Deep Learning Approaches to Segment Eucalyptus Tree Images

Mário de Araújo Carvalho

SERVIÇO DE PÓS-GRADUAÇÃO DA FACOM-UFMS

Data de Depósito:

Assinatura:_____

Deep Learning Approaches to Segment Eucalyptus Tree Images

Mário de Araújo Carvalho

Orientador: *Profº Drº Wesley Nunes Gonçalves*

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Faculdade de Computação, mantido pela Universidade Federal de Mato Grosso do Sul, para a Defesa de Mestrado, como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação.

UFMS - Campo Grande
Fevereiro/2022

To God.

To my parents,
Welison and Maria.

In memoriam *of my maternal grandparents,*
Nonato and Esmerinda.

To my paternal grandparents,
Doralice and Valderico.

To my beloved girlfriend,
Thalia.

To my friends.

Acknowledgments

A Deus pelas coisas maravilhoras que tem feito em minha vida e por me guiar em cada passo que eu dou.

À minha querida avó, Esmerinda. Infelizmente, não há nenhuma palavra ou quantidade de texto que possa expressar a profundidade da minha gratidão por tudo o que você significou para mim em minha vida. Você foi meu porto seguro, minha amiga, minha conselheira e meu exemplo de amor incondicional. Seu sorriso sempre me acalmou e suas palavras sempre me confortaram. Obrigado por tudo o que você fez por mim, por ter me amado tanto e por ter sido a avó mais incrível que alguém poderia ter. Sua memória sempre estará viva em meu coração, bobó.

Gostaria de expressar minha profunda gratidão e reconhecimento aos meus queridos pais, Welison e Maria, e aos meus irmãos Railson, Mateus, Samuel, Natielly e Wellington. Vocês são a minha base, a minha fortaleza, a minha família. Agradeço por terem sempre estado ao meu lado, apoiando e incentivando cada passo que eu dei ao longo da minha jornada. Meus pais, suas palavras de sabedoria e seu exemplo de determinação e amor incondicional foram um farol que me guiou em momentos de incerteza e dificuldade. Vocês são os meus heróis e a razão do meu sucesso. Meus irmãos, vocês são meus amigos, confidentes e companheiros de aventuras. Agradeço por cada risada, cada conversa, cada conselho e cada momento compartilhado. Embora às vezes eu tenha sido um filho e irmão ausente, saibam que eu os amo mais do que as palavras podem expressar.

Minha mais profunda gratidão ao professor Wesley Nunes Gonçalves, meu orientador, por me conduzir durante toda a elaboração desta dissertação de mestrado. Ele sempre esteve disposto a me motivar e me mostrar o caminho a seguir, e eu lhe agradeço por sua paciência, por nunca ter deixado que eu desistisse, e por sua dedicação na orientação.

Aos meus professores e amigos Amaury Antônio de Castro Junior e José Marcato Junior, por todo apoio e dedicação em meu crescimento acadêmico

e pessoal, quero agradecer do fundo do meu coração. Vocês são uma fonte constante de inspiração e sabedoria, e seu conhecimento e apoio incansável foram fundamentais para minha formação. Eu aprendi muito com vocês e serei eternamente grato pelo impacto positivo que tiveram em minha vida. Obrigado por serem grandes professores e amigos.

À minha amada namorada, Thalia. Gostaria de dizer que é impossível conter as emoções ao expressar minha gratidão por todo o amor incondicional, dedicação incansável, paciência inabalável e compreensão incomparável que você demonstrou durante os momentos difíceis em que me dediquei aos estudos e estive distante fisicamente. Sua presença constante e seu apoio incondicional foram a bússola que me guiou durante esta jornada árdua e eu nunca poderia ter chegado tão longe sem você. Você é meu porto seguro, minha rocha, minha âncora e minha musa inspiradora, e sou eternamente grato por ter você ao meu lado. Muito obrigado, minha amada Thalia, por ser a luz que ilumina o caminho da minha vida.

Com muita emoção, gostaria de expressar a minha imensa gratidão à família Medina Braga. Foi um privilégio contar com a amizade, apoio e incentivo de pessoas tão incríveis durante toda a minha jornada acadêmica até aqui. Não poderia deixar de mencionar o querido amigo Ezoir, cuja sabedoria e perseverança nos guiam mesmo após sua partida desse plano material. Seu legado será sempre lembrado e honrado por todos aqueles que tiveram a sorte de conhecê-lo. Lia, sua bondade e paciência foram fundamentais para mim. Sua presença sempre foi muito acolhedora e gentil desde o início. Luana, sua força e determinação são uma inspiração constante para mim. Seu exemplo me motiva a buscar sempre o melhor em mim mesmo e a enfrentar os desafios com coragem e determinação. Gabriel, sua amizade e apoio incondicional me deram a confiança necessária para seguir em frente mesmo nos momentos mais difíceis. Sua presença foi uma fonte de conforto e incentivo durante todo o percurso.

Agradeço aos demais colegas, professores e servidores da Faculdade de Computação da UFMS. Aos amigos Calebe Lemos, Dênis Cardoso, Franklin Barbosa, Gabriel Escobar, Hernanes Almeida, Jhonatan Froeder e Ronaldo da Silva. Ao Laboratório de Geomática da FAENG-UFMS, especialmente por me proporcionar amigos como: Geazy, José Augusto, Márcio, Maurício e Pedro. Agradeço aos membros da banca de mestrado Camilo Carromeu e Celso Soares Costa pelas contribuições valiosas durante a defesa da minha dissertação.

Gostaria de expressar meu profundo agradecimento à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro proporcionado pela minha bolsa de mestrado durante parte do meu curso.

Abstract

Agribusiness is one of Brazil's primary sources of wealth and employment, representing a significant portion of the national Gross Domestic Product (GDP). In 2021, the agribusiness sector reached 27.4% of the Brazilian GDP, the highest share since 2004, when it reached 27.53%. The forest-based industry is an important segment of agribusiness, as it provides vital inputs for various industrial sectors, such as wood products, furniture, and paper. Planted forests play an essential role in carbon capture and other ecosystem services, with eucalyptus being the most used tree, with 7.3 million hectares of eucalyptus forests in 2021. Tree mapping is vital for the economy and environment, and artificial intelligence-based solutions are valuable decision support tools in agriculture and tree mapping. Consequently, there is a strong incentive to look for more comprehensive solutions that use advanced deep learning technologies for this area. Thus, this work aims to evaluate efficient deep learning convolutional neural networks for image segmentation of eucalyptus trunks and present a specific segmentation proposal for eucalyptus trunks that can benefit agricultural applications or decision support tools for tree mapping. This work was divided into two main steps to evaluate the segmentation networks and create a post-processing technique. The first stage of this study evaluated the efficiency of deep learning networks in the semantic segmentation of eucalyptus trunks in panoramic images in RGB colors captured at ground level. The deep learning networks FCN, GCNet, ANN, and PointRend were evaluated in this step for image segmentation of eucalyptus trunks. Training and evaluation of the networks were performed using a five-step cross-validation approach, using a dataset composed of manually annotated images of a eucalyptus forest. The initial dataset was created using a spherical field of view camera. It included a variety of eucalyptus trees with distinct characteristics, such as variations in distances between trunks and changes in curvature, sizes, and diameters of trunks, which pose significant challenges for deep learning methods in semantic segmentation tasks. For

the first stage of this study, the FCN model presented the best performance, with pixel precision of 78.87% and mIoU of 70.06%, in addition to obtaining a good inference time. The GCNet and ANN networks also performed similarly to the FCN but with negative impacts on their ability to generalize tasks in specific contexts. The study concludes that the FCN was the most robust, among the evaluated methods, for semantic segmentation of images of trees in panoramic images. This assessment of segmentation networks can be a crucial step toward developing other relevant tools in forest management, such as estimating trunk height and diameter. The second step of this work was to create and evaluate a post-processing technique for RGB-D images to improve the results of current semantic networks for segmentation in eucalyptus images. We created a new dataset image using images obtained from a stereo camera, which captured not only the color information (RGB) but also the depth information, which allowed an even more complete view of the eucalyptus forest. After the construction of the new image bank, its annotation was carried out by specialists. The next stage of this study was the evaluation of six image semantic segmentation networks and the comparison with results before and after applying the post-processing technique. We trained, evaluated, and tested the FCN, ANN, GCNet, SETR, SegFormer, and DPT networks on the annotated images. The post-processing technique significantly improved the results of the tested image segmentation networks, with a significant gain of 24.13% in IoU and 13.11% in F1-score for convolution-based networks and 12.49% for IoU and 6.56% in F1-score for transformer-based networks. The SegFormer network obtained the best results in all tests before and after applying the technique. The technique also effectively corrected segmentation flaws, erosion, and dilation errors, resulting in more accurate edges and better-delimited trunks. The average computational cost of the technique was 0.019 seconds, indicating that it can be applied in segmentation networks without compromising performance. The results obtained by applying the post-processing technique propose an innovative approach with low computational cost and significant improvements to existing segmentation networks.

Keywords: Deep Learning, Segmentation, Eucalyptus tree

Resumo

O agronegócio é uma das principais fontes de riqueza e emprego do Brasil, representando uma parcela significativa do Produto Interno Bruto (PIB) nacional. Em 2021, o setor do agronegócio atingiu 27,4% do PIB brasileiro, a maior participação desde 2004, quando atingiu 27,53%. A indústria de base florestal é um importante segmento do agronegócio, pois fornece insumos vitais para diversos setores industriais, como produtos de madeira, móveis e papel. As florestas plantadas desempenham um papel essencial na captura de carbono e outros serviços ecossistêmicos, sendo o eucalipto a árvore mais utilizada, com 7,3 milhões de hectares de florestas de eucalipto em 2021. O mapeamento de árvores é vital para a economia e o meio ambiente, e as soluções baseadas em inteligência artificial são valiosas ferramentas de apoio à decisão em agricultura e mapeamento de árvores. Consequentemente, há um forte incentivo para buscar soluções mais abrangentes que utilizem tecnologias avançadas de aprendizado profundo para essa área. Assim, este trabalho tem como objetivo avaliar redes neurais convolucionais de aprendizado profundo eficientes para segmentação de imagens de troncos de eucalipto e apresentar uma proposta de segmentação específica para troncos de eucalipto que pode beneficiar aplicações agrícolas ou ferramentas de apoio à decisão para mapeamento de árvores. Este trabalho foi dividido em duas etapas principais para avaliar as redes de segmentação e criar uma técnica de pós-processamento. A primeira etapa deste estudo avaliou a eficiência de redes de aprendizado profundo na segmentação semântica de troncos de eucalipto em imagens panorâmicas em cores RGB capturadas no nível do solo. As redes de aprendizado profundo FCN, GCNet, ANN e PointRend foram avaliadas nesta etapa para segmentação de imagens de troncos de eucalipto. O treinamento e a avaliação das redes foram realizados usando uma abordagem de validação cruzada de cinco etapas, usando um conjunto de dados composto por imagens anotadas manualmente de uma floresta de eucalipto. O conjunto de dados inicial foi criado usando um campo de visão esférico da câmera. Ele

incluiu uma variedade de eucaliptos com características distintas, como variações nas distâncias entre os troncos e mudanças na curvatura, tamanhos e diâmetros dos troncos, que representam desafios significativos para métodos de aprendizado profundo em tarefas de segmentação semântica. Para a primeira etapa deste estudo, o modelo FCN apresentou o melhor desempenho, com precisão de pixel de 78,87% e mIoU de 70,06%, além de obter um bom tempo de inferência. As redes GCNet e ANN também tiveram desempenho semelhante ao FCN, mas com impactos negativos em sua capacidade de generalizar tarefas em contextos específicos. O estudo conclui que o FCN foi o mais robusto, dentre os métodos avaliados, para segmentação semântica de imagens de árvores em imagens panorâmicas. Essa avaliação das redes de segmentação pode ser um passo crucial para o desenvolvimento de outras ferramentas relevantes no manejo florestal, como a estimativa de altura e diâmetro do tronco. A segunda etapa deste trabalho foi criar e avaliar uma técnica de pós-processamento de imagens RGB-D para melhorar os resultados das redes semânticas atuais para segmentação em imagens de eucalipto. Criamos uma nova imagem de conjunto de dados a partir de imagens obtidas de uma câmera estéreo, que capturou não apenas as informações de cor (RGB), mas também as informações de profundidade, o que permitiu uma visão ainda mais completa da floresta de eucalipto. Após a construção do novo banco de imagens, sua anotação foi realizada por especialistas. A próxima etapa deste estudo foi a avaliação de seis redes de segmentação semântica de imagens e a comparação com os resultados antes e depois da aplicação da técnica de pós-processamento. Treinamos, avaliamos e testamos as redes FCN, ANN, GCNet, SETR, SegFormer e DPT nas imagens anotadas. A técnica de pós-processamento melhorou significativamente os resultados das redes de segmentação de imagens testadas, com um ganho significativo de 24,13% em IoU e 13,11% em F1-score para redes baseadas em convolução e 12,49% para IoU e 6,56% em F1-score para redes baseadas em transformadores. A rede SegFormer obteve os melhores resultados em todos os testes antes e após a aplicação da técnica. A técnica também corrigiu com eficácia falhas de segmentação, erosão e erros de dilatação, resultando em bordas mais precisas e troncos mais bem delimitados. O custo computacional médio da técnica foi de 0,019 segundos, indicando que ela pode ser aplicada em redes de segmentação sem comprometer o desempenho. Os resultados obtidos pela aplicação da técnica de pós-processamento propõem uma abordagem inovadora com baixo custo computacional e melhorias significativas para as redes de segmentação existentes.

Palavras-chave: Aprendizado Profundo, Segmentação, Eucalipto

Contents

List of Figures

List of Tables

List of Abbreviations

GDP Gross Domestic Product

AI Artificial Intelligence

ML Machine Learning

DL Deep Learning

CNN Convolutional Neural Networks

CV Computer Vision

SIS Semantic image segmentation

FCN Fully Convolutional Network

GCNet Context Guided Network

PointRend Point-based Rendering

ANN Asymmetric Non-local Neural Network

DPT Vision Transformer for Dense Prediction

SETR Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers

SegFormer Simple and Efficient Design for Semantic Segmentation with Transformers

RGB Red, Green and Blue

RGB-D Red, Green, Blue and Depth

SNL Non-Local Networks

SE Squeeze Excitation

APNB Asymmetric Pyramid Non-local Block

AFNB Asymmetric Fusion Non-local Block

IoU Intersection over Union

GT Ground Truth

DBH Diameter at Breast Height

TP True Positive

FN False Negative

FP False Positive

TN True Negative

CHAPTER 1

Introduction

1.1 Contextualization

Agribusiness is one of the most critical sectors of Brazil's economy, contributing significantly to the national Gross Domestic Product (GDP). In 2021, agribusiness reached its highest peak since 2004, reaching 27.4% of Brazilian GDP, surpassing the previous result of 27.53%[CEPEA, 2021]. In 2018, the forestry sector contributed 1.3% of the Brazilian GDP and had an estimated export value of US$ 11.3 billion in 2019, making it a significant economic activity [Daniel Feffer, 2019]. In addition, planted forests play an essential role in carbon capture and other ecosystem services, with *Eucalyptus* being the most commonly used tree in this sector [Daniel Feffer, 2019]. In 2021, Brazil had 9.5 million hectares of planted forests, 7.3 million hectares of *Eucalyptus*, and 1.8 million hectares of pine [IBGE, 2021]. Maintaining the quality and productivity of this sector requires constant monitoring of planted areas, which is an operational challenge. In addition, there is a need for accurate measurements and quantification of production for better resource management and maximization. In the past, models that related climate variables and production were standard in the literature [Santana et al., 2008; Porter and Semenov, 2005; White et al., 2011]. However, they needed more robust methods to support this economic sector.

Researchers have increasingly explored machine learning (ML) techniques to extract valuable information from the forest. According to Rodrigues de Oliveira et al. [2021], vegetation indices and spectral bands are two valuable data sources that can extract essential forest characteristics. This data can

be analyzed using machine learning algorithms, including supervised and unsupervised techniques. Supervised machine learning can be instrumental in forest segmentation, as these algorithms require specific information to learn and predict correctly. Thus, combining machine learning techniques and vegetation and spectral data can provide valuable information about the forest, including the estimation of volume and biomass per unit area, which is fundamental to assessing and managing forest resources efficiently.

Precise estimation of volume and biomass per unit area in *Eucalyptus* trees is crucial in assessing forest resources and is essential for their proper management. Various mathematical models and image processing techniques based on remote sensing data, such as satellite images or radar data, are widely used to achieve this accuracy. However, data quality, spatial resolution, cloud cover, and the model's capacity can significantly affect the accuracy of [Mendes et al., 2020] estimates. To estimate the biomass and nutrients of the *Eucalyptus* species, mathematical models use information such as diameter at breast height, height, and age of the tree, resulting in a more accurate and efficient estimate of these values. Some studies have explored the application of mathematical models for this purpose, as described in [Valadão et al., 2020].

New research has shown the combination of remote sensing and machine learning as a solution for the agricultural sector [Yu et al., 2021; Ferreira et al., 2020; Zhao et al., 2019]. Machine learning algorithms efficiently identify complex patterns from data, providing valuable information for evaluations and predictions [LeCun et al., 2015]. Remote sensing, in turn, allows the collection of large amounts of data at different scales, including orbit, air, and land. Several studies have integrated remote sensing data and machine learning algorithms in several areas, including agriculture [Liakos et al., 2018], urban planning [Fathi et al., 2020; Chaturvedi and de Vries, 2021], soil and biomass [Ali et al., 2015; Padarian et al., 2020; Torre-Tojal et al., 2022], and forest [Singh et al., 2016; Maxwell et al., 2018].

Deep learning (DL) techniques have highlighted precision agriculture [Osco et al., 2019, 2020, 2021]. In this context, deep learning models based on convolution neural networks (CNN) for the semantic image segmentation (SIS) of trees have gained prominence [Nogueira et al., 2019; Martins et al., 2019, 2021a]. Advances in deep learning have encouraged the emergence of several methods, such as the Fully Convolutional Network (FCN) [Long et al., 2015] and current methods that consider long-distance dependencies such as Context Guided Network (GCNet) [Cao et al., 2020] and Asymmetric Non-local Neural Network (ANN) [Zhu et al., 2019]. Furthermore, Point-based Rendering (PointRend) [Kirillov et al., 2020] improves the segmentation quality around

object edges by treating this problem as a rendering issue and adapting classic computer graphics ideas.

Several studies have mapped *Eucalyptus* trees using aerial images [Dias et al., 2020; Firigato et al., 2021; Ferreira et al., 2012; Khan et al., 2021], but few have used ground-level panoramic images. Although aerial images can cover larger areas, they have the disadvantage of not showing a side view of the tree. Various purposes, such as diameter estimation, height estimation, biomass estimation, and disease detection, can be accomplished using this side view of the trunk. Ground-level panoramic images can be used to provide a side view of the trunk and ground. Some studies have explored these types of images for the segmentation of plantations and trees [Vepakomma et al., 2011; Darwin et al., 2021], tree detection and segmentation [Li et al., 2017], disease identification [Zhang et al., 2019; Syarief and Setiawan, 2020] and agricultural production evaluation [Yalcin, 2019]. However, these images still represent a gap in the context of mapping *Eucalyptus* trees. Ground-level panoramic images can represent a powerful tool for monitoring *Eucalyptus* forests. These images offer a large field of view compared to regular images, allowing them to cover a large area. However, panoramic images exhibit strong distortion, making the segmentation task more challenging. Hence, one way is to evaluate the ability of new methods based on deep learning to segment targets in panoramic ground-level images. Methods that automatically segment *Eucalyptus* trees into RGB images can represent a leading and low-cost tool in forest inventory and management. In addition, remote sensing allows for collecting a large amount of data at different scales, such as orbital, aerial, and terrestrial. RGB images only have color information, while RGB-D images include depth information. The additional depth information of images could be helpful in many contexts, including planted *Eucalyptus* forests. Using RGB-D images can result in significantly improved segmentation of trees and greater accuracy in detecting issues such as disease or insect infestations.

RGB-D imaging technology allows for obtaining three-dimensional information, including depth, through image capture. These images are valuable in different contexts, including the study of *Eucalyptus*-planted forests, where they can be used to measure the physical characteristics of the trees, such as height and diameter, to identify problems, such as diseases and infestations. The forest manager can use RGB-D images to assess the wood quality and identify potential trees for high-quality wood production. They can also help monitor forest conditions and identify areas vulnerable to external events such as climate change or deforestation. Research around RGB-D images is an area in constant evolution, with the creation of ever more precise and efficient algorithms for the analysis of three-dimensional images [Xing et al.,

2020; Jianbo Jiao, 2019; Seichter et al., 2021].

The project aims to develop a new approach to the semantic segmentation of images in the field of artificial intelligence. The approach proposes an image post-processing process that aims to improve the results of existing networks, using information about the depth of RGB-D images. As a test example, we will use images of tree trunks with color (RGB) and depth (D) information. The ZED stereo cameras [Tadic et al., 2022] can collect RGB-D images, as they capture image data in 4 channels of information, including the depth. Research in AI considers semantic segmentation of images crucial, as it demands high precision in the results to ensure the dependability of technical systems based on semantic segmentation.

This work intends to evaluate and discuss the scientific challenge of improving post-processing techniques for systems that work with RGB-D images in computer vision (CV). Currently, image segmentation algorithms using only RGB images may have some limitations and reduce the accuracy of the inference results. Therefore, implementing a new post-processing approach can increase the accuracy of these methods and, consequently, allow the advancement of this area. Developing this new technique can also help produce wood, paper, and cellulose, as it will improve image segmentation techniques. We will only use images of *Eucalyptus* trunks in this work to evaluate the segmentation models before and after applying the technique. However, there is the possibility of exploring the technique's performance in other species of trees. Current SIS methods present unsatisfactory results to meet the needs of the proposed application, as many of them do not consider the image depth [Long et al., 2015; Cao et al., 2020; Zhu et al., 2019; Kirillov et al., 2020; Ranftl et al., 2021; Zheng et al., 2021; Xie et al., 2021]. The development of this work will advance scientific knowledge in artificial intelligence (AI) and can be used to build new technologies for the agribusiness sector. In addition, it presents new challenges for the semantic segmentation of images, enabling the improvement and development of new techniques.

1.2 Hypothesis and Objectives

1.2.1 Hypothesis

In this work, we have the following hypotheses:

1. The current image segmentation algorithms based only on RGB images suffer from flaws in the segmentation of objects, overlapping objects, and holes in the segmented image, which impairs the final result and decreases the accuracy of the models.

2. The implementation of the proposed approach can reduce investment losses in sectors that use AI and image segmentation, increasing the accuracy of current methods and allowing the advancement of new studies in the area.

3. The availability of accurate depth sensors in modern cameras is an opportunity to advance classical segmentation approaches.

4. The development of post-processing techniques for semantic segmentation of images with depth will result in segmentation models with better performance.

5. The development of this technique can help develop intelligent systems for producing wood, pulp, and paper, expanding the potential productive capacity of the sector.

6. The use of the technique in *Eucalyptus* trunks at ground level will allow testing the proposed method, but we believe that in future works, it can be explored in other tree species.

7. The development project for this technique will advance scientific knowledge in AI and may also be used to build new technologies for the agribusiness sector.

1.2.2 Objectives

The main objectives of this work are to evaluate segmentation methods using deep learning applied to *Eucalyptus* RGB images and to create a post-processing technique based on RGB-D images to improve the results of current segmentation networks. The following specific objectives were defined to achieve the general objective proposed by this dissertation:

1. **Creation and annotation of a data set of *Eucalyptus* images**: It will be necessary to go to the field and collect a set of *Eucalyptus* images and annotate each image with the *Eucalyptus* trees;

2. **Definition of image segmentation algorithms**: This set of deep learning algorithms for image segmentation must be tested and validated empirically in the *Eucalyptus* image data set;

3. **Training, validation, and testing of algorithms**: Train each algorithm using modern pre-processing techniques, analysis of Loss curves and accuracy, cross-validation and variation of optimizers and parameters;

4. **Development of a post-processing technique to improve the results**: From the trained segmentation models, developing a technique to improve the segmentation of the *Eucalyptus* images, to reduce errors such as holes in divided trunks.

1.3 Dissertation Text Organization

This section presents the organization of this dissertation proposal. We organized this dissertation into four main chapters. Chapters 2 and 3 are scientific articles developed during the implementation of the objectives of this work. Chapter 1 contains all the sub-sections that involve the initial parts of the work, such as introduction 1, contextualization 1.1, hypothesis, and objectives 1.2 of this work. Chapter 2 presents the parts of the texts related to the evaluation process of semantic segmentation networks in RGB ground-level paranoid images. In this chapter [2], it is possible to find all the methodology of this evaluation process of the segmentation network, such as the study area, data acquisition, image annotation process, segmentation methods, experimental protocol, and results obtained. Chapter 3 presents the parts of the texts related to the creation, development, and testing of the post-processing technique, as well as a comparison of the results obtained after the application of the technique to the results of the segmentation networks. In this chapter [3], we provide all the theoretical basis, methodology, description of the image datasets used, description of the algorithm of the developed post-processing technique, the final results of the experiments of this work, including performance analysis of the methods compared to the technique, the computational cost of the methods and technique, qualitative analysis and discussion of the results. Finally, Chapter 4 presents the conclusions of this work, including a parallel between the objectives of this dissertation and the results obtained, some limitations of the proposed solutions, and future works.

CHAPTER

2

Semantic Segmentation of *Eucalyptus* Tree in Panoramic RGB Ground-level Images

A mapping *Eucalyptus* trees is a demand in forest inventory and management sectors. Integrating remote sensing images and deep learning algorithms may be a robust approach for it, but still little explored. This task is particularly challenging when considering the *Eucalyptus* trees mapping in panoramic imagery. Although they offer a great field of view, they have a high geometric distortion. This work evaluates the performance of novel deep learning methods for semantic segmentation of *Eucalyptus* trees in panoramic RGB ground-level images. Four (FCN, GCNet, ANN, and PointRend) deep learning methods were evaluated using a five-fold cross-validation approach. A spherical field view camera recorded videos of a *Eucalyptus* tree forest. Videos were captured from a *Eucalyptus* trees forest using a spherical field of view camera. The video frames were converted to a panoramic image format, and we manually annotated 100 images, separating the *Eucalyptus* trees from the background. Additionally, we used data augmentation and generated images, aiming to improve the training of networks. Our dataset comprises *Eucalyptus* trees with different characteristics, such as variation in distances between trunks, curvature alterations, sizes, and trunks of different diameters, bringing more challenges for deep learning methods in semantic segmentation tasks. The FCN model presented the highest performance (pixel accuracy of 78.87%, and mIoU of 70.06%) while a good inference time to deal with several *Eucalyptus* tree characteristics. GCNet and ANN networks also presented

similar performance to the FCN, but for specific contexts, negatively impacting their ability to generalization in the task. We conclude that FCN is the most robust of all the evaluated methods for the segmentation of trees in panoramic imagery, being an initial step for developing other relevant tools in forest management, like height and trunk diameter estimations.

2.1 Introduction and Motivation

The forest sector is an important economic activity, representing 1.3% of Brazil's Gross Domestic Product in 2018 and an estimated exportation value of US$ 11.3 billion in 2019 [Daniel Feffer, 2019]. Planted forest plays a relevant role in carbon sequestration, among other ecosystem services, and *Eucalyptus* is the most commonly used tree in this sector [Daniel Feffer, 2019]. In 2021, the area of planted forests in Brazil reached 9.5 million hectares, especially *Eucalyptus*, with 7.3 million hectares, and pine, with 1.8 million hectares [IBGE, 2021]. Quality and productivity maintenance of the forest sector depends on continuous monitoring of the planted areas, resulting in an operational challenge. It comes imbued with the necessity for fine-scale measures and quantification of the production because it impacts managing the resources to maximize the production of planted areas. In the past, researchers often found models working the relationship between climate variables and production in the literature [Santana et al., 2008; Porter and Semenov, 2005; White et al., 2011]. However, they required more robust methods to support this sector of the economy.

Recent research has seen approaches proposed to attend the agricultural sector that integrates remote sensing and machine learning (ML) [Yu et al., 2021; Ferreira et al., 2020; Zhao et al., 2019]. The machine learning algorithms can extract complex patterns of a dataset, providing a valuable source of models for measurements and predictions [LeCun et al., 2015]. While remote sensing allows the acquisition of a large volume of data in different scales like orbital, aerial, and terrestrial levels. Several study cases in agriculture [Liakos et al., 2018], urban planning [Fathi et al., 2020; Chaturvedi and de Vries, 2021], soil and biomass [Ali et al., 2015; Padarian et al., 2020; Torre-Tojal et al., 2022], forest [Singh et al., 2016; Maxwell et al., 2018] have integrated remote sensing data and machine learning algorithms.

More recently, a sub-field of the ML named deep learning has gained attention in the precision agriculture sector [Osco et al., 2019, 2020, 2021]. Researchers have highlighted deep learning models mainly based on convolutional neural networks for semantic segmentation of trees in this context [Nogueira et al., 2019; Martins et al., 2019, 2021a]. Advances in deep learn-

ing have encouraged the emergence of several methods [Long et al., 2015; Wang et al., 2018; Cao et al., 2020; Bowley et al., 2019]. Fully Convolutional Network (FCN) [Long et al., 2015] is one of the first works with relevant results using deep learning for segmentation. Recent studies [Wang et al., 2018; Cao et al., 2020] have indicated that the accuracy can be improved if long-range dependencies are considered. For example, features of distant pixels are considered when extracting features for a given pixel. Methods based purely on convolution are limited in capturing these long-range dependencies due to the local receptive field. Thus, methods that consider features of all positions in an image have been proposed, such as GCNet [Cao et al., 2020] and ANN [Zhu et al., 2019]. Additionally, to improve the quality of segmentation around the edges of objects, PointRend [Kirillov et al., 2020] treat this task as a rendering problem and adapts classic computer graphics ideas.

Several studies [Dias et al., 2020; Firigato et al., 2021; Ferreira et al., 2012; Khan et al., 2021] mapped *Eucalyptus* trees, but most of them used aerial images instead of panoramic ground-level images. Although aerial images can generally cover larger areas, they have the disadvantage of presenting a lateral view of the tree. The trunk's lateral view can achieve multiple purposes, such as diameter, height and biomass estimation, and disease detection. Once they provide a lateral view of the trunk and the ground, one may use panoramic ground-level images. Some studies have explored these images for plantations and tree segmentation [Vepakomma et al., 2011; Darwin et al., 2021]. However, in the context of *Eucalyptus* tree mapping, the use of these images still represents a gap. Panoramic ground-level images can represent a powerful tool for *Eucalyptus* tree forest monitoring. These images offer a great field of view compared to regular images, allowing them to cover a large area with these images. Nonetheless, panoramic images present a strong distortion, making the segmentation task more challenging. Therefore, the capability of novel deep learning based-methods should be evaluated to segment targets in panoramic ground-level images. Methods that automatically segment *Eucalyptus* trees in RGB images can represent a leading and low-cost tool in forest inventory and management.

This work assesses the performance of novel deep-learning methods for *Eucalyptus* tree segmentation in panoramic RGB ground-level images. We applied four state-of-the-art models: the FCN, GCNet, ANN, and PointRend, using a challenging dataset composed of *Eucalyptus* trees with variation in distances between trunks, curvature, sizes, and trunks of different diameters. The four semantic segmentation methods were trained and evaluated in five-fold cross-validation. A quantitative-qualitative analysis is presented, along with a discussion about the advantages and limitations of each CNN applied.

2.2 Methodology

2.2.1 Study Location Area

A case study was conducted in the municipality of Jaraguari, Mato Grosso do Sul State, Brazil (Figure 2.1). The selected area is composed of a *Eucalyptus* tree forest of 1,05 ha ($70m \times 150m$) planted in 2000 (Figure 2.1). According to the Köppen-Geiger climate classification [Beck et al., 2018], this region is classified as a savanna climate (Aw/As), presenting a dry season in winter (Aw) or summer (As), with monthly average temperatures above 18°C during every month of the year.

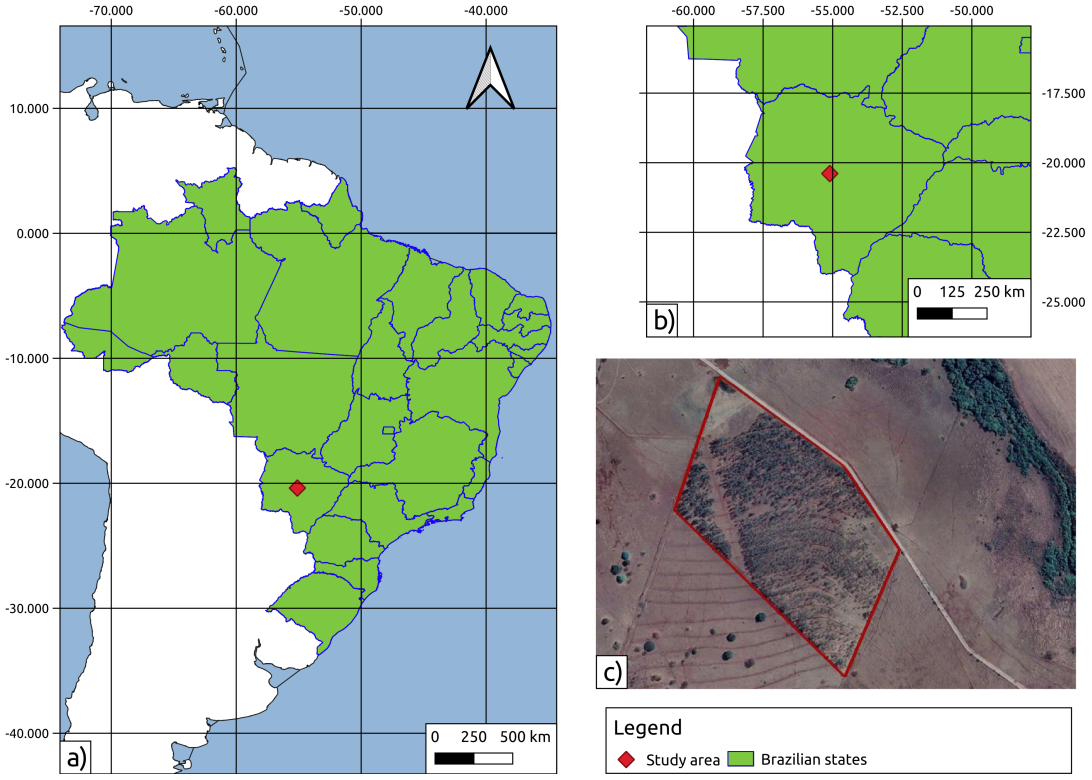


Figure 2.1: Study area in (a) South America and Brazil, (b) Mato Grosso do Sul, and (c) Orthoimage of the area.

2.2.2 Data Acquisition and Image Annotation

The dataset acquisition and data pre-processing consisted of 3 main steps. Firstly, we produced a video recording of *Eucalyptus* trees forest on November 24, 2020, at an approximate height of 2 meters above ground level. We used a spherical field of view camera with a resolution of 3k (3000x1504 pixels) and 60 frames per second (fps). Secondly, the video was cut into frames at five frames per second, totaling 2012 frames, and thirdly the frames were transformed to a panoramic format to correct the fisheye effect. Figure 2.2 shows the image collection process.

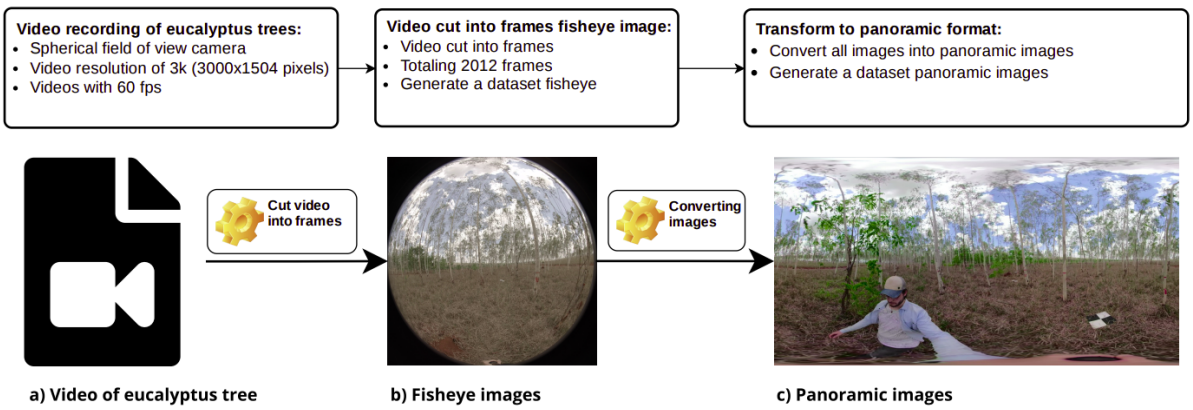


Figure 2.2: Overview of the workflow.

The captured images were manually annotated by specialists using the LabelMe open annotation tool software[1] [Wada, 2018]. In this process, the images of the trees were annotated with polygons around the trunk and labeled as *Eucalyptus*. We decided to annotate only the trees closest to the camera, as the trees in the background are small. After the image annotation process, the images were exported to the segmentation CNNs input format, a binary mask. Figure 2.3 shows an annotated image sample. As a result of this process created a dataset with 100 annotated images, with an average of 6 polygons per image, totaling around 600 annotations. The description of how this dataset of images was partitioned for training, validation, and testing is described in Section 2.2.4, including details of cross-validation and data augmentation.

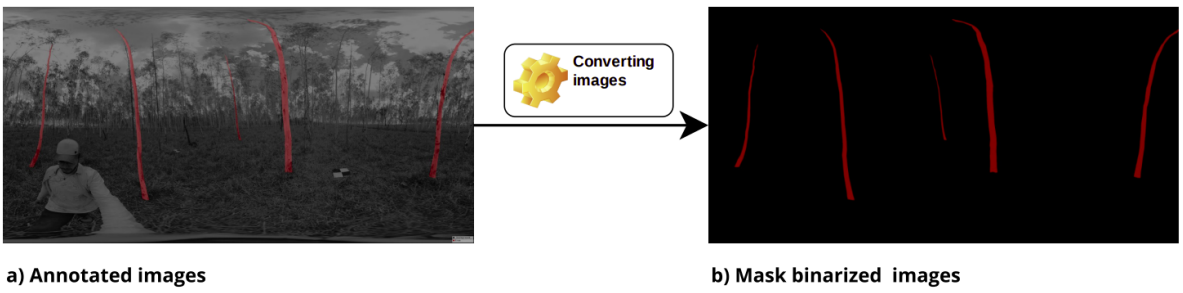


Figure 2.3: Overview of the annotation process.

2.2.3 Semantic Segmentation Methods

This section briefly introduces segmentation methods used in our work which are FCN [Long et al., 2015], Context Guided Network (GCNet)[Cao et al., 2020], Asymmetric Non-local Neural Network (ANN) [Zhu et al., 2019], and Point-based Rendering (PointRend) [Kirillov et al., 2020].

[1] http://labelme.csail.mit.edu/

2.2.3.1 Fully Convolutional Network (FCN)

FCN [Long et al., 2015] is a semantic segmentation approach that proposes a backbone to extract a low-resolution feature map from the input image. A convolutional layer assigns a score to the classes at each pixel. As the resolution is lower than the input image, an upsampling layer is employed to increase the resolution of the output view. To refine the prediction pixel by pixel, FCN combines the prediction with shallower layers by adding the predictions and applying a softmax function.

2.2.3.2 Global Context Network (GCNet)

Convolution layers in convolutional neural networks learn a relationship between pixels in a local neighborhood but do not effectively consider long-range dependency. To overcome this issue, non-local networks (SNL) were proposed to include global context via a self-attention mechanism. However, these blocks are expensive in terms of time and space complexity. Thus GCNet [Cao et al., 2020] proposed a simpler version of non-local networks and Squeeze Excitation (SE) to include global context. GCNet uses only one attention map for all pixels to model long-range dependency, while the SE block has a light computational cost. Multiple layers build the GCNet architecture using GC blocks.

2.2.3.3 Asymmetric Non-local Neural Network (ANN)

ANN [Zhu et al., 2019] is a non-local network incorporating pyramid sampling and non-local blocks to extract semantic features at different scales. This method proposes to include global context in a module called Asymmetric Pyramid Non-local Block (APNB) to reduce the computational cost of standard non-local blocks. For this, this block incorporates a pyramid sampling module in non-local blocks. Furthermore, ANN proposes an adaptation of the APNB called Asymmetric Fusion Non-local Block (AFNB). AFNB combines the features of different stages of the network under sufficient consideration of long-range dependencies. Finally, ANN is an FCN that incorporates the APNB and AFNB modules.

2.2.3.4 Point-based Rendering (PointRend)

PointRend [Kirillov et al., 2020] was proposed to improve the quality of segmentation around the edges of objects, as it renders problems and adapts classic computer graphics ideas. For this, are proposes a new module composed of three stages in this network. The first step is to select a set of pixels for label prediction, avoiding excessive computation for all pixels. Based on

the Adaptive Subdivision technique of computer graphics, pixels are selected in regions with a high probability of the label being different from its neighbors. This selection is iteratively performed from a coarse to fine output, thus being able to refine the predicted regions. The second step extracts a feature vector for each pixel selected in the previous step. Finally, the third step consists of training and predicting the label for each pixel.

2.2.4 Experimental Setup Environment

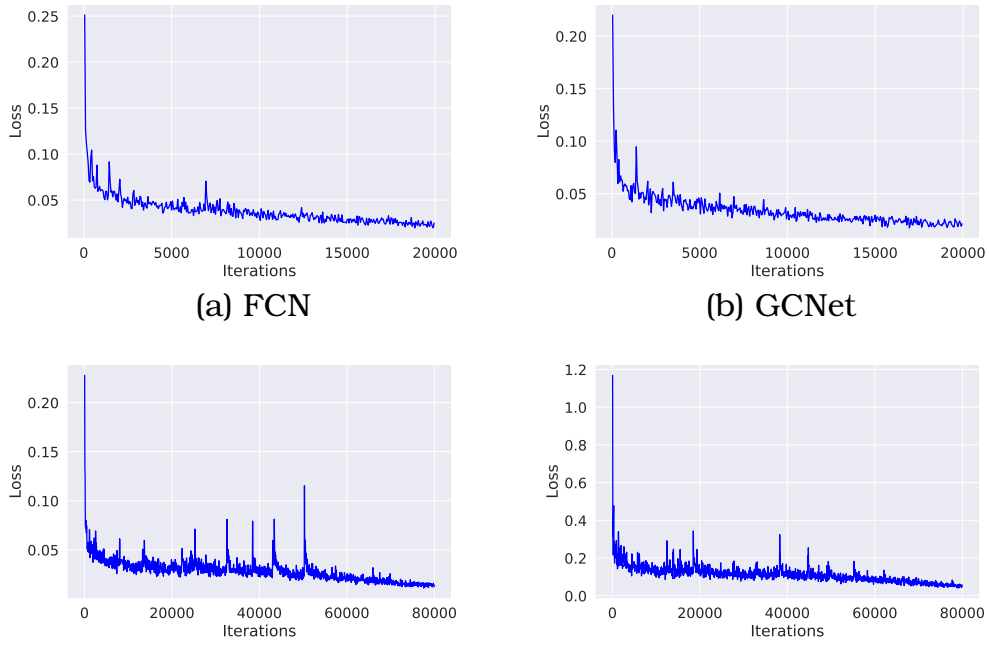
In our experiments, images manually annotated were randomly divided into training (70%), validation (20%), and testing (10%), considering five-fold cross-validation to conduct a robust analysis. *Cross-validation* is a statistical strategy used to split the dataset into folds randomly. Thus, one fold is for testing, and the remaining fold is for training the methods [Arlot and Celisse, 2010]. In a five-fold cross-validation process, this is repeated five times, using each fold one time as the test set. All the methods were implemented using the MMSegmentation [MMSegmentation, 2020].

Additionally, we used the data augmentation technique during the training to improve the generalization capability of tested models. As our data augmentation strategy, we used random cropping, random flip, photometric distortion, and normalization. We used the Stochastic Gradient Descent Optimizer [Ruder, 2016] to train the methods with a learning rate of 0.01, a momentum of 0.9, and a weight decay of 0.0005 for 20k or 80k iterations. Table 2.1 shows the settings of the methods, backbone, and iterations for each network during training. Figure 2.4 shows loss of the methods decreased and stabilized rapidly after a few iterations, indicating that the training procedure was adequate.

Method	Backbone	Iterations
FCN	ResNet50	20000
GCNet	ResNet50	20000
ANN	ResNet101	80000
PointRend	ResNet101	80000

Table 2.1: Setup settings of the methods to semantic segmentation

A workstation computer equipped with an Intel®Xeon CPU E3-1270 @ 3.80 GHz, 250GB SSD with 64 GB of RAM, Titan V graphics card with 12 GB of graphics memory, CUDA version 10.2, and Ubuntu 20.04 operating system, conducted the training, evaluation, and tests of all CNNs methods available in this work.



(a) FCN (b) GCNet

(c) ANN (d) PointRend

Figure 2.4: Loss curves for (a) FCN, (b) GCNet, (c) ANN, and (d) PointRend. All the curves decrease rapidly after a few iterations and stabilize, indicating the converging of CNN methods.

2.2.5 Performance Metrics and Statistical Analysis

We evaluated the methods using pixel accuracy (Equation 2.1) and the Intersection over Union (IoU) (Equation 2.2). The IoU, also known as the Jaccard Index, is the ratio between the intersection and the union between the ground truth (GT) and the prediction masks. We calculated the average pixel accuracy and IoU by averaging the five rounds after performing a five-fold cross-validation.

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

$$IoU = \frac{|GT \cap Prediction|}{|GT \cup Prediction|} \tag{2.2}$$

We also processed an ANOVA test [Box, 1953] with the five-fold cross-validation results for the pixel accuracy to assess whether the mean accuracy between the methods was statically different. We adopt a significance level of 5%. Finally, Tukey's post hoc test was applied to identify the statistical differences between each pair of methods. The results were also analyzed using the *boxplot* graphic to verify the convergence of model losses and accuracy.

2.3 Results

This section presents the results of the experimental evaluation of the semantic segmentation methods regarding pixel accuracy and IoU. In Sections 2.3.1, 2.3.2, and 2.3.3, we present a quantitative, computational, and qualitative analysis, respectively.

2.3.1 Performance Evaluation

We considered only the metrics of the target class (*Eucalyptus* tree) to obtain a more precise analysis of the domain of the proposed problem when analyzing these results. Therefore, the results of the background class were disregarded, as it is a majority class and does not contribute to the objective analysis of the results.

The pixel accuracy of the methods in each round of cross-validation (R1-R5) and the result of the ANOVA test are presented in Table 2.2. The last column presents the average of the rounds, with FCN providing the highest pixel accuracy average. GCNet with 78.32% and ANN with 76.44% have the second and third-best pixel accuracy averages, respectively. The PointRend method had the lowest pixel accuracy among the other methods, with 75.34%. The results of the ANOVA test indicated a p-value of 0.1366, suggesting that the differences between the means are not statistically significant. Therefore, the methods can be categorized into a single group (a).

Table 2.2: Pixel accuracy for *Eucalyptus* tree segmentation in five cross-validation rounds (R1–R5).

Method	**R1**	**R2**	**R3**	**R4**	**R5**	[1]**Average Acc (std)**
FCN	78.45	81.05	81.91	79.9	73.04	[a]**78.87 (±3.51)**
GCNet	79.51	79.28	79.52	77.58	75.72	[a]78.32 (±1.66)
ANN	75.34	78.63	77.56	75.42	75.26	[a]76.44 (±1.63)
PointRend	74.53	71.53	81.83	73.45	75.34	[a]75.34(±3.90)

[1]The same lowercase letter in this column indicates that there are no significant differences by ANOVA test (p-value >0.05).

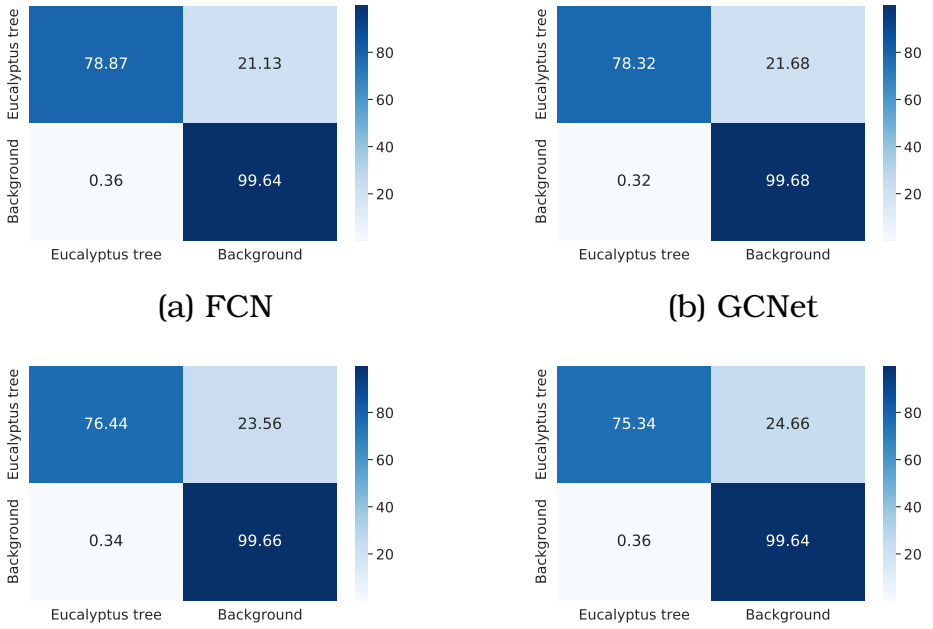
Table 2.3 presents the mean Intersection over Union (mIoU) in the five cross-validation rounds (R1–R5) of the methods for *Eucalyptus* tree class. All methods reached mIoU greater than 50%, thus demonstrating a good performance for the problem. FCN had the highest mIoU of 70.06%, followed by GCNet with 69.86%. ANN and PointRend had the lowest mIoU.

Figure 2.5 presents the confusion matrix of all segmentation methods used in this work. The confusion matrix presents the results of the *Eucalyptus* tree

Table 2.3: Mean Intersection over Union for *Eucalyptus* tree segmentation in five cross-validation rounds (R1–R5).

Method	**R1**	**R2**	**R3**	**R4**	**R5**	**Mean IoU (std)**
FCN	71.61	72.44	72.61	69.59	64.07	**70.06 (±3.56)**
GCNet	71.05	70.05	70	69.5	68.68	69.86 (±0.86)
ANN	66.79	67.07	72.04	65.82	61.81	66.71 (±3.66)
PointRend	67.18	68.07	73.04	66.82	62.81	67.58(±3.65)

and the background classes. As expected, the background class is well delimited, while the most significant confusion occurs for the *Eucalyptus* tree class. Approximately 20-25% of *Eucalyptus* pixels are classified as background. We believe these errors occur mainly near the edges, where labeling is complex.



(a) FCN (b) GCNet

(c) ANN (d) PointRend

Figure 2.5: Confusion matrix for (a) FCN, (b) GCNet, (c) ANN, and (d) PointRend.

Figure 2.6 shows a boxplot with the performance achieved by each method, including the range of performance variation. When analyzing the distribution of the metric by the boxplot, it is observed that FCN has less dispersion around the median and the presence of outliers. The GCNet and ANN showed lower performance than FCN but with similar dispersion around the median. PointRend presented the highest dispersion, including the presence of outliers.

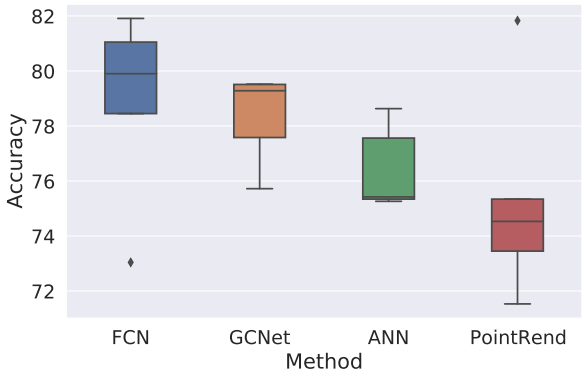


Figure 2.6: Boxplot comparing the performance of methods using Accuracy.

2.3.2 Computational complexity

Table 2.4 presents the mean inference time and standard deviation in seconds and several parameters for each method. The number of parameters of each network was obtained for an input image size of 1024x1024 pixels. Each method's complexity correlates with its inference time. For example, ANN is more profound than the others and has a longer inference time. On the other hand, PointRend is the minor complex, providing the lowest inference time.

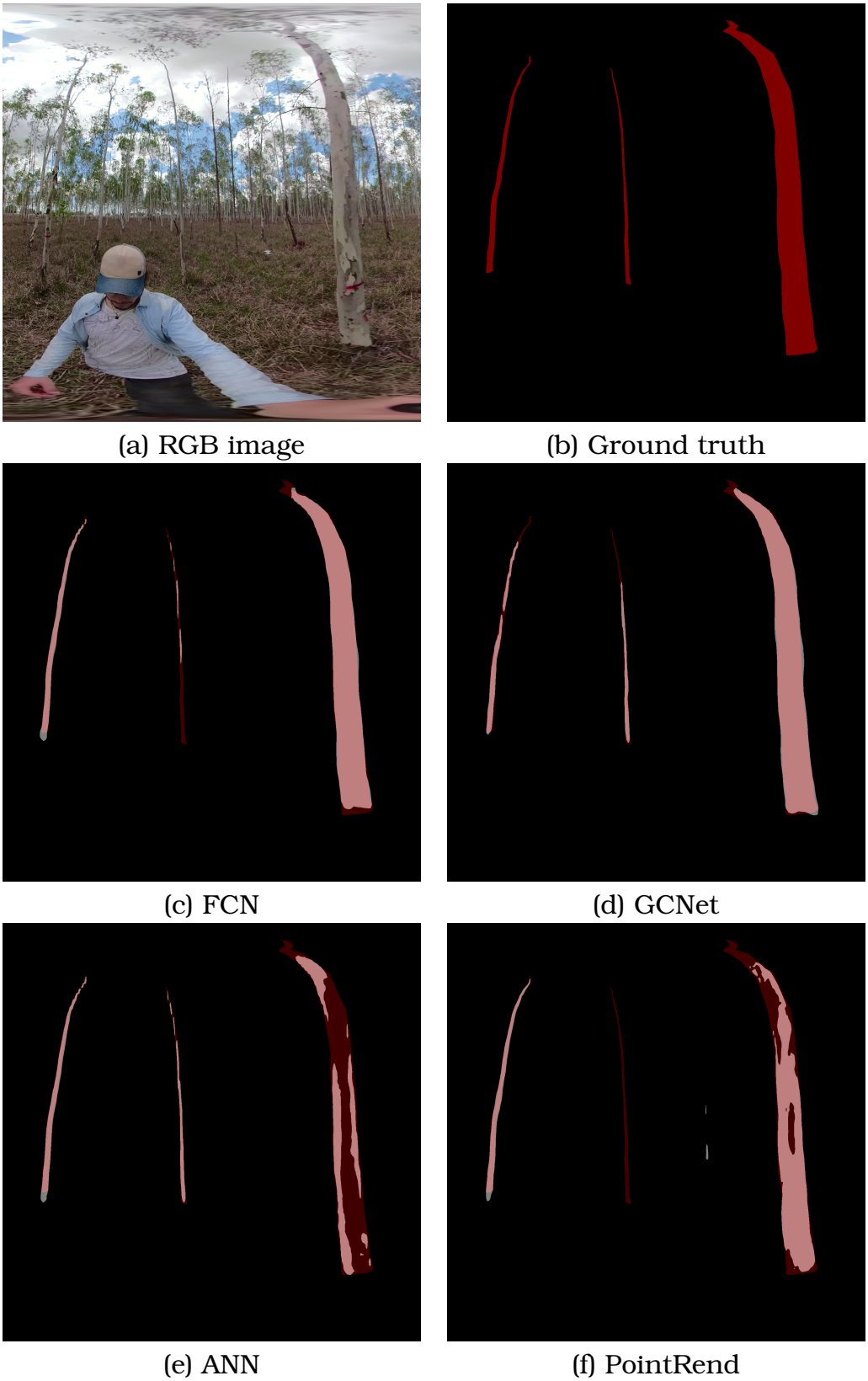
Method	**Inference Time (std)**	**Parameters**
FCN	0.157 (0.043)	49.48 M
GCNet	0.157 (0.043)	49.62 M
PointRend	**0.125 (0.039)**	**47.71 M**
ANN	**0.241 (0.040)**	**65.21 M**

Table 2.4: Results regarding the inference time and the number of parameters of the methods. The inference time represents the time spent by each method to predict an image.

2.3.3 Visual Analysis

This section discusses the qualitative results of inferences made on the test set. Figures 2.7 and 2.8 present segmentation examples of each method. These images were chosen to represent the visual segmentation because they present common situations that the methods can find when segmenting the trunk of an *Eucalyptus* tree. The two images have different scenarios, such as the *Eucalyptus* trees nearby, in the middle distance, and further away. According to the results in Figure 2.7, FCN and GCNet could better segment the edges and the interior of the trunks of the closest *Eucalyptus* trees. We can also observe that ANN and PointRend contain many errors regarding closer trees. These observations highlight that if the Diameter at Breast Height (DBH)

of the *Eucalyptus* trees were calculated, the FCN and GCNet would be better than the ANN and PointRend.

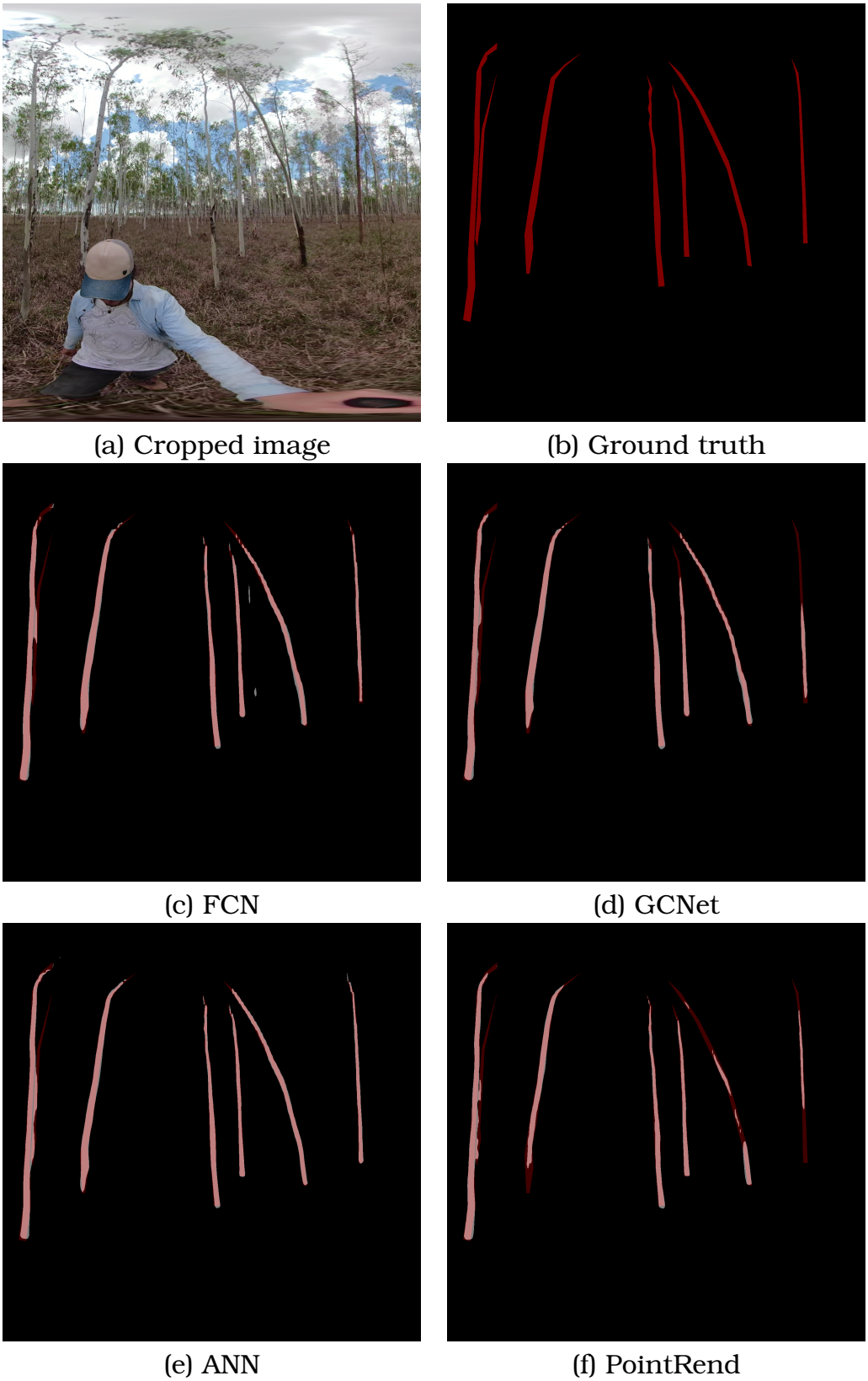


(a) RGB image (b) Ground truth

(c) FCN (d) GCNet

(e) ANN (f) PointRend

Figure 2.7: Visual results of the inference process. Areas in light red are true positives (TP), areas in dark red are false negatives (FN), areas in light gray are false positives (FP), and dark areas are true negatives (TN). Source: The author, 2022.



(a) Cropped image (b) Ground truth

(c) FCN (d) GCNet

(e) ANN (f) PointRend

Figure 2.8: Visual results of the inference process. Areas in light red are true positives (TP), areas in dark red are false negatives (FN), areas in light gray are false positives (FP), and dark areas are true negatives (TN). Source: The author, 2022.

Figure 2.8 presents a challenging scenario with several trees, some of which are curved. In this scenario, FCN and ANN give good tree segmentations.

GCNet and PointRend showed errors at the edges of the trees, resulting in the disconnection of areas. These disconnections could cause problems in tree counting applications. In both scenarios, FCN performed better than the other methods, demonstrating that the quantitative results in Section 2.3.1 are also valid for visual analysis.

2.4 Discussion

Few studies [Dias et al., 2020; Firigato et al., 2021; Ferreira et al., 2012; Khan et al., 2021] related to the segmentation of *Eucalyptus* trees in images have been carried out up to the writing moment, and most of them used aerial imagery. Here, we contribute to fulfilling this gap by evaluating both visual quality and quantitative performance of state-of-the-art deep learning methods to segment *Eucalyptus* trees in panoramic RGB images captured at ground level. The results indicated that the investigated methods performed somewhat similarly in this task, returning an average accuracy between 78.87% (FCN) and 78.32% (GCNet) and IoU between 70.06% (FCN) and 69.86% (GCNet). It is difficult to emphasize an overall better approach when evaluating the segmentation of the trunk *Eucalyptus* tree obtained with each method. Nonetheless, we have a quantitative advantage for the FCN. This method also presented a satisfactory visual result with little noise and false-positives rates regardless of tree segmentation.

The FCN and GCNet methods returned proximal inference time for both tests. PointRend presented the lowest inference time, being the fastest method among all for prediction. However, an estimate of this inference time per *Eucalyptus* trunk area demonstrates how quickly these methods can segment *Eucalyptus* trees in a given dataset once they are trained. The average time for the predictions of all tested methods was 0.17 seconds, with a standard deviation of 0.05 seconds, showing that even the worst time can still be a considerable time. This information is essential for precision image segmentation tasks since this answer can be incorporated into decision-making strategies during the development of applications in the area of counting *Eucalyptus* trees, calculating Diameter at Breast Height (DBH) and Total Height (Ht), generating 3D models of *Eucalyptus* forests, and creating different masks to help extract valuable features for involving the bark of *Eucalyptus* trees. It should also be noted that the times informed here to consider the system used to evaluate these methods (see Section 2.2.4).

Many of the problems the investigated methods face are related to distant, isolated, and small trees, which has already been observed in the literature by [Martins et al., 2021b]. To mitigate these issues, the dataset needs to be well

annotated manually by more than one specialist and contain images that differ from each other, with variations in brightness, shape, height, and distance, in addition to increasing the dataset using techniques of data augmentation, as presented in the Section 2.2.4. Most segmentation methods had problems segmenting the most distant or close to each other or curved trees. A possible approach to solving these problems would be using segmentation methods based on Transformers [Ranftl et al., 2021; Xie et al., 2021; Zheng et al., 2021], which are the most recent semantic segmentation methods in the literature. This work contributes to the literature by evaluating the potential of segmentation methods in the context of segmentation of *Eucalyptus* trees, which is a tree that has a significant environmental and socioeconomic value, as it serves as raw material in various sectors of the paper industry and used in silvopastoral system [Schettini et al., 2021; De Vechi and Júnior, 2021]. For future work, we intend to evaluate the exploration of methods based on Transformers [Ranftl et al., 2021; Xie et al., 2021; Zheng et al., 2021] in the mentioned context since they are the most recent semantic segmentation methods.

2.5 Conclusion

This work demonstrated the ability of four novel deep learning methods (FCN, GCNet, ANN, and PointRend) for *Eucalyptus* tree segmentation in panoramic RGB ground-level images. The FCN network is the most robust model to deal with several *Eucalyptus* tree characteristics, such as trees with variation in distance between trunks or curvature and trees of different heights and sizes. GCNet and ANN networks also presented similar performance to the FCN, but for specific contexts. GCNet and FCN (Figure 2.7) were identical for tree segmentation in a closed field of view, showing fewer errors on the edges and inside of the *Eucalyptus* trunks. While ANN and FCN networks (Figure 2.8) were similar for tree segmentation at a medium field of view, curved and closer to each other, providing results without disconnection in the segments. We also noted that the FCN method presents a proper time for segmenting an image, around 0.157 seconds, as tree segmentation is a complex problem that needs optimization in its computational cost. Our approach contributes to the development of applications in the area, such as *Eucalyptus* tree counting, estimation of parameters like DBH and Ht, 3D models generating of *Eucalyptus* forests, creation of masks to help extract valuable features for the bark of *Eucalyptus* trees, and also forest inventory management. We recommend exploring Transformers based-methods in the mentioned context.

CHAPTER

3

Improving Semantic Segmentation of *Eucalyptus* Trunk using RGB-D Images

Using new modern technologies is extremely important for the agribusiness sector in Brazil. This sector represented 27% of the Brazilian Gross Domestic Product in 2020. The forestry sector stood out with the advance of agribusiness production. Developing new technological solutions can contribute to increased productivity and ensure improvements in the planting process, production, and management of wood. Using artificial intelligence and deep learning technologies can be a possible path. In this sense, this work's objective was to develop and evaluate a post-processing technique to improve the results of current semantic image segmentation (SIS) networks. A stereo camera was used to assemble a robust, high-quality dataset of vertical *Eucalyptus* tree trunks. Each image contains information about the visible color spectrum and its depth to the camera. After creating and annotating the dataset, the SIS algorithms FCN, ANN, GCNet, SETR, SegFormer, and DPT were trained, evaluated, and tested on the images that a specialist duly annotated. The developed post-processing technique significantly improved the results of image segmentation networks. For metrics analysis, IoU and F1-score performance metrics were considered. Before applying the post-processing technique, convolution-based networks (FCN, ANN, and GCNet) averaged 78.89% IoU and 87.36% F1-score, while transformer-based networks (SETR, SegFormer, and DPT) averaged 86.96% IoU and 92.73% F1-score. After applying the technique to the results of the networks, a significant gain was observed in all networks. The gain in networks based on traditional convolution increased to 97.93% in IoU and 98.81% in F1-score, representing a

significant gain of 24.13% in IoU and 13.11% in the F1-score. On the other hand, transformer-based networks reached 97.82% IoU and 98.81% F1-score after applying the technique, representing a significant gain of 12.49% for IoU and 6.56% for F1-score. Transformer-based networks performed well before the application of the technique. However, the technique still brought significant improvements in their results. The inference time was also analyzed. Nevertheless, it was observed that the technique only added 0.019 seconds on average to the final time of the networks, representing a low amount to pay in favor of gains in performance. The SegFormer network achieved the best results in all tests before and after applying the technique, obtaining the best IoU, F1-score, and inference time values. In addition, a post-processing technique proved effective in correcting segmentation failure, erosion, and dilation errors, resulting in more accurate edges and better-delimited trunks. The work evaluated both the developed methods' visual quality and quantitative performance. It contributed to enriching the discussion on the segmentation of *Eucalyptus* trees by proposing an innovative approach.

3.1 Introduction

Agribusiness is an economic sector that encompasses all activities related to the production, processing, storage, marketing, and distribution of agricultural, livestock, forestry, and agro-industrial products. Agribusiness is a sector of great importance in many countries, as it provides food and other essential products for the population. In addition, agribusiness is vital for a country's economy, as it generates jobs and income for many people and contributes to economic growth. Agribusiness also plays an essential role in preserving the environment, as it promotes the conservation of natural resources and the production of food sustainably. Due to its economic and social importance, agribusiness is an area in constant evolution and development, with the increasing use of advanced technologies to increase efficiency and food production. The agricultural industry is one of the main segments of the Brazilian economy, responsible for around 27% of Brazil's Gross Domestic Product (GDP) in 2020 [CNA, 2022]. According to a survey carried out by the Brazilian Association of the Wood Industry (Ibá), the wood, cellulose, and paper sectors stood out with the advance of agricultural production [UOL, 2022]. Cellulose production in Brazil grew by 4.9% in the third quarter of 2021, around 5.6 million tons of pulp, which showed that the apparent pulp consumption increased by 16% compared to 2020 [UOL, 2022]. Planted forests play an essential role in carbon sequestration, among other ecological services, and *Eucalyptus* is the most commonly used tree in this sector

[Daniel Feffer, 2019]. *Eucalyptus* is a species of tree of great economic importance for agribusiness in Brazil. Since it was introduced in the country in the 19th century, *Eucalyptus* has been widely planted throughout Brazil, mainly in regions with temperate and tropical climates. Currently, *Eucalyptus* is one of Brazil's most important forest species, with a fundamental role in producing wood, cellulose, and paper, in addition to other industrial applications.

Eucalyptus wood is used in several applications, such as construction, furniture, and paper. In addition, *Eucalyptus* is a fast-growing tree that can be planted in degraded areas and contributes to forest restoration in many regions of the country. Due to its economic and environmental importance, *Eucalyptus* continues to be a species of great value to Brazilian agribusiness. In 2021, the national territory recorded an expansion of 9.5 million hectares (ha) covered by cultivated forests [IBGE, 2021]. *Eucalyptus* species dominated with 7.3 million hectares planted, followed by pine, with an area of 1.8 million hectares [IBGE, 2021]. The wood, pulp, and paper sectors are exposed to various challenges and factors that can reduce their productivity [ABIMCI, 2018]. Wood production is a fundamental process for the forest sector and is responsible for providing the primary raw material for the wood, pulp, and paper industries. However, this process is exposed to several challenges and factors that can affect its efficiency and productivity. If the wood production process is affected or poorly managed, it could result in possible economic losses for the industry. This could be due to internal factors such as logistical issues, poor management, or external factors such as adverse weather conditions or changes in market demands. Therefore, the wood production process must be carefully managed to minimize risks and maximize benefits for the forestry sector. Quality and productivity maintenance of the forest sector depends on continuous monitoring of the planted areas, resulting in an operational challenge. It comes imbued with the necessity for fine-scale measures and quantification of the production because it impacts managing the resources to maximize the production of planted areas. Models that address the relationship between climatic variables and production are frequently found in the literature [Santana et al., 2008; Porter and Semenov, 2005; White et al., 2011], but it is demanded more robust methods to support this sector of the economy.

SIS is a sub-area of image processing and artificial intelligence (AI). Image processing is an area of AI dedicated to developing algorithms and techniques to analyze and extract useful information from images. SIS is an analysis process that tries to divide an image into different regions or segments and assign a semantic tag to each one, allowing the classification of the parts of the image according to their meaning. Bringing the application of this technique to

the forest context, the semantic segmentation of an image of a forest can divide the image into segments with tags such as a tree, soil, and sky. In *Eucalyptus* plantations, semantic segmentation can be helpful to monitor the growth and development of trees, identify problems such as diseases or insect attacks, assess the wood quality, and identify trees with the potential to produce high-quality wood. Furthermore, it can be helpful to identify forest areas potentially affected by external factors such as climate change or deforestation. In short, SIS can help manage *Eucalyptus* plantations, providing information on tree growth, development, and health.

To increase productivity and ensure improvements in the process of wood management is necessary to develop new technological solutions to support the current management of production, for example, technologies for automatic tree counting, the measurement from Diameter to Breast Height (DBH), measurement of carbon sequestered in the trunk and automatic detection of diseases in trunks or leaves. Several studies have applied the techniques of convolutional neural networks (CNN) to work with the detection and segmentation of trees [Li et al., 2017], disease detection [Zhang et al., 2019; Syarief and Setiawan, 2020] and crop field yield estimation [Yalcin, 2019]. In recent years, approaches based on the integration between remote sensing and machine learning (ML) have been proposed to attend the agricultural sector [Yu et al., 2021; Ferreira et al., 2020; Zhao et al., 2019]. The machine learning algorithms can extract complex patterns of a dataset, providing a valuable source of models for measurements and predictions [LeCun et al., 2015]. While remote sensing allows the acquisition of a large volume of data in different scales like orbital, aerial, and terrestrial levels. Several study cases in agriculture [Liakos et al., 2018], urban planning [Fathi et al., 2020; Chaturvedi and de Vries, 2021], soil and biomass [Ali et al., 2015; Padarian et al., 2020; Torre-Tojal et al., 2022], forest [Singh et al., 2016; Maxwell et al., 2018] have integrated remote sensing data and machine learning algorithms. CNNs are a class of machine learning models widely used in several areas of AI, including SISs. CNNs can learn patterns and features from input data, such as images, and use them to perform classification and prediction tasks. In SIS, CNNs are trained to divide an image into different regions or segments and assign a semantic label to each segment. CNNs can be used in both RGB and RGB-D images. RGB images contain only color information, while RGB-D images contain depth information. Adding depth information can be useful in several contexts, including *Eucalyptus*-planted forests. For example, using RGB-D images can allow for better segmentation of trees and greater precision in identifying problems such as diseases or insect attacks.

RGB-D images allow measuring the distance of each point in the image to

the sensor that captured it. These images are beneficial in many contexts, allowing for more significant image analysis and interpretation accuracy. In the context of *Eucalyptus* planted forests, RGB-D images can be used to measure the height and diameter of trees, in addition to allowing the identification of problems such as diseases or insect pests. In addition, RGB-D images can also be used to assess the wood quality and identify trees with the potential to produce high-quality wood. RGB-D images can also help identify forest areas that can be affected by external factors such as climate change or deforestation. The use of RGB-D images is an active area of research in the field of Artificial Intelligence, with the development of increasingly accurate and efficient algorithms for analyzing these types of images [Xing et al., 2020; Jianbo Jiao, 2019; Seichter et al., 2021].

This work proposes constructing, testing, and evaluating a new approach for SIS. For this, it is intended to develop an image post-processing technique that improves the results of current networks using depth information from RGB-D images. Using as a test case images of tree trunks with a color aspect (RGB) at ground level that considers the depth (D) of the images. With this, we will use images with four dimensions of information, which will be collected by a specific camera that captures the depth of the [Tadic et al., 2022] images. These images have four dimensions of information (RBG-D), so they will be indicated in this project as RGB-D images. It is understood that the semantic segmentation of images is a vital area for AI, as it requires a high degree of precision in its results. This precision will guarantee that developing modern technological systems based on semantic segmentation will be more reliable.

The development and improvement of post-processing techniques for SISs, which work with RGB-D images, are significant scientific challenges for the computer vision (CV) area, mainly because the current image segmentation algorithms, which use only RGB images, suffer from segmentation faults, overlapping objects and holes in the segmented image of the same object, which significantly degrades the expected final result, reducing the accuracy of the models. Implementing the proposed approach can reduce investment losses in several sectors that work with AI and image segmentation. The proposed method can increase the accuracy of current image segmentation methods and, consequently, allow the advancement of new studies in this area. The materialization of precise depth sensors in modern cameras allows the advancement of classic segmentation approaches. New post-processing techniques for the semantic segmentation of images can bring good results and better performances. The development of a post-processing technique in RGB-D images results in an approach that can help in the development of intelligent systems for forest management, as it brings improvements to mod-

ern SIS techniques, expanding the potential productivity of the sector of this vital agribusiness sector. For the proposed test case, a set of RGB-D images will consist of *Eucalyptus* trunks at ground level. Only images of *Eucalyptus* trees will be manipulated during this study, as it is the most planted tree in Brazil and the most in several sub-sectors, such as the wood, cellulose, and coal industry. Future work may explore the technique developed in other tree species.

Although the areas of CV and SIS present methods with satisfactory results for image segmentation, such methods still do not produce satisfactory results for the challenges imposed by the proposed application since most of them do not consider the depth of the images [Long et al., 2015; Cao et al., 2020; Zhu et al., 2019; Kirillov et al., 2020; Ranftl et al., 2021; Zheng et al., 2021; Xie et al., 2021]. Through this development of this work, both areas of semantic segmentation of images and agribusiness will benefit, as the expected results of the project advance scientific knowledge in AI. They can also be used to construct new technologies for the agribusiness sector. In this way, new challenges are presented to SIS methods, enabling the improvement and development of new post-processing techniques and methods to meet the needs and increase the success rate of current models.

3.2 Materials and Methods

3.2.1 Study area

The images will be captured in Jaraguari town (Zone 1) and Embrapa Gado de Corte (Zone 2) in Campo Grande, Mato Grosso do Sul, Brazil. Based on the Köppen-Geiger climate classification [Beck et al., 2018], this area is categorized as a savanna climate (Aw/As), characterized by a lack of precipitation during either the winter (Aw) or summer (As) season and monthly mean temperatures that stay above 18°C all year round. Zone 1 is composed of a *Eucalyptus* tree forest of 1.05 ha ($70m \times 150m$) planted in the 2000s (Figure 3.1 (a)). Zone 2 has trees planted following the Integrated Crop-Livestock-Forestry Systems (ICLFS) [Embrapa, 2022] protocol, located at -20.4450317º, -54.7256457º, with trees planted in the year 2012 (Figure 3.1 (b)). Both areas are planted with *Eucalyptus*, and a camera with a depth capture sensor (more details in the next subsection 3.2.2) will be used to collect the images.

3.2.2 Data acquisition

The acquisition and pre-processing of the dataset involved three main steps. To begin with, we recorded videos in two *Eucalyptus* forests, Area 1 and Area

(a) Jaraguari (b) Embrapa Gado de Corte

Figure 3.1: Zones where captures will be taken: (a) Zone 1, (b) Zone 2. All images will be taken using a camera with a depth sensor. Source: The author, 2022.

2, at a height of approximately 2 meters (m) above the ground, with the camera at two different angles, 0º and 45º. For this, we used a ZED 2 Camera [Tadic et al., 2022] with a depth capture sensor, with a resolution of 2k (2048x1080 pixels), 15 frames per second (fps), and a depth capture distance of 0.2m to 20m. We extracted the images from the captured video, sampling the video frames at 30 fps. Each frame provided an image with two pieces: the visual aspect information (RGB image) and the frame depth matrix (Depth). Finally, the frames were used to generate training, validation, and test sets, each containing the information of the RGB image, depth matrix, and its annotation (more details on the annotation process were presented in the following subsection 3.2.3), as illustrated in Figure 3.2.

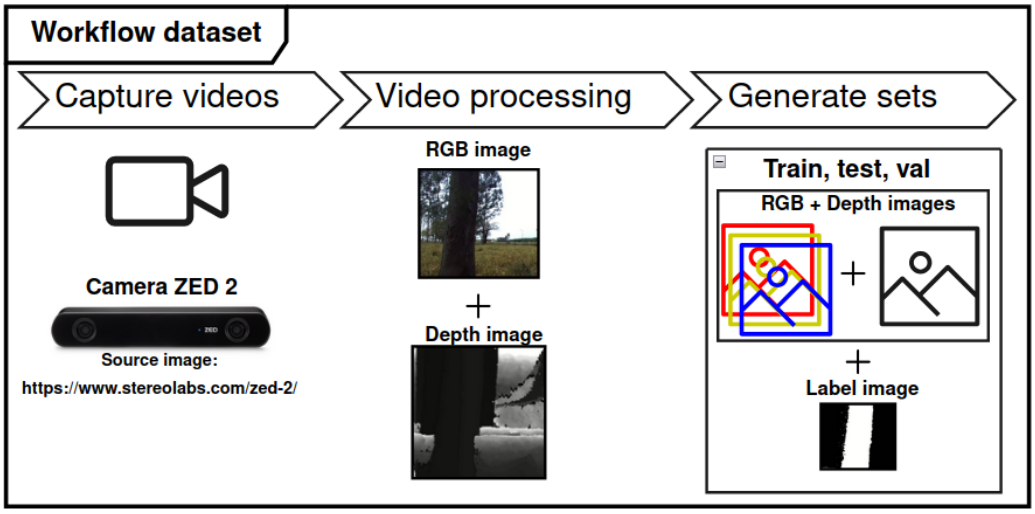
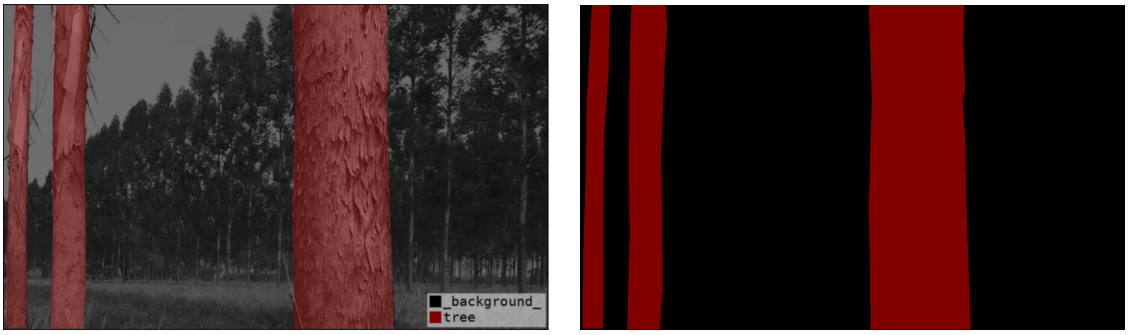


Figure 3.2: Overview of the workflow for data acquisition and data processing. Source: The author, 2022.

3.2.3 Image annotation

The captured images were manually annotated by specialists using the open annotation software LabelMe[1][Wada, 2018] and with the aid of the RGB-D depth matrix. In this process, the tree images were annotated with polygons around the trunk and labeled *Eucalyptus*. Only the trees closest to the camera were annotated, as the trees in the background were too small to annotate accurately. An illustration of the annotation pictures is presented in Figure 3.3. After the image annotation process, the images were exported to the input format of the segmentation CNNs, which was a binary mask. As a result of this process created a data set with three parts of information for each image, namely the RGB image, the image annotation, and the depth matrix. The final dataset has 2611 annotated images, averaging 1 to 3 polygons. The description of how this image dataset was partitioned for training, validation, and testing are described in Section 3.2.7.



(a) Annotated image (b) Binary mask image

Figure 3.3: Representation of annotated images (a) and the corresponding binary mask (b). The class of interest represents the red color, and the other regions are considered background. Source: The author, 2022.

3.2.4 Semantic Image Segmentation Methods

This section presents the modern segmentation methods that perform image segmentation and their paradigms for comparison with the method that this work intends to develop. The segmentation methods that work only with the RGB information of the images use different paradigms to perform segmentation on images, such as the encoder-decoder paradigm [Ronneberger et al., 2015], and atrous paradigm [Chen et al., 2016; Yu and Koltun, 2015]. The following subsections will present more details about each of these paradigms.

[1]http://labelme.csail.mit.edu/

3.2.4.1 Paradigms of image segmentation methods

This sub-subsection briefly describes the image segmentation paradigms, the encoder-decoder, and the atrous paradigms.

- **Encoder-decoder paradigm**: The models considering the encoder-decoder paradigm [Ronneberger et al., 2015] are composed of two main modules: the encoder and decoder. Each module plays a distinct role throughout the learning process, but both are connected. In this paradigm, the encoder, called the backbone [He et al., 2016], extracts the features that will be passed on to the decoder. The decoder uses this received information to reconstruct the semantic segmentation map. This process is prevalent in methods based on Transformers [Ranftl et al., 2021; Xie et al., 2021; Zheng et al., 2021].

- **Atrous paradigm**: The atrous [Chen et al., 2016] paradigm is based on the atrous convolution [Yu and Koltun, 2015] multiscale context. The atrous convolution works similarly to modern convolutions. However, it adds an extra parameter called the dilation map. This dilation map determines the values in the convolution core to extract from gaps during the convolution process. This paradigm maintains the high resolution of the extracted features.

3.2.4.2 Semantic segmentation methods in RGB images

This sub-subsection briefly presents the semantic segmentation methods of RGB images that will be tested in our work, which are Fully Convolutional Networks for Semantic Segmentation (FCN) [Long et al., 2015], Context Guided Network (GCNet)[Cao et al., 2020], Asymmetric Non-local Neural Network (ANN) [Zhu et al., 2019], Vision Transformer for Dense Prediction (DPT) [Ranftl et al., 2021], Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers (SETR) [Zheng et al., 2021], Simple and Efficient Design for Semantic Segmentation with Transformers (SegFormer) [Xie et al., 2021].

- **FCN**: Deep learning methods for image segmentation, such as the Fully Convolutional Network [Long et al., 2015], have been widely used in the literature. These networks consist of a series of layers of convolution operations that allow the extraction of features from images, followed by up-sampling layers that increase the resolution of the output to approximate the resolution of the input. The goal is to transform the input into a segmented image that classifies each pixel in the image according to the class it belongs. This approach has proven to be effective in many

segmentation tasks, but there are still challenges to be overcome, such as preserving spatial resolution and segmentation accuracy.

- **GCNet**: The Context Guided Network [Cao et al., 2020] is a deep network architecture that focuses on leveraging the context of images to improve targeting. It combines global image information with detailed local information to produce accurate results. GCNet was designed to be able to handle complex segmentation problems and has been successfully applied in a variety of image segmentation tasks. Compared to other approaches, GCNet has proven to be an effective and efficient solution for segmentation, providing accurate and high-quality results.

- **ANN**: The Asymmetric Non-local Neural Network [Zhu et al., 2019] is an image-processing neural network that uses non-local processing concepts to improve image segmentation. This network combines two non-local processing approaches, Asymmetric Pyramid Non-local Block (APNB) and Asymmetric Fusion Non-local Block (AFNB), to achieve improved results compared to conventional methods. The combination of these non-local blocks allows ANN to analyze the relationship between different parts of the image, generated in a more precise and detailed segmentation. In summary, the influence of APNB and AFNB on ANN is supported to obtain superior results compared to conventional image segmentation methods.

- **DPT**: Most non-transformer-based networks use CNNs as a backbone, especially architectures that work with the encoder-decoder paradigm. The DPT [Ranftl et al., 2021] is a dense convolution network that uses transformers as the backbone. This dense convolution approach using transformers assembles tokens at various stages of the transformer with different resolutions to generate more faithful representations of the images. Then these representations are progressively combined into full-resolution predictions. For this, a convolutional decoder is used. The DPT backbone has a global field that receives information from the stages, so this backbone constantly processes the representations. With this, the predictions are more coherent globally, meaning a gain compared to networks that use convolutional backbones. For a large amount of training data, the DPT proves to be accurate and efficient, improving performance by up to 28% compared to modern networks that work only with convolutions.

- **SETR**: SETR [Zheng et al., 2021] is a powerful segmentation model that uses a pure transformer approach to encode images. First, these images are encoded in a sequence of patches and combined with a global context.

Therefore, the encoder transformer passes the extracted information to a simpler decoder. This approach has proved to be very accurate compared to CNNs that work with the encoder-decoder format. Inserting a transformer in the encoder is the crucial point of SETR. The method SETR has reached good results in the famous image bases ADE20K [Zhou et al., 2017], where it obtained an average of 50.28% of the IoU (mIoU), Pascal Context [Mottaghi et al., 2014], where it obtained an average of 55.83% of the mIoU, in addition to good results in the Cityscapes [Cordts et al., 2016] image set.

- **SegFormer**: The SegFormer semantic segmentation works with the unification of transformers with multilayer perception (MLP) decoders. In this approach, the encoder is a hierarchically structured transformer to produce multiscale features. SegFormer decoders are simple and lightweight. This decoder carries the information from the layers to work with local and global attention more efficiently. SegFormer has a simple and lightweight design yet achieves robust results in image segmentation. The SegFormer framework contains a series of models with various parameters. The SegFormer-B4 model achieved a state-of-the-art result in the ADE20K image set, with 50.3% mIoU. SegFormer's best model, SegFormer-B5, achieved an excellent result in the validation set of the Cityscapes-C dataset, with 84.0% mIoU.

3.2.5 Approach to Post-Processing Image RGB-D

Semantic segmentation of *Eucalyptus* trunk images using RGB-D images is a challenging CV problem. This work proposes an image post-processing approach to improve the output of current segmentation networks. This approach uses RGB-D images to increase segmentation accuracy. First, the SIS networks are trained and evaluated. Next, segmentation networks are run on the dataset of images, and each output image is refined using a post-processing technique developed in this work. This technique consists of an algorithm that uses each image's depth map to improve the segmentation network's output. This image post-processing approach can significantly improve the output of current segmentation networks, increasing segmentation accuracy. The technique's main consequence is eliminating segmentation faults, holes, dilatations, and erosions in the segmented images. Figure 3.4 contains the workflow with the illustration of this process.

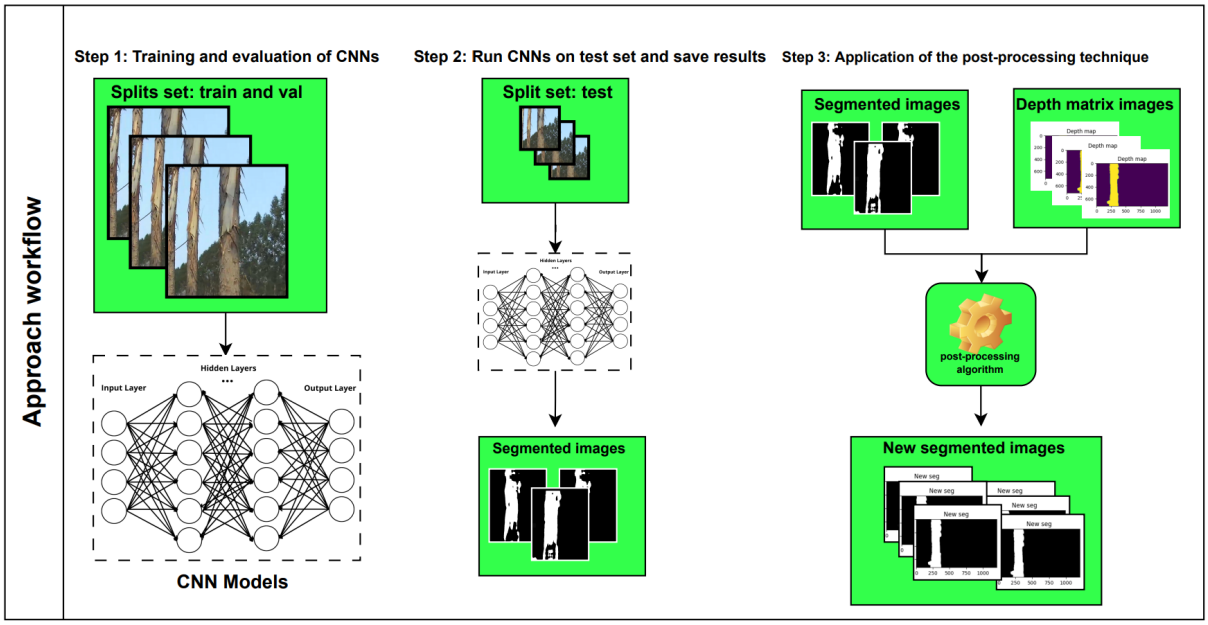


Figure 3.4: Overview of the workflow of approach. Source: The author, 2022.

3.2.5.1 Finding Connected Components in the Depth Map

The first stage of the post-processing technique consists of creating an algorithm capable of identifying all connected components in the depth map of RGB-D images using region growth. The depth matrices contain information on the distance of objects to the camera. The greater the value in the depth matrix at a given coordinate, the closer the object is to the camera. The 3.2.5.1 algorithm presents the pseudo-code to find the connected components in the depth map to create a new matrix filled with the values of the connected objects in the depth matrix, separated by class. This algorithm aims to group all connected objects in the depth matrix and enumerate them with individual classes by growing regions. The depth information will be used, as different objects at the same depth are expected to belong to the same object. This approach may be helpful, as it will provide a matrix filled with all objects connected from the depth matrix, therefore allowing the possibility of assigning a pixel-by-pixel vote on the pixels in the image segmented by the SIS networks. For this approach to work, we will need to consider some essential points about depth maps, such as the acceptable tolerance limit for two objects to be of the same class when creating the new depth matrix filled with the connected objects. For growing regions and defining classes of connected objects, we start by assigning negative classes to all connected values of the depth map. This approach to negative class values will be helpful in our work. It will help the algorithm's performance, as it prevents the same value from being reclassified again, given that during the execution of the algorithm, only positive values will be calculated. In the proposed algorithm to search for

connected components in the depth matrix, we start by defining class -1 for the background in the depth matrix. The other segments will be classified in descending order from the value -2, as shown in the algorithm. The growing regions were then cultivated using the *flood_fill* function from the *Scikit-image* library [van der Walt et al., 2014]. The *flood_fill* function fills all values close to the source value based on an acceptable tolerance. The new value filled in the matrix differs from the origin point value, which is the negative class passed as a parameter in our case. This process is repeated until all matrix points are visited and enumerated with classes with negative values. The region's growth step ends when there are no more positive points to visit. At the end of the algorithm, we multiply the final matrix values by -1 to make the class values positive and return a new depth matrix filled with the values of the connected components.

1: **Algorithm 1:** Growing Region of Post-Processing Approach
Require: DepthMatrix, Tolerance
Ensure: FillImage
2: **if** *DepthMatrix* there is no pixel != 0 **then**
3: **return** *DepthMatrix*
4: **end if**
5: *Index* $\leftarrow$ -2
6: *seed_point* $\leftarrow$ first point != -1 of *DepthMatrix*
7: Set the value -1 for a background in *DepthMatrix*
8: **while** there are pixels to grow **do**
9: # Grows current pixel based on tolerance
10: $DepthMatrix \leftarrow flood\_fill(DepthMatrix, seed\_point, Index, Tolerance)$
11: **if** *DepthMatrix* there is no pixel > 0 **then**
12: break
13: **end if**
14: $Index \leftarrow Index - 1$
15: *seed_point* $\leftarrow$ next point > 0 of *DepthMatrix*
16: **end while**
17: *FillImage* = $\leftarrow$ ((*DepthMatrix*)*(-1))
18: **return** *FillImage*

3.2.5.2 Algorithm for Improve Results

The last stage of our approach was the development of an algorithm to improve the results of images segmented by SIS networks. In this method, we combine the pixels of the image segmented by the SIS networks with the values of a new matrix of connected components obtained from the algorithm

of the previous subsection. This algorithm voted for each pixel of the image segmented by the SIS network, validating them with the connected components matrix. This poll compared the pixel values of the segmented image with the neighboring pixels of the depth matrix to overcome problems such as segmentation faults, holes, erosions, and segmentation errors in *Eucalyptus* trunks. The Algorithm pseudo-code 3.2.5.2 presents how this process works. First, we extracted the new array of components. This array provided a set of components that belonged to the same class. Then, we loaded the segmented image and extracted the objects with the value 255, as they represented the segmented *Eucalyptus* trunks of interest. We use the x and y coordinates of the segmented image to index the array of connected components, resulting in a list of objects that we use to calculate the labels and frequencies for comparison. We calculated the frequency of each pixel in the image object to determine how many times a given class appeared in the indexed image. This returned us the unique objects and the number of times each object appeared in the indexed array of objects. The next step was to find objects with a volume greater than 10% of the total volume of the image. For this, we divided the number of times each object appeared in the depth matrix by the total number of pixels in the image, resulting in a matrix with the percentage of each object. Finally, we replaced the values of objects with a volume greater than 10% of the total image volume with the value 255, representing the segmented trunks. We expected this approach to improve the outputs of current SIS networks significantly.

1: **Algorithm 2:** Improve Results of Networks Outputs
Require: ImageSegmeted, DepthMatrixImage, Tolerance
Ensure: NewImageSegmented
2: $FillDepthImage \leftarrow$ grown regions of $DepthMatrixImage$
3: $Points_{XY} \leftarrow$ new empty array of points
4: **for** each *pixel* in *ImageSegmeted* **do**
5: **if** *pixel* is equal to 255 **then**
6: add *pixel* in the $Points_{XY}$
7: **end if**
8: **end for**
9: $Objects \leftarrow$ new empty array
10: **for** each *pixel* in *FillDepthImage* **do**
11: **if** $Points_{XY}$ contains *pixel* **then**
12: add *pixel* in the $Objects$
13: **end if**
14: **end for**

Find unique values in *Objects* and set *bins* to those values
Count the number of occurrences of each bin and set the *counts* to those values
Divide *counts* by the sum of all values in *counts*
Find all bins with a count greater than 0.1 and set *new_objects* to those bins
for each *obj* in *new_objects* **do**
    if *obj* is not equal to 1 **then**
        Set *FillDepthImage* to 255 where *FillDepthImage* is equal to *obj*
    end if
end for
NewImageSegmented ← *FillDepthImage*
return *NewImageSegmented*

3.2.6 Performance Evaluation

The efficacy of the tested and developed methods was validated using the pixel F1-score metrics (Equation 3.4) and the Intersection over Union (IoU) (Equation 3.5). The results of the F1-score equation were used, as it represents the harmonized average between precision and sensitivity. Precision is the proportion of correct predictions made, while sensitivity is the fraction of true positives correctly identified. We did not use the results of the accuracy equation (Equation 3.3) because, in image segmentation problems, accuracy is not a good evaluation metric, as it is affected by class imbalance, which means that if there are more instances of one class than another, the accuracy will be higher. The IoU, also referred to as the Jaccard Index, was the ratio of the intersection and union between the ground truth (GT) and predicted masks. In the equations 3.1, 3.2, and 3.5, the true positivity (TP) are pixels correctly classified as part of objects of interest, false positives (FP) are pixels incorrectly classified as part of objects of interest, true negatives (TN) are pixels correctly classified as not part of objects of interest, and false negatives (FN) are pixels incorrectly classified as not part of the objects of interest.

$$Precision = \frac{TP}{(TP+FP)} \tag{3.1}$$

$$Recall = \frac{TP}{(TP+FN)} \tag{3.2}$$

$$Accuracy = \frac{TP+TN}{(TP+TN+FP+FN)} \tag{3.3}$$

$$F1-score = \frac{2*(Precision*Recall)}{(Precision+Recall)} \tag{3.4}$$

$$IoU = \frac{TP}{(TP+FP+FN)} \tag{3.5}$$

3.2.7 Experimental Setup

In our experiments, the annotated images were randomly divided into training (60%), validation (20%), and testing (20%) sets. Table 3.1 presents the information about the partitioned dataset. Due to the large number of images obtained and annotated, cross-validation techniques proposed by [Arlot and Celisse, 2010] were not used, as the dataset was robust. All RGB image segmentation methods were implemented using MMSegmentation[2] [MMSegmentation, 2020], which is an artificial intelligence algorithms benchmark that uses PyTorch[Paszke et al., 2019] libraries, taking advantage of its strong GPU acceleration for model training. Potential GPUs were used to train, evaluate and test the models in all cases.

Split folder	**Number of imagens**	**Size**
Train	1.566	2,8 GB
Val	522	944,8 MB
Test	523	953,8 MB

Table 3.1: Dataset information about folders, including the folder name, the number of images contained in it, and the total size.

The Table 3.2 presented shows the configuration of different SIS networks used in this study, including the name of the method, the backbone used, and the number of iterations during training and validation. The table's first column presents the segmentation method's name, such as FCN, GCNet, ANN, SETR, SegFormer, and DPT. Each of these methods is a different approach to image segmentation. The second column of the table shows the type of backbone used for each method. The backbone is the basic structure of a neural network that can extract features and feature vectors from images. In this case, all methods use CNNs as a backbone. The specific backbones used are ResNet 50 [He et al., 2015], ResNet 101 [He et al., 2015], ViT-L [Dosovitskiy et al., 2020], MIT-B0 [Xie et al., 2021] and ViT-B [Dosovitskiy et al., 2020]. The third column shows the number of iterations performed during training and validation for each method. The number of iterations is important because it is directly related to the processing time to train the SIS networks. It can be seen that SIS networks based on transformers, such as SETR, SegFormer, and DPT, were trained for the longest time, with 160000 iterations, while the other SIS networks, such as FCN, GCNet, and ANN, were trained with fewer

[2]https://github.com/open-mmlab/mmsegmentation

iterations, with 20000 or 80000 iterations. This can be explained by the fact that transformer-based networks contain more complex decoders [Xie et al., 2021], which require longer training to adjust the weights correctly.

Method	**Backbone**	**Number of Iterations**
FCN	ResNet 50	20000
GCNet	ResNet 50	20000
ANN	ResNet 101	80000
SETR	ViT-L	160000
SegFormer	MIT-B0	160000
DPT	ViT-B	160000

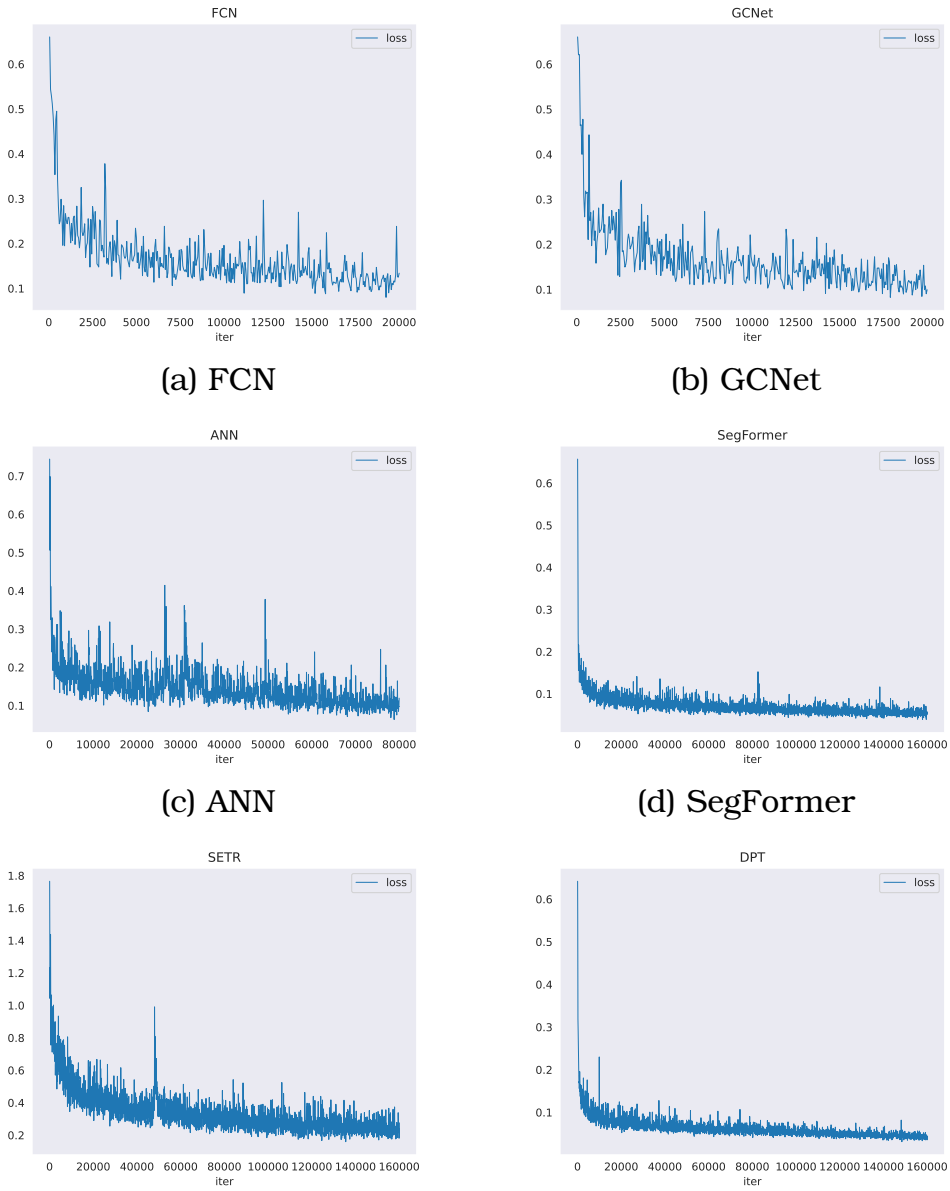
Table 3.2: Table of SIS network configurations, including name, backbone, and number of interactions during training and validation.

In addition, data augmentation strategies will be applied during training to improve the generalization of the tested models. Specifically, we will apply random clipping, random flipping, photometric distortion, and normalization. The stochastic gradient descent optimizer [Ruder, 2016] was used to train the methods with a learning rate of 0.01, a momentum of 0.9, and a decay weight of 0.0005 for 20k, 80k, and 160k iterations. As is evident from Figure 3.5, the loss of the methods dropped quickly after a few iterations and leveled off, suggesting that the training process was successful. Figure 3.6 shows the convergence progress of the IoU result during the training of SIS networks. As seen in each graph, all SIS networks showed convergence, indicating that the networks' performance increased and stabilized throughout the training. Convergence of the IoU result is a positive indication of the training process and provides confidence in the model's ability to perform well in future tasks. Empirical experiments were conducted to validate the loss and to adjust the previously determined thresholds to assess the performance of the methods.

The process of training, evaluating, and testing the CNN models was carried out on a workstation with an Intel®Xeon E3-1270 CPU @ 3.80 GHz, 250 GB SSD with 64 GB of RAM, an NVIDIA Titan V graphics card with 12 GB memory dedicated graphics, CUDA 10.2 [NVIDIA et al., 2020] and open-source operating system GNU/Linux Ubuntu 22.04.

3.3 Results

In this session, the results of the experimental evaluation of the semantic segmentation methods in terms of pixel precision and IoU will be presented, as well as the comparison of the results obtained with the post-processing technique developed during this study. In Sections 3.3.1, 3.3.2, and 3.3.3, we



(a) FCN (b) GCNet

(c) ANN (d) SegFormer

(e) SETR (f) DPT

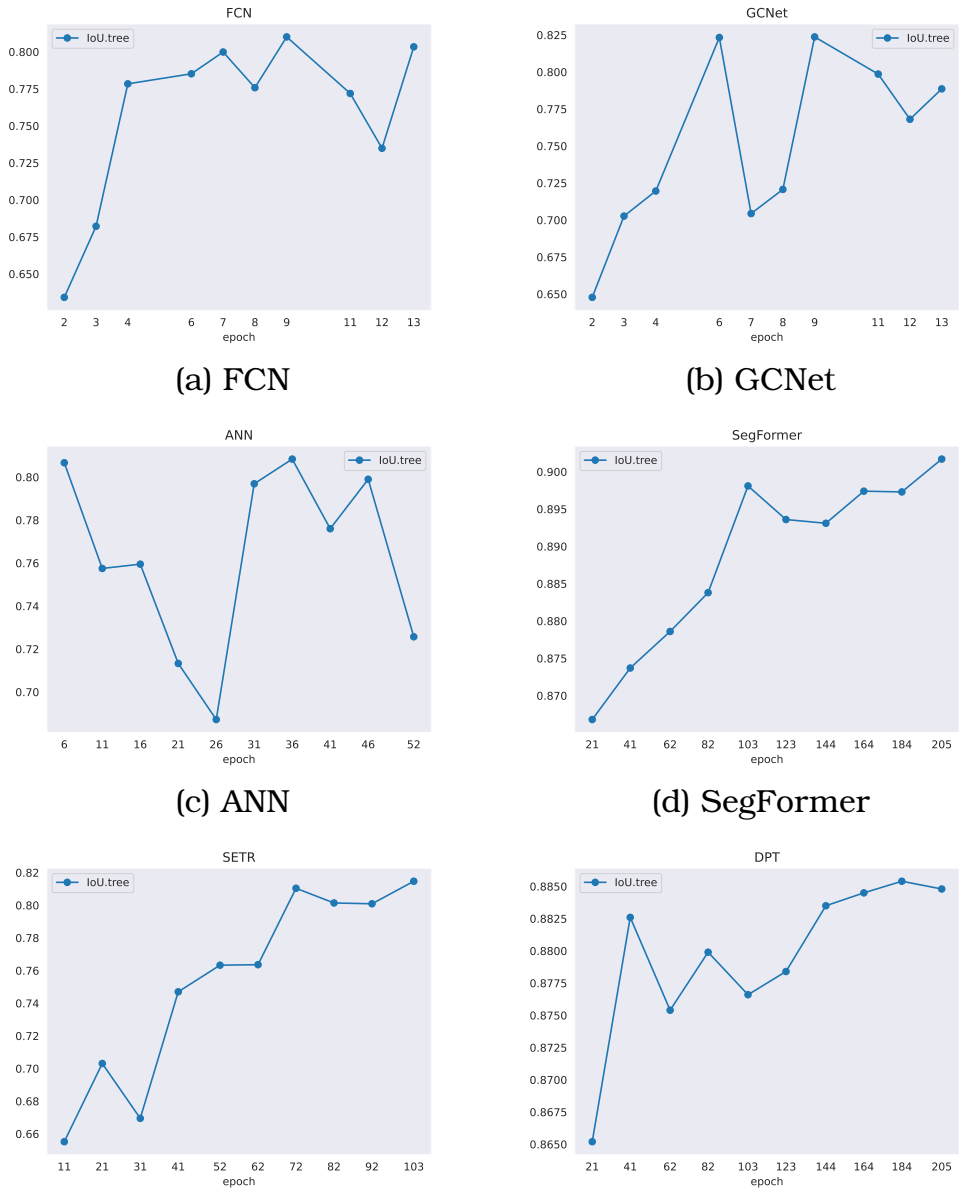
Figure 3.5: Loss curves during training for (a) FCN, (b) GCNet, (c) ANN, (d) SegFormer, (e) SETR, and (f) DPT. The curves quickly decline after a few iterations and become steady, suggesting that the techniques were effectively trained.

provide quantitative analysis, computational complexity analysis, and qualitative analysis, respectively.

3.3.1 Quantitative analysis

Only the metrics of the target class (*Eucalyptus* tree) were considered for evaluating these outcomes. As the background class is a majority class and does not contribute to the accurate analysis of the results, its data was disregarded.

Table 3.3 presents the F1-score results of six different networks (FCN,



(a) FCN (b) GCNet

(c) ANN (d) SegFormer

(e) SETR (f) DPT

Figure 3.6: IoU curves during training for (a) FCN, (b) GCNet, (c) ANN, (d) SegFormer, (e) SETR, and (f) DPT. The curves gradually increase after a few iterations and become stable, suggesting that the techniques were effectively trained.

ANN, GCNet, SETR, SegFormer, and DPT) and the results after applying the post-processing technique. Regarding the actual results of the networks, SegFormer presented the best performance, with a 94.51% F1-score. On the other hand, ANN presented the worst result, with 86.16% of the F1-score. The other networks presented intermediate results, with an average value of F1-score of 90.04%. After applying the post-processing technique, all networks showed significant improvements in the F1-score. ANN showed the most remarkable improvement, with an increase of 14.37% in its result. The FCN and GCNet networks also had powerful performances, with gains of 12.71% and 12.29%,

respectively. SegFormer, which already had the best results before the technique, also performed well, with a gain of 4.79%, which raised its F1-score to 99.04%, while DPT presented a gain of 4.71%, changing from 94.12% to 98.55 %. The other networks also significantly improved, with average gain values of 9.74%.

Method	**F1-Score CNN**	**F1-Score Post-process**	**Gain**
FCN	87.73%	98.88%	12.71%
ANN	86.16%	98.54%	**14.37%**
GCNet	88.18%	99.02%	12.29%
SETR	89.55%	98.84%	10.37%
SegFormer	**94.51%**	**99.04%**	4.79%
DPT	94.12%	98.55%	4.71%

Table 3.3: Percentage of Pixel Accuracy (F1-Score) results for *Eucalyptus* tree segmentation.

The results demonstrate that the post-processing technique successfully enhanced the performance of the F1-score metric for the evaluated networks, particularly for the ANN, FCN, and GCNet networks, which experienced the most significant gains, with an average improvement of 13.11%, compared to the actual results. Transformers-based networks also had significant gains, although smaller than the other networks. However, it is essential to remember that these results are contextual and were obtained from segmenting the *Eucalyptus* trunks in RGB-D images at ground level.

Presented in Table 3.4 are the results of the IoU of evaluating the SIS networks and applying the developed post-processing technique. The table includes the name of the method, the result obtained by the machine learning network, the result obtained by the network after applying the post-processing technique, and the percentage gained from applying the technique. Regarding the results for the IoU metric, the SegFormer network presented the best result before applying the post-processing technique, with an IoU of 89.86%. The FCN network presented the worst result before the technique, with an IoU of 79.38%. The post-processing technique showed a significant gain in the ANN network, increasing its initial result from 77.5% to 97.74%, equivalent to an increase of 26.12%. This result indicates that the post-processing technique was very effective in improving the accuracy of the ANN network. The more than 25% increase in accuracy represents a considerable improvement and could significantly impact ANN network applications.

The SegFormer network presented an initial IoU of 89.86%, and after the application of the post-processing technique, it presented an increase to 98.17%, resulting in a gain of 9.25%. Compared to the other networks, SegFormer already had high results before the technique was applied, which may explain

Method	IoU CNN	IoU Post-process	Gain
FCN	79.38%	97.90%	23.33%
ANN	77.50%	97.74%	**26.12%**
GCNet	79.79%	98.14%	23.00%
SETR	81.82%	97.97%	19.74%
SegFormer	**89.86%**	**98.17%**	9.25%
DPT	89.19%	97.31%	9.10%

Table 3.4: Percentage of IoU results for segmentation of *Eucalyptus* trunks.

the lower gain compared to other networks. However, even with already high results, the SegFormer still significantly improved after applying the post-processing technique. It was observed that all networks showed significant IoU gains, with gains ranging from 9.10% (DPT) to 26.12% (ANN). The general average of the results after applying the technique was 97.87%, while the average before the technique was 82.92%. The developed post-processing technique showed an overall positive impact on the performance of SIS networks. This can be observed by the significant gains in the FCN, GCNet, and ANN networks, which presented performance gains of 23.33%, 23.00%, and 26.12%, respectively. Although all networks showed significant gains, the FCN, ANN, and GCNet networks were the ones that most benefited from the technique. The post-processing technique effectively improves the overall performance of the IoU of image segmentation networks.

After applying the post-processing technique, the results show that convolution-based networks (ANN, FCN, GCNet) had a superior gain over transformer-based networks (SETR, SegFormer, DPT). Specifically, convolution-based networks achieved an average increase of 24.13% in IoU and 13.11% in the F1-score, while networks based on transformers showed increases of 12.49% in IoU and 6.56% in the F1-score. Although transformer-based networks are already considered very good, based on the results before applying the post-processing technique, it is essential to highlight that the application of the technique still brought significant gains in its performance. This suggests that the technique can be beneficial even for already highly optimized models and can further improve the performance of the networks in question.

3.3.2 Computational Cost Analysis

The computational analysis evaluated the inference time of the CNN networks, the post-processing time, and the total time (CNN time + post-processing time), presented in Table 3.5. The inference time analysis allowed identifying the most suitable SIS networks for real-time applications to be quantified in seconds using the post-processing technique. The table indicates the system's

overall performance and highlights methods with a good relationship between processing time and result quality. SegFormer had the lowest network inference time, with 0.062 seconds (standard deviation of 0.0302), followed by GCNet with 0.216 seconds (standard deviation of 0.129). SETR, on the other hand, presented the longest inference time, with 1.55 seconds (standard deviation of 0.0586). The time spent by the post-processing technique was minimal to the network inference time and did not significantly affect the total time. In summary, SegFormer performed the best, followed by GCNet, while SETR performed the worst. The results showed that applying the post-processing technique did not harm the final performance of the networks.

Method	**CNN Time (std)**	**Post-process (std)**	**Time Total**
FCN	0.444 (0.1290)	0.0195 (0.007)	0.4635
ANN	0.354 (0.0370)	0.0196 (0.007)	0.3736
GCNet	0.216 (0.0323)	0.0193 (0.007)	0.2353
SETR	1.550 (0.0586)	0.0194 (0.007)	1.5697
SegFormer	**0.062 (0.0302)**	**0.0184 (0.006)**	**0.0804**
DPT	0.408 (0.0425)	0.0186 (0.006)	0.4266

Table 3.5: The results include the inference time of the SIS networks, the processing time of the post-processing technique, and the total processing time. The total time represents the period required for each SIS network to complete an inference, including the time spent in the post-processing technique.

After evaluating the mean network inference times before and after applying the post-processing technique, it is observed that the time spent by the technique is relatively tiny. According to the data analyzed, the mean time spent by the technique was around 0.019 seconds. The mean original time for inferring the networks was 0.51 seconds (with a standard deviation of 0.05). The mean time after applying the technique was 0.529 seconds, representing an absolute error of only 0.019 seconds and a percentage error of 3.72%. Suggesting that the post-processing technique did not negatively affect the performance of the networks, the results indicate that its application did not harm the final performance of the networks, hence highlighting its importance.

3.3.3 Qualitative Analysis and Visual Discussion

In this section, we will evaluate the visual results of the segmentation performed before and after the post-processing technique on the test set. In our analysis, we will highlight the technique's improvements and limitations. To illustrate the improvements, we will consider cases where the SIS network presented segmentation failures, such as holes and erosions in the resulting image. We will choose representative examples from different scenarios to dis-

cuss and examine common issues faced by the SIS network during the final segmentation process.

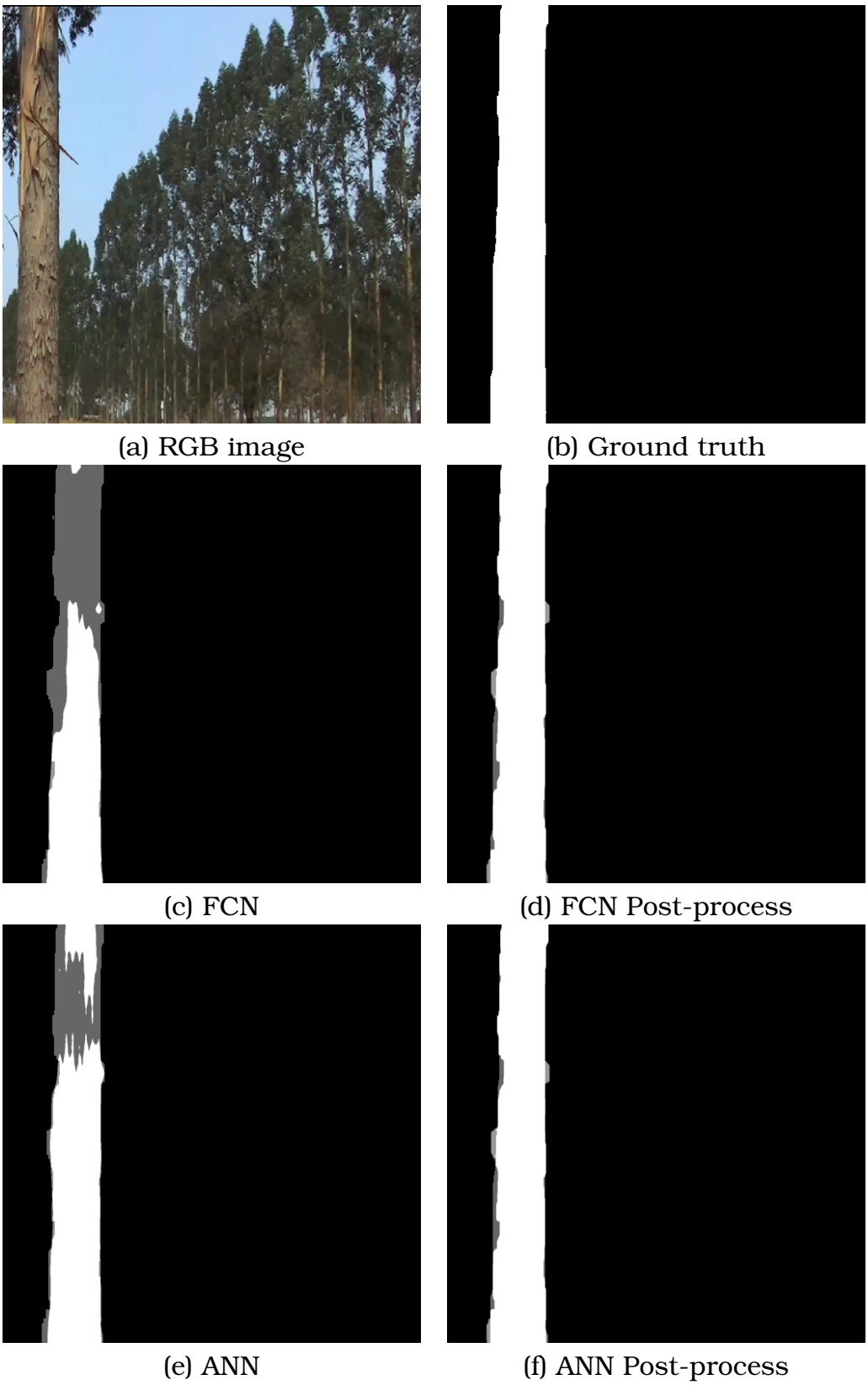
3.3.3.1 Improving Object Segmentation Edges

This subsection examined how the post-processing technique improved segmentation faults in SIS networks. The presence of segmentation errors at the edges of trees can cause disconnections in the segmentation of areas, which can impair the accuracy of solutions that rely on accurate tree counts. These disconnects can negatively affect the efficiency of applications requiring accurate information about the number of trees in an image. Therefore, it is essential to guarantee the precision of the segmentation of the edges of the trees to avoid errors that could harm the final result. We selected images that depict common mistakes made by the SIS networks evaluated during the segmentation of *Eucalyptus* trunks in the test set. The application of the technique resulted in a notable improvement in segmentation accuracy and error correction, as illustrated in Figure 3.7. After applying the technique, both FCN and ANN had segmentation failures, as evidenced by the gray areas in images (c) and (e) of Figure 3.7. However, after applying the technique, the accuracy in identifying the trunks increased significantly, as seen in images (d) and (f) of Figure 3.7, which resulted in more accurate results and closer to the true position of the trunks (ground truth). The comparison between the images before and after the application of the technique suggests that this can be a valuable addition to the pipeline of current image segmentation networks. The results suggest that the technique effectively corrected segmentation errors and segmentation failures of the objects of interest, providing more accurate and reliable visual results.

Fixing segmentation faults on objects is effective in improving segmentation accuracy. Specifically, about *Eucalyptus* trunks, a significant improvement in segmentation accuracy was observed after applying the technique. This improvement can be precious in evaluating the amount of wood, and the quality of trees in a forest since the accuracy in measuring the diameter at breast height (DBH) is fundamental for this evaluation. Correcting segmentation faults can also be helpful in applications that require precise information about the position and size of objects in an image.

3.3.3.2 Improving Object Segmentation Erosions and Dilatations

This section will analyze how the post-processing technique improved erosion and dilatation errors in SIS networks' segmentation of *Eucalyptus* trunks. We will verify the results before and after applying the technique and evaluate its effectiveness in reducing these errors. The SIS networks faced a cru-



(a) RGB image (b) Ground truth

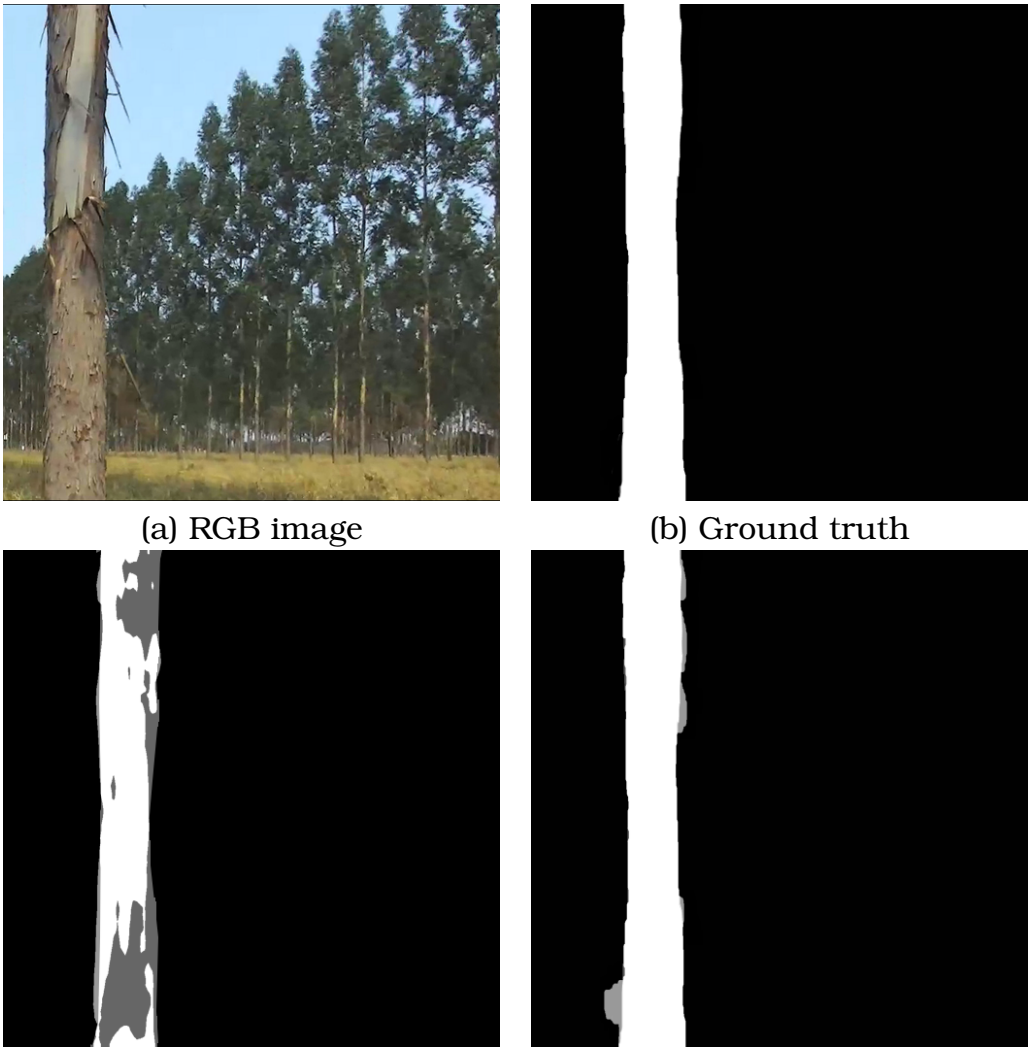
(c) FCN (d) FCN Post-process

(e) ANN (f) ANN Post-process

Figure 3.7: Visual results of the inference process before and after applying the technique. The white areas are the pixels where the network is segmented correctly, while the gray areas are the image pixels where segmentation failure occurred.

cial challenge in accurately delimiting the *Eucalyptus* trunks in the evaluated images. The images obtained by SIS network inference showed erosions and dilatations at the edges of the trunks, resulting in partially disconnected, eroded, or dilated segments. These erosions and dilatations can negatively affect technological solutions in the forest area, such as estimating tree biomass or calculating carbon stock, which depends on precise segmentation and delimitation of the edges of the trunks.

The post-processing technique effectively increased trunk boundaries' accuracy and minimized erosion errors, as shown in Figure 3.8 for the FCN. The figure illustrates a challenging scenario for the FCN, where erosions on the edges are significant, almost causing disconnection in the image, as we can see in image (c). Applying the post-processing technique effectively corrected these problems, greatly improving the accuracy of erosion errors.

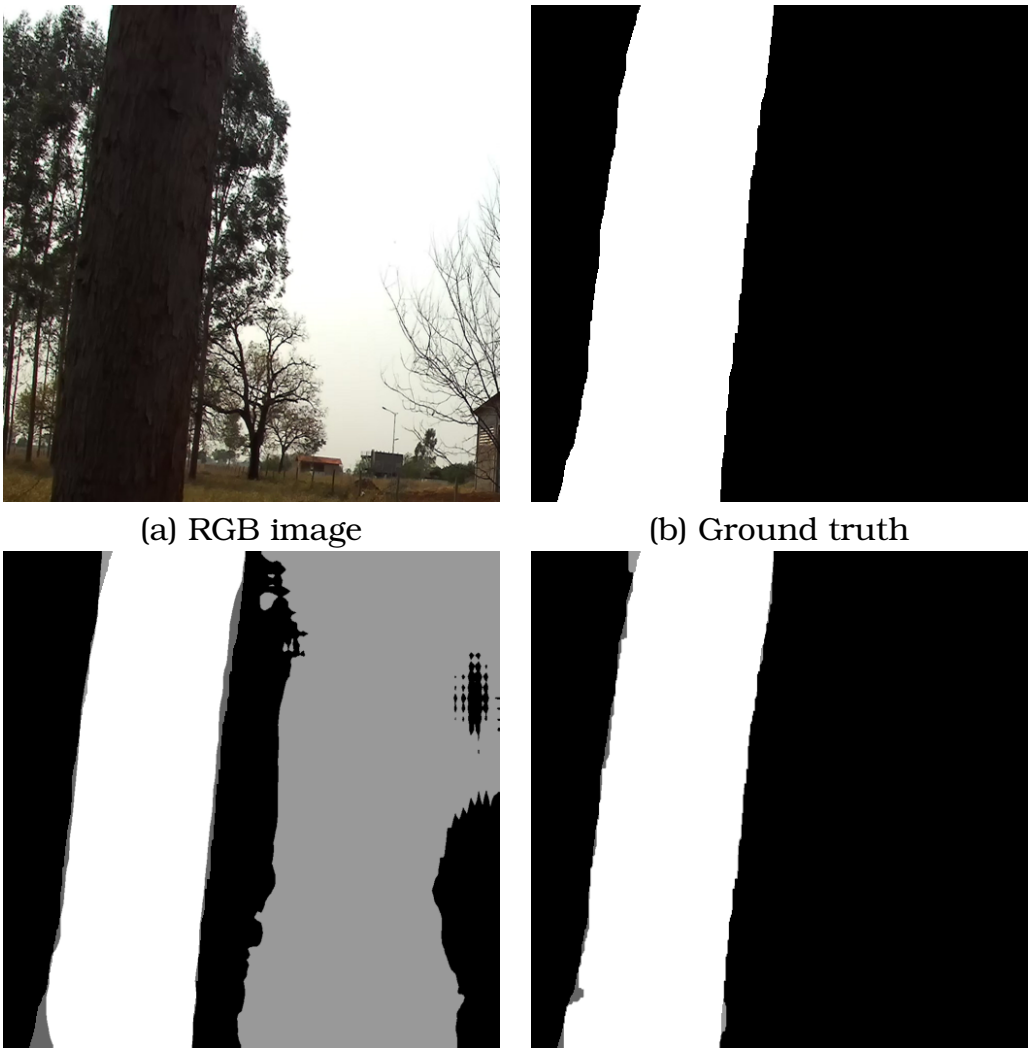


(a) RGB image (b) Ground truth

(c) FCN (d) FCN Post-process

Figure 3.8: Visual results of the inference process before and after applying the technique. The white areas are the pixels where the network is segmented correctly, while the gray areas are the image pixels where segmentation failure occurred.

In addition to the segmentation faults involving erosions, the presence of dilatations in the inferences of the SIS networks was noted, resulting in enlarged and partially disconnected segments. These dilation errors can harm applications that work with tree biomass estimation or carbon stock calculation. Figure 3.9 shows a challenging scenario for the FCN, with significant dilatations at the edges to the point of creating a partial disconnection in the image, as can be seen in image (c). In this scenario, the FCN network generated many false positives after the trunk boundary region, which caused the emergence of a large misclassified region, resulting in enlarged, partially disconnected trunks and new regions. The post-processing technique proved effective in minimizing and correcting problems related to segmentation dilation, resulting in a significant improvement in dilation errors, as shown in image (d).



(a) RGB image (b) Ground truth

(c) FCN (d) FCN Post-process

Figure 3.9: Visual results of the inference process before and after applying the technique. The white areas are the pixels where the network is segmented correctly, while the gray areas are the image pixels where segmentation failure occurred.

3.4 Discussion

The discussion about *Eucalyptus* segmentation is a current research topic, with several studies exploring using CNNs as a [Dias et al., 2020; Firigato et al., 2021; Ferreira et al., 2012; Khan et al., 2021] solution. However, these studies have limitations, as they do not take into account the geographic information of the depth of the images, focusing on aerial images captured by Unmanned Aerial Vehicles (UAV) or satellite images. This work complements this discussion by designing, developing, and evaluating state-of-the-art deep learning methods based on RGB-D images captured at ground level. In addition, we evaluated both the visual quality and the quantitative performance of the methods, allowing us to assess their effectiveness in solving this problem (see Section 3.3).

Our results indicate that including depth geographic information can greatly value the segmentation of *Eucalyptus* trees. Furthermore, our approach provides a basis for future investigations on applying the developed technique in other segmentation tasks since the technique is flexible and not limited to *Eucalyptus* trees. It can be easily adapted for other problems, as it was designed to be applied at the end of image segmentation networks (see Subsection 3.2.5). The post-processing technique developed proved to be effective in correcting errors such as segmentation failures, erosion, and dilation, thus confirming the surveys discussed in previous stages of this work. These errors negatively affect forestry applications, such as tree biomass estimation and carbon stock calculation, which depend on accurate segmentations and correct delimitation. When comparing the results obtained before and after the implementation of this technique in SIS networks, its significant value in improving the results produced by SIS networks was evident. The technique helped to produce more accurate edges, trunks without segment failures, and well delimited, resulting in more accurate trunk segmentations.

It is essential to emphasize that, before applying the post-processing technique, the SIS networks faced significant challenges in precisely segmenting the precise segmentation of *Eucalyptus* trunks, mainly networks based on traditional convolution (FCN, ANN, GCNet). On the other hand, transformer-based SIS networks (SETR, SegFormer, and DPT) performed better. However, transformer-based networks also have challenges, such as the need for large amounts of data for training, higher training time, and computational complexity compared to convolution networks. After applying the technique, all networks started to have an average performance of 97.87% in the IoU, and 98.81% in the F1-score metric, which represented a significant increase in both metrics, thus suggesting the excellent functioning of the technique. The

discussion of the results shows that the post-processing technique is a valuable tool to improve the accuracy of the segmentation of *Eucalyptus* trunks by SIS networks and can be helpful in similar solutions in other areas of imaging technology.

One of the main limitations of the technique is the dependence on a stereo camera that captures depth information. The acquisition of three-dimensional images can be a complex and expensive process, and using low-quality stereo cameras can negatively affect the results' accuracy. In addition, the technique can also be affected by limitations related to the accuracy of the stereo camera used, such as distortions, deviations, and uncertainties, which can affect the quality of the captured information. However, these limitations can be circumvented as new depth capture technologies are developed, as new depth capture technologies improve the quality of transmitted information, allowing greater accuracy in the technique. This can result in more accurate results, which is essential for applications where accuracy is critical. The advancement of new stereo cameras is a positive aspect of the post-processing technique, as it can make it more accessible and accurate, expanding its applicability and potential. It will be interesting to follow how these new technologies will impact the technique's evolution over time, paving the way for further studies on RGB-D image segmentation.

Despite adding a layer of complexity to the segmentation of *Eucalyptus* trunks, the post-processing technique does not represent a high cost in terms of time. According to the results, the average time to perform the technique is approximately 0.019 seconds, which is considered low for the benefits this technology can bring to segmenting the trunks. Thus, it is possible to conclude that the technique is a viable addition and does not preclude using the technology in practice. Furthermore, it is essential to highlight that this average time may be even lower with the advancement of technology and computer processing capacity, making the technique even more attractive in terms of execution time. In image (d) of Figure 3.8, we can observe slight noises and dilatation at the edges of the trunks. These noises are caused by the inaccuracy of the cameras when capturing the depth at the edges of the *Eucalyptus* trunks. This results in minor dilation errors or noise, which somewhat impair the efficiency of the post-processing technique. However, it is essential to highlight that the evolution of computer vision technologies can offer increasingly accurate and reliable solutions to this problem.

This work contributes to the literature by evaluating the potential of combining segmentation methods with post-processing techniques based on RGB-D images of *Eucalyptus*. This tree has significant environmental and socioeconomic value since it serves as raw material in several sectors of the log-

ging industry. New research can act by evaluating the use of more advanced computer vision techniques, combined with the use of more precise stereo cameras, to find possible ways to improve the results reported by this work. Thus, it may be possible to exploit the potential of the developed algorithm even more efficiently, maximizing its ability to accurately segment the edges of *Eucalyptus* trunks. Another possibility for future research is the development of new post-processing techniques based on the presented approach. This can help correct dilation or noise errors and solve other issues, such as optimizing processing time, reducing computational costs, and improving precision and accuracy. In addition, it is possible to explore the application of the technique in other sectors, such as agriculture, botany, or urban sectors, which work with solutions based on segmentation. We believe that in the future many solutions will be able to take advantage of the results developed in this work, such as technological solutions for measuring the height of trees, calculating carbon stock, measuring diameter at breast height, detecting pests in trunks, creating 3D models of forests or development of intelligent systems in the forest sector. For future works, it is intended to evaluate, explore and develop CNNs that work with the depth information of RGB-D images even during the training and learning process in the convolution layers [Xing et al., 2020; Jianbo Jiao, 2019; Seichter et al., 2021], bringing the concept transformer attention to these RGB-D segmentation networks [Ranftl et al., 2021; Xie et al., 2021; Zheng et al., 2021]. Therefore, several future works can be developed from the presented post-processing technique to expand its scope and applicability and improve the precision and accuracy of RGB-D image segmentation.

3.5 Conclusion

In this work, a post-processing technique for RGB-D images was developed and evaluated, which significantly improved the segmentation results of *Eucalyptus* trunks for six different SIS networks (FCN, ANN, GCNet, SETR, SegFormer, DPT). Average gains for all networks represented an 18.03% increase for IoU and a 9.74% gain for the F-1 score. The technique was applied at the end of each network and only added 0.019 seconds of image inference time, which suggests a low cost to pay for the significant gains obtained. The SegFormer network was the most robust model to deal with the segmentation of *Eucalyptus* trunks, as it obtained the best results in all evaluations, before and after applying the technique, with the best final results for the IoU and F1-score metrics, in addition to the lowest average time of inference. Convolution-based networks (FCN, GCNet, and ANN) performed

worse than Transformers-based networks (SETR, SegFormer, and DPT) before applying the post-processing technique (Tables 3.3 and 3.4). However, this difference in performance was mitigated by applying the post-processing technique, which meant that all networks started to have similar performance in the IoU and F1-score metrics. The results of the complexity analysis suggest that the post-processing technique did not improve the complexity of the networks in terms of inference time. However, it did not significative add time to the existing inference time results (Table 3.5). These results suggest that the developed technique effectively improved the performance of SIS networks and had a low computational cost to be applied. Among the improvements obtained, the post-processing technique proved to be effective in helping to correct grotesque dilation errors, erosions in the trunks, and segment failures in the segmentations, resulting in fewer errors on the edges and inside the *Eucalyptus* trunks, and without disconnection. in segments (see Subsection 3.3.3). The limitations of the technique are directly related to the precision of the stereo cameras, which capture information about the depth of the trees. These limitations concern small noises at the edges of the segmentations after their application (see Section 3.4). Because it is a limitation directly related to the hardware used to capture the RGB-D images, we believe this limitation will be eliminated with the advancement of stereo camera technology. Our approach has the potential to contribute to the development of new technological applications in the forestry area. In the context of *Eucalyptus* trees, our work can help in some sectors of forest inventory management that use image-based technologies, such as tree counting systems, trunk biomass calculation, DBH or tree height estimation, creation of models 3D of *Eucalyptus* forests and more accurate extraction of tree bark images.

CHAPTER

4

Conclusions and Future Work

This study was divided into two main phases. The first stage aimed to evaluate the efficiency of different semantic segmentation networks in segmenting *Eucalyptus* trunks from panoramic images acquired at ground level. This analysis was fundamental to starting the second phase, which involved the creation, development, and evaluation of a post-processing technique in improving the performance of these networks. During the first phase, a rigorous analysis was carried out to identify the potential of segmentation networks, analyzing and discussing the values of the IoU and F1-score metrics. The evaluation was performed using a cross-validation approach with five replications and four deep-learning methods (FCN, GCNet, ANN, and PointRend). The dataset included *Eucalyptus* trees with varied characteristics, such as variations in the distance between trunks, changes in curvature, and different sizes and diameters, making the task challenging for deep learning algorithms. The results of the first phase showed that the FCN model presented the best performance, with a pixel precision of 78.87% and mIoU of 70.06%. The GCNet and ANN networks also showed promising results but with limitations in the ability to generalize to different contexts. With the results of the first phase, it was possible to take an essential step towards the development of other tools in forest management, discussing and analyzing the results to seek corrections in the next phase. In addition, the need to evaluate more complex networks, such as networks based on transformers (SETR, SegFormer, and DPT), and expand the image dataset to obtain better quality became evident.

The second stage of the work's main objective was developing and evaluating a post-processing technique to improve the results of current image semantic segmentation networks. A stereo camera was used to create a new

robust and high-quality dataset of *Eucalyptus* trunks. The newly captured images had visible spectrum information and depth information. SIS algorithms were trained, evaluated, and tested with RGB images that an expert annotated. The developed post-processing technique significantly improved the results of the image segmentation networks, with a gain of up to 24.13% in IoU and 13.11% in the F1-score in the best cases, as discussed in the section 3.4. The average processing time of the technique is speedy, adding only 0.019 seconds to the final time of the networks. This represents a small amount to pay in favor of performance gains. Although processing time may be necessary in some applications, the results evaluated in the second stage of this project indicated that the technique would not add a significant overhead to the processing times of network inferences. The work evaluated both the visual quality and the quantitative performance of the developed methods, and it was found that the SegFormer had the most favorable results in all evaluations. In addition, the post-processing technique effectively corrected flaws in segmentation, erosion, and dilation, providing sharper edges and better-defined trunks. This study contributed to enriching the debate on the segmentation of *Eucalyptus* trees by presenting an innovative approach that considers the depth information of RGB-D images.

4.1 Contributions

Despite the challenges, the work presents contributions that can assist both the scientific community and the agribusiness sector that works with forestry management. As such, it is expected that the development of this post-processing technique will assist future work involving more precise *Eucalyptus* tree segmentation. In this way, this work contributes to the following:

- **Development of a post-processing technique**: The work presented the creation and evaluation of a post-processing technique to improve the segmentation results of currently used images. The technique was developed to correct segmentation, erosion, and dilation errors, resulting in more accurate edges and better-delimited trunks.

- **Significant increase in accuracy**: The application of the post-processing technique resulted in a significant increase in the accuracy of image segmentation networks, both in convolution-based and transformer-based networks. The gain in accuracy was measured by the IoU and F1-score metrics.

- **Fast processing time**: Although the technique significantly improved accuracy, the average additional processing time was only 0.019 seconds,

which is considered a low cost compared to the gains in performance.

- **Contribution to the forest sector**: This work contributed to the forest sector, enriching the discussion on the segmentation of *Eucalyptus* trees and proposing an innovative approach.

- **Use of advanced technologies in the agricultural sector**: The work highlighted the importance of using advanced technologies, such as artificial intelligence and deep learning, in the agricultural sector, especially in the forestry sector. The use of these technologies can contribute to increasing productivity and improve the planting process, production, and management of wood.

4.2 Limitations

Although the post-processing technique developed has shown significant improvements in the quality of segmentation of *Eucalyptus* trunk images, some limitations need to be considered:

- **Accuracy of the camera stereo used**: The quality of the images captured by the camera is critical for the performance of the image segmentation system. If the camera is not accurate enough, this can result in poor images that impair the segmentation system's ability to perform its task accurately.

- **Application to other cases use**: The post-processing technique was developed and evaluated specifically for the segmentation of *Eucalyptus* tree trunks. Therefore, it is possible that the application of the technique to other types of images or objects may not have the same performance and accuracy. New studies and tests with the technique in other domains and problems would be necessary.

In summary, the post-processing technique presented in this work is an essential contribution to the field of image segmentation of *Eucalyptus* trunks. However, some limitations still need to be considered before its large-scale application. More research and testing must address these limitations before the technique is widely adopted.

4.3 Future works

The research carried out in this work represents the beginning of a journey towards more efficient and precise solutions for the segmentation of eucalyp-

tus trees. Some areas can be explored and benefited from this post-processing approach. Some possibilities include the following:

- Application of the post-processing technique to other species of trees and objects, in addition to eucalyptus trees.
- Improvements to detection performance in adverse conditions such as shadows, reflections, and weather variations.
- Integration of computer vision methods to further automate the segmentation process.
- Adding extra layers of artificial intelligence, such as deep neural networks, to improve segmentation accuracy and robustness.
- Large-scale testing of the post-processing technique on other objects of interest and applying it to real test cases to validate the effectiveness of the solution in production environments.
- Evaluation of image segmentation networks specialized in RGB-D images [Xing et al., 2020; Jianbo Jiao, 2019; Seichter et al., 2021].
- Creation of new datasets of RGB-D images with other stereo cameras.

Future work aims to make the solution proposed in this work even more efficient and accurate so that the agribusiness sector can use it on a large scale. Thus, it can contribute to increased productivity and guarantee improvements in the planting process, production, and management of wood.

Bibliography

ABIMCI (2018). Desafios para a indústria de madeira: aumentar consumo interno e melhorar produtividade. http://bit.ly/3DT2gTt. Cited on page 25.

Ali, I., Greifeneder, F., Stamenkovic, J., Neumann, M., e Notarnicola, C. (2015). Review of machine learning approaches for biomass and soil moisture retrievals from remote sensing data. *Remote Sensing*, 7(12):16398–16421. Cited on pages 2, 8, and 26.

Arlot, S. e Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79. Cited on pages 13 and 38.

Beck, H. E., Zimmermann, N. E., McVicar, T. R., Vergopolan, N., Berg, A., e Wood, E. F. (2018). Present and future Köppen-Geiger climate classification maps at 1-km resolution. *Scientific Data*, 5(1):180214. Cited on pages 10 and 28.

Bowley, C., Mattingly, M., Barnas, A., Ellis-Felege, S., e Desell, T. (2019). An analysis of altitude, citizen science and a convolutional neural network feedback loop on object detection in unmanned aerial systems. *Journal of Computational Science*, 34:102–116. Cited on page 9.

Box, G. E. P. (1953). *Departures from Independence and Homoskedasticity in the Analysis of Variance and Related Statistical Analysis (1953)*. PhD thesis, University of London. Cited on page 14.

Cao, Y., Xu, J., Lin, S., Wei, F., e Hu, H. (2020). Global context networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1. Cited on pages 2, 4, 9, 11, 12, 28, 31, and 32.

CEPEA, ESALQ, U. (2021). PIB do Agronegócio Brasileiro. https://cepea.esalq.usp.br/br/pib-do-agronegocio-brasileiro.aspx. Cited on page 1.

Chaturvedi, V. e de Vries, W. T. (2021). Machine learning algorithms for urban land use planning: A review. *Urban Science*, 5(3):68. Cited on pages 2, 8, and 26.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., e Yuille, A. L. (2016). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. Cited on pages 30 and 31.

CNA, B. (2022). Panorama do Agro. https://cnabrasil.org.br/cna/panorama-do-agro. Cited on page 24.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., e Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Cited on page 33.

Daniel Feffer, Horacio Lafer Piva, P. H. (2019). Industria brasileira de arvores, relatório 2019. Relatório técnico. Cited on pages 1, 8, and 25.

Darwin, B., Dharmaraj, P., Prince, S., Popescu, D. E., e Hemanth, D. J. (2021). Recognition of bloom/yield in crop images using deep learning models for smart agriculture: A review. *Agronomy*, 11(4). Cited on pages 3 and 9.

De Vechi, A. e Júnior, C. A. D. O. M. (2021). Avaliação dos aspectos ambientais do cultivo do eucalipto, relato de caso em goioerê-paraná: Uma perspectiva para a educação ambiental. *UNICIÊNCIAS*, 25(1):57–64. Cited on page 21.

Dias, D., Dias, U., Menini, N., Lamparelli, R., Le Maire, G., e Torres, R. d. S. (2020). Image-based time series representations for pixelwise eucalyptus region classification: A comparative study. *IEEE Geoscience and Remote Sensing Letters*, 17(8):1450–1454. Cited on pages 3, 9, 20, and 49.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., e Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929. Cited on page 38.

Embrapa (2022). Integração Lavoura Pecuária Floresta - Portal Embrapa. https://www.embrapa.br/tema-integracao-lavoura-pecuaria-floresta-ilpf. Cited on page 28.

Fathi, S., Srinivasan, R., Fenner, A., e Fathi, S. (2020). Machine learning applications in urban building energy performance forecasting: A systematic review. *Renewable and Sustainable Energy Reviews*, 133:110287. Cited on pages 2, 8, and 26.

Ferreira, M. P., de Almeida, D. R. A., de Almeida Papa, D., Minervino, J. B. S., Veras, H. F. P., Formighieri, A., Santos, C. A. N., Ferreira, M. A. D., Figueiredo, E. O., e Ferreira, E. J. L. (2020). Individual tree detection and species classification of amazonian palms using uav images and deep learning. *Forest Ecology and Management*, 475:118397. Cited on pages 2, 8, and 26.

Ferreira, M. P., La Rosa, L. E. C., Happ, P. N., Theobald, R. B., e Queiroz, R. (2012). Mapping eucalyptus plantations and natural forest areas in landsat-tm images using deep learning. *Remote sensing*, page 4. Cited on pages 3, 9, 20, and 49.

Firigato, J. O. N., Junior, J. M., Gonçalves, W. N., e Bacani, V. M. (2021). Deep learning and google earth engine applied to mapping eucalyptus. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 4696–4699. IEEE. Cited on pages 3, 9, 20, and 49.

He, K., Zhang, X., Ren, S., e Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385. Cited on page 38.

He, K., Zhang, X., Ren, S., e Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. Cited on page 31.

IBGE (2021). Produção da Extração Vegetal e da Silvicultura | IBGE. https://www.ibge.gov.br/estatisticas/economicas/agricultura-e-pecuaria/9105-producao-da-extracao-vegetal-e-da-silvicultura.html?=&t=resultados. Cited on pages 1, 8, and 25.

Jianbo Jiao, Yunchao Wei, Z. J. H. S. R. L. T. S. H. (2019). Geometry-aware distillation for indoor semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Cited on pages 4, 27, 51, and 56.

Khan, A., Asim, W., Ulhaq, A., Ghazi, B., e Robinson, R. W. (2021). Health assessment of eucalyptus trees using siamese network from google street and ground truth images. *Remote Sensing*, 13(11):2194. Cited on pages 3, 9, 20, and 49.

Kirillov, A., Wu, Y., He, K., e Girshick, R. (2020). Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9799–9808. Cited on pages 2, 4, 9, 11, 12, and 28.

LeCun, Y., Bengio, Y., e Hinton, G. (2015). Deep learning. *Nature*, 521:436–44. Cited on pages 2, 8, and 26.

Li, W., Fu, H., Yu, L., e Cracknell, A. (2017). Deep learning based oil palm tree detection and counting for high-resolution remote sensing images. *Remote Sensing*, 9(1). Cited on pages 3 and 26.

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., e Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors*, 18(8):2674. Cited on pages 2, 8, and 26.

Long, J., Shelhamer, E., e Darrell, T. (2015). Fully convolutional networks for semantic segmentation. Cited on pages 2, 4, 9, 11, 12, 28, and 31.

Martins, J., Junior, J. M., Menezes, G., Pistori, H., Sant´Ana, D., e Gonçalves, W. (2019). Image segmentation and classification with slic superpixel and convolutional neural network in forest context. In *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 6543–6546. Cited on pages 2 and 8.

Martins, J. A. C., Nogueira, K., Osco, L. P., Gomes, F. D. G., Furuya, D. E. G., Gonçalves, W. N., Sant'Ana, D. A., Ramos, A. P. M., Liesenberg, V., dos Santos, J. A., de Oliveira, P. T. S., e Junior, J. M. (2021a). Semantic segmentation of tree-canopy in urban environment with pixel-wise deep learning. *Remote Sensing*, 13(16). Cited on pages 2 and 8.

Martins, J. A. C., Nogueira, K., Osco, L. P., Gomes, F. D. G., Furuya, D. E. G., Gonçalves, W. N., Sant'Ana, D. A., Ramos, A. P. M., Liesenberg, V., dos Santos, J. A., de Oliveira, P. T. S., e Junior, J. M. (2021b). Semantic segmentation of tree-canopy in urban environment with pixel-wise deep learning. *Remote Sensing*, 13(16). Cited on page 20.

Maxwell, A. E., Warner, T. A., e Fang, F. (2018). Implementation of machine-learning classification in remote sensing: An applied review. *International Journal of Remote Sensing*, 39(9):2784–2817. Cited on pages 2, 8, and 26.

Mendes, T. R., Miguel, E. P., Vasconcelos, P. G., Valadao, M. B., Rezende, A. V., Matricardi, E. A., Angelo, H., Gatto, A., e Nappo, M. E. (2020). *Australian Journal of Crop Science*, 14(2):286–294. Cited on page 2.

MMSegmentation (2020). MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation. Cited on pages 13 and 38.

Mottaghi, R., Chen, X., Liu, X., Cho, N.-G., Lee, S.-W., Fidler, S., Urtasun, R., e Yuille, A. (2014). The role of context for object detection and semantic segmentation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Cited on page 33.

Nogueira, K., Dalla Mura, M., Chanussot, J., Schwartz, W. R., e Dos Santos, J. A. (2019). Dynamic multicontext segmentation of remote sensing images based on convolutional networks. *IEEE Transactions on Geoscience and Remote Sensing*. Cited on pages 2 and 8.

NVIDIA, Vingelmann, P., e Fitzek, F. H. (2020). Cuda, release: 10.2.89. Cited on page 39.

Osco, L. P., de Arruda, M. d. S., Marcato Junior, J., da Silva, N. B., Ramos, A. P. M., Moryia, É. A. S., Imai, N. N., Pereira, D. R., Creste, J. E., Matsubara, E. T., Li, J., e Gonçalves, W. N. (2020). A convolutional neural network approach for counting and geolocating citrus-trees in UAV multispectral imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*. Cited on pages 2 and 8.

Osco, L. P., dos Santos de Arruda, M., Gonçalves, D. N., Dias, A., Batistoti, J., de Souza, M., Gomes, F. D. G., Ramos, A. P. M., Jorge, L. A. C., Liesenberg, V., Li, J., Ma, L., Junior, J. M., e Gonçalves, W. N. (2021). A cnn approach to simultaneously count plants and detect plantation-rows from uav imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 174:1–17. Cited on pages 2 and 8.

Osco, L. P., Ramos, A. P. M., Pereira, D. R., Moriya, é. A. S., Imai, N. N., Matsubara, E. T., Estrabis, N., de Souza, M., Junior, J. M., Gonçalves, W. N., Li, J., Liesenberg, V., e Creste, J. E. (2019). Predicting canopy nitrogen content in citrus-trees using random forest algorithm associated to spectral vegetation indices from UAV-imagery. *Remote Sensing*. Cited on pages 2 and 8.

Padarian, J., Minasny, B., e McBratney, A. B. (2020). Machine learning and soil sciences: A review aided by machine learning tools. *Soil*, 6(1):35–52. Cited on pages 2, 8, and 26.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., e Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc. Cited on page 38.

Porter, J. R. e Semenov, M. A. (2005). Crop responses to climatic variation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1463):2021–2035. Cited on pages 1, 8, and 25.

Ranftl, R., Bochkovskiy, A., e Koltun, V. (2021). Vision transformers for dense prediction. *CoRR*, abs/2103.13413. Cited on pages 4, 21, 28, 31, 32, and 51.

Rodrigues de Oliveira, B., Pereira da Silva, A., Ribeiro, L., Azevedo, G., Azevedo, G., Baio, F., Sobrinho, R., Silva Junior, C. A., e Teodoro, P. (2021). Eucalyptus growth recognition using machine learning methods and spectral variables. *Forest Ecology and Management*, 497:119496. Cited on page 1.

Ronneberger, O., Fischer, P., e Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Navab, N., Hornegger, J., Wells, W. M., e Frangi, A. F., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham. Springer International Publishing. Cited on pages 30 and 31.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*. Cited on pages 13 and 39.

Santana, R. C., Barros, N. F. d., Leite, H. G., Comerford, N. B., e Novais, R. F. d. (2008). Estimativa de biomassa de plantios de eucalipto no brasil. *Revista Árvore*, 32(4):697–706. Cited on pages 1, 8, and 25.

Schettini, B. L. S., Jacovine, L. A. G., Torres, C. M. M. E., de Oliveira Neto, S. N., da Rocha, S. J. S. S., Villanova, P. H., Alves, E. B. B. M., e Rufino, M. P. M. X. (2021). Sistemas silvipastoris com eucalipto: estocagem de carbono em diferentes espaçamentos e clones. *Ciencia Florestal*, 31(3):1047–1062. Cited on page 21.

Seichter, D., Köhler, M., Lewandowski, B., Wengefeld, T., e Gross, H.-M. (2021). Efficient rgb-d semantic segmentation for indoor scene analysis. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 13525–13531. Cited on pages 4, 27, 51, and 56.

Singh, A., Thakur, N., e Sharma, A. (2016). A review of supervised machine learning algorithms. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 1310–1315. Ieee. Cited on pages 2, 8, and 26.

Syarief, M. e Setiawan, W. (2020). Convolutional neural network for maize leaf disease image classification. *Telkomnika*, 18:1376–1381. Cited on pages 3 and 26.

Tadic, V., Toth, A., Vizvari, Z., Klincsik, M., Sari, Z., Sarcevic, P., Sarosi, J., e Biro, I. (2022). Perspectives of realsense and zed depth sensors for robotic vision applications. *Machines*, 10(3). Cited on pages 4, 27, and 29.

Torre-Tojal, L., Bastarrika, A., Boyano, A., Lopez-Guede, J. M., e Graña, M. (2022). Above-ground biomass estimation from lidar data using random forest algorithms. *Journal of Computational Science*, 58:101517. Cited on pages 2, 8, and 26.

UOL, E. (2022). Produção de celulose no Brasil cresce 4,9% no 3º tri, mostra Ibá. https://economia.uol.com.br/noticias/reuters/2021/11/23/producao-de-celulose-no-brasil-cresce-49-no-3-tri-mostra-iba.htm. Cited on page 24.

Valadão, M. B. X., Carneiro, K. M. S., Ribeiro, F. P., Inkotte, J., Rodrigues, M. I., Mendes, T. R. S., Vieira, D. A., Matias, R. A. M., Lima, M. B. O., Miguel, E. P., e Gatto, A. (2020). Modeling Biomass and Nutrients in a Eucalyptus Stand in the Cerrado. *Forests*, 11(10):1097. Cited on page 2.

van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., Yu, T., e the scikit-image contributors (2014). scikit-image: image processing in Python. *PeerJ*, 2:e453. Cited on page 35.

Vepakomma, U., St-Onge, B., e Kneeshaw, D. (2011). Response of a boreal forest to canopy opening: assessing vertical and lateral tree growth with multi-temporal lidar data. *Ecological Applications*, 21(1):99–121. Cited on pages 3 and 9.

Wada, K. (2018). Labelme: Image polygonal annotation with python. https://github.com/wkentaro/labelme. Cited on pages 11 and 30.

Wang, X., Girshick, R., Gupta, A., e He, K. (2018). Non-local neural networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7794–7803. Cited on page 9.

White, J. W., Hoogenboom, G., Kimball, B. A., e Wall, G. W. (2011). Methodologies for simulating impacts of climate change on crop production. *Field Crops Research*, 124(3):357–368. Cited on pages 1, 8, and 25.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., e Luo, P. (2021). Segformer: Simple and efficient design for semantic segmentation with

transformers. *CoRR*, abs/2105.15203. Cited on pages 4, 21, 28, 31, 38, 39, and 51.

Xing, Y., Wang, J., e Zeng, G. (2020). Malleable 2.5d convolution: Learning receptive fields along the depth-axis for RGB-D scene parsing. *CoRR*, abs/2007.09365. Cited on pages 3, 27, 51, and 56.

Yalcin, H. (2019). An approximation for a relative crop yield estimate from field images using deep learning. *2019 8th International Conference on Agro-Geoinformatics (Agro-Geoinformatics)*, pages 1–6. Cited on pages 3 and 26.

Yu, F. e Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. Cited on pages 30 and 31.

Yu, R., Luo, Y., Zhou, Q., Zhang, X., Wu, D., e Ren, L. (2021). Early detection of pine wilt disease using deep learning algorithms and uav-based multispectral imagery. *Forest Ecology and Management*, 497:119493. Cited on pages 2, 8, and 26.

Zhang, S., Zhang, S., Zhang, C., Wang, X., e Shi, Y. (2019). Cucumber leaf disease identification with global pooling dilated convolutional neural network. *Computers and Electronics in Agriculture*, 162:422–430. Cited on pages 3 and 26.

Zhao, Q., Yu, S., Zhao, F., Tian, L., e Zhao, Z. (2019). Comparison of machine learning algorithms for forest parameter estimations and application for forest quality assessments. *Forest Ecology and Management*, 434:224–234. Cited on pages 2, 8, and 26.

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P. H., e Zhang, L. (2021). Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*. Cited on pages 4, 21, 28, 31, 32, and 51.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., e Torralba, A. (2017). Scene parsing through ade20k dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130. Cited on page 33.

Zhu, Z., Xu, M., Bai, S., Huang, T., e Bai, X. (2019). Asymmetric non-local neural networks for semantic segmentation. Cited on pages 2, 4, 9, 11, 12, 28, 31, and 32.