9, can be used as a plug-in, where different attention techniques such as spatial attention or channel attention can be applied, as described by the authors (PENG *et al.*, 2022)

The spatial attention module exploits the inter-spatial relationship to produce a weight that represents the importance of each pixel in the input features. It performs mean and max operations along the channel axis to generate four features, which are then concatenated and processed through convolution and sigmoid operations to produce the final attention map. On the other hand, the channel attention module leverages the inter-channel relationship to generate a weight indicating the importance of each channel in the input features. It uses average-pooling and max-pooling operations to squeeze the spatial dimension, followed by convolution and sigmoid operations to produce the channel attention map. Both mechanisms enhance feature representation by focusing on significant parts of the input data.

2.3.3.2 Simple Pyramid Pooling Module (SPPM)

PP-LiteSeg incorporates SPPM to aggregate contextual information from different regions of the input image (PENG *et al.*, 2022). The SPPM performs global average pooling at various scales and uses these pooled features to augment the feature maps processed by the decoder.

This module is designed to capture a comprehensive context without the computational complexity typically associated with pyramid pooling architectures.

Pyramid pooling architectures, such as Spatial Pyramid Pooling (SPP) (HE *et al.*, 2015) and Pyramid Scene Parsing Network (PSPNet) (ZHAO *et al.*, 2017), address the limitation of fixed input size constraints in CNNs by performing pooling operations at multiple levels or scales. This approach involves dividing the feature map into several sub-regions or bins, performing pooling within each bin, and then concatenating these pooled features. By capturing information at various scales, pyramid pooling allows the network to maintain important spatial hierarchies and context, enhancing its ability to perform accurate recognition and segmentation tasks. This multi-level pooling strategy enables the network to incorporate both fine-grained and coarse contextual information, leading to more robust feature representations and improved performance on tasks involving complex scenes and objects of varying sizes.

Figure 2.10 shows a diagram of the multi-scale feature processing pipeline used in SPPM. In this module, features are pooled and then passed through parallel convolutional (Conv) layers. Each branch is resized to match dimensions before being combined through an addition operation. The aggregated feature map then undergoes a final convolutional transformation to produce the output.

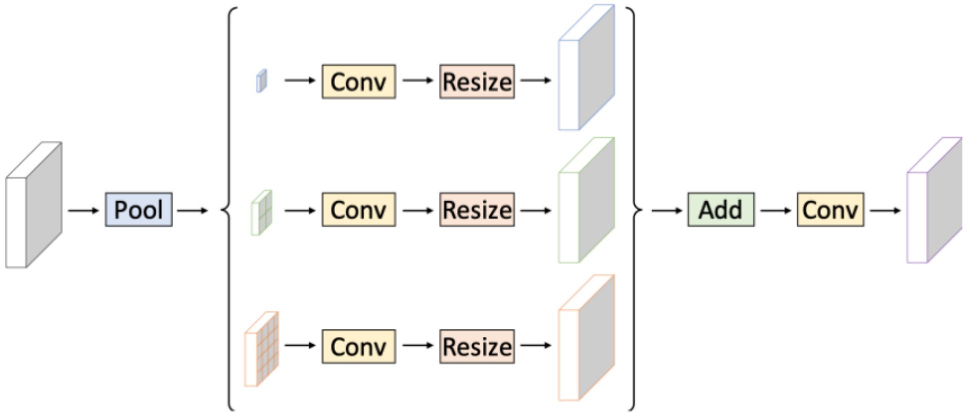


Figure 2.10. Illustration of SPPM. Source: PP-LiteSeg: A Superior Real-Time Semantic Segmentation Model (PENG *et al.*, 2022)

2.3.3.3 Overview of PP-LiteSeg

As highlighted by the authors (PENG *et al.*, 2022), the architecture is balanced to provide high segmentation accuracy while operating efficiently on standard hardware. The use of techniques such as depthwise separable convolutions and simplified pooling modules enables PP-LiteSeg to achieve fast inference times, making it suitable for applications like mobile and embedded systems where computational resources are constrained.

The combination of these features allows PP-LiteSeg to deliver robust performance in real-time semantic segmentation tasks, providing a balance between speed and accuracy. This makes it a reasonable choice for deployment in scenarios where both performance and computational efficiency are critical. Based on these points, the architecture was chosen to be used in this study, given the hardware constraints and training time available.

2.3.4 EfficientFormerV2

EfficientFormerV2 (LI *et al.*, 2022b) is engineered to match the efficiency of lightweight CNNs like MobileNet in terms of size and speed, while still delivering robust performance. It rethinks Vision Transformers (ViTs) to create a supernet characterized by its low latency and high parameter efficiency.

As shown in Figure 2.11, the network architecture evolves through four stages, increasing incrementally in complexity and depth while spatial resolution decreases. Initially, it processes features locally (a) using pooling and convolution layers. Then it transitions to a unified Feed Forward Network (FFN) design (b) that integrates depth-wise convolutions. This is followed by a Multi-Head Self-Attention (MHSA) block (c) that enhances locality and employs 'Talking Head' attention, thus refining feature interaction. The flow of the architecture (d) is defined by subsampling stages that merge local and global processing. For handling higher-resolution features, it utilizes a strategy (e) involving attention with downsampling. Lastly, the dual-path attention downsampling technique (f) combines static and learnable local downsampling with global attention, allowing for context-aware reduction in feature dimensionality.

Multi-head self-attention is a key component of transformer architectures, which enables the model to focus on different parts of the input sequence simultaneously (LI *et al.*, 2022b). The

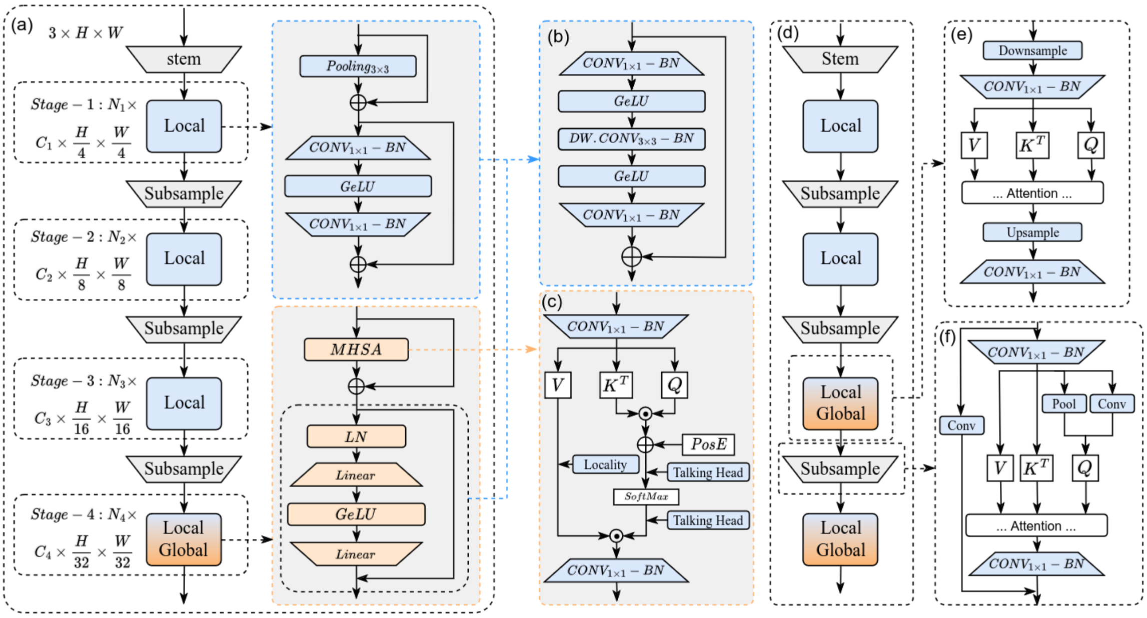


Figure 2.11. EfficientFormerV2 architecture with local and global processing blocks across four stages, unified FFN design, MHSA enhancements, and dual-path attention downsampling. Source: Rethinking Vision Transformers for MobileNet Size and Speed (LI *et al.*, 2022b)

mechanism works by first computing self-attention scores, which determine the importance of each element in the input relative to all other elements. This is achieved by projecting the input sequence into three distinct vectors: queries, keys, and values. Attention scores are computed as the dot product of queries and keys, scaled by the square root of the dimensionality of the keys, and passed through a softmax function to obtain normalized weights.

These attention weights are then used to compute a weighted sum of the values, producing the output for each head. In a multi-head setup, several self-attention mechanisms run in parallel, each with its own set of learned projections for queries, keys, and values. This allows the model to capture diverse patterns and relationships in the input data. The outputs of all heads are concatenated and projected to form the final output.

In general, the design of EfficientFormerV2 involves a deliberate refinement of the network search space (LI *et al.*, 2022b), focusing on a configuration that is both deeper and narrower to enhance precision while reducing the number and latency of parameters. The original article, (LI *et al.*, 2022b), introduces improvements to MHSA by integrating local information and facilitating communication between heads, improving performance without additional cost.

Attention to higher resolution, a known challenge for mobile efficiency due to its complexity, is adeptly handled through Stride Attention, reducing latency while preserving accuracy.

Moreover, the dual-path attention downsampling approach outperforms conventional methods by employing both static and dynamic strategies to downsample features in a way that is cognizant of the context, ensuring efficient performance in mobile settings. Together, these innovations establish EfficientFormerV2 as a transformer model that not only rivals but potentially surpasses lightweight CNNs in mobile efficiency.

2.3.5 SegFormerB0

SegFormer (XIE *et al.*, 2021), illustrated in Figure 2.12, introduces a transformative approach to semantic segmentation, merging Transformer efficiency with an All-MLP decoder. This segmentation model stands out with its hierarchically structured Transformer encoder that outputs multi-scale features without the need for positional encoding, which is particularly beneficial when testing resolutions differ from training ones. Additionally, the model's MLP decoder eschews complex decoders in favor of a simple design that aggregates multi-layer information, thereby integrating both local and global attention to produce powerful representations. This leads to an exceptionally lightweight and efficient model suitable for real-time applications in high-resolution images, as detailed in the SegFormer paper (XIE *et al.*, 2021).

The encoder employs overlap patch embeddings and progressive stages of transformer blocks, each followed by merging for multi-scale feature extraction. The decoder fuses the multiresolution features using MLP layers and upsamples to match the original image resolution for pixel-wise classification. With a focus on model size, run-time, and accuracy, SegFormerB0 demonstrates superior performance on various datasets, offering a substantial improvement in terms of both efficiency and accuracy compared to other models. It achieves this by leveraging a series of Mix Transformer encoders, ranging from MiT-B0 for rapid inference to MiT-B5 for peak performance.

The encoder consists of several transformer blocks. Each block includes an Efficient Self-Attention mechanism and a Mix-FeedForward Network (Mix-FFN). The Efficient Self-Attention mechanism reduces computational complexity by concentrating on the most significant regions of the input features. The Mix-FFN incorporates a 3x3 convolution within the feed-forward network, eliminating the need for positional encoding. This design choice enhances the model's

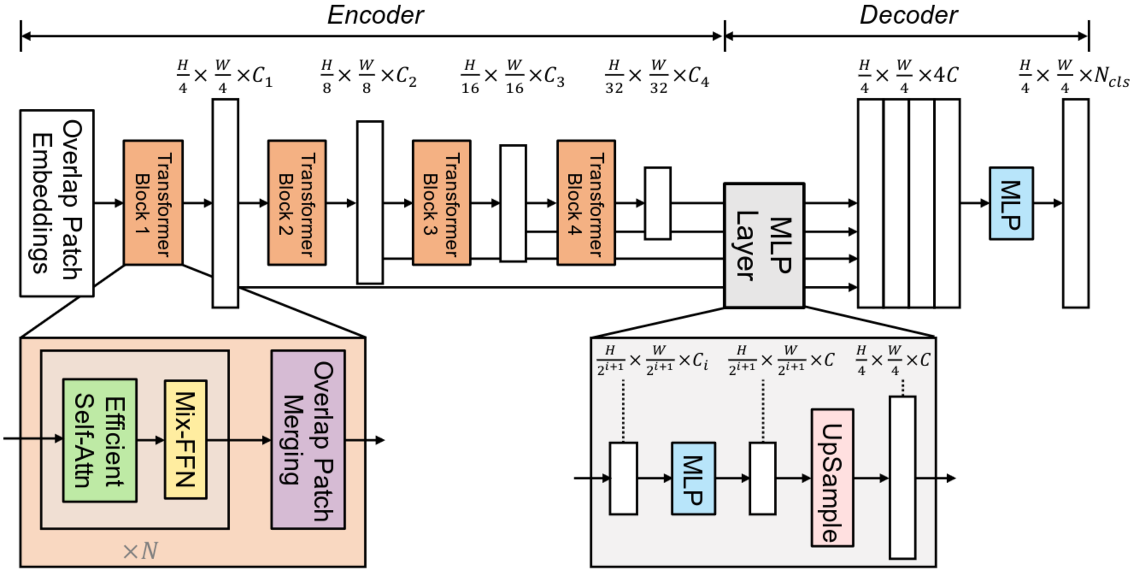

Encoder
Decoder
$\frac{H}{4}\times\frac{W}{4}\times C_1$
$\frac{H}{8}\times\frac{W}{8}\times C_2$
$\frac{H}{16}\times\frac{W}{16}\times C_3$
$\frac{H}{32}\times\frac{W}{32}\times C_4$
$\frac{H}{4}\times\frac{W}{4}\times 4C$
$\frac{H}{4}\times\frac{W}{4}\times N_{cls}$
Overlap Patch Embeddings
Transformer Block 1
Transformer Block 2
Transformer Block 3
Transformer Block 4
MLP Layer
MLP
Efficient Self-Attn
Mix-FFN
Overlap Patch Merging
×N
$\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times C_i$
$\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times C$
$\frac{H}{4}\times\frac{W}{4}\times C$
MLP
UpSample

Figure 2.12. Architecture of SegFormerB0, featuring an encoder with hierarchically structured transformer blocks and a simple MLP decoder. Source: SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers(XIE *et al.*, 2021)

ability to handle various input resolutions during both training and inference, thus improving overall robustness and accuracy.

Between transformer blocks, the architecture utilizes Overlap Patch Merging. This module merges overlapping patches to progressively reduce the spatial dimensions while increasing the channel dimensions. This hierarchical reduction allows the network to efficiently manage computational resources and learn more complex features at different levels of the hierarchy.

At the end of the encoder, feature maps are refined through an MLP layer before being upsampled. This refinement ensures that the high-dimensional features are appropriately scaled and merged. The decoder then reconstructs the high-resolution segmentation map from these refined features. Using a series of MLP layers and upsampling operations, the decoder incrementally increases the spatial resolution of the feature maps. The final MLP layer maps these high-resolution features to the desired number of classes, producing the segmentation output.

By combining self-attention mechanisms and MLP layers, the SegFormer architecture achieves a balance between model complexity and segmentation accuracy. This makes it particularly suitable for real-time applications such as autonomous driving and video surveillance.

Algorithm 1 Calculation of TP, TN, FP, FN

```
Require: Predicted matrix $P$, Ground truth matrix $G$
Initialize $TP \leftarrow 0$, $TN \leftarrow 0$, $FP \leftarrow 0$, $FN \leftarrow 0$
for each pixel $i$ in $P$ do
    if $P[i] = 1$ and $G[i] = 1$ then
        $TP \leftarrow TP + 1$
    else if $P[i] = 0$ and $G[i] = 0$ then
        $TN \leftarrow TN + 1$
    else if $P[i] = 1$ and $G[i] = 0$ then
        $FP \leftarrow FP + 1$
    else if $P[i] = 0$ and $G[i] = 1$ then
        $FN \leftarrow FN + 1$
    end if
end for
return $TP$, $TN$, $FP$, $FN$
```

2.4 METRICS USED TO EVALUATE THE MODELS

To verify the quality of the results obtained, the following metrics were chosen: jaccard index, recall, precision, and F1 score, as they are reference metrics used in state-of-the-art semantic segmentation tasks (WANG *et al.*, 2023b; CHEN *et al.*, 2023). Since both the model output and the ground are binary matrices, the process described in algorithm 1 was used to calculate the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) for each pair predicted/ground truth mask.

The Jaccard index, also known as the Intersection over Union (IoU), is a significant metric in semantic segmentation and is defined by the following equation:

$$\text{Jaccard Index} = \frac{TP}{TP + FP + FN}. \tag{2.3}$$

The Jaccard index highlights the ratio of the intersection to the union of predicted and actual segmentation areas. This metric is effective for evaluating segmentation because it provides a straightforward measure of how closely the predicted segmentation aligns with the ground truth, which makes it particularly suitable for tasks where the scale of objects varies greatly (BEERS *et al.*, 2019).

Recall in semantic segmentation is defined as the proportion of actual positive pixels that were correctly identified. It is defined by the following equation:

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{2.4}$$

It focuses on the model's ability to capture all relevant cases in the image, which is crucial in medical imaging and other applications where missing a positive case can have significant consequences. This metric is highly valued in scenarios where the cost of false negatives is high, such as not detecting a possible condition in a medical diagnosis (SHOAIB *et al.*, 2022).

Finally, precision measures the proportion of predicted positive pixels that are truly positive, given by the following equation:

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{2.5}$$

It emphasizes the model's ability to deliver accurate positive predictions without many false positives. It is particularly important in contexts where the cost of a false positive is significant, such as in automated surveillance systems (PATRIKAR; PARATE, 2022).

Finally, the F1 score is defined by the harmonic mean between precision and recall:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{2.6}$$

In practical terms, the F1 score provides a single metric that balances both precision (the accuracy of positive predictions) and recall (the ability to capture all positive instances). This is particularly useful in scenarios where the class distribution is imbalanced or where both false positives and false negatives carry significant costs. Considering both precision and recall, the F1 score ensures that a model does not overly favor one metric at the expense of the other, thus offering a more comprehensive evaluation of the model's performance in identifying the target class (RIYANTO *et al.*, 2023).

2.4.1 Cross-Entropy

In this study, the segmentation task was treated as a pixel-by-pixel classification problem, thus, the model output will be a matrix where each pixel corresponds to the probability that the pixel, in the same position in the original image, represents deforestation. Consequently, each pixel in the prediction matrix and the ground-truth matrix can be seen as a probability function, therefore, it is correct to take the minimization of cross-entropy as a goal to maximize the models' accuracy rate. This is a common practice present in academic literature in semantic segmentation tasks using binary masks (WEI *et al.*, 2016)

The cross-entropy function measures the difference between two probability distribution functions over the same set of events. This measure is often used to quantify the error in classification problems. Cross entropy is defined as:

$$H(p,q) = -\sum_{x} p(x) \log q(x),$$

where p is the actual probability distribution of the data (ground-truth) and q is the probability distribution estimated by the model (model predictions). In the context of image segmentation, each pixel is treated as a separate event and $p(x)$ and $q(x)$ represent the probability of each class for that pixel. Minimizing cross-entropy means adjusting the model in such a way that the distribution q approaches p as closely as possible, thereby reducing the discrepancy between the model's predictions and the true values. This is crucial for enhancing the model's accuracy in correctly classifying whether a pixel belongs to a deforestation area or not.

CHAPTER 3

ONLINE SAR IMAGE LABELING EXPERIMENT

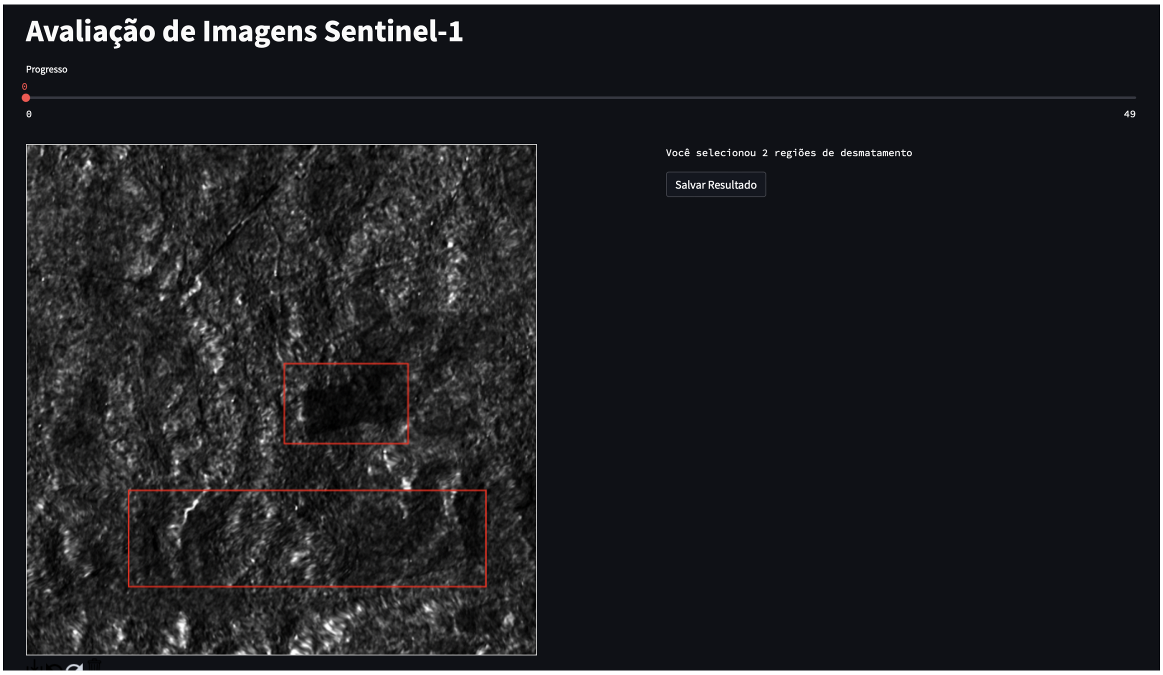
To compare the performance between automatic models and humans in the deforestation detection task, an online experiment was proposed to create a benchmark based on human performance in the task. The study was conducted online and promoted via email lists and social media platforms. The study included 24 volunteers who had different levels of experience in remote sensing. The assessment gathered deforestation bounding box annotations for 50 segments of pre-treated SLC Sentinel-1 pictures, where only the Vertical Transmit-Vertical Receive Polarisation (VV) band was used to produce a grayscale image. Besides the experiment, an artificial intelligence automatic model was used as the baseline. This model was based on a UNet architecture (RONNEBERGER *et al.*, 2015) due to its ease of training and proven effectiveness in satellite image segmentation(MCGLINCHY *et al.*, 2019; SHIRVANI *et al.*, 2023).

This chapter of the work was published in the XLI Brazilian Symposium on Telecommunications and Signal Processing, presented on October 10, 2023, at the São José dos Campos Technology Park (CORREIA *et al.*, 2023).

3.1 DEVELOPED PLATFORM

Although there are good annotation applications, such as CVAT(AL., 2020), we chose to develop a custom annotation system, which is simpler and has a more agile interface. Figure 3.1 shows a screenshot of the developed system. The system is based on Streamlit[1], an open source platform that allows the quick and efficient creation of web applications. The programming language used in this project was Python version 3.7. We adapted the code from Streamlit, implementing functions to collect annotations and save the results in a SQLite relational database. The platform was served online with the NGINX reverse proxy in a domain belonging

[1] Available on https://streamlit.io/ Accessed on May 2024



Avaliação de Imagens Sentinel-1

Progresso

Você selecionou 2 regiões de desmatamento

Figure 3.1. Screenshot showing the interface of the developed system. The volunteer must mark the bounding boxes using click' and drag with the mouse'. To finish, the user must click on the 'Save Result' button. Software originally available at sar.igorbispo99.com, currently offline.

to the author of this work. Using this online platform, participants viewed and annotated the deforested areas in SAR images.

3.2 VOLUNTEER PREPARATION

Participants were invited through email lists and social media. The invitation contained a link to a form, which contained a video[2] with an introduction to the problem of deforestation detection, a description of the image labeling process, and a demonstration of the developed software. In this video, examples of ground truth and desirable boxes shown in Figure 3.2 were also presented. These areas marked in green are the regions that contain deforestation according to the database used. Participants were instructed to mark the deforested regions using the smallest rectangle possible for each non-contiguous deforested area. On the left is the original image that the volunteers would see during the marking process, in the center is the same image with the ground-truth masks superimposed, and further to the right, the desired bounding boxes. These images were shown to volunteers as a reference for optimal labeling.

[2] Video available on https://www.youtube.com/watch?v=ijZRCheIXro

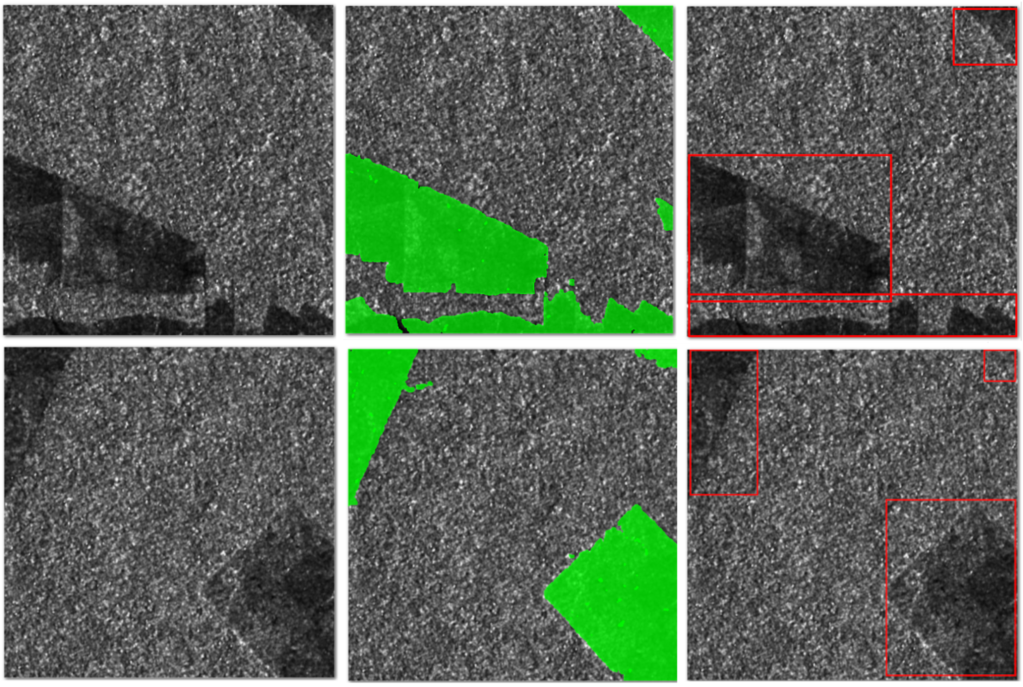


Figure 3.2. Composition of cuts from a Sentinel-1 scene with visible deforestation.

After viewing the video, volunteers were asked to complete a form on Google Forms [3] with five demographic statistical questions: Name, Gender, Age, Occupation, Level of Education; and two questions to gauge self-reported experience: **What is your level of experience with the topic of remote sensing? (0 to 4)** and **What is your level of experience with the topic of synthetic aperture radar? (0 to 4)**. The last two questions were later correlated with the results obtained in the box-checking stage for each of the participants, as shown in Table 3.1.

3.3 DATA PREPROCESSING

In total, 50 images were made available to each participant. The images correspond to 512×512 cuts extracted from an SLC capture already processed with radiometric calibration(PONZONI *et al.*, 2015), orbit correction, multilook (YAGÜE-MARTÍNEZ *et al.*, 2016), and anti-speckle filter(CHOI; JEONG, 2019) of the Sentinel-1A radar. The capture was made on 07/27/2021 and can be obtained through the Open Access Hub of the European Space Agency[4]. For the experiment, only the VV band of the signal was used, both for neural net-

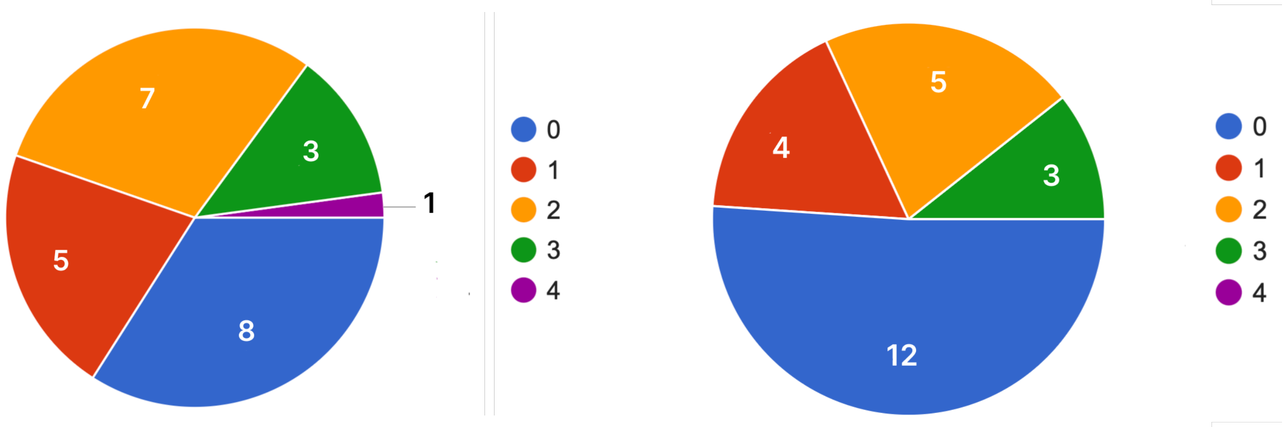
[3] Available on https://forms.gle/3Lc3w7WY6VuPThUW6

[4] https://scihub.copernicus.eu

work classification and for the experiment with volunteers.

3.4 SUBJECTIVE EXPERIMENT RESULTS

In total, 24 individuals participated in the experiment. The participants had different levels of experience. At the beginning of the experiment, participants reported their level of experience in remote sensing and synthetic aperture radars using a scale from 0 to 4, where 4 indicates that the participant is an expert and 0 indicates that they have no experience with the subject. Figures 3.3 (a) and (b) show graphs that illustrate the distribution of participants' experience in remote sensing and synthetic aperture radars, respectively. From these graphs, it can be observed that the experience of the participants in the field of remote sensing varied greatly, with 34% having no experience, 21% having little experience, 29.8% having some experience, 12.8% having a good level of experience, and 2.1% being experts on the subject. In synthetic aperture radars, about half (51.1%) of the participants had no experience, 17% had little experience, 21.3% had some experience, and 10.6% had a good level of experience.



(a)

(b)

Figure 3.3. Distribution of participants' self-declared experience in the field of: (a) remote sensing and (b) SAR.

To analyze the results of the subjective experiment, we grouped the demarcation annotations given by different volunteers for the same image. For each deforestation demarcation given by each participant for a specific image, the value 1 is added to the deforestation mask. Then, each pixel of the mask is divided by the number of participants who annotated that image, in order to normalize the mask values between 0 and 1. Finally, a threshold of 0.5 is applied to the image, so that values below 0.5 are considered negative and values above 0.5 are considered

Algorithm 2 Aggregate Predictions in Image M

```
Require: $N$ images
Require: $P$ people
Require: $B$ bounding boxes
$M_{512,512}$ matrix
for $i = 1$ to $N$ do
    $M_{512,512} \leftarrow 0$
    for $j = 1$ to $P$ do
        for all $b \in B_j$ do
            Add 1 to $M_{512,512}$ beneath $b$
        end for
    end for
    $M_{512,512} \leftarrow M_{512,512}/P$
    for $k$,$l$ in $M_{512,512}$ do
        if $M_{k,l} < 0.5$ then
            $M_{k,l} \leftarrow 0$
        else if $M_{k,l} > 0.5$ then
            $M_{k,l} \leftarrow 1$
        end if
    end for
end for
```

positive. An overview of the process is described in Algorithm 2.

Figure 3.4 shows all the rectangles marked by the participants for two of the test images and the corresponding mask generated with the proposed strategy. Boxes of the same color were marked by the same participant. This composition consists of four Sentinel-1 image crops submitted for evaluation. On the left, the VV polarization band of each crop is displayed, annotated with bounding boxes. Each box, color-coded to represent a different evaluator, is overlaid on the imagery. In the center, the mask created through the consolidation process is presented, as outlined in Algorithm 2 On the right, the corresponding ground truth mask is depicted.

Using the ground-truth markings of the images that show confirmed deforestation areas as a reference, we analyzed the performance of the participants and the model by calculating the following metrics: precision, F1 (the harmonic mean between precision and recall), intersection over union (IOU), and area under the curve (AUC). These performance metrics are considered standard in the literature for evaluating detection and segmentation models (SHIRVANI *et al.*, 2023; TOVAR *et al.*, 2021). It is worth noting that the AUC calculation is done on the non-binarized mask, i.e., without the application of the threshold. Table 3.1 presents the results obtained.

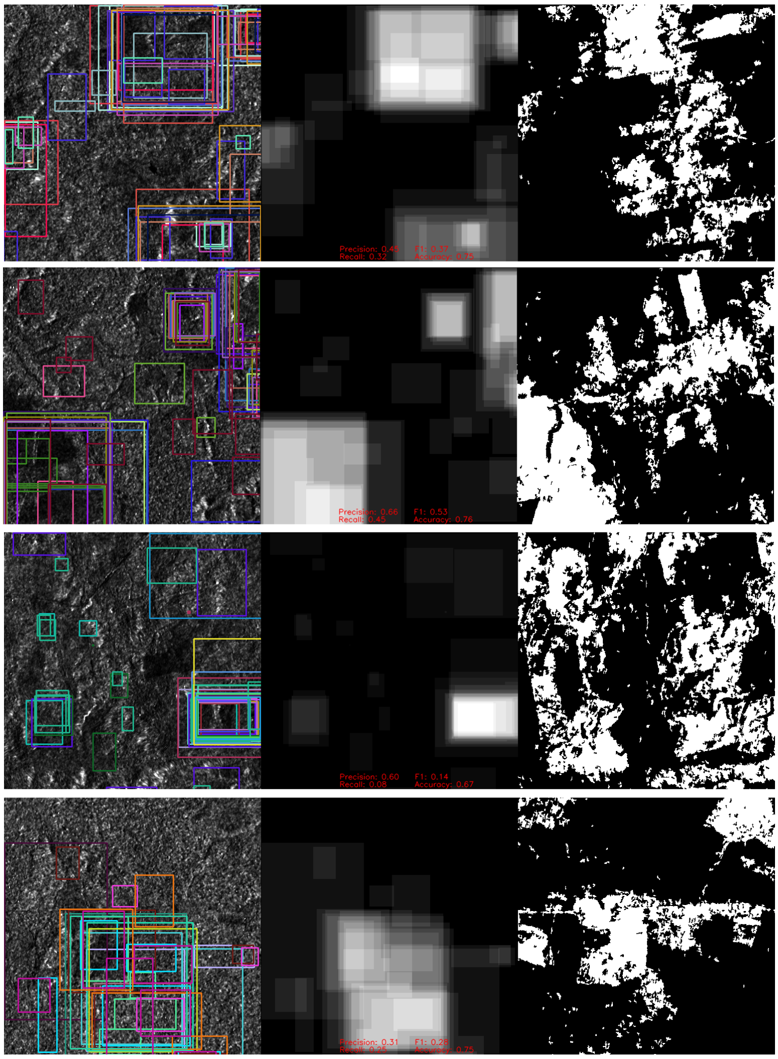


Figure 3.4. Composition with: Sentinel-1 crops, bounding boxes marked by the evaluators, the respective predicted mask and ground-truth.

3.5 AUTOMATIC BASELINE UNET-MODEL

The automatic model for deforestation detection is based on the UNet architecture (RONNEBERGER *et al.*, 2015). In the experiments conducted, UNet not only achieved performances close to those obtained by state-of-the-art models, but also surpassed human labelers in accuracy and processing time in the detection of deforestation in SAR images. This demonstrated superiority makes UNet a practical and efficient choice for the task at hand.

The UNet network implementation was done in Python 3.7 with PyTorch version 1.3. The model was trained in a database containing 1,279 512x512 cuts extracted from 4 SLC Sentinel-

Table 3.1. F1, IOU, AUC, and Precision metrics, grouped according to the participant's self-declared experience in Remote Sensing and SAR.

Self-Declared Experience	F1%	IOU%	AUC%	Precision%
Remote Sensing				
0	5.5	3.2	51.0	8.3
1	11.0	6.3	51.4	14.8
2	6.4	3.8	52.0	10.0
3	0.9	0.5	49.7	2.4
4	15.1	9.1	52.5	21.5
SAR				
0	6.6	3.8	51.1	10.7
1	10.9	6.2	51.6	13.4
2	2.4	1.5	52.3	3.5
3	7.9	4.6	51.8	11.8

Table 3.2. F1, IOU, ROC, and Precision metric values for the UNet-based deforestation detection model.

F1%	IOU%	AUC%	Precision%
9.5	27.5	67.0	37.8

1A scenes from the Amazon forest, captured on 07/27/2019, 08/04/2020, 08/13/2020, and 08/30/2021. The images were pre-processed following the same operation chain used to generate the images submitted to the online labeling system. It should be noted that there is no intersection between the areas of the scenes used in neural network training and those used for comparative tests. The model was trained for 5 epochs, with a batch size of 16 and a regressive dynamic learning rate, starting with a value of 0.00001. RMSProp was used as the gradient optimizer. The model training was done on a Windows 11 machine with an i7-11700K processor and an RTX 3080 GPU.

Table 3.2 presents the F1, IOU, AUC, and Precision values for the results obtained with UNet. When comparing these values with those of Tables 3.1, a clear difference in performance is observed. More specifically, the UNet-based detection model showed better performance than that obtained by the participants in the experiment. It should be noted that the participants' performance on this task was low, as demonstrated by the AUC. In this metric, a result of 50% corresponds to the performance of a random model. On the other hand, the UNet-based model showed an AUC of 67%, indicating moderate performance in the detection of deforestation.

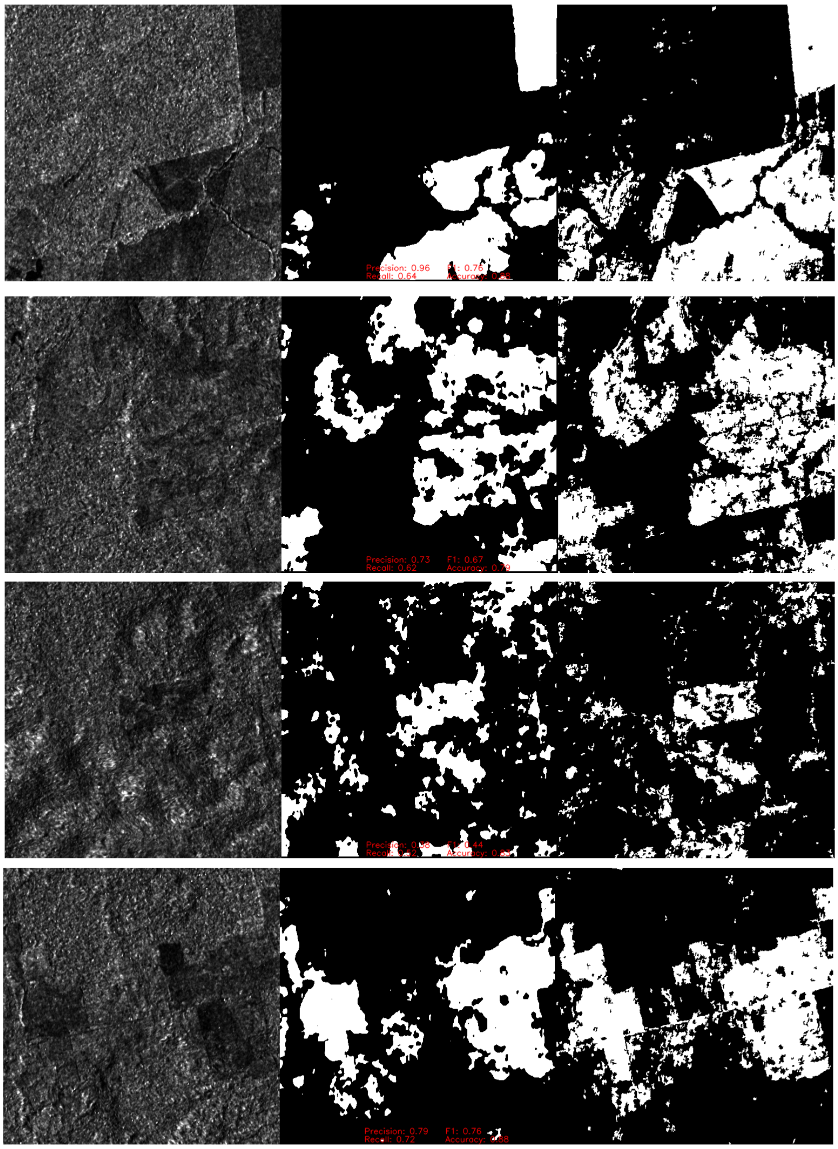


Figure 3.5. Composition of four Sentinel-1 image crops submitted for evaluation. On the left, the VV polarization band of each crop is displayed. In the center, the mask predicted by the UNet-based model is presented. On the right, the corresponding ground truth mask is depicted.

Finally, the qualitative visual results shown in Figure 3.5 confirm the metrics seen in Table 3.2 and show a greater match with groundtruth in the UNet-based model.

3.6 CONCLUSIONS

In this chapter, we investigated the relationship between the level of self-declared experience of individuals and the quality of annotations produced for deforested areas in SAR images, as well as the ability of an automatic system to detect these areas. The results suggest a low correlation between self-declared experience and the quality of annotations, indicating that greater declared knowledge in remote sensing and SAR does not necessarily lead to higher quality annotations.

Furthermore, the study showed that the accuracy of an automatic model, based on UNet, to detect deforested areas exceeded that obtained with participants. These findings demonstrate the potential for the use of deep learning in deforestation detection. However, this should be verified in future work, given the low number of expert participants and the limitations of the box annotation tool.

It is crucial to emphasize that the method chosen for collecting bounding box annotations was designed to optimize practicality and speed during the collection process, thereby enhancing participation rates in the experiment. It is reasonable to assume that this annotation method may not provide a fair comparison to the pixel-precise assessments performed by UNet. A more suitable alternative to enhance the accuracy of human annotators would involve the adoption of polygon-based annotations or even the use of a variable-sized brush tool, which would allow for pixel-precise annotations. In the next chapters, the task of detecting deforestation in SAR images will be expanded with the introduction of data fusion and state-of-the-art segmentation models.

CHAPTER 4

DEFORESTATION DETECTION USING DATA FUSION

This chapter outlines the methodologies for data processing and the development of machine learning models. The methods use a data fusion technique of optical and SAR data sourced from Sentinel-1 (see Section 2.1.2.1) and Sentinel-2 (see Section 2.1.1.1) satellites. In this scenario, data fusion refers to the combination of two distinct datasets to produce a more comprehensive and detailed informational aggregate.

4.1 DATA FUSION

The data fusion process consists of synchronizing and combining different types of data, allowing the complementary information of each sensor to be maximized (DOBLAS *et al.*, 2020). The RGB images from Sentinel-2 provide details about the color and texture of the land cover, while the SAR images from Sentinel-1 offer relevant information about the structure and moisture of the soil, regardless of atmospheric conditions or light levels.

Integrating these two types of data results in a series of benefits for machine learning applications, especially in the context of deforestation detection. Firstly, the combination of optical and radar data contributes to the robustness and reliability of the model, overcoming the individual limitations of each type of sensor(JOSHI *et al.*, 2016b). Furthermore, the resulting dataset is enriched with a greater diversity of features, allowing a more complete profile of the terrain.

In addition, the fusion of optical and SAR data can significantly improve the accuracy and sensitivity of the model in identifying deforested areas (DOBLAS *et al.*, 2020), which is particularly valuable in regions where deforestation may be subtle or obscured by cloud cover. This approach also offers flexibility to be applied in different scenarios and geographic conditions, making the model adaptable to a variety of landscapes. Therefore, the strategy of fusing optical and SAR data establishes itself as an effective approach in the automatic detection of deforestation, taking advantage of the unique capabilities of each type of sensor to create a more efficient and accurate monitoring system.

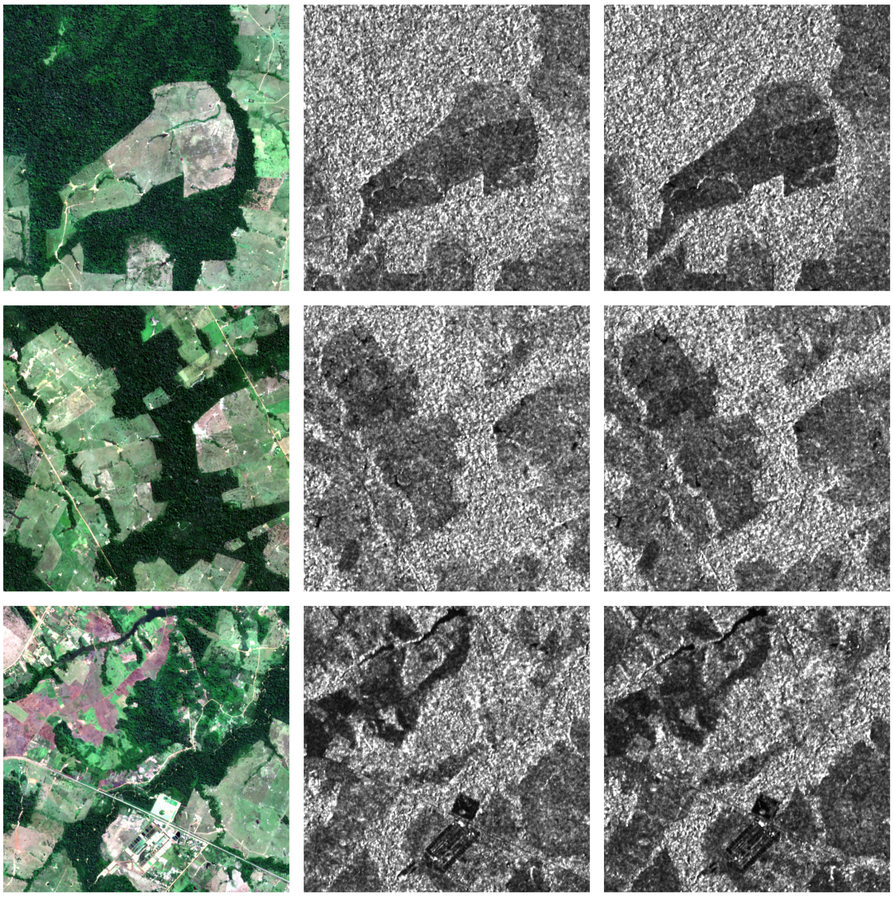


Figure 4.1. Composition with 3 Sentinel-1 and Sentinel-2 captures of the same scene.

In this study, the optical and SAR captures were fused using the scene coregistration, more detailed in Section 4.2.1, technique followed by stacking the images in the channel dimension. In this way, a Sentinel-2 RGB image containing three channels and a Sentinel-1 SAR capture containing two channels are composed into a resulting five-channel image that was used in the training and inference of the AI models. Figure 4.1 shows examples of cut-outs from scenes S1 and S2 before they were stacked and used for training[1]. The images on the left corresponds to the RGB bands of the Sentinel-2 stacked, the middle image is the Sentinel-1A band in vertical polarization (VV) and, further to the right, the Sentinel-1A band in horizontal polarization (VH)

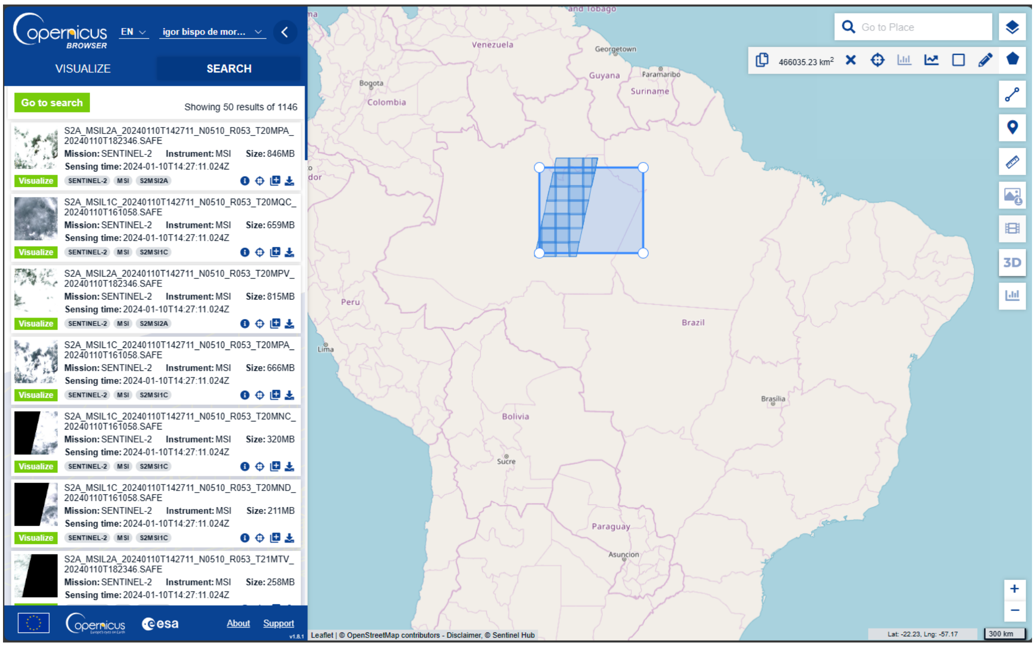
[1] The stacking order chosen was: R, G, B, VV and VH

4.2 DATA ACQUISITION AND PRE-PROCESSING

The acquisition of optical data from Sentinel-2, carried out through the public access portal of the European Space Agency (ESA), was manually coordinated to match the dates and times of the Sentinel-1 SLC images, as described in previous chapters. This temporal alignment is crucial for the data fusion process and for the effectiveness of automatic deforestation detection techniques in satellite images. A screenshot showing the interface of the Copernicus system can be seen in Figure 4.2.

The process of selecting Sentinel-2 optical images on the ESA portal involved searching for specific dates and times that matched exactly, or as closely as possible, the moments of the Sentinel-1 SLC image captures used as reference. This correspondence is essential to ensure the synchrony between the data sets, enabling a more accurate and detailed analysis of deforestation areas since, if there is a large temporal mismatch between SAR and optical capture, the deforestation labels may not be completely trustworthy.

After downloading, Sentinel-2 images were processed and aligned with Sentinel-1 radar data through the co-registration process, using the SNAP software, as shown in Figure 4.3. This process initially involved the preprocessing of the images, including atmospheric corrections on the optical images and the calibration of backscatter data on the SAR images.



Showing 50 results of 1146

S2A_MSIL2A_20240110T142711_N0510_R053_T20MPA_20240110T182346.SAFE
Mission: SENTINEL-2 **Instrument:** MSI **Size:** 846MB
Sensing time: 2024-01-10T14:27:11.024Z
SENTINEL-2 MSI S2MSI2A

S2A_MSIL1C_20240110T142711_N0510_R053_T20MQC_20240110T161058.SAFE
Mission: SENTINEL-2 **Instrument:** MSI **Size:** 659MB
Sensing time: 2024-01-10T14:27:11.024Z
SENTINEL-2 MSI S2MSI1C

S2A_MSIL2A_20240110T142711_N0510_R053_T20MPV_20240110T182346.SAFE
Mission: SENTINEL-2 **Instrument:** MSI **Size:** 815MB
Sensing time: 2024-01-10T14:27:11.024Z
SENTINEL-2 MSI S2MSI2A

S2A_MSIL1C_20240110T142711_N0510_R053_T20MPA_20240110T161058.SAFE
Mission: SENTINEL-2 **Instrument:** MSI **Size:** 666MB
Sensing time: 2024-01-10T14:27:11.024Z
SENTINEL-2 MSI S2MSI1C

S2A_MSIL1C_20240110T142711_N0510_R053_T20MNC_20240110T161058.SAFE
Mission: SENTINEL-2 **Instrument:** MSI **Size:** 320MB
Sensing time: 2024-01-10T14:27:11.024Z
SENTINEL-2 MSI S2MSI1C

S2A_MSIL1C_20240110T142711_N0510_R053_T20MND_20240110T161058.SAFE
Mission: SENTINEL-2 **Instrument:** MSI **Size:** 211MB
Sensing time: 2024-01-10T14:27:11.024Z
SENTINEL-2 MSI S2MSI1C

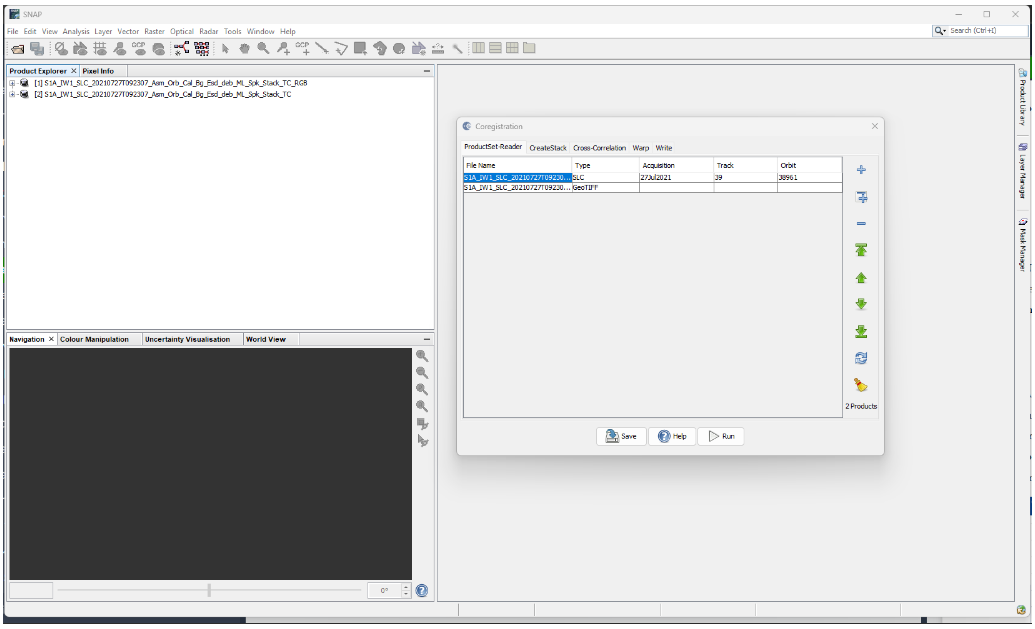
S2A_MSIL2A_20240110T142711_N0510_R053_T21MTV_20240110T182346.SAFE
Mission: SENTINEL-2 **Instrument:** MSI **Size:** 258MB
Sensing time: 2024-01-10T14:27:11.024Z

Figure 4.2. Interface of the Copernicus Browser, the portal where Sentinel-2 optical images were acquired.

4.2.1 Coregistration

The co-registration process is essential when dealing with captures from different sensors or at different times (ZITOVá; FLUSSER, 2003). In this process a reference image is first selected based on its clarity and comprehensive coverage of the area of interest, then ground control points (GCPs), which are specific, easily identifiable geographic landmarks, are pinpointed in both the reference and the secondary images. These GCPs are crucial as they anchor the alignment process, ensuring that each point corresponds accurately across both datasets.

Using these GCPs, the SNAP tool calculates a geometric transformation that mathematically adjusts the pixel coordinates of the secondary image to match those of the reference image. This adjustment is performed through algorithms that minimize spatial discrepancies, effectively overlaying the images with high precision.



[1] S1A_IW1_SLC_20210727T092307_Asm_Orb_Cal_Bg_Esd_deb_ML_Spk_Stack_TC_RGB
[2] S1A_IW1_SLC_20210727T092307_Asm_Orb_Cal_Bg_Esd_deb_ML_Spk_Stack_TC

File Name	Type	Acquisition	Track	Orbit
S1A_IW1_SLC_20210727T09230...	SLC	27Jul2021	39	38961
S1A_IW1_SLC_20210727T09230...	GeoTIFF			

Figure 4.3. SNAP software interface displaying the application of the coregistration function to align Sentinel-1 and Sentinel-2 images.

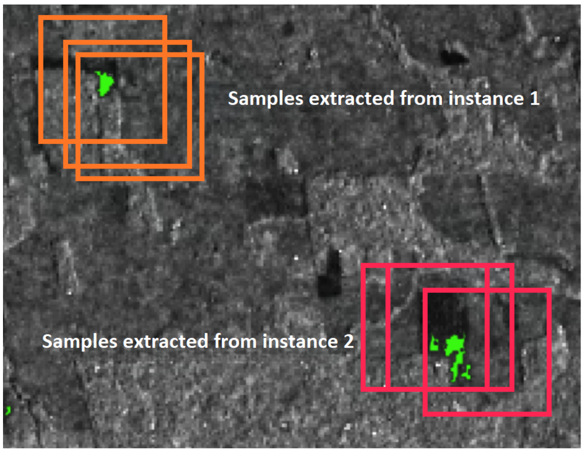
After mapping, the secondary image undergoes a resampling procedure, where its pixels are interpolated or extrapolated to fit the geometric framework of the reference image. This step is vital to ensure that the corresponding pixels in both images align perfectly, thus maintaining the integrity of geographic information when the images are integrated for analysis. The quality of coregistration is visually evaluated to ensure that the alignment is precise and free of any spatial or dimensional distortions.

From the accurately aligned optical and radar data, scenes with five channels are generated. The first three channels correspond to the red, green, and blue bands (bands 4, 5, and 6) of Sentinel-2, while the last two channels are dedicated to the VV and VH bands of Sentinel-1, as shown in Figure 4.1. This multiband alignment allows for comprehensive analysis across different spectral signatures and radar responses, enabling detailed assessments of the observed areas.

4.3 DATA PROCESSING FOR TRAINING

Considering that each of the compositions generated in the previous step has vertical and horizontal resolutions in the order of tens of thousands of pixels, training with the entire image at once is not feasible or effective due to memory and convergence constraints. Because of this, a strategy was developed to generate samples for training.

The pre-processing process is done as follows: for each instance of deforestation, 10 regions of size 512x512 are extracted from the fused images, each with slight random offsets. The same respective region is then cropped from the binary mask containing deforestation to generate the sample's ground-truth. A graphical representation of the process and pseudocode of the algorithm can be seen in Figure 4.4.


Samples extracted from instance 1
Samples extracted from instance 2

```
samples = {}
samples_per_region = 10

for each image I:
    for each deforested region R in I:
        for each i in [0, samples_per_region):
            crop := random_region_containing(R,
                                                size=(512, 512))
            deforestations := deforestations_in(crop)
            samples.add((deforestations, deforestations))
```

Figure 4.4. On the left: visual representation of sample extraction from an instance, where three samples were extracted per deforestation. On the right: the pseudo-algorithm used for cropping to create training batches.

Next, statistical processing was applied to remove outliers and normalize the input channels since the SAR (VV and VH) and optical (R, G, B) bands are in different magnitude ranges. After processing, all bands were scaled to fall within the range of 0 and 1. Specifically, outliers were managed by calculating the 2nd and 98th percentiles of each band, and values outside this range were clipped to the respective percentile values. This step effectively mitigated the impact of extreme values that could distort the analysis.

Subsequently, each band was normalized by subtracting the minimum and dividing by the range (maximum minus minimum) of the clipped data, ensuring a standardized scale. Finally, to facilitate visual inspection and potential further processing, the normalized data were scaled to the 8-bit range (0-255) by multiplying by 255 and converting to an unsigned 8-bit integer format. This transformation maintains the relative differences within the data while making them suitable for conventional image processing techniques and visualization tools.

After this process, 10000 deforestation samples were obtained, of which 80% was reserved for training and 20% for testing. It was certified that none of the training samples intersected with the test samples, even in cases where both came from the same Sentinel capture. As many of the samples had small regions of deforestation with a few square pixels and discontinuous regions caused by the automatic process used to generate labels, morphological opening and closing operations were applied. In Figure 4.5, is shown a composition with three examples of labels used to train the four segmentation architectures. On the left is the original noisy mask generated by the automatic process described in the 3.3 section, in the center is the same image after the opening process has been applied and on the right is the image after the closing process

The masks initially contained pixels with a binary grayscale value of 0 or 255, where 0 corresponds to no deforestation and 255 corresponds to deforestation. To be used in training, the masks were converted to 512x512 matrices in one-hot encoding [2] where each pixel corresponded to a two-dimensional vector with binary values. In all trainings, the task was treated as a pixel-by-pixel classification or semantic segmentation problem with cross-entropy loss.

In an initial test, the training and testing process was carried out with samples from the same capture. However, it was identified that this process, even if the disjunction between the training and testing regions is ensured, generated bias in the result. As a result, a new training process was carried out in which captures from different days and regions were used for training and testing.

[2] [1,0] for non-deforestation and [0,1] for deforestation

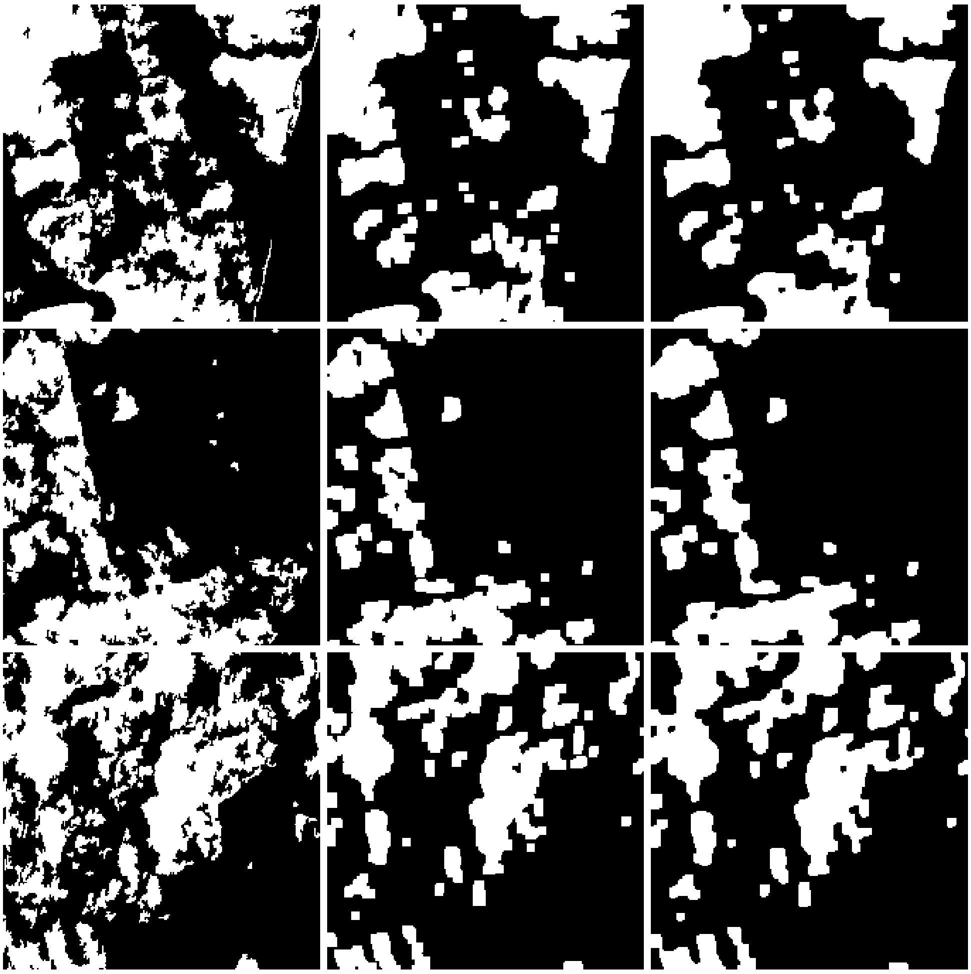


Figure 4.5. Composition with three examples of labels used to train the four segmentation architectures.

4.4 TRAINING

Each model was trained with the hyperparameters displayed in Table 4.1. The objective task was pixel-by-pixel classification using cross-entropy. The training phase of each model followed a rigorous approach, as indicated in Table 4.1, we chose a batch size of 16 for SegFormer and PP-LiteSeg to take advantage of the benefits of large batch training, which include stable gradients and efficient GPU utilization. For EfficientFormer, a smaller batch of 8 was necessary to accommodate its architectural complexity within the GPU memory constraints.

The number of iterations was chosen using as a reference the original articles of each of the models (XIE *et al.*, 2021; LI *et al.*, 2022b; PENG *et al.*, 2022) and using the early stopping method (CAWLEY; TALBOT, 2010). SegFormer and UNet underwent 160,000 and 40,000 iterations, respectively, allowing SegFormer's deeper architecture more time to learn intricate features. PP-LiteSeg required only 10,000 iterations, reflecting its efficiency in quickly reaching

Table 4.1. Hyperparameters for Segformer, Efficientformer, PP-LiteSeg, and UNet models

	SegFormer	**EfficientFormer**	**PP-LiteSeg**	**UNet**
Batch Size	16	8	16	16
Iterations	160,000	40,000	10,000	40,000
		Optimizer		
Type	AdamW	AdamW	SGD	Adam
Beta1	0.9	-	-	0.9
Beta2	0.999	-	-	0.999
Momentum	-	-	0.9	-
Weight Decay	0.01	0.0001	0.00004	0
		Learning Rate Scheduler		
Type	Polynomial Decay	Polynomial Decay	Polynomial Decay	Polynomial Decay
Learning Rate	0.00006	0.0006	0.01	0.001
End LR	-	0.000001	0	0
Power	1	0.9	0.9	0.9
		Loss		
Types	CrossEntropyLoss	CrossEntropyLoss	CrossEntropyLoss	CrossEntropyLoss

performance saturation.

The AdamW optimizer was chosen for training the SegFormer and EfficientFormer models using the original works of these architectures as reference (XIE *et al.*, 2021; LI *et al.*, 2022b). According to the original authors, AdamW aligns with the need to scale the weight decay independently of the learning rate, which is critical for transformer models that are sensitive to the scale of updates (LOSHCHILOV; HUTTER, 2019). PP-LiteSeg used a different optimizer, following the original article's authors (PENG *et al.*, 2022) who employed SGD for its momentum component, aiding convergence in regions with subtle gradient changes.

The learning rate scheduler was uniformly set to Polynomial Decay across models, ensuring a controlled learning rate reduction and helping in fine-tuning the models towards the end of the training (MISHRA; SARAWADEKAR, 2019). The learning rate was selected following the authors' directives, with SegFormer starting at a conservative rate to prevent divergence given its complexity, while PP-LiteSeg's aggressive rate to capitalize on its swift learning capabilities (XIE *et al.*, 2021; PENG *et al.*, 2022).

Each model was optimized against the pixel-by-pixel classification task using cross-entropy loss, a standard choice for such segmentation tasks, which directly aligns with maximizing pixel-classification accuracy.

Table 4.2. Number of Parameters for Segformer, Efficientformer, PP-LiteSeg, and UNet models

	SegFormer	**EfficientFormer**	**PP-LiteSeg**	**UNet**
Number of Parameters	3.7M	7.9M	8.1M	13.4M

Table 4.2 shows the number of trainable weights for each of the architectures used in this work. Although a greater number of parameters, given a similar architecture, generally show a positive correlation with performance in computer vision tasks, as observed by (LIU *et al.*, 2015) and (TZELEPIS *et al.*, 2019), the results obtained in this study suggest that a more modern architecture, with fewer parameters, can overcome a heavier model.

4.5 EXPERIMENTAL RESULTS

During the evaluation of training and inference on the models, it was noted that when the training and testing datasets are disjoint subsets of the same capture session, specifically, from the same date and geographical region, but involving nonoverlapping crops, the performance of all evaluated architectures significantly improved. In contrast, when the training data are sourced from a capture on one specific day and the testing data is derived from a capture on a different day, there is a noticeable deterioration in performance metrics.

This performance disparity can be attributed to potential changes in lighting at different times of the day, which particularly affects Sentinel-2 optical images, and variations in the image acquisition angle and calibration, which can create shadows impacting Sentinel-1 captures. Therefore, the outcomes of this analysis will be detailed in two distinct scenarios to reflect these variations in data handling: training and inference in the same image and training and inference on different images. Note that even in cases where training and inference were conducted on the same image, there was no overlap of areas in these two sets.

As indicated in Table 4.3, the SegFormerB0 architecture demonstrates superior average performance across all metrics evaluated for tasks involving training and inference within the same scene. Conversely, in scenarios where training and inference occur across different temporal captures, the transformer-based model PP-LiteSeg outperforms other models in all metrics except for average Recall, where SegFormerB0 maintains a higher score. We attribute this to its optimized Transformer design, which excels in capturing complex patterns. UNet's per-

Table 4.3. Metric results for models trained and tested on the same and different images. The highest values for each metric are highlighted in bold.

	Model			
Metric	**UNet**	**PPLiteSeg**	**EfficientFormerV2**	**SegFormerB0**
Trained and inference in the same image				
Average Jaccard	0.94	0.9519	0.9704	**0.9788**
Average F1	0.96	0.9753	0.9850	**0.9893**
Average Precision	0.95	0.9756	0.9851	**0.9894**
Average Recall	0.97	0.9752	0.9849	**0.9892**
Training and inference on different images				
Average Jaccard	0.46	**0.6596**	0.6571	0.6500
Average F1	0.63	**0.7789**	0.7776	0.7726
Average Precision	0.52	**0.7485**	0.7454	0.7393
Average Recall	0.81	0.8352	0.8430	**0.8513**

formance drop on different images may be result of its architecture's limited generalizability, which could be improved by integrating more robust feature extraction methods that are used in more modern architectures.

The Transformer-based models, namely PPLiteSeg and EfficientFormerV2, showcased their strong suit in generalization, which may be related to the Transformer's self-attention mechanism that captures long-range dependencies effectively. The close scores of EfficientFormerV2 and SegFormerB0, despite the difference in complexity, suggests that efficiency does not necessarily compromise accuracy.

In Figure 4.6, it can be seen that all models showed good performance, even in cases where the ground truth contained a reasonable amount of noise. This noise occurs because ground-truth labels were generated using an automatic random forest-based process, as described in Section 2.2, which, although supervised, is not ideal and may have been harmed in some cases by the morphological operation process.

Furthermore, in Figure 4.7 it is possible to see that the SegFormer-based model was able to better adjust the deforested area seen in the image, including being able to capture details and small deforestation, such as those seen in the capture of the third row. In contrast, Figure 4.8 shows some cases in which the models confused the clouds visible in the optical layers with false positives for deforestation, indicating that the models may not be robust to occlusion of the optical bands, even if the corresponding SAR bands are visible. This may suggest that the trained models are relying excessively on the RGB bands to make the prediction.

In general, the results, as shown by the objective metrics Tables 4.3 and visual inspection of the predictions, show that the models were able to predict the deforested region satisfactorily. In some cases, the groundtruth did not appear to visually correspond exactly to the deforested region visible in the RGB bands. In these cases of label noise, the models tended to adjust more faithfully to the visible area than the groundtruth. One last factor that may contribute to the mismatching of labels is the fact that Sentinel-1 and Sentinel-2 images were matched using the closest possible time. However, in some cases, it was not possible to achieve this temporal matching perfectly and there was a lag of up to a week between the two captures.

Lastly, the disparity in results between scenarios where training and inference are performed on the same image versus different images suggests that the models' generalization capability might be limited. It is recommended to train the models on a more diverse database in various weather conditions, times of day, and regions. The model with fewer parameters and the best performance in the first test scenario, SegFormerB0, experienced a significant drop in performance when the task changed to different images. This may indicate that the lighter model adjusted more closely to the specific characteristics of the training image, such as weather conditions (visible in the optical bands) and shadows caused by the SAR acquisition angle, rather than learning more generic patterns. In this latter scenario, the transformer network model with more parameters, PP-LiteSeg, displays the best metrics, possibly because it is a heavier model in terms of parameters, more modern, and may have a better capacity for generalization.

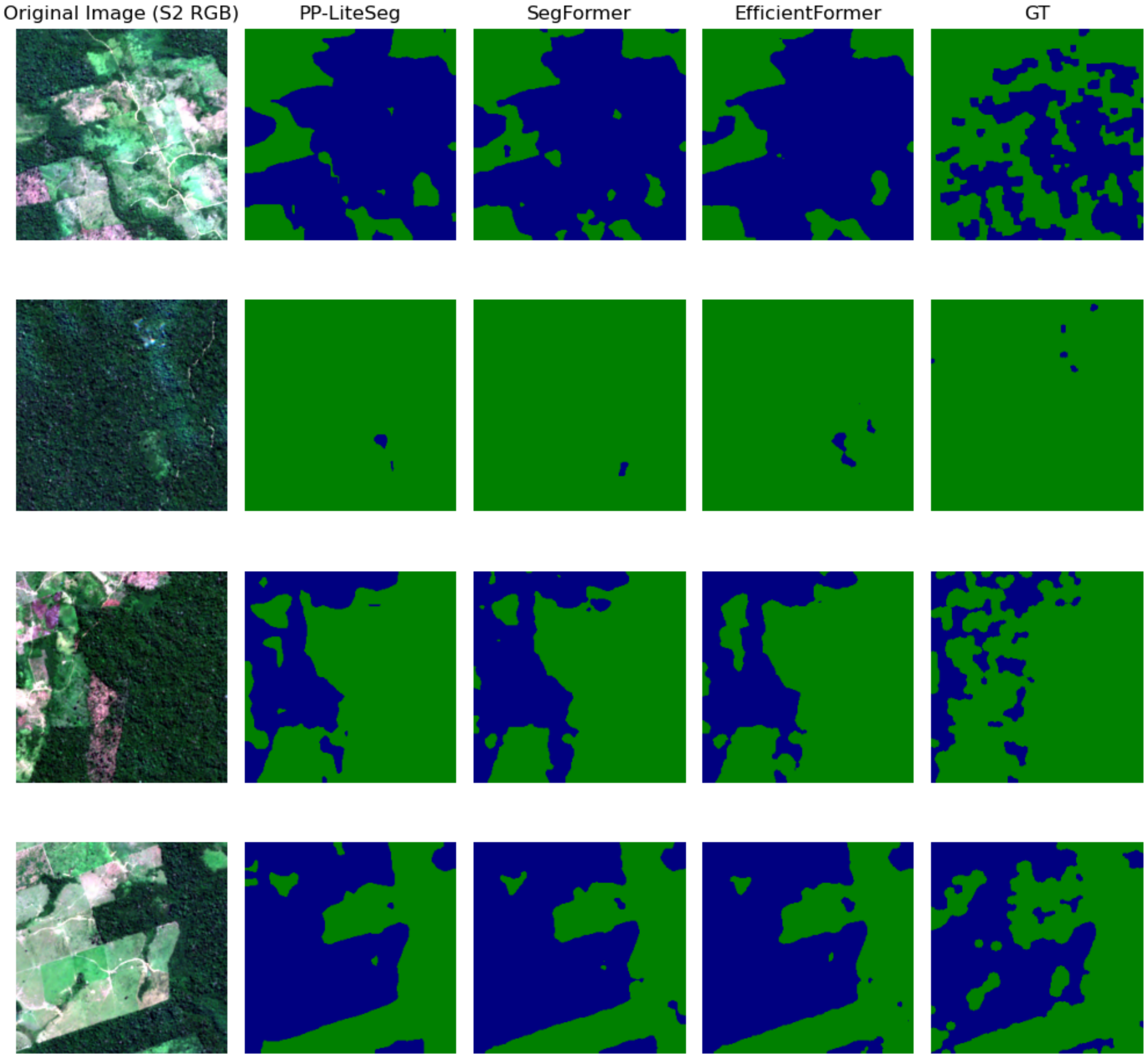


Figure 4.6. Composition with the prediction results obtained with the trained models, PPLiteSeg, SegFormerB0 and EfficientFormerV2 along with ground truth (last column). Green means forest and blue means deforestation.

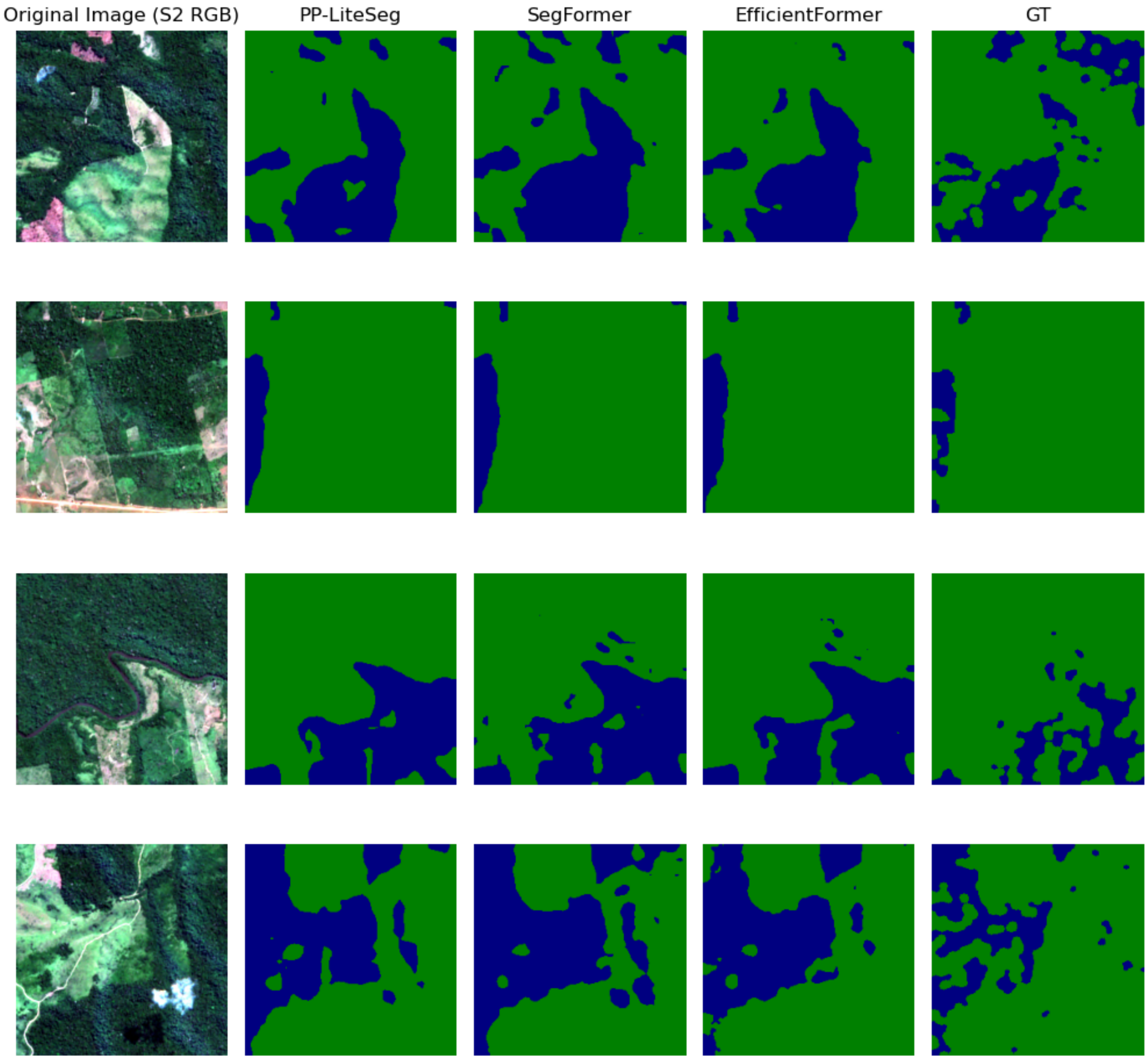


Figure 4.7. Composition with the prediction results obtained with the trained models, PPLiteSeg, SegFormerB0 and EfficientFormerV2 along with ground truth (last column). Green means forest and blue means deforestation.

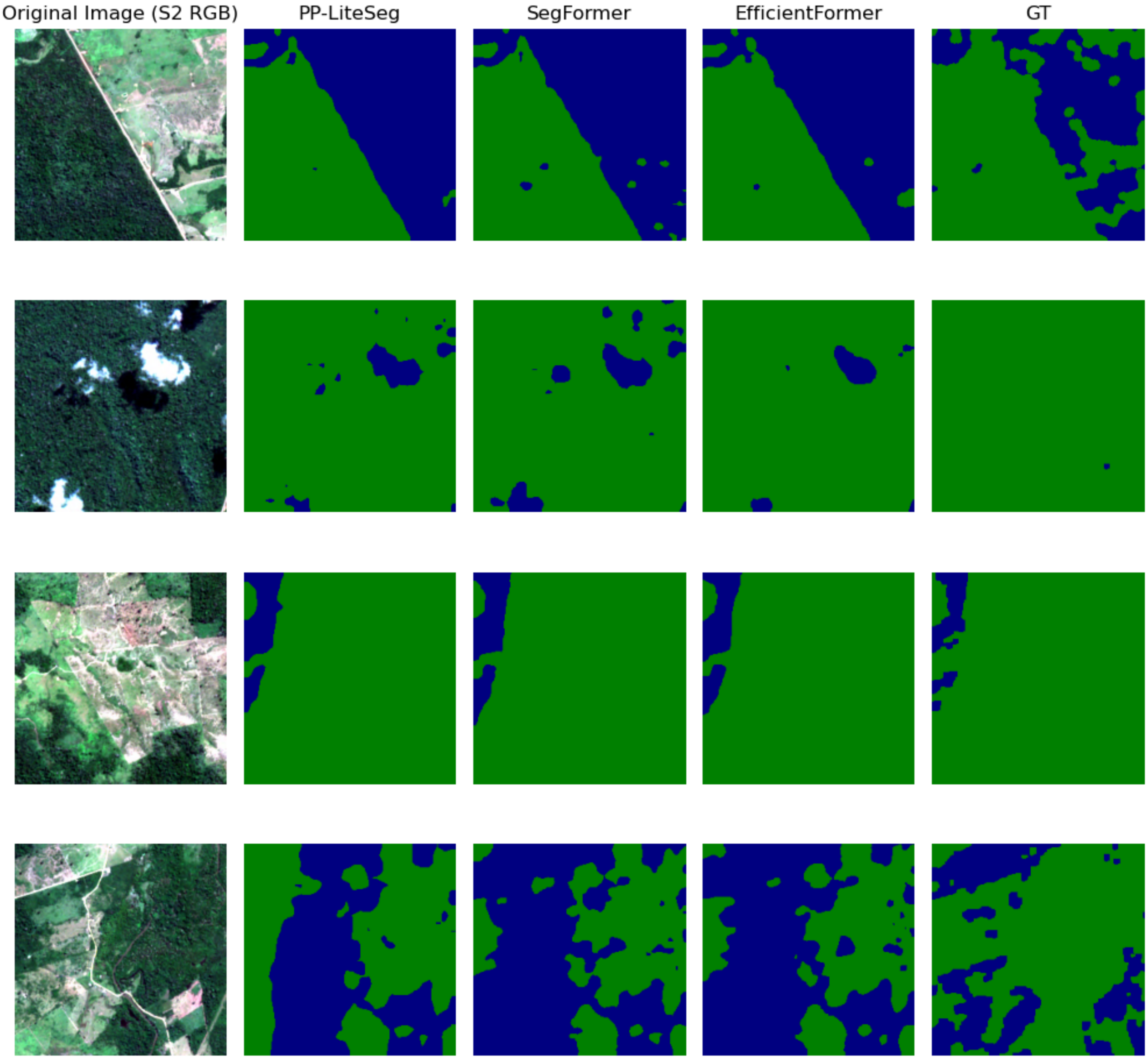


Figure 4.8. Composition with the prediction results obtained with the trained models, PPLiteSeg, SegFormerB0 and EfficientFormerV2 along with ground truth (last column) Green means forest and blue means deforestation.

CHAPTER 5

CONCLUSIONS

Throughout this study, a comprehensive examination of manual (Chapter 3) and automatic (Chapter 4) deforestation detection methods was carried out, revealing insights into the efficacy and practicality of each approach. The investigation of the relationship between the levels of self-declared experience of individuals and the quality of their annotations of deforested areas in the SAR images indicated a weak correlation, as in Section 3.1. This finding may challenge the assumption that higher self-proclaimed expertise in remote sensing and SAR leads to superior annotation quality, however, future studies should be conducted, using a larger number of participants and more precise annotation techniques, such as pixel-by-pixel, to investigate this fact.

In contrast, the performance of an automated model based on the UNet architecture surpassed the accuracy of human-generated annotations, Section 3.2. This success underscores the advantage and potential of deep learning models in the recognition of deforested areas, particularly when human expertise does not guarantee improved detection. This result, although interesting and enlightening, should be approached with a degree of skepticism because, for practical reasons, the human annotators did not have access to all the tools necessary to perform pixel-precise annotation.

Furthermore, the comparative analysis of segmentation models, including transformer-based SegFormerB0 (Section 2.3.5), EfficientFormerV2 (Section 2.3.4), and PP-LiteSeg (Section 2.3.3), along with conventional UNet, showcased the remarkable ability of transformer models to generalize. The superior performance of SegFormerB0, attributed to its optimized Transformer design, highlights the critical role of advanced architectural features, such as self-attention, in handling complex patterns and achieving high accuracy. These results were particularly interesting because this model had the fewest trainable parameters despite being the most modern architecture. This improved performance might be attributed to potential detriment to generalization ability in the scenario where training and inference were performed on the same image but in different regions. In the task where training and testing were performed on different images, the heavier transformer-based model exhibited better performance.

The results of the tests on different images revealed that, when models are trained and tested on the same images, we likely have a very optimistic and unrealistic performance, given that, in a real-world context, training would be done on a historical database and inference would be made on future data, so that, almost always, the training and inference scenes would be from very different climatic moments and situations.

In addition, it was identified that the ground truth generation method, using random forest and semi-human supervision, contained extremely small deforestation markings which hindered the training of the models. To mitigate this, we propose that a complete review of the ground truth database be carried out with manual editing of all masks used, if possible, by experts in remote sensing and comparing with other data sources beyond Sentinel-1. Furthermore, advanced data augmentation techniques such as CutMix (YUN *et al.*, 2019) and MixUp (ZHANG *et al.*, 2017) can be implemented to regularize the models and reduce overfitting. Lastly, considering that ground truth labels may not be completely accurate, it would be advisable to use a label smoothing technique (MÜLLER *et al.*, 2019) to reduce overconfident predictions.

This study sets the stage for future exploration in the detection of deforestation by advocating for the fusion of data and cutting-edge segmentation models. The advances in automatic deforestation detection showcased here promise a significant step forward in environmental monitoring and the preservation of our forests.

Regarding the labeling experiment part of this work, future studies should be conducted with a larger number of participants, especially with the participation of more experts in SAR image analysis, to verify the apparent lack of correlation between the quality of annotations and experience self-declared. Furthermore, it is important to provide labelers with more precise annotation methods for a finer delimitation of the deforested area.

Regarding detection with fused data, it is worth carrying out a more detailed study, in future work, of different deforestation segmentation scenarios, such as: segmentation using only optical data, only SAR data and both fused data, maintaining the same architectures and base of data to assess whether the merger is, in fact, beneficial, and identify possible situations of low performance in each of the scenarios.

Finally, it is interesting to investigate the use of simpler segmentation techniques using color thresholds, random forest and more modest classifiers to verify whether the use of complex architectures, such as transformers, is justified, as simple methods also show good performance in this task as observed in the literature (FU *et al.*, 2018; HOLLOWAY *et al.*, 2019).

BIBLIOGRAPHY

ADARME, M. O.; FEITOSA, R.; HAPP, P.; ALMEIDA, C.; GOMES, A. Evaluation of deep learning techniques for deforestation detection in the brazilian amazon and cerrado biomes from remote sensing imagery. *Remote. Sens.*, v. 12, p. 910, 2020. Cited on page 6.

AL., B. S. et. *opencv/cvat: v1.1.0.* Zenodo, 2020. Disponível em: <https://doi.org/10.5281/zenodo.4009388>. Cited on page 37.

ALI, R.; CHUAH, J. H.; TALIP, M. S. A.; MOKHTAR, N.; SHOAIB, M. A. Structural crack detection using deep convolutional neural networks. *Automation in Construction*, Elsevier, v. 133, p. 103989, 2022. Cited on page 7.

ALLISON, R.; JOHNSTON, J.; CRAIG, G.; JENNINGS, S. Airborne optical and thermal remote sensing for wildfire detection and monitoring. *Sensors (Basel, Switzerland)*, v. 16, 2016. Cited on page 9.

ALMEIDA, C. A. de; MAURANO, L. E. P.; VALERIANO, D. de M.; CAMARA, G.; VINHAS, L.; GOMES, A. R.; MONTEIRO, A. M. V.; SOUZA, A. A. de A.; RENNÓ, C. D.; SILVA, D. E. *et al.* Methodology for forest monitoring used in prodes and deter projects. *CEP*, v. 12, n. 010, 2021. Cited on page 17.

AZIMI, S.; HENRY, C.; SOMMER, L.; SCHUMANN, A.; VIG, E. Skyscapes fine-grained semantic understanding of aerial scenes. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, p. 7392–7402, 2019. Cited on page 21.

BAATZ, M.; SCHäPE, A. Multiresolution segmentation: An optimization approach for high quality multi-scale image segmentation. In: *Angewandte Geographische Informationsverarbeitung XII.* Heidelberg, Germany: Wichmann, 2000. p. 12–23. ISBN 0273-9615. Cited on page 18.

BEERS, F. V.; LINDSTRöM, A.; OKAFOR, E.; WIERING, M. Deep neural networks with intersection over union loss for binary image segmentation. p. 438–445, 2019. Cited on page 34.

BOERNER, W. Recent advances in polarimetric-interferometric sar theory and technology and its application. *13th International Conference on Microwaves, Radar and Wireless Communications. MIKON - 2000. Conference Proceedings (IEEE Cat. No.00EX428)*, v. 3, p. 212–229 vol.3, 2000. Cited on page 14.

BROWNLEE, J. *Deep Learning with Convolutional Neural Networks.* [S.l.]: Machine Learning Mastery, 2018. Cited on page 22.

BROWNLEE, J. A gentle introduction to pooling layers for convolutional neural networks. *Machine Learning Mastery*, 2019. Cited on page 23.

CAWLEY, G. C.; TALBOT, N. L. C. A systematic review on overfitting control in shallow and deep neural networks. *Journal of Machine Learning Research*, v. 11, p. 2079–2107, 2010. Disponível em: <https://link.springer.com/article/10.1007/s10462-010-9163-5>. Cited on page 52.

CHATZIANTONIOU, A.; PETROPOULOS, G. P.; PSOMIADIS, E. Co-orbital sentinel 1 and 2 for lulc mapping with emphasis on wetlands in a mediterranean setting based on machine learning. *Remote Sensing*, MDPI, v. 9, n. 12, p. 1259, 2017. Cited on page 17.

CHEN, Z.; DUAN, Y.; WANG, W.; HE, J.; LU, T.; DAI, J.; QIAO, Y. Vision transformer adapter for dense predictions. In: *International Conference on Learning Representations*. [S.l.: s.n.], 2023. Cited on page 34.

CHOI, H.; JEONG, J. Speckle noise reduction technique for sar images using statistical characteristics of speckle noise and discrete wavelet transform. *Remote Sensing*, MDPI, v. 11, n. 10, p. 1184, 2019. Cited on page 39.

CLEMENT, M. T.; CHI, G.; HO, H. C. Urbanization and land-use change: A human ecology of deforestation across the united states, 2001-2006. *Sociological inquiry*, Blackwell Publishing Ltd, HOBOKEN, v. 85, n. 4, p. 628–653, 2015. ISSN 0038-0245. Cited on page 1.

CLOUDE, S.; PAPATHANASSIOU, K. Polarimetric sar interferometry. *IEEE Trans. Geosci. Remote. Sens.*, v. 36, p. 1551–1565, 1998. Cited on page 15.

CORREIA, I. B. M. C.; FARIAS, M.; HUNG, E.; GUIMARãES, U. S.; JR, H. V.; RODRIGUES, T. B. Estudo comparativo da detecção de desmatamento em cenas sentinel-1 da floresta amazônica. In: *XLI Simpósio Brasileiro de Telecomunicações e Processamento de Sinais (SBrT2023)*. [s.n.], 2023. Disponível em: <https://biblioteca.sbrt.org.br/articles/4561>. Cited on page 37.

DIAMANTIS, D.; IAKOVIDIS, D. Pooling in convolutional neural networks for medical image analysis: a survey and an empirical study. *Neural Computing and Applications*, Springer, v. 32, n. 23, p. 16477–16492, 2020. Cited on page 22.

DOBLAS, J.; SHIMABUKURO, Y.; SANT'ANNA, S.; CARNEIRO, A.; ARAGãO, L.; ALMEIDA, C. Optimizing near real-time detection of deforestation on tropical rainforests using sentinel-1 data. *Remote Sensing*, v. 12, n. 23, 2020. ISSN 2072-4292. Disponível em: <https://www.mdpi.com/2072-4292/12/23/3922>. Cited on page 46.

DOSOVITSKIY, A.; BEYER, L.; KOLESNIKOV, A.; WEISSENBORN, D.; ZHAI, X.; UNTERTHINER, T.; DEHGHANI, M.; MINDERER, M.; HEIGOLD, G.; GELLY, S.; USZKOREIT, J.; HOULSBY, N. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. Disponível em: <https://arxiv.org/abs/2010.11929>. Cited on page 24.

ESA. *About Copernicus*. 2023. Disponível em: <https://www.copernicus.eu/en/about-copernicus>. Cited on page 15.

ESA. *SENTINEL-1 MISSION GUIDE*. 2023. Disponível em: <https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-1>. Cited 4 times on pages iii, 7, 15, and 16.

ESA. *SENTINEL-2 MISSION GUIDE*. 2023. Disponível em: <https://sentinel.esa.int/web/sentinel/missions/sentinel-2>. Cited 3 times on pages iii, 10, and 11.

FERNANDES, G. W.; OLIVEIRA, H. F. M.; BERGALLO, H. G.; BORGES-JUNIOR, V. N. T.; COLLI, G.; FERNANDES, S.; FONSêCA, N. C.; GARDA, A. A.; GRELLE, C. E. V.; NUNES, A. V.; PERILLO, L. N.; ROCHA, T. C.; RODRIGUES, D. J.; SILVEIRA-FILHO, R. R. da; STREIT, H.; TOMA, T. S. P.; VIANA, P. L.; ROQUE, F. O. Hidden costs of europe's deforestation policy. *Science (American Association for the Advancement of Science)*, United States, v. 379, n. 6630, p. 341–342, 2023. ISSN 0036-8075. Cited on page 1.

FONSECA, I. F. d.; LINDOSO, D. P.; BURSZTYN, M. Deforestation (lack of) control in the brazilian amazon: from strengthening to dismantling governmental authority (1999-2020). *Sustainability in Debate*, v. 13, n. 2, p. 12–31, Aug. 2022. Disponível em: <https://periodicos.unb.br/index.php/sust/article/view/44532>. Cited on page 3.

FU, H.; SHEN, Y.; LIU, J.; HE, G.; CHEN, J.; LIU, P.; QIAN, J.; LI, J. Cloud detection for fy meteorology satellite based on ensemble thresholds and random forests approach. *Remote. Sens.*, v. 11, p. 44, 2018. Cited on page 62.

GEUDTNER, D.; TORRES, R.; SNOEIJ, P.; DAVIDSON, M.; ROMMEN, B. Sentinel-1 system capabilities and applications. In: *2014 IEEE Geoscience and Remote Sensing Symposium*. [S.l.: s.n.], 2014. p. 1457–1460. Cited on page 3.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016. ISBN 9780262035613. Cited 2 times on pages 21 and 22.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*. [S.l.]: IEEE, 2015. v. 37, n. 9, p. 1904–1916. Cited on page 29.

HOLLOWAY, J.; HELMSTEDT, K.; MENGERSEN, K.; SCHMIDT, M. A decision tree approach for spatially interpolating missing land cover data and classifying satellite images. *Remote. Sens.*, v. 11, p. 1796, 2019. Cited on page 62.

JING, J.; WANG, Z.; RÄTSCH, M.; ZHANG, H. Mobile-unet: An efficient convolutional neural network for fabric defect detection. *Textile Research Journal*, SAGE Publications Sage UK: London, England, v. 92, n. 1-2, p. 30–42, 2022. Cited on page 7.

JOHN, D.; ZHANG, C. An attention-based u-net for detecting deforestation within satellite sensor imagery. *International Journal of Applied Earth Observation and Geoinformation*, v. 107, p. 102685, 2022. ISSN 1569-8432. Cited on page 7.

JOSHI, N.; BAUMANN, M.; EHAMMER, A.; FENSHOLT, R.; GROGAN, K.; HOSTERT, P.; JEPSEN, M. R.; KUEMMERLE, T.; MEYFROIDT, P.; MITCHARD, E. T. *et al.* A review of the application of optical and radar remote sensing data fusion to land use mapping and monitoring. *Remote Sensing*, MDPI, v. 8, n. 1, p. 70, 2016. Cited on page 17.

JOSHI, N.; BAUMANN, M.; EHAMMER, A.; FENSHOLT, R.; GROGAN, K.; HOSTERT, P.; JEPSEN, M. R.; KUEMMERLE, T.; MEYFROIDT, P.; MITCHARD, E. T. A.; REICHE, J.; RYAN, C. M.; WASKE, B. A review of the application of optical and radar remote sensing data fusion to land use mapping and monitoring. *Remote Sensing*, v. 8, n. 1, 2016. ISSN 2072-4292. Disponível em: <https://www.mdpi.com/2072-4292/8/1/70>. Cited on page 46.

LASSALLE, P.; INGLADA, J.; MICHEL, J.; GRIZONNET, M.; MALIK, J. A scalable tile-based framework for region-merging segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, IEEE, v. 53, n. 11, p. 5473–5485, 2015. Cited on page 18.

LEITE-FILHO, A. T.; SOARES-FILHO, B.; DAVIS, J.; ABRAHãO, G.; BöRNER, J. Deforestation reduces rainfall and agricultural revenues in the brazilian amazon. *Nature Communications*, v. 12, 2021. Cited on page 15.

LEYVA-MAYORGA, I.; MARTINEZ-GOST, M.; MORETTI, M.; PéREZ-NEIRA, A.; V'AZQUEZ, M. A.; POPOVSKI, P.; SORET, B. Satellite edge computing for real-time and very-high resolution earth observation. *IEEE Transactions on Communications*, v. 71, p. 6180–6194, 2022. Cited on page 10.

LI, M.; RUI, J.; YANG, S.; LIU, Z.; REN, L.; MA, L.; LI, Q.; SU, X.; ZUO, X. Method of building detection in optical remote sensing images based on segformer. *Sensors (Basel, Switzerland)*, v. 23, 2023. Cited on page 7.

LI, X.; CHENG, Y.; FANG, Y.; LIANG, H.; XU, S. 2dsegformer: 2-d transformer model for semantic segmentation on aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, v. 60, p. 1–13, 2022. Cited on page 7.

LI, Y.; HU, J.; WEN, Y.; EVANGELIDIS, G.; SALAHI, K.; WANG, Y.; TULYAKOV, S.; REN, J. Rethinking vision transformers for mobilenet size and speed. *arXiv preprint arXiv:2212.08059*, 2022. Cited 7 times on pages iv, 7, 24, 30, 31, 52, and 53.

LIU, B.; WANG, M.; FOROOSH, H.; TAPPEN, M.; PENSKY, M. Sparse convolutional neural networks. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 806–814, 2015. Cited on page 54.

LIU, Y.; CHU, L.; CHEN, G.; WU, Z.; CHEN, Z.; LAI, B.; HAO, Y. *PaddleSeg: A High-Efficient Development Toolkit for Image Segmentation*. 2021. Cited 2 times on pages 7 and 24.

LOSHCHILOV, I.; HUTTER, F. Decoupled weight decay regularization. In: *International Conference on Learning Representations*. [s.n.], 2019. Disponível em: <https://openreview.net/forum?id=Bkg6RiCqY7>. Cited on page 53.

MACEDO, K. D. de; MASALIAS, G.; COCCIA, A.; META, A. Recent l-c- and x-band metasensing airborne sar campaigns for emerging applications. *2020 17th European Radar Conference (EuRAD)*, p. 190–193, 2021. Cited on page 14.

MADHUGIRI, D. Convolutional neural networks. *KnowledgeHut*, 2020. Cited on page 23.

MARQUES, A.; MARTINS, I. S.; KASTNER, T.; PLUTZAR, C.; THEURL, M. C.; EISENMENGER, N.; HUIJBREGTS, M.; WOOD, R.; STADLER, K.; BRUCKNER, M.; CANELAS, J.; HILBERS, J.; TUKKER, A.; ERB, K.; PEREIRA, H. Increasing impacts of land-use on biodiversity and carbon-sequestration driven by population and economic growth. *Nature ecology and evolution*, v. 3, p. 628 – 637, 2019. Cited on page 2.

MARZANO, F.; MORI, S.; PIERDICCA, N.; PULVIRENTI, L.; WEINMAN, J. Characterization of atmospheric precipitation effects on spaceborne synthetic aperture radar response at x, ku, ka band. *European Journal of Remote Sensing*, p. 73–88, 2009. Cited on page 14.

MCCOY, R. D.; TANENHAUS, M. E. Synthetic aperture radar. *Inverse Synthetic Aperture Radar Imaging with MATLAB® Algorithms*, 1992. Cited on page 13.

MCCULLUM, N. The relu and pooling layers in convolutional neural networks. *Nick McCullum*, 2019. Cited on page 23.

MCGLINCHY, J.; JOHNSON, B.; MULLER, B.; JOSEPH, M.; DIAZ, J. Application of unet fully convolutional neural network to impervious surface segmentation in urban environment from high resolution satellite imagery. In: *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*. [S.l.: s.n.], 2019. p. 3915–3918. Cited 2 times on pages 25 and 37.

MISHRA, P.; SARAWADEKAR, K. Polynomial learning rate policy with warm restart for deep neural network. In: IEEE. *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*. [S.l.], 2019. p. 2087–2092. Cited on page 53.

MOREIRA, A.; PRATS-IRAOLA, P.; YOUNIS, M.; KRIEGER, G.; HAJNSEK, I.; PAPATHANASSIOU, K. P. A tutorial on synthetic aperture radar. *IEEE Geoscience and Remote Sensing Magazine*, v. 1, n. 1, p. 6–43, 2013. Cited 4 times on pages iii, vi, 13, and 14.

MÜLLER, R.; KORNBLITH, S.; HINTON, G. E. When does label smoothing help? *CoRR*, abs/1906.02629, 2019. Disponível em: <http://arxiv.org/abs/1906.02629>. Cited on page 61.

NASA. *Landsat*. 2023. Disponível em: <https://www.nasa.gov/mission_pages/landsat/main/index.html>. Cited on page 3.

NIU, R.; ZHI, X.; JIANG, S.; ZHANG, W.; GONG, J. Development of new ultra-large aperture optical remote sensing imaging technology. v. 12752, p. 127520H – 127520H–7, 2023. Cited on page 10.

NOBRE, C. A.; SAMPAIO, G.; SALAZAR, L. Mudanças climáticas e amazônia. *Ciência e Cultura*, Sociedade Brasileira para o Progresso da Ciência, v. 59, n. 3, p. 22–27, 2007. Cited on page 1.

PAJANKAR, A.; JOSHI, A. *Convolutional Neural Networks: A Comprehensive Guide*. [S.l.]: Springer, 2022. ISBN 978-1-4842-7921-2. Cited on page 22.

PATRIKAR, D. R.; PARATE, M. R. Anomaly detection using edge computing in video surveillance system: review. *International Journal of Multimedia Information Retrieval*, v. 11, n. 2, p. 85–110, 6 2022. ISSN 2192-662X. Disponível em: <https://doi.org/10.1007/s13735-022-00227-8>. Cited on page 35.

PENG, J.; LIU, Y.; TANG, S.; HAO, Y.; CHU, L.; CHEN, G.; WU, Z.; CHEN, Z.; YU, Z.; DU, Y.; DANG, Q.; LAI, B.; LIU, Q.; HU, X.; YU, D.; MA, Y. *PP-LiteSeg: A Superior Real-Time Semantic Segmentation Model*. 2022. Cited 10 times on pages iv, 7, 24, 26, 27, 28, 29, 30, 52, and 53.

PEPE, M.; FREGONESE, L.; SCAIONI, M. Planning airborne photogrammetry and remote-sensing missions with modern platforms and sensors. *European Journal of Remote Sensing*, v. 51, p. 412 – 436, 2018. Cited on page 9.

PERUMAL, B.; KALAIYARASI, M.; DENY, J.; MUNEESWARAN, V. Forestry land cover segmentation of sar image using unsupervised ilkfcm. *Materials Today: Proceedings*, 2021. Cited on page 6.

PIMENTA, G. A.; DALLAQUA, F. B. J. R.; FAZENDA, Á. L.; FARIA, F. Neuroevolution-based classifiers for deforestation detection in tropical forests. *2022 35th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, v. 1, p. 13–18, 2022. Cited on page 6.

PONZONI, F. J.; ANTUNES, M. A. H.; PINTO, C. T.; LAMPARELLI, R. A. C.; JUNIOR, J. Z. *Calibração de sensores orbitais*. [S.l.]: Oficina de Textos, 2015. Cited on page 39.

RIYANTO, S.; SITANGGANG, I. S.; DJATNA, T.; ATIKAH, T. D. Comparative analysis using various performance metrics in imbalanced data for multi-class text classification. *International Journal of Advanced Computer Science and Applications*, 2023. Disponível em: <https://consensus.app/papers/analysis-using-performance-metrics-imbalanced-data-riyanto/02695651a869586eb546aa16e714737c/?utm_source=chatgpt>. Cited on page 35.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. Disponível em: <http://arxiv.org/abs/1505.04597>. Cited 4 times on pages 7, 24, 37, and 42.

SEN, R.; GOSWAMI, S.; CHAKRABORTY, B. Jeffries-matusita distance as a tool for feature selection. In: IEEE. *2019 International Conference on Data Science and Engineering (ICDSE)*. [S.l.], 2019. p. 15–20. Cited on page 19.

SHIMABUKURO, Y.; DUARTE, V.; SANTOS, J. D. dos; BATISTA, G. Mapping and monitoring deforestation areas in amazon region using semi-automatic classification of landsat thematic mapper images. *IGARSS 2000. IEEE 2000 International Geoscience and Remote Sensing Symposium. Taking the Pulse of the Planet: The Role of Remote Sensing in Managing the Environment. Proceedings (Cat. No.00CH37120)*, v. 5, p. 1999–2001 vol.5, 2000. Cited on page 5.

SHIRVANI, Z.; ABDI, O.; GOODMAN, R. C. High-resolution semantic segmentation of woodland fires using residual attention unet and time series of sentinel-2. *Remote Sensing*, v. 15, n. 5, 2023. Cited 2 times on pages 37 and 41.

SHOAIB, M.; LAI, K.; CHUAH, J. H.; HUM, Y.; ALI, R.; DHANALAKSHMI, S.; WANG, H.; WU, X. Comparative studies of deep learning segmentation models for left ventricle segmentation. *Frontiers in Public Health*, v. 10, 2022. Cited on page 35.

SINGH, P.; SHREE, R. Analysis and effects of speckle noise in sar images. *2016 2nd International Conference on Advances in Computing, Communication, and Automation (ICACCA) (Fall)*, p. 1–5, 2016. Cited on page 4.

TAVARES, P. A.; BELTRãO, N. E. S.; GUIMARãES, U. S.; TEODORO, A. C. Integration of sentinel-1 and sentinel-2 for classification and lulc mapping in the urban area of belém, eastern brazilian amazon. *Sensors*, v. 19, n. 5, p. 1140, 2019. Cited 4 times on pages vi, 16, 17, and 18.

TEAM, D. An introduction to convolutional neural networks. *Deepgram*, 2020. Cited on page 22.

THAI, D. H.; FEI, X.; LE, M. T.; ZUFLE, A.; WESSELS, K. Riesz-quincunx-unet variational autoencoder for unsupervised satellite image denoising. *IEEE Transactions on Geoscience and Remote Sensing*, v. 61, p. 1–19, 2022. Cited on page 7.

TOVAR, P.; ADARME, M.; FEITOSA, R. Deforestation detection in the amazon rainforest with spatial and channel attention mechanisms. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, Copernicus GmbH, v. 43, p. 851–858, 2021. Cited on page 41.

TOWNSEND, P. Relationships between forest structure and the detection of flood inundation in forested wetlands using c-band sar. *International Journal of Remote Sensing*, v. 23, p. 443 – 460, 2002. Cited on page 14.

TZELEPIS, G.; ASIF, A.; BACI, S.; ÇAVDAR, S.; AKSOY, E. Deep neural network compression for image classification and object detection. *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, p. 1621–1628, 2019. Cited on page 54.

UNSALAN, C.; BOYER, K. Remote sensing satellites and airborne sensors. p. 7–15, 2011. Cited on page 9.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L.; POLOSUKHIN, I. Attention is all you need. *CoRR*, abs/1706.03762, 2017. Disponível em: <http://arxiv.org/abs/1706.03762>. Cited on page 23.

VIEIRA, I. C. G.; JUNIOR, R. A. O. S.; TOLEDO, P. M. d. Dinâmicas produtivas, transformações no uso da terra e sustentabilidade na amazônia. *COCST*, Banco Nacional de Desenvolvimento Econômico e Social, 2014. Cited on page 1.

WANG, H.; CHEN, Y.; CAI, Y.; CHEN, L.; LI, Y.; SOTELO, M.; LI, Z. Sfnet-n: An improved sfnet algorithm for semantic segmentation of low-light autonomous driving road scenes. *IEEE Transactions on Intelligent Transportation Systems*, v. 23, p. 21405–21417, 2022. Cited on page 21.

WANG, W.; DAI, J.; CHEN, Z.; HUANG, Z.; LI, Z.; ZHU, X.; HU, X.; LU, T.; LU, L.; LI, H.; WANG, X.; QIAO, Y. *InternImage: Exploring Large-Scale Vision Foundation Models with Deformable Convolutions*. 2023. Cited on page 25.

WANG, W.; DAI, J.; CHEN, Z.; HUANG, Z.; LI, Z.; ZHU, X.; HU, X.; LU, T.; LU, L.; LI, H.; WANG, X.; QIAO, Y. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2023. p. 14408–14419. Cited on page 34.

WANG, Y.; QIU, Y.; CHENG, P.; ZHANG, J. Hybrid cnn-transformer features for visual place recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, v. 33, p. 1109–1122, 2023. Cited on page 21.

WEI, Y.; LIANG, X.; CHEN, Y.; JIE, Z.; XIAO, Y.; ZHAO, Y.; YAN, S. Learning to segment with image-level annotations. *Pattern Recognit.*, v. 59, p. 234–244, 2016. Cited on page 35.

XIE, E.; WANG, W.; YU, Z.; ANANDKUMAR, A.; ALVAREZ, J. M.; LUO, P. Segformer: Simple and efficient design for semantic segmentation with transformers. In: *Advances in Neural Information Processing Systems*. [S.l.: s.n.], 2021. Cited 7 times on pages iv, 7, 24, 32, 33, 52, and 53.

YAGÜE-MARTÍNEZ, N.; PRATS-IRAOLA, P.; GONZALEZ, F. R.; BRCIC, R.; SHAU, R.; GEUDTNER, D.; EINEDER, M.; BAMLER, R. Interferometric processing of sentinel-1 tops data. *IEEE transactions on geoscience and remote sensing*, IEEE, v. 54, n. 4, p. 2220–2234, 2016. Cited 2 times on pages 4 and 39.

YANG, K.; HU, X.; BERGASA, L.; ROMERA, E.; WANG, K. Pass: Panoramic annular semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, v. 21, p. 4171–4185, 2020. Cited on page 21.

YUN, S.; HAN, D.; OH, S. J.; CHUN, S.; CHOE, J.; YOO, Y. *CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features.* 2019. Cited on page 61.

ZHANG, H.; CISSÉ, M.; DAUPHIN, Y. N.; LOPEZ-PAZ, D. mixup: Beyond empirical risk minimization. *CoRR*, abs/1710.09412, 2017. Disponível em: <http://arxiv.org/abs/1710.09412>. Cited on page 61.

ZHANG, H.; LI, J.; WANG, T.; LIN, H.; ZHENG, Z.; LI, Y.; LU, Y. A manifold learning approach to urban land cover classification with optical and radar data. *Landscape and Urban Planning*, Elsevier, v. 172, p. 11–24, 2018. Cited on page 17.

ZHANG, W.; LIU, S. chao. Applications of the small satellite constellation for environment and disaster monitoring and forecasting. *International Journal of Disaster Risk Science*, v. 1, p. 9–16, 2010. Cited on page 10.

ZHAO, H.; SHI, J.; QI, X.; WANG, X.; JIA, J. Pyramid scene parsing network. In: IEEE. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* [S.l.], 2017. p. 2881–2890. Cited on page 29.

ZHENG, S.; JI, Z. *et al.* Semantic segmentation using vision transformers: A survey. *arXiv preprint arXiv:2101.03546*, 2021. Cited on page 21.

ZHENG, T.; WANG, J.; LEI, P. Deep learning based target detection method with multi-features in sar imagery. In: IEEE. *2019 6th Asia-Pacific Conf. on Synthetic Aperture Radar (APSAR).* [S.l.], 2019. p. 1–4. Cited on page 6.

ZHOU, Y.; WANG, P.; CHEN, Z.; ZHAO, Q.; WANG, W.; ZHANG, L.; YU, W.; DENG, Y.-K. Very-high-resolution sar imaging with dgps-supported airborne x-band data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, PP, p. 1–1, 06 2020. Cited 2 times on pages iii and 4.

ZHU, W.; ZHANG, Y.; QIU, L.; FAN, X. Research on target detection of sar images based on deep learning. In: SPIE. *Image and Signal Processing for Remote Sensing XXIV.* [S.l.], 2018. v. 10789, p. 581–588. Cited on page 6.

ZITOVá, B.; FLUSSER, J. Image registration methods: a survey. *Image and Vision Computing*, v. 21, n. 11, p. 977–1000, 2003. ISSN 0262-8856. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0262885603001379>. Cited on page 49.