Transformer-Based Automatic Weed Segmentation in Multispectral UAV Imagery

Jocival Dantas Dias Junior



Universidade Federal de Uberlândia
Faculdade de Computação
Programa de Pós-Graduação em Ciência da Computação

Uberlândia
2025

Jocival Dantas Dias Junior

Transformer-Based Automatic Weed Segmentation in Multispectral UAV Imagery

Tese de doutorado apresentada ao Programa de Pós-graduação da Faculdade de Computação da Universidade Federal de Uberlândia como parte dos requisitos para a obtenção do título de Doutor em Ciência da Computação.

Área de concentração: Ciência da Computação

Prof. Dr. André Ricardo Backes
Prof. Dr. Maurício Cunha Escarpinati

Uberlândia
2025

Ficha Catalográfica Online do Sistema de Bibliotecas da UFU
com dados informados pelo(a) próprio(a) autor(a).

D541 Dias Júnior, Jocival Dantas, 1994-
2025 Transformer-Based Automatic Weed Segmentation in Multispectral UAV Imagery [recurso eletrônico] / Jocival Dantas Dias Júnior. - 2025.

Orientador: Andre Ricardo Backes.
Coorientador: Mauricio Cunha Escarpinati.
Tese (Doutorado) - Universidade Federal de Uberlândia, Pós-graduação em Ciência da Computação.
Modo de acesso: Internet.
Disponível em: http://doi.org/10.14393/ufu.te.2025.146
Inclui bibliografia.
Inclui ilustrações.

1. Computação. I. Backes, Andre Ricardo,1981-, (Orient.). II. Escarpinati, Mauricio Cunha,1976-, (Coorient.). III. Universidade Federal de Uberlândia. Pós-graduação em Ciência da Computação. IV. Título.

CDU: 681.3

Bibliotecários responsáveis pela estrutura de acordo com o AACR2:

Gizele Cristine Nunes do Couto - CRB6/2091
Nelson Marcos Ferreira - CRB6/3074

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
Coordenação do Programa de Pós-Graduação em Computação
Av. João Naves de Ávila, 2121, Bloco 1A, Sala 243 - Bairro Santa Mônica, Uberlândia-MG, CEP 38400-902
Telefone: (34) 3239-4470 - www.ppgco.facom.ufu.br - cpgfacom@ufu.br

ATA DE DEFESA - PÓS-GRADUAÇÃO

Programa de Pós-Graduação em:	Ciência da Computação				
Defesa de:	Tese, 02/2025, PPGCO				
Data:	28 de Fevereiro de 2025	Hora de início:	08:30	Hora de encerramento:	11:40
Matrícula do Discente:	12013CCP005				
Nome do Discente:	Jocival Dantas Dias Júnior				
Título do Trabalho:	Transformer-Based Automatic Weed Segmentation in Multispectral UAV Imagery				
Área de concentração:	Ciência da Computação				
Linha de pesquisa:	Ciência de Dados				
Projeto de Pesquisa de vinculação:	----------				

Reuniu-se por videoconferência, a Banca Examinadora, designada pelo Colegiado do Programa de Pós-graduação em Ciência da Computação, assim composta: Professores Doutores: Bruno Augusto Nassif Travençolo - FACOM/UFU , Marcelo Zanchetta do Nascimento - FACOM/UFU, Ricardo José Ferrari – CCET/UFSCar, Dalcimar Casanova - DAINF/UTFPR e André Ricardo Backes - DC/UFSCar, orientador do candidato.

Os examinadores participaram desde as seguintes localidades: André Ricardo Backes - São Carlos/SP, Ricardo José Ferrari - São Carlos/ SP, Fátima de Dalcimar Casanova - Pato Branco/PR . Os outros membros da banca e o aluno participaram da cidade de Uberlândia.

Iniciando os trabalhos o presidente da mesa, Prof. Dr. André Ricardo Backes, apresentou a Comissão Examinadora e o candidato, agradeceu a presença do público, e concedeu ao Discente a palavra para a exposição do seu trabalho. A duração da apresentação da Discente e o tempo de arguição e resposta foram conforme as normas do Programa.

A seguir o senhor presidente concedeu a palavra, pela ordem sucessivamente, aos examinadores, que passaram a arguir ao candidato. Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando o candidato:

Aprovado

Esta defesa faz parte dos requisitos necessários à obtenção do título de Doutor.

O competente diploma será expedido após cumprimento dos demais requisitos,

conforme as normas do Programa, a legislação pertinente e a regulamentação interna da UFU.

Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente ata que após lida e achada conforme foi assinada pela Banca Examinadora.

Documento assinado eletronicamente por **André Ricardo Backes**, **Usuário Externo**, em 28/02/2025, às 14:43, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Bruno Augusto Nassif Travençolo**, **Professor(a) do Magistério Superior**, em 28/02/2025, às 15:22, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Ricardo José Ferrari**, **Usuário Externo**, em 28/02/2025, às 17:04, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Marcelo Zanchetta do Nascimento**, **Professor(a) do Magistério Superior**, em 07/03/2025, às 13:00, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Dalcimar Casanova**, **Usuário Externo**, em 10/03/2025, às 16:35, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **6098166** e o código CRC **E18312EF**.

Referência: Processo nº 23117.009240/2025-24 SEI nº 6098166

Dedico este trabalho a todos que sempre me ampararam, em especial aos meus pais Sandra Divina de Jesus Urzeda e Jocival Dantas Dias, meus avós Joseva, Sinval (in memoriam), Iracema (in memoriam) e Jose Balduíno (in memoriam) e aos meus irmãos John e Lenne.

Agradecimentos

À Deus, por ter me dado o privilégio da vida e a oportunidade de trilhar o caminho que me levou a conclusão de mais uma etapa.

Aos meus pais pela sabedoria, ensinamentos e educação que me levaram a ser a pessoa que sou hoje e aos meus irmãos que sempre se dedicaram em me apoiar e orientar em todos os momentos.

Ao meu orientador Prof. Dr. Andre Ricardo Backes e coorientador Prof. Dr. Maurício Cunha Escarpinati pela orientação, amizade, críticas, ensinamentos e oportunidades que me foram dadas.

Aos Profs. Drs. Bruno Augusto Nassif Travençolo, Fabiano Azevedo Dorça e Ronaldo de Castro Oliveira por terem contribuído para minha formação profissional e pessoal.

Aos Profs. da Faculdade de Computação que tive a oportunidade de ser discente. E a todos os técnicos que sempre estiveram a disposição para me atender.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

"Per aspera ad astra"
(Sêneca).

Resumo

O manejo inadequado de plantas daninhas causa perdas anuais significativas à agricultura brasileira, com estimativas que alcançam bilhões de reais. Dada a importância vital da agricultura na economia do Brasil, o manejo eficiente dessas plantas é essencial. Nos últimos anos, os avanços tecnológicos, como o uso de sensores multiespectrais e hiperespectrais acoplados a Veículos Aéreos Não Tripulados (VANTs), aprimoraram as estratégias de manejo de plantas daninhas, possibilitando monitoramento e tomada de decisão precisos. Contudo, o grande desafio no desenvolvimento de sistemas automáticos de manejo de plantas daninhas está em segmentá-las corretamente das culturas. Essa tarefa é intrinsecamente complexa devido a diversos fatores, como a similaridade fenotípica entre culturas e plantas daninhas, oclusões, sombras, variações nas condições climáticas, diferenças nos estágios de crescimento das plantas e inconsistências durante a captura de imagens por VANTs, incluindo borrões de movimento e ruídos. Este trabalho propõe uma nova abordagem de segmentação baseada em *vision transformers*, utilizando índices de vegetação e espaços de cor alternativos para alcançar uma segmentação precisa de plantas daninhas em diversos cenários e *datasets*. A metodologia proposta integra informações espectrais, espaciais e contextuais por meio de estratégias de fusão inicial e tardia, aprimorando significativamente a precisão da segmentação. Os resultados experimentais demonstram a superioridade da abordagem em relação ao estado da arte. No conjunto de dados *Sugarcane* e *RedEdge*, estratégias de fusão tardia alcançaram um *F1-score* de $0,840$ e $0,857$ respectivamente. Para o desafiador conjunto de dados *Sequoia*, a metodologia alcançou um *F1-score* de $0,736$, destacando sua robustez em cenários agrícolas complexos.

Palavras-chave: TRANSFORMERS, APRENDIZADO PROFUNDO, PLANTAS INFESTANTES, AGRICULTURA DE PRECISAO, ÍNDICES DE VEGETAÇÃO.

Transformer-Based Automatic Weed Segmentation in Multispectral UAV Imagery

Jocival Dantas Dias Junior



Universidade Federal de Uberlândia
Faculdade de Computação
Programa de Pós-Graduação em Ciência da Computação

Uberlândia
2025

Abstract

Weed mismanagement causes significant annual losses to Brazilian agriculture, with some estimates reaching billions of reais. Given the vital role agriculture plays in Brazil's economy, efficient weed management is essential. Over the years, advances in technology, such as the use of multispectral and hyperspectral sensors coupled with Unmanned Aerial Vehicles (UAVs), have enhanced weed management strategies, enabling precise monitoring and decision-making. However, the critical challenge in developing automated weed management systems lies in accurately segmenting weeds from crops. This task is inherently complex due to various factors, such as the phenotypic similarity between weeds and crops, occlusions, shadows, variations in weather conditions, differences in plant growth stages, and image capture inconsistencies during UAV flights, including motion blur and noise. This research proposes a novel segmentation framework based on vision transformers architectures, leveraging vegetation indices and alternative color spaces to achieve precise weed segmentation across diverse scenarios and datasets. The proposed methodology integrates spectral, spatial, and contextual information using early and late fusion strategies, significantly enhancing segmentation accuracy. Experimental results demonstrate the superiority of the approach compared to state-of-the-art methods. On the Sugarcane and RedEdge datasets, late fusion strategies achieved F1-scores of 0.840 and 0.857, respectively. For the challenging Sequoia dataset, the methodology achieved an F1-score of 0.736, highlighting its robustness in complex agricultural scenarios.

Keywords: Weed Segmentation, Transformer Models, Precision Agriculture, Vegetation Indices, Multispectral Imaging, UAV-Based Monitoring.

List of Figures

Figure 1 – Variability of scenarios in precision agriculture. 20
Figure 2 – Comparison between the population growth and the needed growth in crop productions to feed population at 2050. 23
Figure 3 – An example of a fixed-wing UAV used in precision agriculture. 25
Figure 4 – Quality factors in remote sensing, from (KHANAL, 2017). 26
Figure 5 – Comparison of vigor map in (a) has a ground sampled distance of 10 meters and (b) has a ground sampled distance of 0.25 meters, from (KHANAL, 2017) 26
Figure 6 – LANDSAT TM and TERRA Aster firsts bands and their spectral resolution, from (OSE; CORPETTI; DEMAGISTRI, 2016) 27
Figure 7 – Example of RGB, multispectral and hyperspectral bands from (XIE; YANG, 2020) 28
Figure 8 – MAPIR Survey2 Camera - Visible Light RGB by MAPIR. 28
Figure 9 – MAPIR Survey3W Camera - red+green+nir (RGN, NDVI) by MAPIR. 29
Figure 10 – MicaSense Red-Edge camera by MicaSense. 29
Figure 11 – FLIR Vue Pro thermal camera by Teledyne Flir. 30
Figure 12 – Example of grayscale image. a) original image b) crop showing the intensity of the values in the region 32
Figure 13 – Example of a color image. a) original image b) crop 10x10 using the RGB color space, from (SANTOS et al., 2022) 32
Figure 14 – Schematic of the RGB color cube. Points along the main diagonal have gray values, from black at the origin to white at point (1, 1, 1), from (WOODS; GONZALEZ, 2008) 33
Figure 15 – Generic structure of a CNN, from (SAHA, 2018) 35
Figure 16 – Example of a convolution operation on two-dimensional input. a) represents the two-dimensional input (I). b) the kernel used (K). c) the result of the operation (S). 36
Figure 17 – Activation functions. 36

Figure 18 – Application of a pooling operation (max and average) of size 2. Decimal values were truncated. 37
Figure 19 – The architecture of the DeepLabV3+ model. Adapted from (CHEN et al., 2018). 38
Figure 20 – Vision Transformer Architecture. From (ALEXEY, 2020) 40
Figure 21 – Segformer Architecture. From (XIE et al., 2021) 41
Figure 22 – Fusion methods: (a) Early fusion, (b) Joint Fusion (intermediate fusion), and (c) Late Fusion. Circles represent raw features, triangles represent neural network features, and squares represent model outputs (logits/predictions). Adapted from (HUANG et al., 2020). 42
Figure 23 – Example of the orthomosaics present in the dataset. Black represents the background, green represents crops, and red represents weeds. Adapted from (SA et al., 2018). 55
Figure 24 – Example of images in the dataset. Black represents the background, green represents crops, and red represents weeds. 56
Figure 25 – Orthomosaic of the SugarCane dataset. In GroundTruth, red represents the background, green represents crops, and yellow represents weeds 56
Figure 26 – Proposed workflow for integrating multiple modalities using early and late fusion strategies in SegFormer-B1. 59
Figure 27 – Calculation of the Overlap Index. (a) The original RGB image is transformed into the color spaces described in this section. (b) Each channel of each color space is processed individually. (c) The Overlap Index (OI) is calculated for background versus weed and crop versus weed classes. 60
Figure 28 – Calculation of the Overlap Index. (a) The RGB or multispectral image is transformed into the vegetation indices described in this section. (b) The OI is computed between background, weed, and crop classes. . . . 60
Figure 29 – Example of color space transformation for images from the RedEdge and Sugarcane datasets. 65
Figure 30 – Example of vegetation index transformation for images from the RedEdge and Sugarcane datasets. 67
Figure 31 – Example of NIR and RedEdge multispectral bands from the RedEdge and Sequoia datasets. 69
Figure 32 – Traning loss, validation loss and F1-score curves for the best combinations of each model and dataset. 72
Figure 33 – Results obtained for the RedEdge dataset. Black represents the background, green represents crops, and red represents weeds. 75

Figure 34 – Results obtained for the Sequoia dataset. Black represents the background, green represents crops, and red represents weeds. 76
Figure 35 – Results obtained for the Sugarcane dataset. Black represents the background, green represents crops, and red represents weeds. 77
Figure 36 – Traning loss, validation loss and F1-score curves for the best combinations of each model and dataset. 79
Figure 37 – Results obtained for the RedEdge dataset. Black represents the background, green represents crops, and red represents weeds. 81
Figure 38 – Results obtained for the Sequoia dataset. Black represents the background, green represents crops, and red represents weeds 82
Figure 39 – Results obtained for the Sugarcane dataset. Black represents the background, green represents crops, and red represents weeds 82

List of Tables

Table 1 – Brazil's position in the world agricultural market in 2022/23, adapted from (AGRICULTURA, 2023). 17
Table 2 – Land use in millions of hectares and availability in Brazil, adapted from (AGRICULTURA, 2019). 18
Table 3 – Main multispectral vegetation indices. B: blue reflectance, G: green reflectance, R: red reflectance, RE: red-edge reflectance, NIR: near-infrared reflectance. Adapted from (RADOČAJ et al., 2023). 31
Table 4 – Main RGB vegetation indices. B: blue reflectance, G: green reflectance, R: red reflectance. Adapted from (RADOČAJ et al., 2023). 31
Table 5 – Details of dataset. Adapted from (SA et al., 2018). 55
Table 6 – Main evaluation metrics. Adapted from (HASAN et al., 2021). TP (true positive) means that the predicted value is positive and it is positive, TN (true negative) means that the predicted value is negative and it is negative, FP (false positive) means that the predicted value is positive but it is negative and FN (false negative) means that the predicted value is negative but it is positive. 57
Table 7 – Weighted Mean of OI obtained for the evaluated color spaces in the RedEdge, Sequoia, and Sugarcane datasets 66
Table 8 – Weighted Mean of OI obtained for the evaluated Vegetation Index (VI) in the RedEdge, Sequoia, and Sugarcane datasets 68
Table 9 – Overlap Index results for multispectral bands. 68
Table 10 – Combination of inputs for each dataset. 71
Table 11 – Results obtained for each combination of dataset, input and model. . . . 73
Table 12 – F1 score per class obtained for each combination of dataset, input, and model. BG stands for background. 74
Table 13 – Results obtained for the best combinations of model, input, and dataset when trained for 200 epochs. 77
Table 14 – Input combination for each encoder in each dataset. 78

Table 15 – Results obtained in late fusion approach for each combination of dataset, input and model. 78
Table 16 – Results per class obtained in the late fusion approach for each combination of dataset, input, and model. BG stands for background. 80
Table 17 – Comparison of our method with the state of the art. BG stands for Background. 83
Table 18 – Comparison of our method with the state of the art for the Sugarcane dataset. BG stands for Background. 84

Acronyms list

AUC Area Under the Curve

ASPP Atrous Spatial Pyramid Pooling

CNN Convolutional Neural Network

CI Chlorophyll Index

CIR Color Infrared

DeiT Data-Efficient Image Transformers

EVI Enhanced Vegetation Index

ExG Excess green index

ExR Excess red index

ExB Excess Blue Index

FCN Fully Convolutional Network

FCL Fully Connected Layer

GDP Gross Domestic Product

GNDVI Green-normalized difference vegetation index

GSD Ground Sample Distance

GDVI Green Difference Vegetation Index

KD Knowledge Distillation

MGRVI Modified Green Red Vegetation Index

MLP Multilayer Perceptron

NDRE Normalized Difference Red Edge Index

NDVI Normalized Difference Vegetation Index

NIR Near-infrared

NGRDI Normalized green-red difference index

NLP Natural Language Processing

NPCI Normalized Pigment Chlorophyll Index

OI Overlap Index

PA Precision Agriculture

RE Red Edge

SAVI Soil-adjusted vegetation index

SSCM Site-Specific Crop Management

SR Simple Ratio

TGI Triangular Greenness Index

UAV Unmanned Aerial Vehicle

VI Vegetation Index

VRI Variable Rate Irrigation

ViT Vision Transformer

VARI Visible Atmospherically Resistant Index

Contents

1 INTRODUCTION 17
1.1 Hypothesis 19
1.2 Contributions 20
1.3 Thesis Organization 20

2 FUNDAMENTALS 23
2.1 Precision Agriculture 23
2.1.1 Unmanned Aerial Vehicles 24
2.1.2 Remote sensing in Precision Agriculture 25
2.1.3 Imaging Sensors 27
2.1.4 Vegetation Indices 29
2.2 Digital Image Processing 30
2.2.1 Color Spaces 32
2.2.2 Image Segmentation 34
2.3 Convolutional Neural Networks 34
2.4 Vision Transformers 38
2.5 Data Fusion in Deep Learning 41

3 RELATED WORKS 45

4 METHODOLOGY 53
4.1 Datasets 54
4.2 Evaluation Metrics 55
4.3 Weed Segmentation using Transformers 57
4.3.1 Analysis of Color Spaces 58
4.3.2 Calculation of Vegetation Indices 58
4.3.3 Fusion Strategies in Transformers Architectures 61

5 EXPERIMENTAL RESULTS 63
5.1 Preprocessing of Color Spaces and Vegetation Indices 63
5.1.1 Color Spaces 64
5.1.2 Vegetation Indices 65
5.1.3 Multispectral Bands 68
5.2 Deep learning weed segmentation 69
5.2.1 Training Process 69
5.2.2 Early Fusion 70
5.2.3 Late Fusion 77
5.2.4 Comparison with the State of the Art 81

6 CONCLUSION 85
6.1 Future Work 86
6.2 Contributions in Bibliographic Production 87

BIBLIOGRAPHY 89

I hereby certify that I have obtained all legal permissions from the owner(s) of each third-party copyrighted matter included in my thesis, and that their permissions allow availability such as being deposited in public digital libraries.

Jocival Dantas Dias Junior

Chapter 1

Introduction

Historically, agriculture has contributed about one-fifth of Brazil's Gross Domestic Product (GDP). In 2022, according to data from the Ministry of Agriculture, Livestock and Food Supply, this sector accounted for 24.8% of GDP, representing in financial terms R$ 675.5 billions (AGRICULTURA, 2023). Agriculture is of great importance for the country's development, having admitted more than 22 million people with a positive balance of 1.9 million jobs from April 2022 to April 2023 (AGRICULTURA, 2023). Considering the global market, Brazil is a leader in producing and exporting several products. See Table 1.

Table 1 – Brazil's position in the world agricultural market in 2022/23, adapted from (AGRICULTURA, 2023).

Main products	**Brazil - World Ranking**	
	Production	**Exports**
Sugar	1^{st}	1^{st}
Coffee	1^{st}	1^{st}
Orange juice	1^{st}	1^{st}
Beef	2^{nd}	1^{st}
Chicken	2^{nd}	1^{st}
Corn	3^{rd}	1^{st}
Soybeans	1^{st}	1^{st}
Soybean meal	3^{rd}	2^{nd}
Soy oil	3^{rd}	2^{nd}
Cotton	4^{th}	2^{nd}
Pork	4^{th}	3^{rd}

Despite the optimistic scenario, the continued expansion of agricultural development faces social and ecological aspects, as the growing demand for land has caused social conflicts and degradation of the environment. Agriculture is an economy of scarce resources. Notice in Table 2 that only 6.7% of the area is available for agricultural expansion (56.6 million hectares) in the Brazilian territory. Therefore, we need to increase productivity in the same planted area.

Table 2 – Land use in millions of hectares and availability in Brazil, adapted from (AGRICULTURA, 2019).

Territorial Distribution (2018)	Hectares	Percentage
Agriculture in production	**244.5**	**28.7%**
Grains	62.5	7.3%
Livestock (Pastures)	158.6	18.6%
Forests Planted with Forest Essences	8.5	1.0%
Sugarcane	8.6	1.0%
Banana, Coffee, Cassava, Cocoa, Citrus, other permanent	6.3	0.7%
Areas Protected by legislation	**5480**	**64.4%**
Conservation Units - UC	133.0	15.6%
Indigenous Lands - TI	121.0	14.2%
Legal Reserve and Permanent Preservation Areas	268.0	31.5%
Cities, Roads, Hydroelectric, other	26.0	3.1%
Areas Available for Agriculture	**56.6**	**6.7%**
Others	**1.9**	**0.2%**
Brazilian territorial area	**851**	**100%**

In this context, of the demand for growth in production and the pressure to minimize social and environmental impacts, Precision Agriculture (PA) proved to be a solution for the sustainable development of agribusiness. With the basic premise of producing more using fewer resources (water, land, and agricultural inputs), PA has become necessary in all stages of production (from family farming to large landowners) (SCHRIJVER; POPPE; DAHEIM, 2016).

Among all PA technologies, weed control is one of the most important, given that these plants cause direct effects on the main crop, such as loss of productivity (through competition and allelopathy), and indirect effects, such as difficulty in harvesting, depreciation of product quality, increase in production costs, and incidence of pests and diseases. Among the various species of weeds, we can highlight, in Brazil, the Hairy Fleabane (*Conyza bonariensis*) and the Sourgrass (*Digitaria insularis*). The estimated losses caused by sourgrass in soybean plantations can reach more than 80% when left uncontrolled (BRAZ et al., 2021). According to a study by the Syngenta (SYNGENTA, 2020), the cost of weeds for Brazilian agriculture is R$9 billion per year. Of Brazil's total planted area of soybeans, 60% of them suffer from these weed plant species. Although this estimate was made only for soybeans, weeds affect several other crops, such as coffee, sugarcane, beans, cotton, rice, potatoes, and tomatoes.

Traditional methods for managing weeds have proven quite expensive and inefficient (HASAN et al., 2021). Based on Computer Vision and Unmanned Aerial Vehicle (UAV), modern methods have shown great promise. The UAV captures images at low/medium altitudes with high spatial resolution in these methods. Computer vision methods are applied to identify and manage these plants correctly. To the detriment of image ac-

quisition with land vehicles, UAVs have several advantages, such as more extensive area coverage in a shorter time and non-compaction of the soil during the acquisition process (HERRMANN et al., 2020).

The essential step to developing an automatic weed management system is correctly segmenting and detecting these weeds. Despite being a well-explored step in the literature, the automatic detection of weeds in UAV images is a complex task due to several reasons that make it difficult to generalize the problem, such as similar phenotypic characteristics between crops and weeds, occlusion, shadows, changes in climatic conditions during capture, changes in plant growth stages, and variations in capturing images during flight (motion blur and noise) (HASAN et al., 2021).

1.1 Hypothesis

This work builds on evidence from various studies demonstrating the superior generalization ability of deep learning models for weed segmentation compared to traditional image processing and machine learning methods (LIU; BRUCH, 2020) (HASAN et al., 2021). Traditional methods often struggle with challenges such as overlapping spectral signatures, varying growth stages, and inconsistent lighting conditions. These challenges can be mitigated by leveraging the richer information provided by multispectral data and vegetation indices. Figure **??** illustrates the variability of scenarios commonly encountered in precision agriculture. Multispectral images capture reflectance beyond the visible spectrum, such as near-infrared and red-edge bands, which are sensitive to vegetation structure, water content, and chlorophyll levels. Vegetation indices, such as NDVI and ExG, emphasize key biophysical properties, enabling better differentiation between crops and weeds. While deep learning has shown promising results on individual datasets, incorporating these additional inputs can further enhance model performance by providing richer spectral and contextual features. This enhancement is particularly valuable in precision agriculture, where accurate weed segmentation can reduce herbicide usage, optimize crop yields, and minimize environmental impacts. Unlike existing approaches that rely solely on RGB inputs, the proposed methodology leverages the synergistic combination of spectral, spatial, and contextual information, setting a new benchmark for weed segmentation performance. Moreover, recent advances in transformer-based architectures have demonstrated superior performance in segmentation tasks by effectively modeling long-range dependencies and capturing global contextual relationships. These models have shown promise in UAV-based weed mapping, outperforming CNN-based approaches in certain scenarios, particularly when dealing with complex backgrounds and variable lighting conditions (ZHAO; BERGE; GEIPEL, 2023). By integrating transformers into our methodology, we aim to further enhance segmentation accuracy and generalization, making weed detection more robust across diverse agricultural environments.

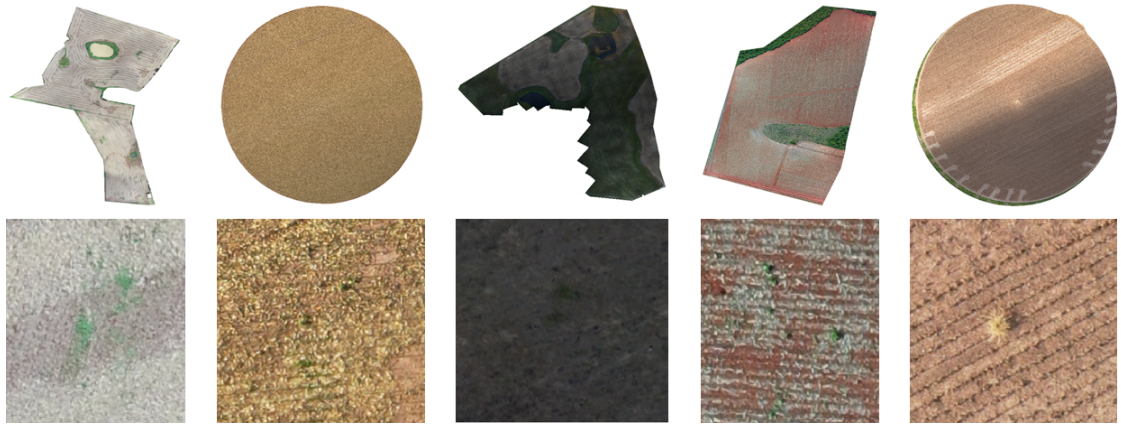


Figure 1 – Variability of scenarios in precision agriculture.

1.2 Contributions

The main contributions of this thesis are:

- ❏ Investigation of color spaces, vegetation indices, and multispectral images for weed analysis;
- ❏ Development and evaluation of a novel late fusion method for integrating vegetation indices, color spaces, and multispectral images.
- ❏ Comparative analysis of deep learning models (CNNs and Transformers) for weed segmentation.
- ❏ Proposal of a new deep learning-based method for automatic weed segmentation.

1.3 Thesis Organization

This thesis is structured into six chapters, each addressing a specific aspect of the research. A brief description of each chapter is provided below:

- ❏ Chapter 02 - Fundamentals: Introduces key concepts in precision agriculture, remote sensing, and digital image processing. It also covers the fundamentals of artificial intelligence and deep learning applied to weed segmentation, providing the necessary theoretical background for this study.
- ❏ Chapter 03 - Related Works: Reviews the state-of-the-art in weed segmentation, focusing on approaches that leverage multispectral and hyperspectral imagery captured by UAVs. It includes a comparative analysis of existing methodologies, highlighting the advancements and challenges in the field.
- ❏ Chapter 04 - Methodology: Describes the datasets used in this research and the preprocessing steps applied to color spaces, vegetation indices, and multispectral

information. It also details the proposed segmentation framework, including feature selection, deep learning models, and fusion strategies, as well as the evaluation metrics employed.

- ❏ Chapter 05 - Experimental Results: Presents the performance analysis of different model architectures (CNNs and Transformers) and fusion strategies across multiple datasets. This chapter discusses quantitative and qualitative results, including comparisons with state-of-the-art methods, and evaluates the impact of different spectral and spatial features on segmentation accuracy.

- ❏ Chapter 06 - Conclusion: Summarizes the key findings, highlighting the contributions of this research to the field of weed segmentation. It discusses the limitations of the proposed approach and outlines potential directions for future work, including improvements in model generalization, dataset expansion, and real-world applications in precision agriculture.

Chapter 2

Fundamentals

2.1 Precision Agriculture

The main discussion of the future of agriculture shows that food production must increase dramatically - probably doubling by 2050 - while using only 5% more land to sustain the growth of the world's population (HUNTER et al., 2017). Projections show that by 2050 the world population is expected to be close to 10 billions of people (HUNTER et al., 2017) (Figure 2). However, with environmental changes, this increase in food production must be made most smartly, that is, without depleting natural resources or leveraging climate change, this is the main challenge in modern agriculture.

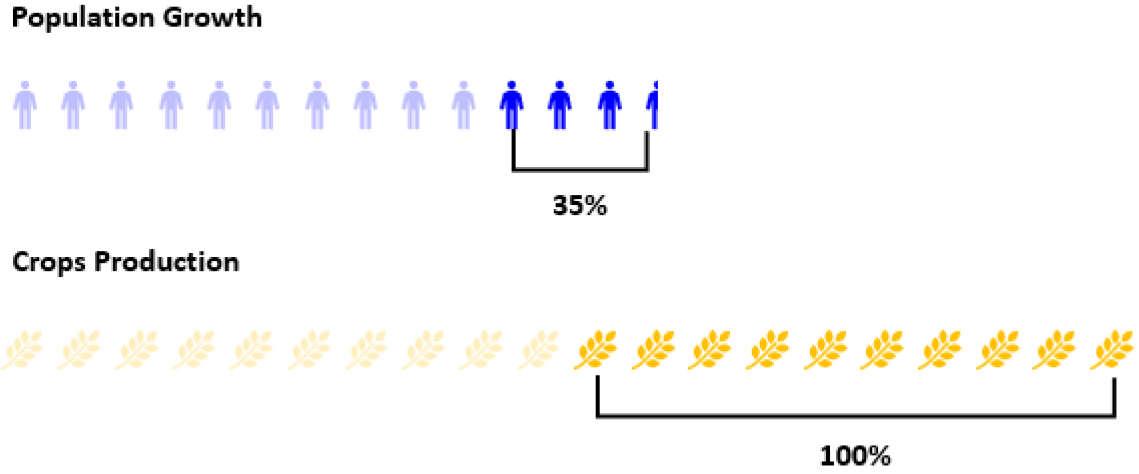


Figure 2 – Comparison between the population growth and the needed growth in crop productions to feed population at 2050.

For this reason, a new movement called PA emerged in the United States in the early 1980s with the purpose to increase the quality and quantity of agricultural production while using less input (energy, fertilizers, pesticides, water, etc.), i.e., producing more and better food while reducing environmental impact and saving costs (SCHRIJVER; POPPE; DAHEIM, 2016). The PA methods rely mainly on a combination of technologies from

remote sensing, automation, robotics, big data, artificial intelligence and the internet of things.

Precision agriculture highlights the paradigm shift in agriculture, taking it to a new level where decisions are based on data and new techniques are applied to productivity and sustainability. Developing new sensors (multispectral and hyperspectral), the Internet of Things, and GPS-based soil sampling have allowed farmers to optimize crop management. Technologies such as Variable Rate Irrigation (VRI) and Site-Specific Crop Management (SSCM) allowed the application of the correct amount and location of inputs such as water and pesticides, increasing production while reducing the cost of production, the use of materials and the environmental impact (LIAKOS et al., 2018).

Unmanned aerial vehicles have revolutionized precision agriculture, offering unprecedented capabilities in high-resolution aerial imagery. These images, combined with image processing techniques and predictive analysis, have been fundamental to detect weeds, monitoring plantation health, and predicting crop productivity (KHANAL; FULTON; SHEARER, 2017). UAV act complementary to sensors on the ground and satellite images, generating a complete view of the state of the plantation. This view allows for the correct application of inputs necessary for the crop, leading to more responsible and efficient agriculture. Combining the UAV with emerging technologies, such as deep learning, has expanded analysis and decision making capabilities (FERNANDEZ-GALLEGO et al., 2018).

2.1.1 Unmanned Aerial Vehicles

Unmanned Aerial Vehicles have developed significantly since their inception, particularly within agriculture, where they have become a fundamental tool for precision agriculture. Modern UAV used in agriculture are typically divided into fixed-wing and multi-rotor, each offering unique qualities. Fixed-wing UAVs cover a large area of ground efficiently, making them ideal for large-scale monitoring and mapping. Multi-rotor drones, such as quadcopters, are used when there is a need for stationary flight over a precise location. They are essential for tasks such as targeted crop spraying, spot-checking, and field-level monitoring. There is also a hybrid category of UAV, which combines vertical takeoff and landing capabilities with the durability and speed of fixed wings, making them a versatile solution to several precision agriculture problems (MOGILI; DEEPAK, 2018).

The use of UAV in precision agriculture is not just due to the variety of models available, but also to their integration with new technologies. Modern UAVs are typically equipped with remote sensing equipment, such as thermal and multispectral cameras, which can identify problems before they become visible to the naked eye, such as nutrient deficiencies, weeds, and water deficiencies. This ability to quickly obtain images and information regarding the plantation allows for a quick intervention in the crop, improving resource use, and increasing productivity (TSOUROS; BIBI; SARIGIANNIDIS, 2019).

The potential of UAVs goes beyond simple plantation monitoring since they can be used in crop management. Typical applications of UAVs in plantations are planting, harvesting, applying agricultural input, and controlling weeds. Automation by UAVs of these tasks can significantly reduce the cost of crop production and the environmental impact of agricultural activities due to conscious application of resources. With the advancement of robotics and automation technologies, UAVs tend to be increasingly integrated with agricultural practice.

For an example of a fixed-wing UAV, see Figure 3.



Figure 3 – An example of a fixed-wing UAV used in precision agriculture.

2.1.2 Remote sensing in Precision Agriculture

Remote sensing can be defined as the acquisition of information about an object without making physical contact, that is, using land platforms, satellites, and/or aircraft (ZARCO-TEJADA et al., 2016). In PA, remote sensing represents a convergence between technological advancement and agricultural techniques, allowing crops to be managed more efficiently, both financially and ecologically. The quality of data in remote sensing is determined by several key factors: spatial, spectral, temporal, and radiometric resolutions. See Figure 4.

Spatial resolution refers to the smallest possible object distinguishable in an image. In precision agriculture, the higher the spatial resolution, the easier it is to detect weeds, planting lines, and crop anomalies. This higher spatial resolution is essential for specific interventions in culture. Figure 5 shows the difference between the generated map of the vigor of a crop with higher and lower resolution.

Spectral resolution can be defined as the length of the continuous wavelength that a sensor can detect in the electromagnetic spectrum. In PA, spectral resolution is essential to differentiate plant species and plant conditions and to help detect soil properties. Each band in the spectrum range provides unique information that is used to monitor crop

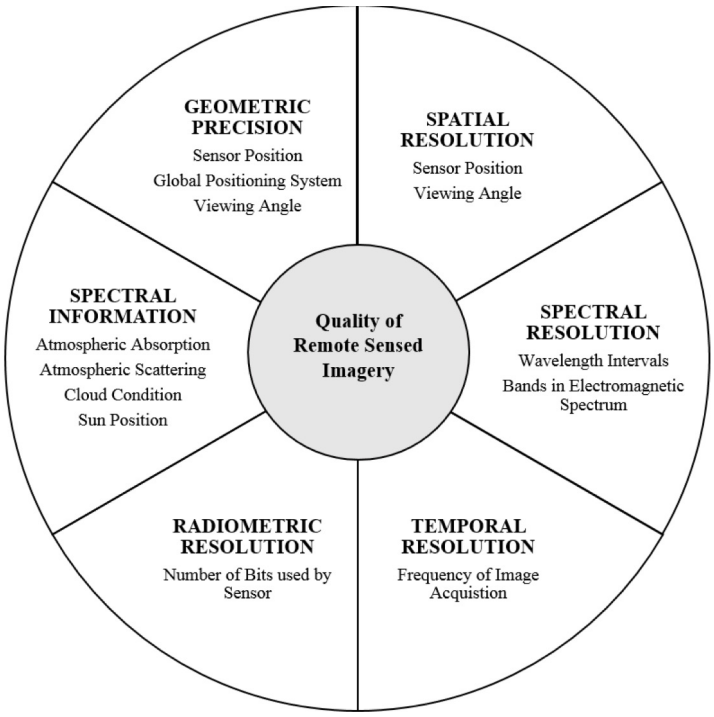
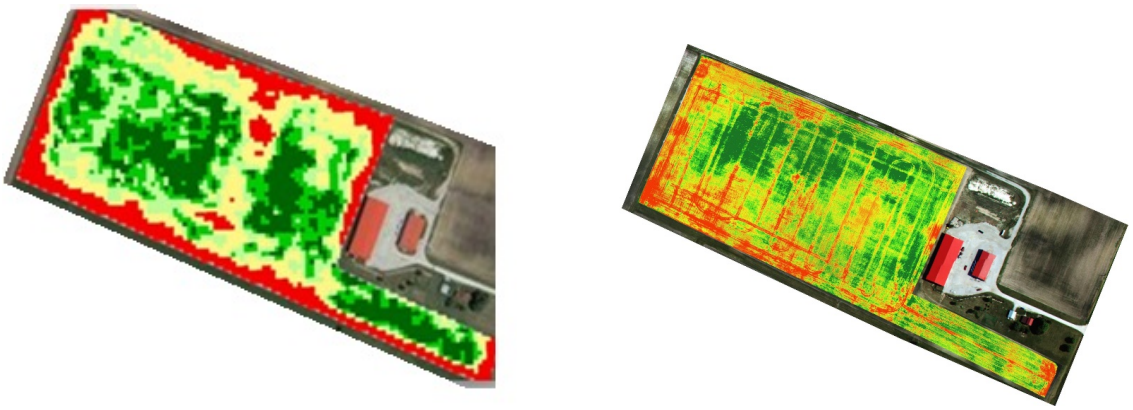


Figure 4 – Quality factors in remote sensing, from (KHANAL, 2017).



(a) Lower Spatial Resolution. (b) Higher Spatial Resolution.

Figure 5 – Comparison of vigor map in (a) has a ground sampled distance of 10 meters and (b) has a ground sampled distance of 0.25 meters, from (KHANAL, 2017)

health. Figure 6 shows an example of the difference in spectral resolution between two satellites, channel 3 of TERRA Aster corresponds to a narrower range that is centered on a different value than the near-infrared of the LANDSAT TM channel 4. For channel 4 of TERRA Aster and 5 of LANDSAT TM, the bandwidth is also narrower. Between 2.05 and 2.35 µm, the sensors have different designs: Terra Aster achieves four channels (5 to 8) and LANDSAT TM integrates the wavelengths in one single spectral measure (channel 7) (OSE; CORPETTI; DEMAGISTRI, 2016).

Temporal resolution represents the frequency at which a sensor collects data over the same area and is crucial for monitoring crop growth stages, phenological events, and the timing of agricultural inputs. A higher temporal resolution is essential to effectively

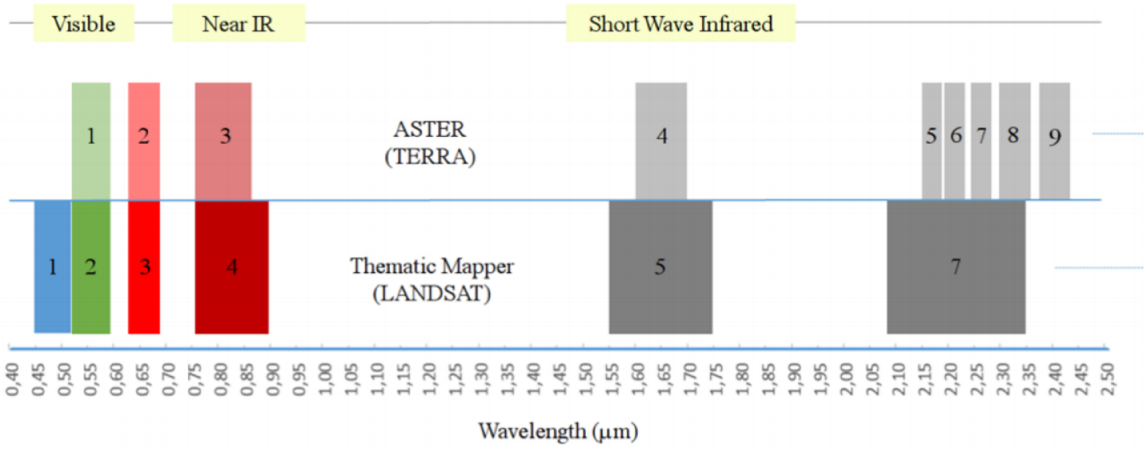


Figure 6 – LANDSAT TM and TERRA Aster firsts bands and their spectral resolution, from (OSE; CORPETTI; DEMAGISTRI, 2016)

capture the dynamic nature of agricultural systems, enabling timely decision-making. Various applications in PA utilize UAVs due to their higher temporal resolution. For example, in soil monitoring, it is common to conduct multiple flights before and after the application of inputs (TOSCANO et al., 2024).

Radiometric resolution defines the sensitivity of a sensor to detect slight differences in reflected or emitted energy, enabling the discrimination of subtle variations in crop vigor and soil moisture. It is essential for precision agriculture applications that require monitoring slight changes in plant health or water stress levels.

Comparing drone and satellite imagery, drones typically provide higher spatial resolution due to their ability to fly at lower altitudes, resulting in more detailed observations and the ability to monitor small-scale variability within fields. However, satellites offer greater temporal resolution with the capability to capture images of large areas regularly, which can be more cost-effective for monitoring expansive agricultural lands. However, satellites often fall short in spectral and radiometric resolution compared to drones, which can carry customized sensors to target specific spectral bands and detect fine-scale differences in reflectance related to plant health and stress.

2.1.3 Imaging Sensors

Radiation that reaches the surface of plants is absorbed, reflected, or transmitted. This reflected radiation is expressed as the electromagnetic spectrum with a range of bandwidths. RGB images contain three bands of the visible spectrum (Red, Green, and Blue), and multispectral images can contain several bands of the visible and Near-infrared (NIR) spectrum. In contrast, hyperspectral images contain dozens or even hundreds of continuous bands in the visible and NIR (XIE; YANG, 2020). Figure 7 shows a diagram differentiating the bands of an RGB, multispectral, and hyperspectral image.

RGB sensors, popularly known as digital cameras, are usually cheaper, with a high

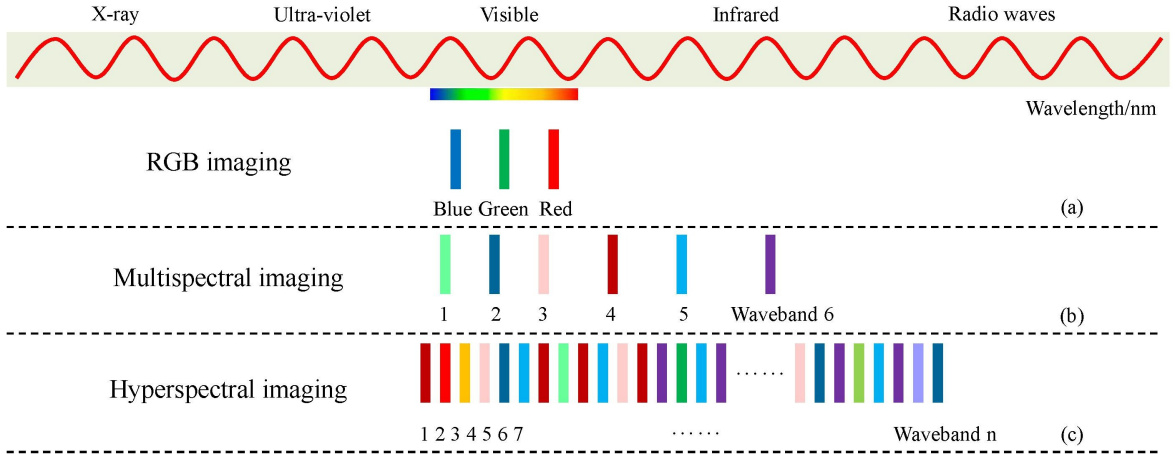


Figure 7 – Example of RGB, multispectral and hyperspectral bands from (XIE; YANG, 2020)

spacial resolution but low spectral resolution. These sensors cover three spectrum regions (Red, Green, and Blue) and are used to calculate Vegetation Indices, high-resolution digital elevation models (DEMs), and vegetation height maps. Figure 8 shows an example of a Visible Light Sensor.



Figure 8 – MAPIR Survey2 Camera - Visible Light RGB by MAPIR.

Some RGB sensors are modified by exchanging the red filter for a near-infrared one, making these modified cameras sensitive to the near-infrared spectrum (NIR) (MAES; STEPPE, 2019). The wavelength of the electromagnetic spectrum between 720 and 1,000nm is fundamental to precision agriculture, since various vegetation indices use this range to calculate the state of vegetation (MAES; STEPPE, 2019). Figure 9 shows an example of an RGN filter.

Compared to RGB cameras, multispectral cameras can provide bandwidths (both in the visible and in the NIR spectrum) that are more useful for precision agriculture. They usually capture up to 10 bands between 400 and 1000 nanometers in the spectrum range.



Figure 9 – MAPIR Survey3W Camera - red+green+nir (RGN, NDVI) by MAPIR.

Most multispectral cameras are a combination of sensors responsible for obtaining red, green, blue, and other spectra such as NIR and red-edge (rededge) (XIE; YANG, 2020). Figure 10 shows an example of a multispectral camera.

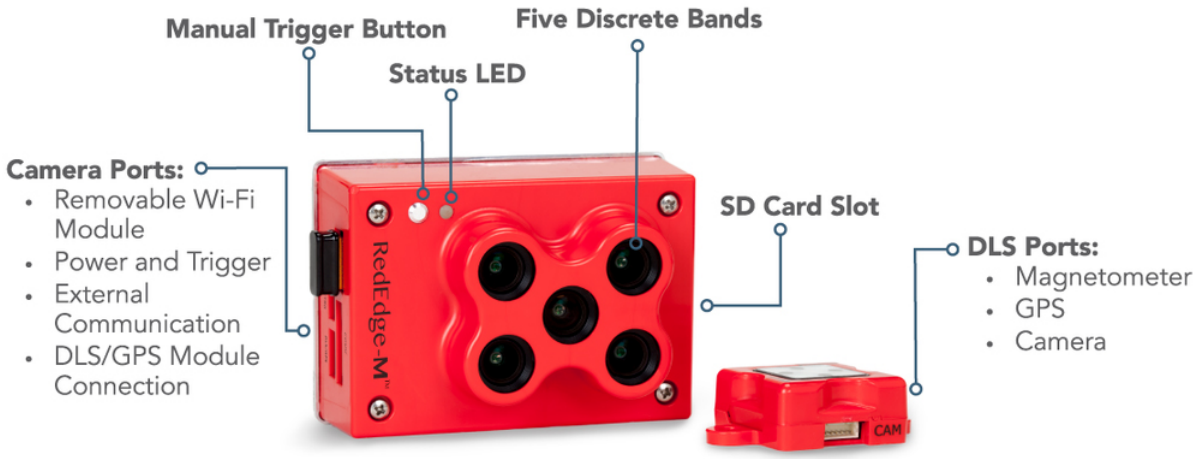


Figure 10 – MicaSense Red-Edge camera by MicaSense.

Hyperspectral sensors combine thousands of narrow wavelength intervals to provide detailed plant information. Compared to multispectral cameras, hyperspectral cameras provide a real spectrum in every image pixel, allowing the calculation of more advanced vegetation indices and the estimation of unperceived crop phenotypes (XIE; YANG, 2020).

The radiation emitted by objects in the long-wave infrared range (7.5 - 13.5μm) is proportional to their surface temperature. This property is used in thermal infrared (TIR) cameras to check for crop problems, such as water stress or detection of soil salinity (XIE; YANG, 2020). Figure 11 shows an example of a thermal camera.

2.1.4 Vegetation Indices

A VI combines one or more spectral bands to highlight some desired characteristic. In precision agriculture, VIs provides various information about plantations such as plant growth, health, soil conditions, and environmental factors (RADOČAJ et al., 2023).



Figure 11 – FLIR Vue Pro thermal camera by Teledyne Flir.

Given its importance, there are more than 500 indexes in the index database (HENRICH et al., 2017). In multispectral images, we have Normalized Difference Vegetation Index (NDVI) as the most popular index among papers published since 2000 (RADOČAJ et al., 2023). It is calculated based on red reflectance and NIR and provides a measure for vegetation's greenness or photosynthesis activity. Its main applications are crop yield estimation (ROZNIK; BOYD; PORTH, 2022), specific nitrogen application (SOZZI et al., 2021), and crop growth monitoring (SINGH; KOMAL; VICTOROVICH, 2020). Another important vegetation index obtained through multispectral images is Enhanced Vegetation Index (EVI). EVI considers the linear relationship between reflectance and vegetation cover (in addition to including blue reflectance in the calculation). These changes make EVI more robust to analyze the vigor and health of vegetation, especially in locations with high atmospheric interference or with a lot of soil in the background (RADOČAJ et al., 2023). Table 3 presents some of the most widely used indices according to (RADOČAJ et al., 2023).

When working with RGB images, the primary index derived from the NDVI is the Normalized green-red difference index (NGRDI), which has been predominantly utilized with RGB sensors over the past decade (RADOČAJ et al., 2023). NGRDI provides a low-cost approach to detect chlorophyll changes in plants, changing the NIR reflectance of NDVI by the green reflectance. Other RGB indices worth highlighting are Excess green index (ExG) and Excess red index (ExR), where ExG provides information about the health of the plant while ExR provides information on biomass. Table 4 presents some of the most used indexes according to (RADOČAJ et al., 2023).

2.2 Digital Image Processing

An image can be defined as a two-dimensional function $f(x, y)$, where x and y are coordinates of a plane, and the amplitude of f in any pair of coordinates is called the intensity or grayscale of the image at that point. We call this image digital if this intensity value is finite and discrete. In this sense, the research field of digital image processing

Table 3 – Main multispectral vegetation indices. B: blue reflectance, G: green reflectance, R: red reflectance, RE: red-edge reflectance, NIR: near-infrared reflectance. Adapted from (RADOČAJ et al., 2023).

Vegetation Index	**Abbreviation**	**Formula**
Normalized difference vegetation index	NDVI	$\frac{\text{NIR}-\text{R}}{\text{NIR}+\text{R}}$
Enhanced vegetation index	EVI	$2.5 \times \frac{\text{NIR}-\text{R}}{\text{NIR}+6\times\text{R}-7.5\times\text{B}+1}$
Green-normalized difference vegetation index	GNDVI	$\frac{\text{NIR}-\text{G}}{\text{NIR}+\text{G}}$
Soil-adjusted vegetation index	SAVI	$1.5 \times \frac{\text{NIR}-\text{R}}{\text{NIR}+\text{R}+0.5}$
Simple ratio	SR	$\frac{\text{NIR}}{\text{R}}$
Normalized difference red-edge index	NDRE	$\frac{\text{NIR}-\text{RE}}{\text{NIR}+\text{RE}}$
Optimized soil-adjusted vegetation index	OSAVI	$1.16 \times \frac{\text{NIR}-\text{R}}{\text{NIR}+\text{R}+0.16}$

Table 4 – Main RGB vegetation indices. B: blue reflectance, G: green reflectance, R: red reflectance. Adapted from (RADOČAJ et al., 2023).

Vegetation Index	**Abbreviation**	**Formula**
Normalized green-red difference index	NGRDI	$\frac{G-R}{G+R}$
Excess green index	ExG	$2 \times G - R - B$
Excess red index	ExR	$1.4 \times R - G$
Visible atmospherically resistant index	VARI	$\frac{G-R}{G+R-B}$
Modified green-red vegetation index	MGRVI	$\frac{G^2-R^2}{G^2+R^2}$
Normalized pigment chlorophyll ratio index	NPCI	$\frac{R-G}{R+G}$
Triangular greenness index	TGI	$G - 0.39 \times R - 0.61 \times B$
Excess blue index	ExB	$1.4 \times B - G$

refers to the processing of these digital images for better transmission, understanding, and representation (WOODS; GONZALEZ, 2008). Figure 12 shows an example of a grayscale image that contains the intensity of its pixels.

In a digital color image, each pixel is associated with a vector value $p(x, y) = (i_1, i_2, ..., i_n)$ where i_n represents the intensity value of that color/dimension while n represents the number of dimensions. Figure 13 shows an example of a color image and its dimensions.

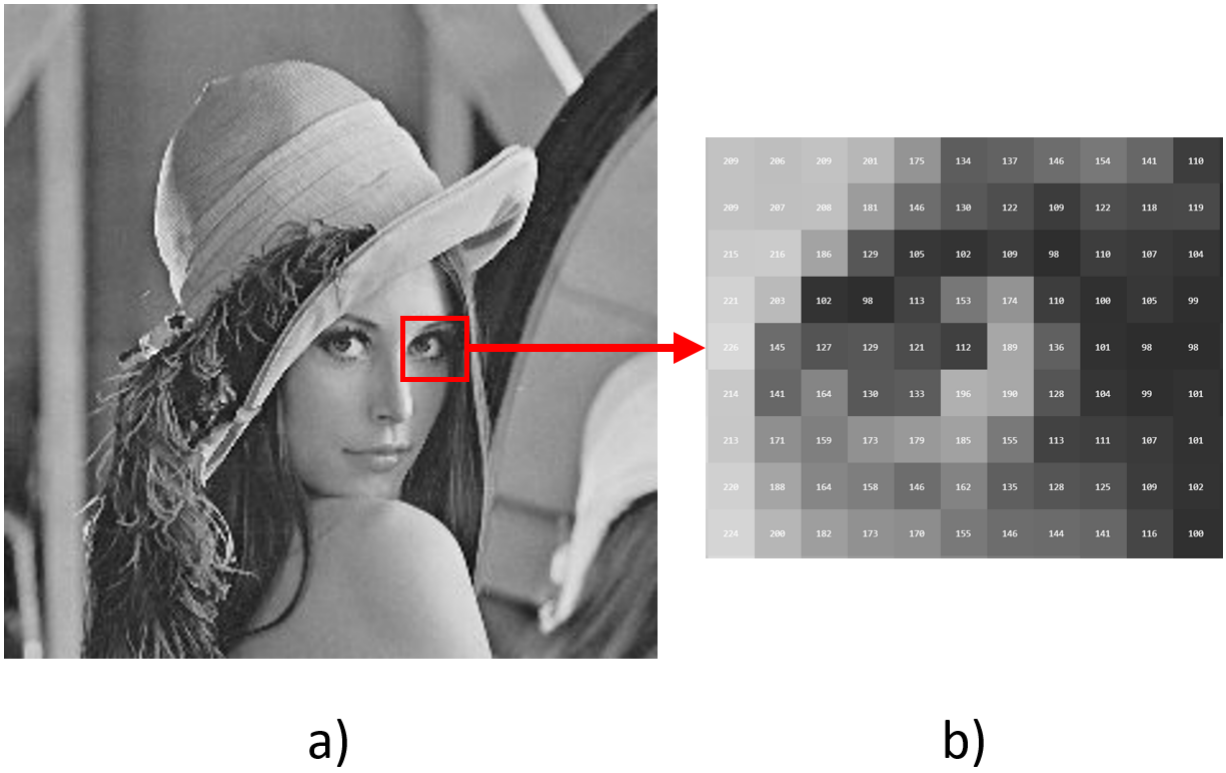


Figure 12 – Example of grayscale image. a) original image b) crop showing the intensity of the values in the region

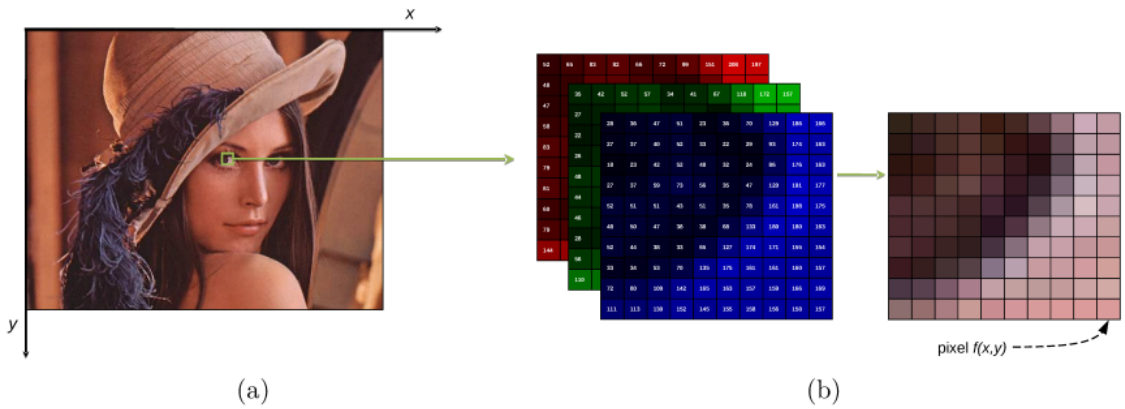


Figure 13 – Example of a color image. a) original image b) crop 10x10 using the RGB color space, from (SANTOS et al., 2022)

2.2.1 Color Spaces

A color space (also called a color system or color model) can be defined as the specification of a coordinate system and a subspace where each point represents a single color. Its primary purpose is to facilitate the standardization of colors in a way accepted by everyone for a given hardware or application (WOODS; GONZALEZ, 2008); for example, RGB is widely used on monitors and digital cameras, while CMY/CMYK is used in color printing, and the HSI best matches the way we describe colors. The primary color spaces

used in the segmentation of weeds are:

- The RGB model, which is based on the Cartesian coordinate system. It defines each color based on the three primary spectral components: red, green, and blue. Normally, RGB images are created using a 24-bit system, with 8 bits for each channel (colors) of the image (WOODS; GONZALEZ, 2008). This model is widely used in segmentation mainly due to the VI calculated based on the red, green and blue components (RADOČAJ et al., 2023). Figure 14 shows the cube formed by the RGB color space.

- The HSV (Hue, Saturation, and value) and HSI (Hue, Saturation, and Intensity) model, despite varying in their transformation calculation starting from an RGB image, aims to approximate the way we describe colors (in terms of Hue, saturation, and brightness). In these models, color is decomposed from intensity/value, making these algorithms ideal for image processing based on color description (WOODS; GONZALEZ, 2008) (MOREIRA et al., 2022).

- The CIE L*a*b* model, also called CIELAB, is a three-dimensional color space where the L* axis is the grayscale, ranging from 0 (black) to 100 (white), a* representing a difference between green and red and b* representing the difference between yellow and blue. Like the HSV models, HSI and CIE L*a*b* also have the advantage of being more robust about the presence of shadows when segmenting plants (WOODS; GONZALEZ, 2008) (KIM et al., 2023).

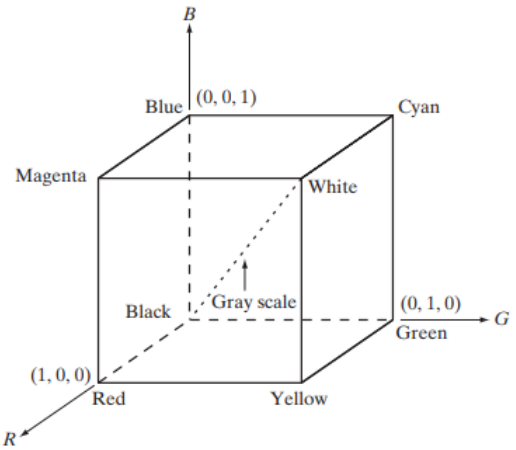


Figure 14 – Schematic of the RGB color cube. Points along the main diagonal have gray values, from black at the origin to white at point (1, 1, 1), from (WOODS; GONZALEZ, 2008)

2.2.2 Image Segmentation

Image segmentation can be defined as the technique that divides an image between regions of interest, each representing some information/object useful for processing (WOODS; GONZALEZ, 2008). Each pixel in a given region is similar to others in the same region based on some computed characteristic or property, such as color, intensity, or texture (JAYANTHI; SHASHIKUMAR, 2017). This segmentation can be done based on the sudden change in appearance in the image (i.e., its discontinuities), a method called edge detection. There are several edge detection methods, but the most famous are the Roberts, Prewitt, and Sobel (WOODS; GONZALEZ, 2008) operators. Another widespread type of segmentation is Thresholding, which is based on the possibility of the existence of one (or more values) that separate the desired object from image background based on intensity. The Otsu (OTSU, 1979) method is one of the most traditional global thresholding methods. It assumes the image contains two classes of pixels following a bimodal distribution in its histogram. With this, it seeks to minimize the intra-class variance (or maximize the between-class variance), thus obtaining the best threshold that separates the classes in the image. Another class of segmentation methods is region growing, a procedure that groups pixels into larger regions given a particular growth criterion. The essential operation consists of adding a set of pixels to be the seed of the algorithm. For each adjacent pixel, we add or not the specific region, usually using characteristics such as intensity and color (WOODS; GONZALEZ, 2008).

Segmentation methods have evolved unprecedentedly in recent years, mainly with Convolutional Neural Networks (CNNs) and deep learning. U-Net (RONNEBERGER; FISCHER; BROX, 2015) is one of the most influential architectures for image segmentation using CNN. Initially designed to segment medical images, its use has been expanded to various fields of research, including precision agriculture (BRILHADOR et al., 2019) (RAMIREZ et al., 2020). Other models, such as Fully Convolutional Network (FCN), have expanded segmentation methods, making them more prepared for a greater number of practical applications (LONG; SHELHAMER; DARRELL, 2015).

2.3 Convolutional Neural Networks

CNNs constitute a specialized category within the broad spectrum of deep neural networks and are highly effective in processing data with a grid-like topology, such as images. Unlike other models, CNNs implement convolutional layers where variable filters are applied to the input data. These filters enable CNNs to learn low-level features, such as edges and textures in the initial layers, and complex features, such as objects in deeper layers. This learning capacity of CNNs allowed its use in various applications such as image recognition, segmentation, and classification (LECUN et al., 1998). Typically, a CNN comprises several convolutional layers interspersed with activation functions and

polling layers to obtain a feature map. This feature map is later sent to a fully connected layer to be classified into the desired categories. This arrangement makes networks more generic to previously unseen data and more efficient due to the consumption of fewer computational resources. Figure 15 shows a generic structure of the convolutional neural network.

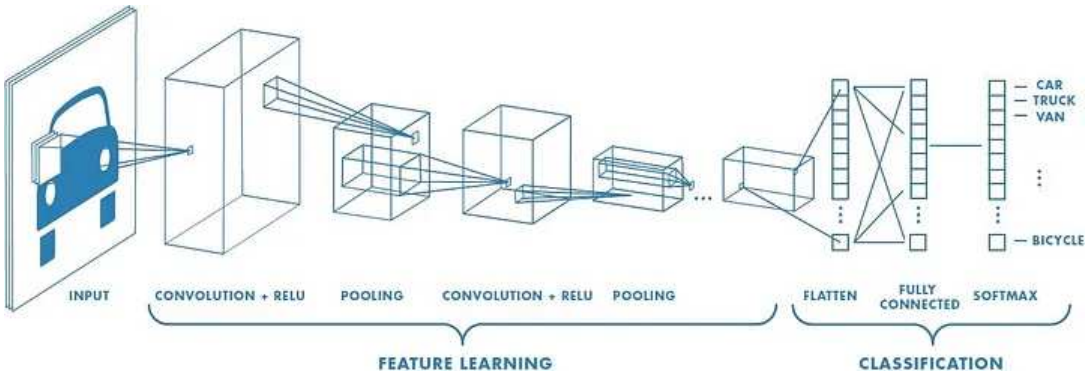


Figure 15 – Generic structure of a CNN, from (SAHA, 2018)

The convolution layer is normally composed of small filters (kernels) that traverse the input by applying the convolution operator. For example, given a two-dimensional image I as input and a two-dimensional kernel K, the convolution operation can be defined as (GOODFELLOW; BENGIO; COURVILLE, 2016):

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(m,n) \cdot K(i-m, j-n). \tag{1}$$

The output of equation 1 is generally known as a feature map. It is worth noting that several implementations of CNNs typically use the cross-correlation operator where the kernel is not inverted:

$$S(i,j) = (K * I)(i,j) = \sum_m \sum_n I(i+m, j+n) \cdot K(m,n). \tag{2}$$

Figure 16 shows an example of a convolution operation on two-dimensional input. Note that the value 18 is obtained by calculating $S(0,0) = I(0,0) * K(0,0) + I(0,1) * K(0,1) + I(1,0) * K(1,0) + I(1,1) * K(1,1) = 9*1 + 1*0 + 4*0 + 9*1 = 18$

One of the problems with convolution, as shown in Figure 16, is that convolution tends to lose edge pixels after its application, which in the image example reduces from a 4×4 input to a 3×3 feature map. This, after multiple layers of convolution, can lead to the disappearance of important information from the edge of the input. To solve this problem, we can apply padding, which consists of adding rows and columns to the edges of the input (generally with the value zero) before the convolution process to preserve the original size of the input. In other cases, we may want to reduce the resulting feature map; in these cases, we apply a Stride that will reduce the convolution feature map to a desired size.

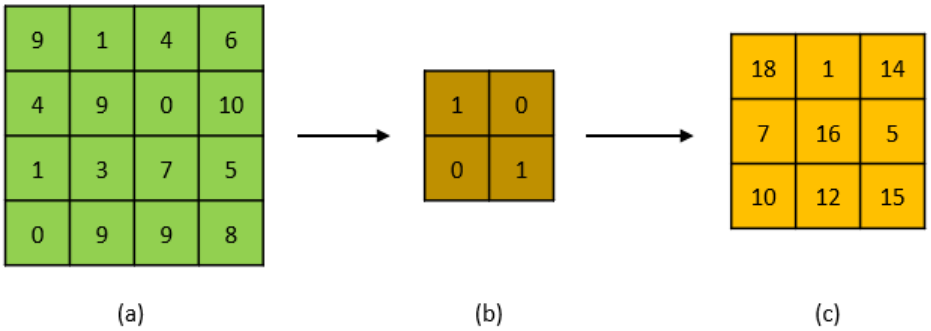


Figure 16 – Example of a convolution operation on two-dimensional input. a) represents the two-dimensional input (I). b) the kernel used (K). c) the result of the operation (S).

After the convolution of each kernel on the image, activation functions are applied to add non-linearity to the model. A model without activation functions becomes linear, which does not allow the learning of complex data relationships. The main activation functions are (LECUN; BENGIO; HINTON, 2015):

- Rectified Linear Unit (ReLU): the ReLU activation function is defined as $f(x) = \max(0, x)$. It is one of the most used in deep learning models when compared to other functions given its computational efficiency.

- Hyperbolic Tangent (tanh): tanh, in turn, is a sigmoidal function that varies in the range from -1 to 1. It is defined by the equation $f(x) = \tanh(x) = \frac{2}{1+e^{-2x}} - 1$. One of its main advantages is that it provides a normalized output, which tends to be useful in certain problems.

- Sigmoid: the sigmoid function is defined by equation $f(x) = \frac{1}{1+e^{-x}}$ and varies in the range of 0 to 1. It is beneficial when the function's output needs to be 0 or 1 and is widely used in binary classification problems.

Figure 17 presents an example of the application of the activation function.

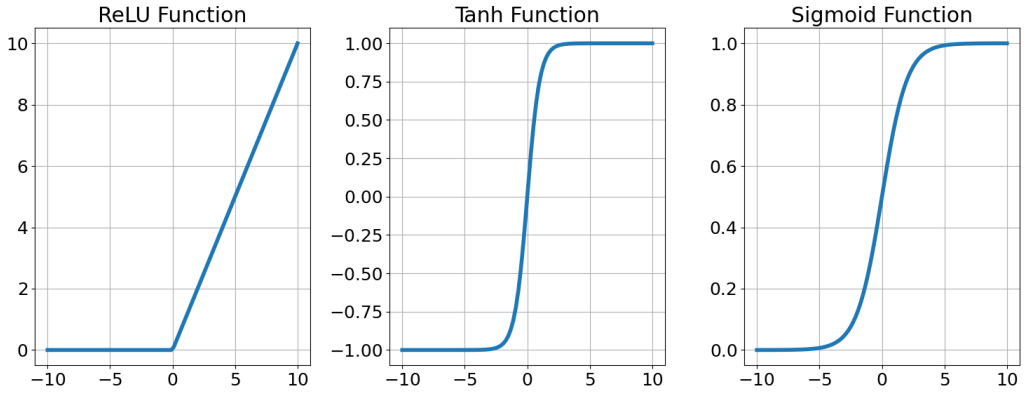


Figure 17 – Activation functions.

After convolution, pooling is applied to transform the feature map into something new, discarding irrelevant information while preserving essential details. Furthermore, pooling helps make the model invariant to changes in object positions and light conditions and provides a compact representation of features. The most common ways to perform pooling is by applying an average or maximum filter to the feature map. Figure 18 shows an example of max pooling and average pooling of size 2.

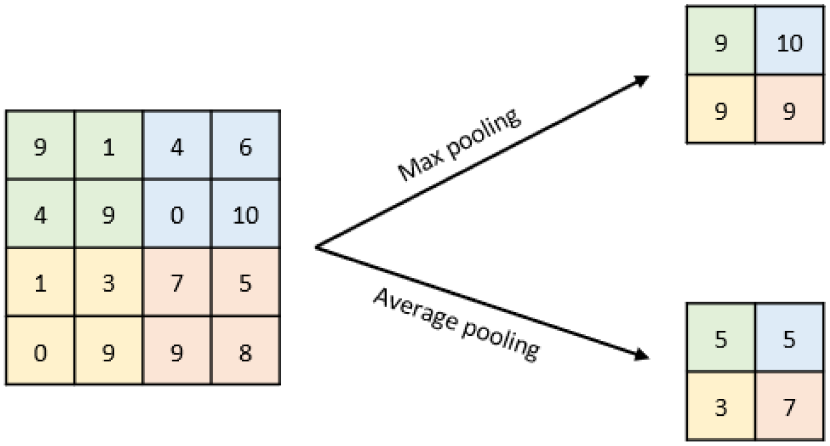


Figure 18 – Application of a pooling operation (max and average) of size 2. Decimal values were truncated.

Finally, after several convolution layers (containing the convolution itself, activation, and pooling functions), the result is passed to an Fully Connected Layer (FCL). In this layer, all neurons are connected to the previous layers. Their objective is to synthesize all the features learned by the network to perform a classification (or segmentation) (LECUN; BENGIO; HINTON, 2015).

A widely used model in the semantic segmentation literature is DeepLabV3 (CHEN, 2017; XU et al., 2023a; CELIKKAN et al., 2023). It was developed to address challenges in accurately segmenting images, particularly for objects with varying scales and complex boundaries. At its core, DeepLabV3 employs Atrous Spatial Pyramid Pooling (ASPP), a mechanism that leverages atrous (dilated) convolutions with different dilation rates to effectively capture multi-scale contextual information. This approach is highly efficient because it allows the model to expand the receptive field of convolutions without increasing the number of parameters or reducing the spatial resolution. Consequently, the model becomes more robust in handling variations in object sizes within images. In addition, the model introduces a global average pooling mechanism that aggregates global contextual information, improving its understanding of the overall image structure. Furthermore, DeepLabV3 utilizes atrous convolutions throughout its backbone architecture, replacing traditional downsampling techniques, which helps maintain higher resolution feature maps.

Building on DeepLabV3, the DeepLabV3+ model (CHEN et al., 2018) extends its capabilities by introducing a decoder module that enhances object boundary delineation

and improves spatial detail reconstruction. One of the main limitations of DeepLabV3 is its simplicity in reconstructing high-resolution outputs. DeepLabV3+ addresses this limitation by incorporating a lightweight decoder that combines high-level semantic features from the encoder with low-level spatial features extracted from earlier network layers. This fusion of features enables DeepLabV3+ to achieve more precise boundary segmentation and enhanced visual fidelity in its final predictions. In addition to the decoder, DeepLabV3+ also incorporates atrous separable convolutions to improve computational efficiency while maintaining accuracy. Its architecture exemplifies a balance between capturing global context through ASPP and preserving fine-grained spatial information through the decoder, making it a robust solution for complex segmentation challenges. Figure 19 illustrates the architecture of the DeepLabV3+ model.

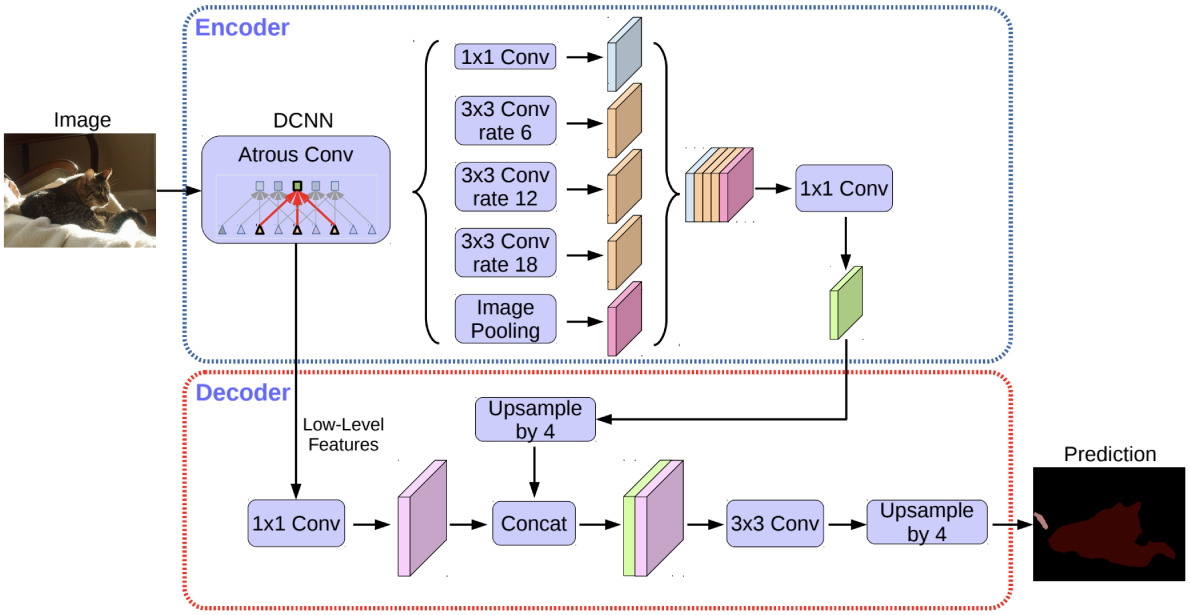


Figure 19 – The architecture of the DeepLabV3+ model. Adapted from (CHEN et al., 2018).

2.4 Vision Transformers

Transformers were originally introduced by (VASWANI, 2017), revolutionizing Natural Language Processing (NLP) by replacing recurrent and convolutional operations with a purely attention-based framework. This architecture is based on the use of parallelizable self-attention mechanisms to capture global dependencies among tokens in a sequence. This approach facilitates and makes the process of learning complex contexts more efficient. Applied to computer vision, researchers have shown that the same attention-centric principle benefits the image analysis process, especially in tasks that require global context understanding. However, the direct application of Transformers to images demanded

novel strategies to represent pixels as input tokens while preserving spatial information (ALEXEY, 2020).

One of the main adaptations of Transformers for images is the transformation of an image into tokens. To achieve this, each image is divided into smaller patches of fixed size, and each patch is treated as a word (token) for the model. For example, a 224×224 image can be divided into 16×16 patches, generating a total of 196 tokens. Each patch is flattened into a 1D vector and passed through a learnable linear projection that transforms it into a high-dimensional embedding. After this process, the Transformer can process images using the same attention mechanisms originally designed for language. This mechanism allows the model to learn the relationships between distant patches and to understand the image. The choice of patch size introduces some trade-offs: smaller patches capture more details but produce a longer token sequence, thereby increasing computational costs; conversely, larger patches reduce the token size but may lose spatial information. The patch-based tokenizer must be carefully tuned to balance representational power against efficiency, particularly for high-resolution images (TOUVRON et al., 2021).

In their original conception, transformers do not store the order of the input sequences. This property is advantageous for recognizing arbitrary relationships; however, when we consider images, the model lacks an understanding of the pixel arrangement. To address this, position encodings are introduced to preserve spatial structure within the patch sequence. For instance, if a patch is located in the upper-left corner of the image, its positional embedding will differ from that of a patch in the lower-right. In the work proposed by (ALEXEY, 2020), a standard learnable 1D positional embedding is added, and the authors explain that there was no significant performance increase when using more advanced 2D-aware position embeddings.

In Transformer-based vision models, self-attention allows each patch (token) to selectively focus on other patches in the sequence. For this purpose, each token is projected onto three different vectors. Query (Q), Key (K) and Value (V). Typically, these are computed as follows:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{X}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{X}\mathbf{W}_V, \tag{3}$$

where $\mathbf{X}$ is the input and $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ are learnable weight matrices. The attention is then computed by scaling the dot product of $\mathbf{Q}$ and $\mathbf{K}$, and applying a softmax function to derive weights for each $\mathbf{V}$:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}, \tag{4}$$

where d_k is the dimension of k vector.

The multi-head self-attention mechanism within each Transformer block allows the model to learn different levels of semantic and spatial relationships among patches. The

model splits Query, Key, and Value across multiple heads. Each head is responsible for capturing a different relationship in the image; for example, one head might focus on local image features (e.g., edges, textures), while another might capture long-range dependencies (e.g., object composition, spatial relationships):

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}\Big(\text{head}_1, \text{head}_2, \ldots, \text{head}_h\Big)\, \mathbf{W}_O, \tag{5}$$

where each head_i is a self-attention operation with its own projection matrices, and $\mathbf{W}_O$ is another learnable weight matrix to combine the heads' outputs.

The original Vision Transformer (ViT) introduced by (ALEXEY, 2020) was the first successful attempt to directly apply a Transformer architecture to images. Its main contribution was to treat an image as a sequence of fixed-size patches that are subsequently projected into one-dimensional vectors. These patch embeddings are then fed into a standard transformer encoder. ViT dispenses with the use of convolutional operations, relying solely on self-attention and feed-forward layers. One of its main advantages is the ability to capture long-range dependencies in the data. However, this model requires large amounts of data (e.g. ImageNet-21k or JFT-300M) to outperform well-optimized CNN models in image classification tasks. Figure 20 shows ViT architecture.

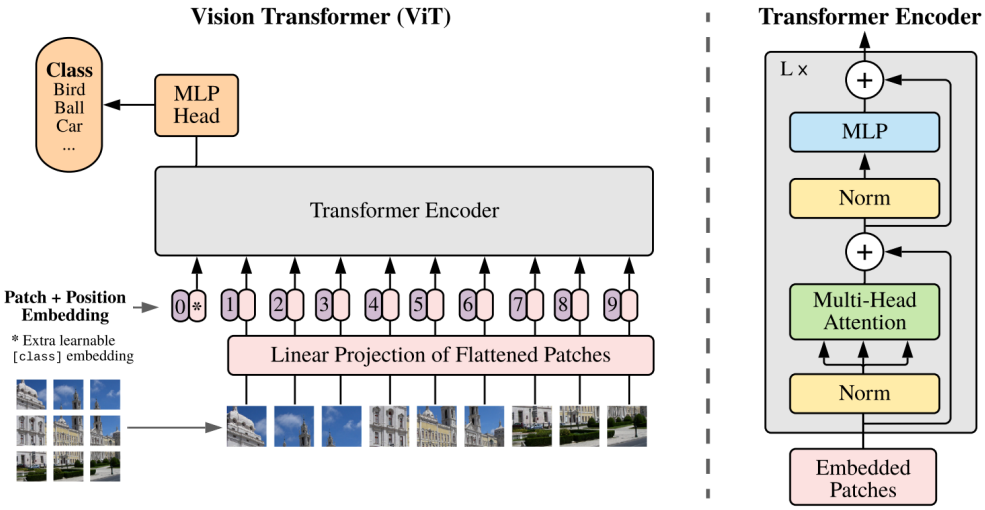


Figure 20 – Vision Transformer Architecture. From (ALEXEY, 2020)

Building on ViT, (TOUVRON et al., 2021) proposed Data-Efficient Image Transformers (DeiT) with the primary objective of improving the training efficiency of Transformers on smaller datasets. DeiT introduced various training methods to achieve competitive performance on ImageNet-1k without relying on massive external datasets. This model used knowledge distillation from a CNN-based teacher and carefully tuned augmentation. Furthermore, DeiT uses the same patch-based tokenization approach as ViT, but it is trained with a specialized distillation token that helps the model absorb knowledge from the teacher network.

Another highly popular model is the Swin Transformer proposed in (LIU et al., 2021), which aimed to reduce the quadratic complexity (with respect to the number of patches) inherent in vanilla ViT. This model introduces a hierarchical structure in which images are processed in stages with progressively smaller feature map resolutions. In each stage, the model applies the self-attention mechanism to localized windows, significantly reducing the computation by focusing attention on spatially neighboring patches. With this approach, Swin Transformers achieve linear complexity with respect to image size, enabling more efficient processing of higher-resolution images.

In this work, we utilized SegFormer, a transformer-based model designed for semantic segmentation, focusing on simplicity, efficiency, and scalability (XIE et al., 2021). Its main distinction from other models lies in the application of a lightweight all-MLP decoder instead of complex decoders or multi-stage pipelines, making it computationally efficient while maintaining high performance. Unlike the ViT, which extracts features using a single resolution, SegFormer generates hierarchical feature maps, enabling better spatial granularity for dense prediction tasks such as segmentation. Another significant modification lies in the self-attention layer, which was replaced by an Efficient Self-Attention mechanism that reduces the sequence (token) size by a ratio R. This modification allows SegFormer to operate on higher resolutions. Figure 21 illustrates the architecture of SegFormer.

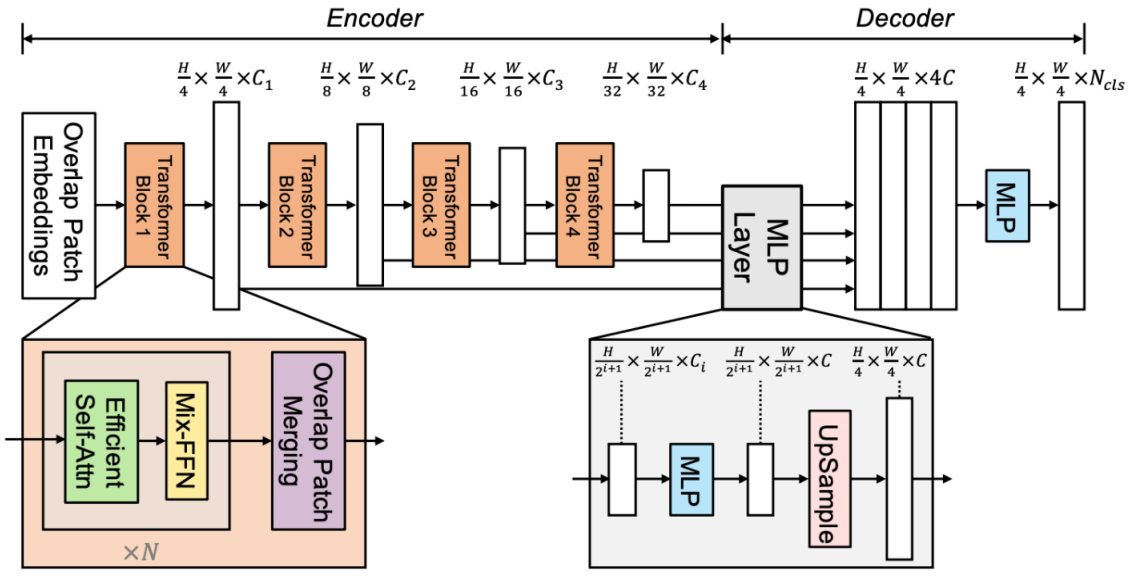


Figure 21 – Segformer Architecture. From (XIE et al., 2021)

2.5 Data Fusion in Deep Learning

Fusion can be defined as the process of integrating different data sources to enhance the performance, robustness, and interpretability of machine learning systems. This technique

is widely applied across various domains, such as medical imaging, precision agriculture, and autonomous systems, where various types of data need to be processed to improve decision-making (ZHOU; RUAN; CANU, 2019; CUNHA et al., 2023; ZHANG et al., 2015). Specifically, in precision agriculture, fusion techniques can be employed for the detection of weeds in multispectral images, vegetation indices, and data from multiple sensors, among others (CUNHA et al., 2023). Fusion can be categorized into three main types: early fusion, intermediate fusion, and late fusion. Figure 22 illustrates how these fusion techniques are commonly applied in deep learning models.

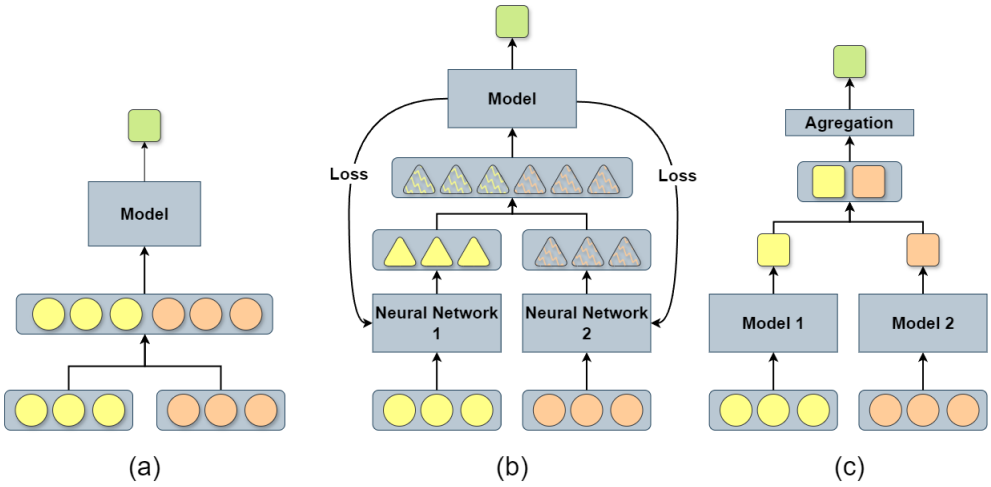


Figure 22 – Fusion methods: (a) Early fusion, (b) Joint Fusion (intermediate fusion), and (c) Late Fusion. Circles represent raw features, triangles represent neural network features, and squares represent model outputs (logits/predictions). Adapted from (HUANG et al., 2020).

Early Fusion combines raw data or low-level features, such as pixel values or spectral bands, before processing them through a model. This is commonly used in CNNs, where raw data from different modalities are concatenated or embedded into a unified input representation. This approach offers advantages such as ease of implementation and the ability to extract relationships between complementary data sources. However, it often requires significant computational resources and can be sensitive to variations in input data formats, making it less suitable for heterogeneous or noisy datasets.

Joint or Intermediate Fusion takes a different approach by integrating features extracted from individual modalities at a mid-level representation. In this method, one or more models produce intermediate outputs that are jointly processed, leveraging Multilayer Perceptrons (MLPs), attention mechanisms, or specialized cross-modal architectures. This joint representation is then further processed to obtain the final result. Intermediate fusion balances computational efficiency with the ability to exploit relationships between modalities, making it suitable for scenarios with moderately diverse data sources.

Late Fusion processes the outputs or predictions of different deep learning models. This fusion method often involves decision rules (e.g., weighted averaging or majority

voting) or trainable methods such as MLPs or attention layers to aggregate the outputs. This type of fusion typically achieves the best results in terms of performance and modularity. Each model can be individually trained to learn features specific to one domain, allowing new data (or domains) to be seamlessly added to the decision-making process.

In conclusion, the fusion process can utilize various mechanisms, such as convolutions, MLPs, and attention-based layers, depending on the stage of fusion and the nature of the data. In the context of weed detection, these mechanisms enable the integration of diverse data sources, such as multispectral images, spatial patterns, and texture features, which are crucial to distinguish weed species from crops. While early and intermediate fusion offer deeper integration and richer feature representations by directly combining low-level or mid-level features extracted from multispectral bands or RGB images, late fusion stands out for its flexibility and robustness in agricultural applications. For example, in weed segmentation tasks, late fusion can combine outputs from models specialized in different data types—such as a model trained on spectral indices and another on spatial features—into a unified decision. This modularity is especially advantageous in precision agriculture, where data quality and availability can vary significantly across environments and crop types. Furthermore, late fusion facilitates the seamless incorporation of additional data modalities, such as soil moisture maps or weather data, to enhance model accuracy and adaptability, making it a preferred approach for managing the diverse and evolving datasets common in weed detection.

Chapter 3

Related works

This chapter details the works that served as the theoretical foundation and comparison basis for this thesis, representing the state-of-the-art in weed segmentation in images captured by UAV. We conducted an exploratory literature search using Google Scholar, applying a publication date filter to consider works published after 2020. The search was performed using the following keyword sets: ["transformers", "uav", "weed", and "segmentation"], ["deep learning", "uav", "weed", "segmentation"], and ["weed", "segmentation"]. Furthermore, we searched for all studies that cited the datasets used in this work (MONTEIRO; WANGENHEIM, 2019) and (SA et al., 2018).

In (SA et al., 2018), the authors proposed a dense semantic segmentation framework to segment weeds in images obtained by UAV. This framework consists of a modified version of SegNet, where each class's frequency of appearance (FoA) is adapted based on our training dataset for better class balancing. This frequency is used in the network loss function to penalize errors in classes with fewer representations. Typically, the proportion of pixels that belong to weeds in plantations is much smaller compared to the crops and the background. Another modification involves adapting the input and output layers to use images with more channels. The authors also presented a new dataset to the community to validate their work. This data set consists of flights carried out in sugar beet plantations in two different locations. The images are multispectral and were obtained with different sensors. A total of 20 tests were carried out, changing the channels sent to the network and, in some cases, the batch size. The best result was obtained in the combination of channels B, CIR, G, NDVI, NIR, R, and RE for the Micasense dataset with an AUC of 0.839, 0.863 and 0.782 for background, crop, and weed, respectively. In the second dataset, the best input combination was G, NDVI, NIR, R, and RE, with an AUC of 0.951, 0.957, and 0.621 for background, crop and weed, respectively. It is essential to note that this work is the basis for several others due to the quality and applicability of the available dataset. The authors also discuss that NDVI significantly increases the performance of discrimination between crops and weeds. However, they warn that simply increasing the amount of input can introduce more ambiguous information into the network, reducing

its accuracy.

The work proposed by (CASTELLANO; MARINIS; VESSIO, 2023b) presents an approach to weed mapping using the lightweight Transformer architecture to process high-resolution aerial images obtained from drones and performs semantic segmentation to distinguish between crops and weeds. The model used as a basis for the work was the one proposed by (YAN; ZHANG; WU, 2022), who presented a ViT for semantic segmentation using a large window attention. (CASTELLANO; MARINIS; VESSIO, 2023b) proposed three variations to this model to add multispectral information (helpful in processing agricultural images). The first variation of the model proposed by Lawin consists of the simple concatenation of multispectral images in the RGB input. The second proposed variation is a more refined model called DoubleLawin, where two Mix Parallel Transformer (MiT) encoders were used, one for visible spectra and the other for multispectral ones. The features produced by both encoders are concatenated using fusion blocks (three different ones were proposed). The third proposed model, called SplitLawin, which, instead of duplicating the encoder, duplicated only the first transformation block, and both resulting features are concatenated using a fusion block. (SA et al., 2018) proposed the dataset used in the work and, as a comparison, already established architectures such as SegNet and DeepLabv3 were used. The method proposed by the authors surpassed the state-of-the-art, obtaining an F1-score of 0.86 (considering all classes) with the SplitLawin architecture on the Rheinbach dataset (0.63 on the Eschikon dataset) and an AUC of 0.96 using the DoubleLawin architecture on the Rheinbach dataset (0.97 in the Eschikon dataset). Despite the promising result, the author discusses the incorrect classification of pixels belonging to weed using the SplitLawin model; of the total pixels, 20.8 are incorrectly classified as background and 1.4% classified as crop. Regarding the plantation, 11.8% are classified as background and 1.7% as weed. Here, it is essential to highlight that 1.7% of crops classified as weeds can lead to considerable losses in the producer's income.

In (CASTELLANO; MARINIS; VESSIO, 2023a), the authors use Transformers to segment weeds in images obtained by UAVs. Unlike the work proposed by (CASTELLANO; MARINIS; VESSIO, 2023b), the authors use a combination of transformers with knowledge distillation in this one. Knowledge distillation is the process of transferring knowledge from one or multiple large models to a single smaller model that can be deployed in real world constraints; in this process the large models are called teachers and smaller students. As a teacher, the authors used the HRNet+OCR+PSA model, which consists of using the HRNetV2 model with the Polarized Self Attention (PSA) block as the attention block. Subsequently, the features obtained by HRNet+PSA are transmitted to the decoder that serves as the OCR (Object Contextual Representation) module. For the student, the model proposed by (YAN; ZHANG; WU, 2022) was used, which uses a Mix Transformer (MiT) as a decoder, and for the decoder, it uses a technique called

Large Window Attention Spatial Pyramid Pooling (LawinASPP). The authors modified the input to support multispectral images, and the Lawin model had its embedding size reduced by half (this modification was made to reduce the computational resource used). Three different types of Knowledge Distillation (KD) were used: vanilla KD which consists of a weighted sum of a task loss and a distillation loss; Teacher Assistant Knowledge Distillation (TAKD) that adds another intermediate model between the teacher and the student, in this case the Lawin-B0 model was used; Annealing Knowledge Distillation which modifies the loss function by introducing a dynamic temperature function to make the student's training gradual and smooth and finally the Continuation Knowledge Distillation which increases the complexity of the objective function during the training. (SA et al., 2018) proposed the dataset used in the work, consisting of two multispectral image datasets (Rheinbach, Eschikon) containing background, crop, and weed. The results of the work demonstrated that the application of KD can increase the performance of simple models obtaining an F1-score of 0.863 (with Continuation KD) in the Rheinbach dataset (the teacher's F1 of 0.868) and, for the Rheinbach dataset, the F1-score was 0.631 (with TAKD) while the teacher's F1-score was 0.656. It should be noted that although the result was promising, for both datasets, the F1-score when we analyzed the weed class was 0.736 and 0.138, demonstrating a deficiency in the models to detect weeds.

In the work proposed by (MOAZZAM et al., 2021), the authors transformed the pixel-by-pixel segmentation problem into a patch-based approach. According to the authors, this is a more realistic way of managing weeds. Furthermore, the authors also highlight the transformation of a 3-class pixel classification problem into a 2-class crop-weed patch classification problem, which increases the solution's performance. In this work, two datasets were used, one that the authors obtained using the Agrocam sensor and evaluated their approach against the public dataset proposed in (SA et al., 2018). Their approach consists of a proposed pre-processing where each image NDVI is calculated, the result is binarized, and a vegetation mask is created. The original image is then remembered in sizes (21×21 and 41×41), and each patch is compared with the vegetation mask; if there is no vegetation in the image, it is discarded. The authors also proposed a modification to the VGG16 CNN model with only 11 convolution layers. In their experiments, the authors tested the proposed model by varying the patch size (21×21 and 41×41) and the combination of input channels in the network. As demonstrated in the work, the best channel configuration was NGB, CIR, and CIR for the Agrocam, Parrot Sequoia, and Red Edge datasets, respectively. Furthermore, the best results, on average, were obtained using the 6-channel configuration (RGB + CIR) and larger patches. Compared with state-of-the-art segmentation networks (UNet and DeeplabV3), the work was also superior, obtaining an average accuracy of 92.4%. It should be noted that, based on the results presented, for the best configuration presented, the error for the patches where the plant is classified as weed is 1.6%.

In (RAMIREZ et al., 2020), the authors carried out a comparative study of different CNN models in the problem of weed segmentation using multispectral images obtained by UAV. In their work, the dataset proposed by (SA et al., 2018) was also used, the SegNet, DeepLabv3, and U-Net models were evaluated, and their approach consists of analyzing the performance of the segmentation when there is overlap between images. To this end, the dataset was evaluated using two different protocols. In the first, there was no overlap; the size of the images was 480×360, and tiles 000, 001, 002, and 004 of the dataset were used for training while tile 003 for testing in the According to the protocol, there was an overlap in the images of 30%, with a size of 512×512, and tiles 000, 001, 002 were used for training. In contrast, 003 and 004 were used for testing. The results were evaluated using the Area Under the Curve (AUC) metric and the F1-score, obtaining the best result with DeepLabv3, an AUC of 0.83, 0.89 and 0.92 and an F1-score of 0.78, 0.81 and 0.92 for the weed classes, crop, and background, respectively. It is important to note that for all models evaluated, the input was modified to 7 dimensions: red, green, blue, near-infrared, color-infrared, red-edge, and NDVI.

In the work (NONG; FAN; WANG, 2022), the authors proposed a semi-supervised learning method for the process of segmenting weeds with varying sizes in complex environments. When applying semi-supervised methods, the need for a large amount of annotated data is considerably reduced, and good results can be achieved. The authors' proposal, called SemiWeedNet, is based on the DeeplabV3+ architecture, which, compared to DeeplabV3, increases the ability to extract multiscale information and encode global structure information, and uses together labeled and unlabeled images to generate a unified segmentation architecture. As the backbone for DeeplabV3, ResNet50 and ResNet101 were used. The authors also created a multiscale enhancement module to integrate the encoded feature with selective kernel attention (SKA). According to the authors, this approach allows the network to better segment pixels even in complex backgrounds. Another addition to the model uses Online Hard Example Mining (OHEM) to refine the labeled data to reduce the similarity between crops and weeds. In their experiments, the authors evaluated their model with two different backbones (ResNet100 and ResNet50) in addition to analyzing the proportion of labeled and unlabeled images provided to the network (2/8, 3/7, 5/5, and fully labeled). Their work obtained results superior to the state of the art, having the best result with the ResNet100 backbone and using fully labeled data (IoU of 0.7). It is essential to note that even in the configuration in which only 20% of the images were labeled, the result was an IoU of 0.692 (very close to the result with fully labeled data). The consistency of the model with other datasets has yet to be evaluated.

In the study proposed by (REEDHA et al., 2022), the authors explore the use of vision transformers (ViT) to classify crops and weeds in drone imagery. The researchers implemented a methodology that involved collecting high-resolution images with UAVs,

followed by data augmentation and transfer learning techniques to boost model performance despite the limited availability of labeled datasets. They carried out experiments using different training and testing image sets, applying a five-fold cross-validation method to assess the model's accuracy and robustness. The original dataset contained 12,844 images, which were split into training sets (66.6%), validation (16.7%) and testing sets (16.7%). The key findings showed that the ViT architecture significantly outperformed conventional convolutional neural networks (CNNs) such as EfficientNet and ResNet, achieving a top F1-score of 99.4% with the ViT B-16 model, even when trained on reduced datasets. Specifically, the ViT B-16 model maintained strong performance with an accuracy of 99.28% on a smaller training set of 6,422 images and 99.63% with only 3,211 images, showcasing its robustness. In contrast, EfficientNet and ResNet experienced a more significant drop in performance as the size of the training dataset was reduced. The results for ResNet50 and EfficientNet models across all dataset sizes highlighted this disparity further. ResNet50 achieved an accuracy of 97.54% with 12,844 images, but its performance decreased to 95. 63% when trained on only 3,211 images, indicating a greater sensitivity to reductions in training data. Similarly, EfficientNet B0 and B1 reached accuracies of 96.53% and 95.91%, respectively, with EfficientNet B1 showing a significant drop in performance to 95.91% under the same reduced dataset conditions. In general, both the ResNet50 and EfficientNet models exhibited notable performance declines as the size of the training dataset decreased, underscoring the limitations of traditional CNN architectures in handling scenarios with limited data. This analysis demonstrated that the self-attention mechanism of the ViT model effectively learned key distinctions between crops and weeds, which is crucial for precision agriculture. The robustness of ViT, even with smaller datasets, marks a promising advance for agricultural image classification, providing a more efficient and reliable alternative to traditional CNN-based approaches.

The paper (OSORIO et al., 2020) presents a comprehensive investigation of the application of advanced imaging and machine learning technologies for precision weed detection. The study used a Mavic Pro drone equipped with a Parrot Sequoia multispectral camera to capture high-resolution images of lettuce fields. The images are composed of red, green, blue and near-infrared spectra, and the authors created a Color Infrared (CIR) image by replacing the blue channel with the near-infrared channel to enhance vegetation features. The research explored three detection methods: (i) a HOG-SVM model, which uses histograms of oriented gradients for feature extraction and Support Vector Machines for classification; (ii) YOLOv3, a deep learning-based object detection model; and (iii) Mask R-CNN, which provides pixel-level segmentation of crops and weeds. To further refine the detection process, a Normalized Difference Vegetation Index (NDVI) was employed to distinguish vegetation from non-vegetation areas. The study found that deep learning models outperformed traditional machine learning approaches, with YOLOv3 and Mask R-CNN achieving F1-scores of 94%, compared to 88% for the HOG-

SVM method. These models demonstrated higher accuracy in detecting and segmenting weeds and were further validated against estimations from three agronomy experts. Using metrics such as Bland–Altman plots and intraclass correlation coefficients, the researchers demonstrated that machine-based estimations reduced the subjectivity and variability inherent in human evaluations. The model output was used to create binary crop masks, which were combined with NDVI-based vegetation masks to calculate weed coverage, providing a precise estimate of the weed density in the fields. This research makes significant contributions to precision agriculture by integrating multispectral imaging based on UAVs with state-of-the-art deep learning models.

The primary objective of the study (HASHEMI-BENI et al., 2022) was to evaluate and compare various deep learning architectures for the pixel-wise segmentation of crops and weeds. Using two open-source datasets, the researchers focused on images acquired from UAVs in the fields of carrots and sugarcane. These datasets were pre-processed through data augmentation techniques like cropping, rotation, and reflection to enhance the training process and address data imbalances inherent in the datasets. The authors tested several state-of-the-art CNN-based segmentation models, including U-Net, SegNet, Fully Convolutional Networks (FCN) in different variants (e.g., FCN-8s, FCN-16s), and DeepLabV3+. Among these, DeepLabV3+ demonstrated the highest accuracy (84.3%) in distinguishing weeds from crops, followed by FCN-8s and U-Net, which achieved competitive results with accuracies of 81.1% and 77.9%, respectively. DeepLabV3+'s advanced architecture, which employs atrous convolutions and spatial pyramid pooling, enabled it to capture multiscale features effectively, making it particularly suited for complex patterns in UAV imagery. However, U-Net excelled in preserving boundary features, which proved useful in delineating crop rows. Models like SegNet and simpler FCN variants performed less effectively, with lower accuracy ranging between 62.6% and 77.2%. The authors also highlighted key challenges in segmenting overlapped vegetation and managing data imbalance, where weeds were often underrepresented in the datasets. Techniques such as median frequency balancing were necessary to mitigate these issues. The findings also emphasized the importance of transfer learning in improving model performance on limited datasets, as well as the potential benefits of integrating vegetation indices such as NDVI to further improve segmentation accuracy.

In the study proposed by (JIANG; AFZAAL; LEE, 2022) the authors used Transformer models to automate weed detection and segmentation in grassy environments. The study evaluated the performance of three state-of-the-art Transformer architectures, Swin Transformer, SegFormer (mit-B0) and Segmenter on a dataset of 1006 images featuring ten classes of weeds. The dataset, collected to represent real-world grass management scenarios, was augmented to address the data requirements of the Transformer models, allowing for improved training and generalization. The researchers employed semantic segmentation techniques to distinguish weeds from non-weed regions, leveraging metrics

like Mean Accuracy (mAcc) and Mean Intersection over Union (mIoU) to measure performance. Among the models tested, SegFormer achieves the highest mAcc of 75.18% and mIoU of 65.74%. It also stood out for its computational efficiency, requiring only 3.7 million parameters compared to the heavier computational demands of other models like the Swin Transformer. This efficiency makes SegFormer particularly suitable for deployment in real-time and resource-constrained agricultural settings.

In (XU et al., 2023b), the authors used drones equipped with RGB, multispectral and thermal cameras to capture high-resolution images of corn fields in Hainan Province, China. These images were analyzed to combine spectral, textural, structural, and thermal measurements to enhance the precision of weed detection. The extracted features included vegetation indices, texture metrics derived from the Gray Level Co-occurrence Matrix (GLCM), plant height estimated from 3D models, and canopy temperature obtained by thermal imaging. Data analysis involved Principal Component Analysis (PCA) to reduce dimensionality and prevent overfitting, while machine learning models such as support vector machines (SVM), Random Forrest (RF) and Nave Bayes Classifier (NBC) were used to classify weeds and corn. The authors found that including thermal data alongside spectral, textural, and structural features significantly improved the accuracy of weed mapping. The combination of textural, structural, and thermal data achieved the highest performance, with an overall accuracy of 96.4% and an F1-score of 0.964. They also discuss that adding thermal images proved particularly valuable, providing unique insights into canopy temperature differences between weeds and corn. When examining the best machine learning model, SVM achieved the highest performance.

The study (CUNHA et al., 2023) the authors investigates how multispectral imagery can enhance image segmentation for agricultural tasks, particularly crop-row detection. The research emphasizes the use of fusion approaches combining information from RGB and NDVI (Normalized Difference Vegetation Index) improves segmentation accuracy and adaptability in agricultural settings. Two main fusion techniques were explored: early fusion, which integrates data at the input level before segmentation, and late fusion, which processes each modality independently and combines their outputs afterward. These methods were applied using classical segmentation techniques, such as Otsu's thresholding, edge detection, and region-based approaches, as well as deep learning (DL) models, including SegNet and DeepLabV3. The researchers used two datasets to evaluate their approaches. The first, a maize crop dataset, was collected using a Parrot Sequoia multispectral camera mounted on a robotic platform. This dataset, recorded during the early growth stage of maize in Coimbra, Portugal, includes RGB and multispectral bands (Green, Red, Red Edge, and Near Infrared). A second dataset, consisting of vineyard images, was used to assess the cross-domain generalization of the proposed methods. The results demonstrate that classical methods remain competitive for tasks such as simple foreground-background separation, where their low computational cost is an advantage.

However, for more complex segmentation tasks that require the detection of spatial patterns, DL-based approaches are better because of their ability to learn intricate features from the data. Among fusion techniques, late fusion techniques are the most robust, consistently achieving superior segmentation accuracy and adaptability across different scenarios.

Chapter 4

Methodology

Existing approaches to weed segmentation span classical CNN-based methods (RAMIREZ et al., 2020; SA et al., 2018) and more recent Transformer powered architectures (CASTELLANO; MARINIS; VESSIO, 2023b; JIANG; AFZAAL; LEE, 2022), incorporating data-fusion schemes (e.g., NDVI, thermal imaging) (XU et al., 2023b) and semi-supervised learning (NONG; FAN; WANG, 2022) to improve robustness and scalability. While some studies adopt patch-level classification (MOAZZAM et al., 2021) or object detection pipelines (OSORIO et al., 2020), others employ knowledge-distillation techniques to compress models without sacrificing accuracy (CASTELLANO; MARINIS; VESSIO, 2023a).

Compared to these solutions, our work introduces a novel framework that combines:

- ❏ A detailed analysis of color spaces and vegetation indices to identify the most discriminative features using the Overlap Index (OI).
- ❏ The integration of pre-selected channels and indices with transformer architectures, leveraging advanced fusion strategies (Early and Late Fusion).
- ❏ A focus on maximizing generalization by testing on diverse datasets, including RGB and multispectral images captured under different environmental and geographical conditions.

This expands the current state of the art by:

- ❏ Providing a systematic evaluation of the separability between classes in multiple feature spaces (color channels and vegetation indices), which has not been explored in depth in previous studies.
- ❏ Proposing architectural improvements to transformer models, such as tailored fusion strategies that enhance segmentation accuracy in complex datasets.

- ❏ Demonstrating robust model performance across diverse datasets, including those from different sensors and altitudes, ensuring applicability in real-world agricultural scenarios.

4.1 Datasets

As demonstrated by (HASAN et al., 2021), there is a large number of datasets for detecting diseases or weeds. However, the majority of these were obtained with sensors attached to vehicles or drones at very low altitudes. To evaluate this work, we will use two public datasets (RGB and multispectral) obtained by UAV in different locations, with different sensors and different resolutions.

The Weed Map Dataset was made available by (SA et al., 2018). This dataset consists of two flights carried out with UAVs, sensors, and at different dates/times. The main crop is sugar beet (Beta vulgaris), whose plantation has the following weeds: *Galinsoga spec.*, *Amaranthus retroflexus*, *Atriplex spec.*, *Polygonum spec.* and *Gramineae*. It is important to highlight that although this crop is not common in Brazil, several of these weeds affect our crops, such as *Gramineae* and *Polygonum convolvulus* that affect the country's southern region. The sizes of crops and weeds exhibited 8–10cm and 5–10cm, respectively.

The first flight, from now on identified as Sequoia (given the sensor used), was carried out in the Eschikon region in Switzerland between 5 and 18 May 2017 around noon. The Mavic Pro drone with a Sequoia sensor attached was used. The flight altitude was 10 meters, resulting in an average Ground Sample Distance (GSD) of 0.95 cm/pixel. This flight covered three tiles named 005, 006, and 007, with a total area of 0.762 hectares. The sensor used on this flight is capable of capturing four spectrum bands: Green (550.40 nm), Red (660.40 nm), Red Edge (735.10 nm) and Near Infrared (790.40 nm).

The second flight, which will be identified as RedEdge, was carried out in the Rheinbach region, Germany, on 18 September 2017, 9:18–40 a.m. The Inspire 2 drone with the RedEdge-M sensor attached was used. The altitude was also 10 meters with an average GSD of 1 cm/pixel. This flight covered five tiles named 000, 001, 002, 003, and 004, with a total area of 0.893 hectares. The sensor used on this flight is capable of capturing five spectrum bands: Blue (475.20 nm), Green (560.20 nm), Red (668.10 nm), Red Edge (717.10 nm), and Near-Infrared (840.40 nm).

The summary of this dataset can be seen in Table 5. Figure 23 illustrates some of the orthomosaics along with their labels, and Figure 24 presents an example of the images available in each dataset.

(MONTEIRO; WANGENHEIM, 2019) proposes the second dataset used in this work, from now on referred to as SugarCane, and consists of a sugar cane plantation in the Northern region of Brazil. This dataset was captured with a single fixed-wing UAV with a Canon G9X RGB sensor. The capture altitude was an average of 162 meters, with a

Table 5 – Details of dataset. Adapted from (SA et al., 2018).

Camera	**RedEdge-M**					**Sequoia**		
Dataset name	000	001	002	003	004	005	006	007
Area covered (ha)	0.31	0.11	0.20	0.13	0.13	0.25	0.33	0.17
GSD (cm)	1.04	0.94	0.96	0.99	1.07	0.85	1.18	0.83
Tile resolution (row/col)	360/480					360/480		
Channels	**5**					4		

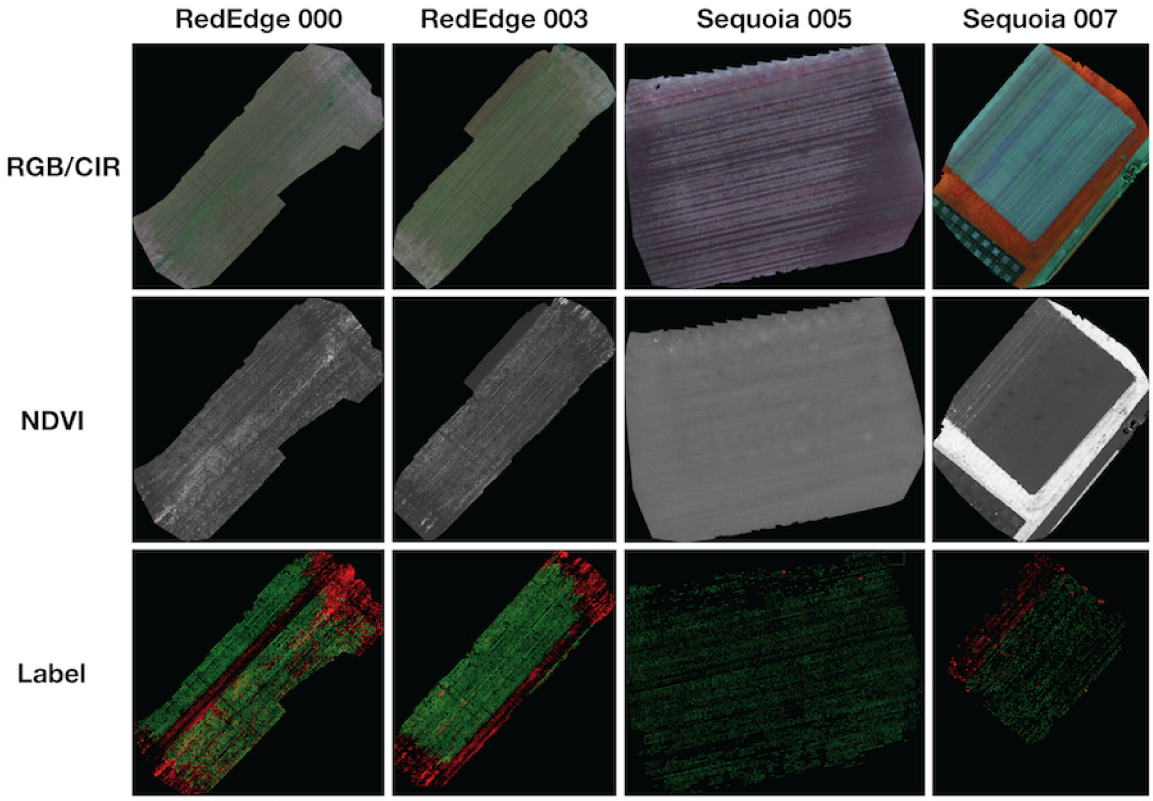


Figure 23 – Example of the orthomosaics present in the dataset. Black represents the background, green represents crops, and red represents weeds. Adapted from (SA et al., 2018).

special resolution of 5 cm/pixel. The weed plant species were not specified, but given the culture and region, we can infer that they may be *Ipomoea spp.*, *Brachiaria plantaginea* and *Digitaria horizontalis*. Figure 25 shows this dataset and its ground-truth.

4.2 Evaluation Metrics

Various metrics have been created to measure the performance of segmentation models and algorithms (HASAN et al., 2021). Typically, these metrics seek to compare the predicted class versus the class present in the ground truth of each pixel. Table 6 presents the main evaluation metrics.

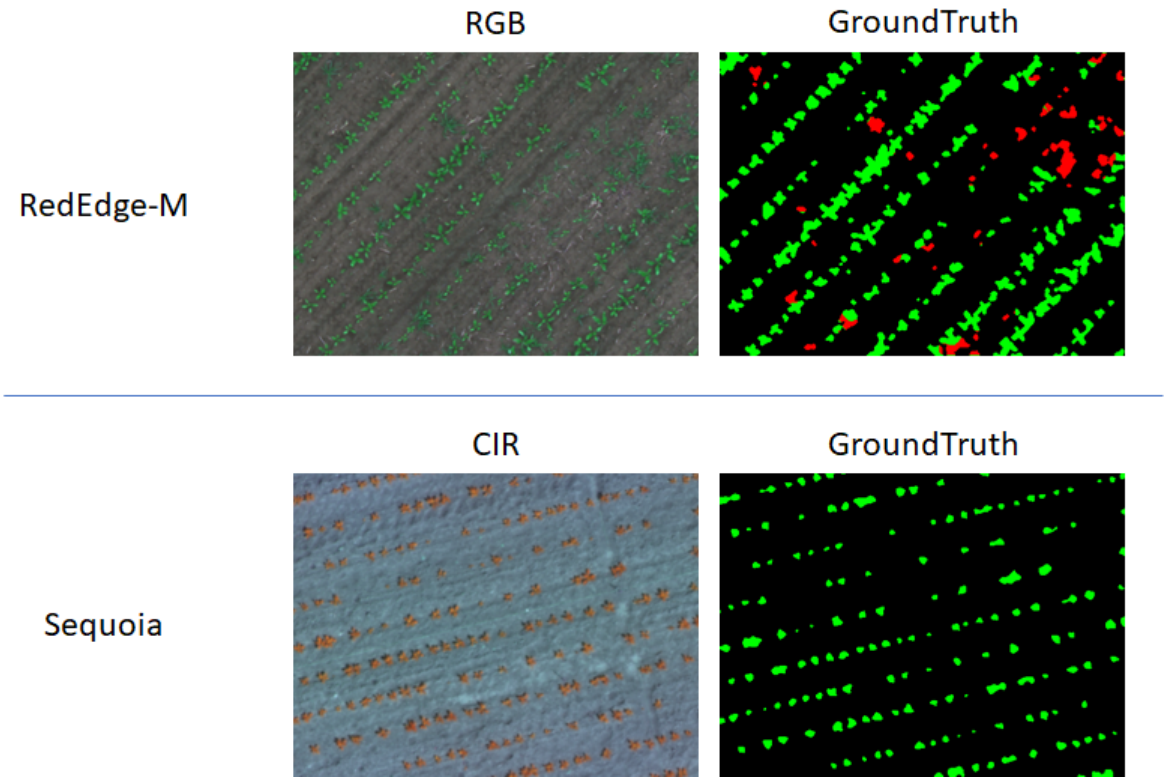


Figure 24 – Example of images in the dataset. Black represents the background, green represents crops, and red represents weeds.

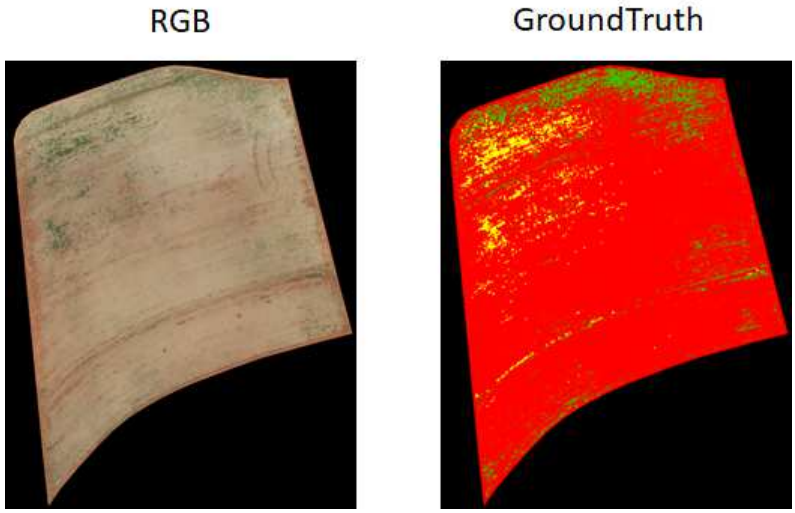


Figure 25 – Orthomosaic of the SugarCane dataset. In GroundTruth, red represents the background, green represents crops, and yellow represents weeds

To identify promising vegetation indices and color spaces, we will use the OI, a metric that quantifies the percentage of overlap between the pixel value intervals of two classes, considering their mean (μ) and standard deviation (σ). It is particularly useful for evaluating the separability between classes in a feature space, such as color channels or vegetation indices. A low OI indicates that the distributions of the classes have little overlap, making them easier to distinguish, whereas a high OI suggests greater confusion

Table 6 – Main evaluation metrics. Adapted from (HASAN et al., 2021).
TP (true positive) means that the predicted value is positive and it is positive, TN (true negative) means that the predicted value is negative and it is negative, FP (false positive) means that the predicted value is positive but it is negative and FN (false negative) means that the predicted value is negative but it is positive.

Metric	Meaning	Equation
Accuracy	The percentage of correct prediction among the input. A model is judged based on how high the value is.	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	The fraction of correct prediction (True Positive) from the total number of relevant results (Sum of True Positive and False Positive). It helps when the value of False Positives is high	$\frac{TP}{TP+FP}$
Specificity	The fraction of True Negative from the sum of False Positive and True Negative.	$\frac{TN}{TN+FP}$
F1-score	The harmonic mean of precision and recall.	$\frac{2 \cdot TP}{2 \cdot TP+FP+FN}$

between the classes. This metric is valuable in image segmentation problems, such as the separation of weeds from crops or background, as it helps identify channels or indices that maximize class discrimination.

$$\text{OI} = \frac{|(\mu_1 - \sigma_1, \mu_1 + \sigma_1) \cap (\mu_2 - \sigma_2, \mu_2 + \sigma_2)|}{|(\mu_1 - \sigma_1, \mu_1 + \sigma_1) \cup (\mu_2 - \sigma_2, \mu_2 + \sigma_2)|} \times 100 \tag{6}$$

where:

- μ_1 and μ_2 are the means of the pixel values for classes 1 and 2, respectively.
- σ_1 and σ_2 are the standard deviations of the pixel values for classes 1 and 2, respectively.
- $(\mu_1 - \sigma_1, \mu_1 + \sigma_1)$ and $(\mu_2 - \sigma_2, \mu_2 + \sigma_2)$ represent the intervals formed by the mean and standard deviation for classes 1 and 2, respectively.
- $|(\mu_1 - \sigma_1, \mu_1 + \sigma_1) \cap (\mu_2 - \sigma_2, \mu_2 + \sigma_2)|$ is the length of the intersection between the two intervals.
- $|(\mu_1 - \sigma_1, \mu_1 + \sigma_1) \cup (\mu_2 - \sigma_2, \mu_2 + \sigma_2)|$ is the length of the union of the two intervals.
- OI represents the Overlap Index, expressed as a percentage.

4.3 Weed Segmentation using Transformers

This work explores the use of transformers in conjunction with different pre-processing techniques and vegetation indices for weed segmentation. Initially, all datasets will be

processed by computing the vegetation indices (VIs) described in Section 2.1.4. As highlighted by (XUE; SU et al., 2017), each VI has its own characteristics, and selecting the most suitable one depends on analyzing factors such as the platform where the image was captured, background, light, and vegetation. An exploratory search of these VIs will provide insights into the most effective combinations, potentially improving generalization across different datasets.

Additionally, we will analyze color spaces. In previous work (FERRAZ et al., 2023), it was demonstrated that specific color space channels can have zero overlap in pixel values between background and vegetation classes (crop and weed). Identifying such channels facilitates segmentation. By the end of this step, several inputs will be defined, which will be further analyzed and combined to achieve optimal results. Finally, we will explore various transformer architectures, highlighting their strengths and weaknesses in weed segmentation to propose architectural improvements for enhanced accuracy. Figure 26 illustrates the complete workflow of the proposed approach, showcasing the integration of multiple data sources through early and late fusion strategies.

4.3.1 Analysis of Color Spaces

The first step of this work involves a detailed analysis of different color spaces to identify channels that contribute to efficient weed segmentation. The color spaces considered will be: **RGB**, **HSV**, **LAB**, **LUV**, **XYZ**, **YCBCR**, **YIQ**, **YPBPR**, and **YUV**. These color spaces were selected due to their popularity and performance in various segmentation studies (LI et al., 2021; RAMADAN; ALY, 2022; TURHAL; DAGDELEN, 2022).

As demonstrated by (FERRAZ et al., 2023), certain color spaces are more suitable for weed segmentation. To identify the most effective channels, we will analyze each channel and compute the OI between background and weed classes, as well as between crops and weeds. This analysis will allow us to select the most promising channels for further processing. Figure 27 outlines the processing steps.

4.3.2 Calculation of Vegetation Indices

The second step calculates vegetation indices (VIs) for all datasets. The multispectral indices include NDVI, EVI, Green-normalized difference vegetation index (GNDVI), Soil-adjusted vegetation index (SAVI), Simple Ratio (SR), Normalized Difference Red Edge Index (NDRE), Green Difference Vegetation Index (GDVI), and Chlorophyll Index (CI). For RGB datasets, indices such as NGRDI, ExG, ExR, Visible Atmospherically Resistant Index (VARI), Modified Green Red Vegetation Index (MGRVI), Normalized Pigment Chlorophyll Index (NPCI), Triangular Greenness Index (TGI), and Excess Blue Index (ExB) will be used. As seen in Section 2.1.4, these indices have been successfully used in plant segmentation and various applications (SAHIN et al., 2023; COX et al., 2023).

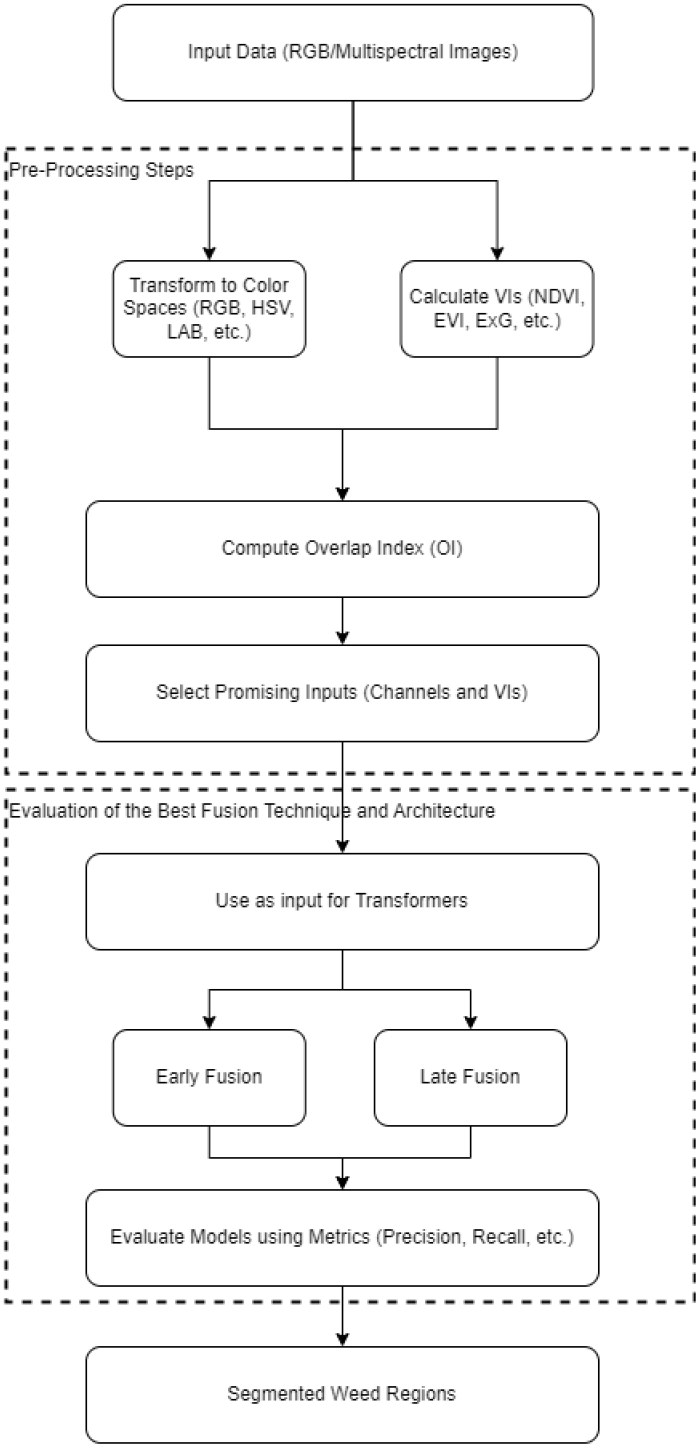


Figure 26 – Proposed workflow for integrating multiple modalities using early and late fusion strategies in SegFormer-B1.

Each index will be computed based on the spectral bands of the dataset images. These indices enhance specific vegetation features, as noted by (XUE; SU et al., 2017). Similar to the color space analysis, an investigation will determine which indices exhibit minimal value overlap between classes, ensuring optimal discrimination.

Figure 28 illustrates the steps.

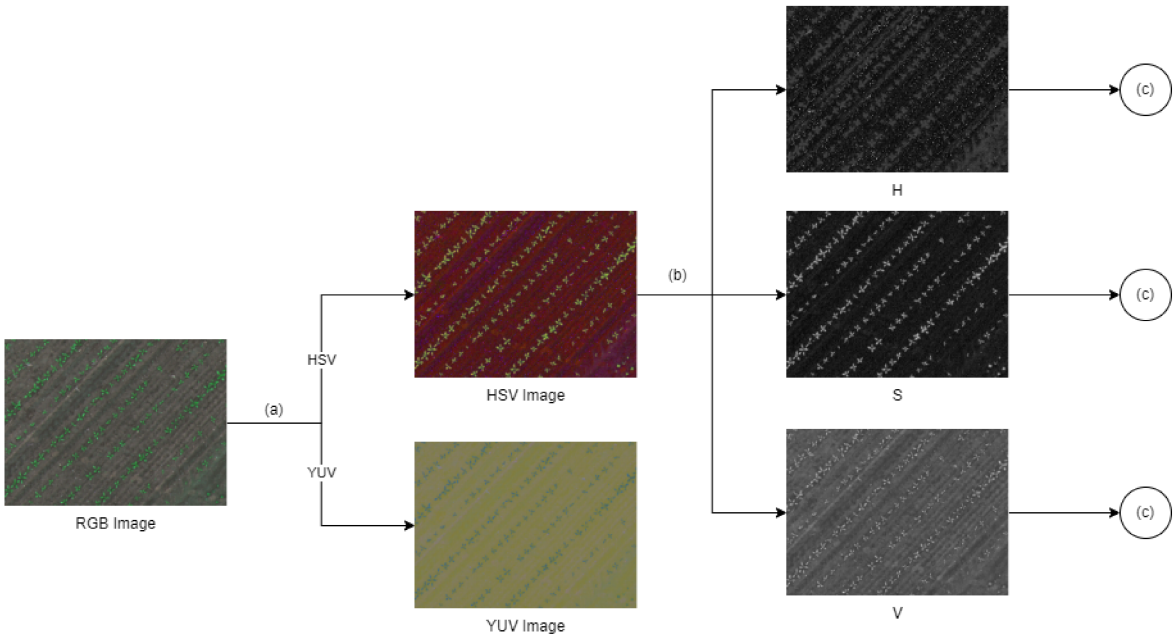


Figure 27 – Calculation of the Overlap Index. (a) The original RGB image is transformed into the color spaces described in this section. (b) Each channel of each color space is processed individually. (c) The OI is calculated for background versus weed and crop versus weed classes.

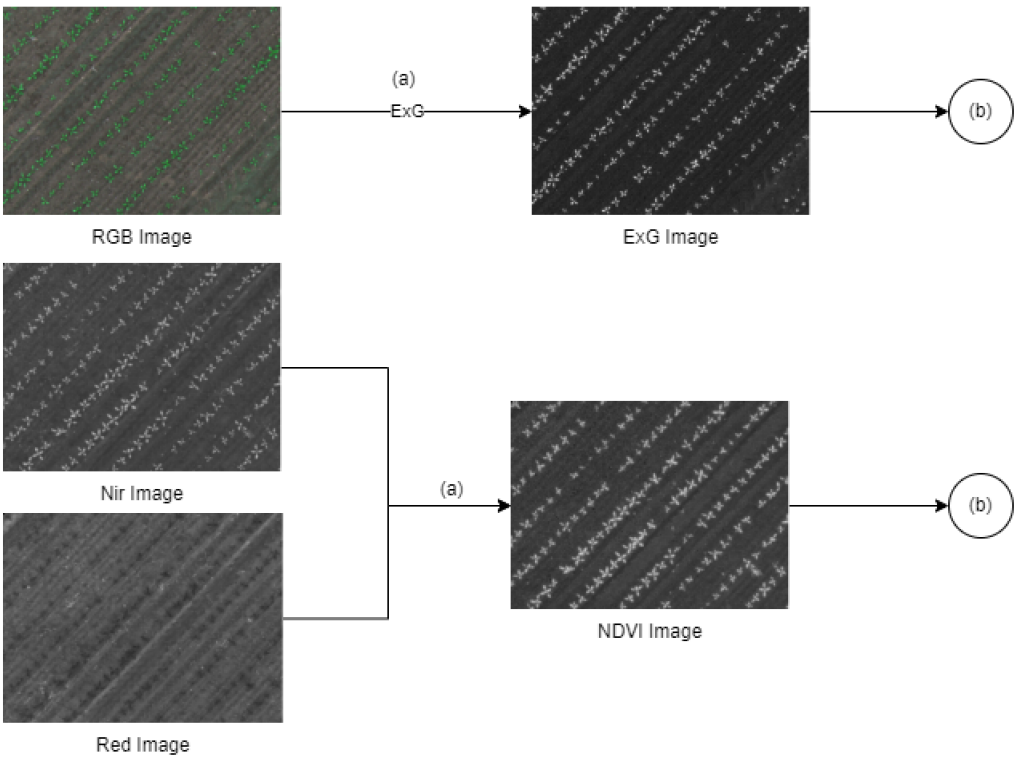


Figure 28 – Calculation of the Overlap Index. (a) The RGB or multispectral image is transformed into the vegetation indices described in this section. (b) The OI is computed between background, weed, and crop classes.

4.3.3 Fusion Strategies in Transformers Architectures

This work explores the use of SegFormer architecture for semantic segmentation of weeds, specifically focusing on the application of fusion techniques to integrate diverse input data sources. These include RGB images, multispectral bands, VIs, and color spaces. The fusion strategies are designed to leverage the complementary nature of these modalities, improving the segmentation accuracy while maintaining computational efficiency. We implemented two primary fusion strategies:

- ❑ **Early Fusion:** Combines all input modalities (e.g., RGB channels, vegetation indices, and multispectral bands) at the network's input stage.
- ❑ **Late Fusion:** Processes each modality independently with separate encoders, combining their outputs in later stages of the network.

In the early fusion approach, the inputs from all modalities are concatenated along the channel dimension at the very beginning of the architecture. The combined input, which contains the combination of several input data (e.g. RGB, multispectral, vegetation indices, and color spaces), is fed into a single SegFormer encoder. This integration allows the encoder to learn joint representations of all modalities from the earliest stage of processing. The hierarchical structure of the encoder enables the extraction of features at multiple scales, capturing both spatial details and global context.

By merging the data at the input level, the model simplifies the fusion process, requiring only one encoder and a standard SegFormer decoder. This strategy ensures computational efficiency while preserving the richness of the multimodal inputs. However, it assumes that the encoder is capable of effectively learning and integrating the diverse characteristics of all input channels.

The late fusion approach combines the final outputs (logits) of separate models at the end of the processing pipeline. This method allows each model to specialize in processing its respective input modality, such as RGB data in one model and multispectral or vegetation indices in the other. By processing the inputs independently, the overall architecture preserves the unique characteristics of each modality until the final stage.

In our implementation, two independent models are used, each producing a final segmentation prediction. The fusion is performed solely on these final outputs, rather than combining intermediate features from each transformer block. The workflow can be summarized as follows:

1. **Processing with Independent Models:** Each model processes its respective input (e.g., RGB data in $inputs_1$ and multispectral or vegetation indices in $inputs_2$) through a hierarchical transformer architecture, extracting features and producing a final segmentation output.

2. **Fusion of Final Outputs:** The final outputs (logits) from each model are first adjusted via a 1×1 convolutional layer to match the desired number of channels (classes). These adjusted logits are then concatenated along the channel dimension, effectively combining the complementary information from the two modalities.

3. **Dimensionality Reduction with Convolutional Fusion:** The concatenated logits are passed through another 1×1 convolutional layer to reduce the channel dimension to the fixed number of classes. This learnable operation enables the model to adaptively fuse the information from both models.

4. **Residual Connections for Enhanced Information Flow:** In addition to the fusion, residual connections incorporating adjusted versions of the original inputs are added to the fused output. These residuals help preserve information from the individual inputs, reinforcing the final segmentation prediction.

This approach preserves the specialization of each model for its respective modality by processing them independently and fusing only their final outputs to generate the final segmentation map.

Chapter 5

Experimental Results

This chapter presents the experimental evaluation of the proposed methodology for automatic weed segmentation in UAV-based multispectral images. The performance of different input configurations, including RGB, vegetation indices, and multispectral data, is analyzed using deep learning models. Additionally, the impact of data fusion strategies, particularly early and late fusion, is assessed across multiple datasets. The results are compared against state-of-the-art approaches to validate the effectiveness of the proposed framework.

5.1 Preprocessing of Color Spaces and Vegetation Indices

To efficiently identify the most relevant color channels and vegetation indices for the task of weed segmentation, a systematic framework was implemented to process and analyze spectral and spatial image data across diverse datasets. The approach leveraged the integration of color spaces, vegetation indices, and statistical measures to extract meaningful features and normalize the data for subsequent analysis.

Initially, the methodology utilized the RGB and multispectral datasets described in Section 4.1, representing various spectral bands, including visible light (Red, Green, Blue) and NIR or Red Edge (RE) channels. A critical step involved the preprocessing of these bands to ensure the completeness and integrity of the data. Since the Sequoia sensor does not capture the blue band, it was estimated as the arithmetic mean of the red and green channels. This approach ensures spectral consistency across datasets, allowing the calculation of vegetation indices that require an RGB input (e.g., ExG, VARI).

To enhance the feature set, we computed multiple vegetation indices. These indices, such as the NDVI, EVI, and GNDVI, provide spectral insights strongly associated with vegetation health and biomass. Each index was calculated using well-defined formulas that combine specific spectral bands. For example, NDVI uses the contrast between the NIR

and red bands to highlight vegetation, while indices like ExG and VARI are tailored for scenarios where RGB data are predominantly available. The selection of indices for each dataset was guided by its spectral composition, ensuring compatibility and maximizing the use of available bands.

Color spaces further enriched the analytical framework. Standard RGB data were transformed into alternative representations, such as HSV (Hue, Saturation, Value), LAB, YCbCr, and others, using established conversion methods. These transformations enabled the extraction of features that emphasize color properties, luminance, and chrominance, providing complementary perspectives on the same data. By employing multiple color spaces, the analysis benefited from a broader range of feature dimensions, facilitating better differentiation between vegetation and non-vegetation classes.

5.1.1 Color Spaces

Analyzing color spaces through the OI metric revealed that several channels from different color spaces performed well in segmenting background-crop and background-weed, as these classes exhibit clear differences in illumination and color. However, very few channels achieved satisfactory performance in the crop-weed segmentation task. Given the critical importance of crop-weed separation in this study, the weights for background-crop (0.15), background-weed (0.15), and crop-weed (0.70) relations were determined empirically to reflect their relative significance. Lower weights were assigned to background-related relations as they are less challenging, while the crop-weed relation was prioritized due to its direct impact on weed detection accuracy. This weighting scheme aligns with the study's objective of emphasizing precise crop-weed differentiation, crucial for practical applications in precision agriculture. Table 7 presents the final results obtained.

It is noteworthy that the components Cb, Pb, and U from the YCbCr, YPbPr, and YUV color spaces, respectively, exhibited identical results due to their mathematical derivation, which shares similar formulas and transformation principles. These components represent chrominance, encoding the color information of an image by quantifying the differences between the blue and red channels relative to luminance (Y). The Cb, Pb, and U components showed the lowest OI when analyzing the weighted average across all datasets. This result can be explained by their ability to capture chromaticity while minimizing the influence of luminance. This feature is crucial in drone imagery, where variations in lighting and shadows are common. By isolating color information, these components reduce sensitivity to soil reflections and variations in light intensity, allowing for greater precision in distinguishing weeds and crops. This property also facilitates the identification of vegetation at different growth stages, as plants' spectral signatures vary in shades of green and other colors depending on their phenological stage. This ability may explain their good performance in separating weeds and crops, as these plants typically present distinct growth stages. Notably, these color spaces have been widely used

in plant segmentation studies (SHIRZI; KERMANI, 2024; PRASETYO et al., 2017).

The U component of the LUV color space, the Q component of the YIQ color space, and the A component of the LAB color space represent chromaticity along the green-magenta axis, explaining their high performance in segmenting weeds and crops from the background in aerial imagery. Soil (background) generally exhibits minimal contrast on this axis, whereas crops and weeds display distinct proportions of green and magenta because of their pigmentation, such as chlorophyll. This shared characteristic allows these components to effectively highlight vegetation and differentiate it from non-vegetative elements like soil. Their sensitivity to variations along the same chromatic axis reinforces their ability to separate crops and weeds, making them particularly valuable for improving segmentation accuracy in agricultural applications. The study by (LU et al., 2022) demonstrated that combinations of channels such as H (in HSV), HS (in HSV), U (in YUV), CrCb (in YCbCr), YU (in YUV) and LA (in LAB) achieved the best results in segmentation of canopy fruit trees. Figure 29 shows the U and V components for the YUV and LUV color spaces, respectively. In these images, it is evident that a simple transformation allows for a clearer distinction between the classes.

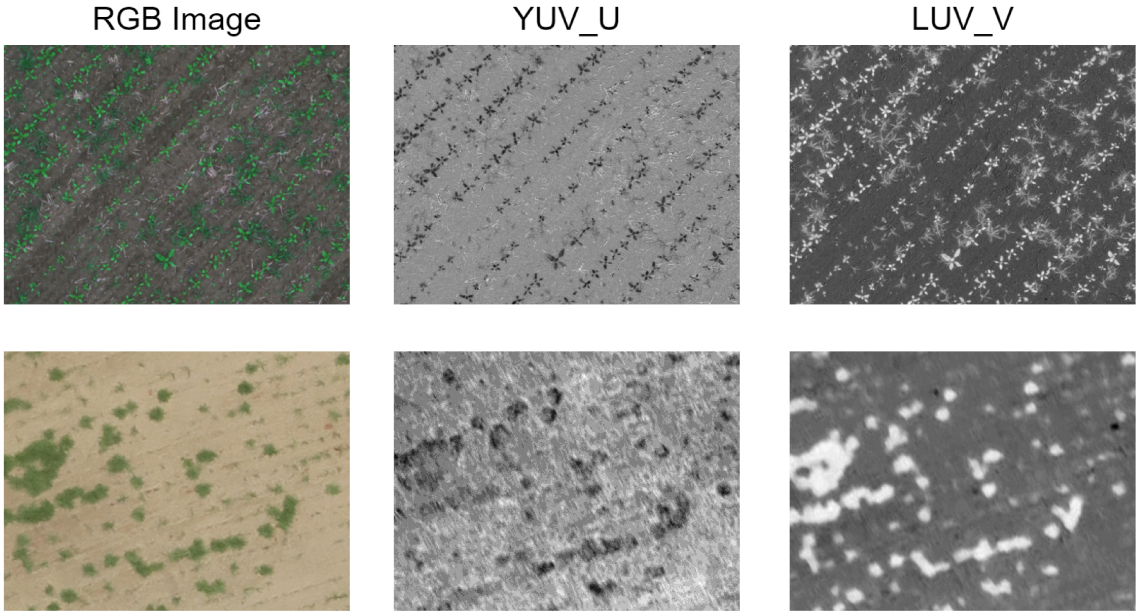


Figure 29 – Example of color space transformation for images from the RedEdge and Sugarcane datasets.

5.1.2 Vegetation Indices

As observed in the color spaces analysis, although the crop-weed segmentation achieves a high OI, it is still lower compared to the results obtained using color spaces. As shown in Table 8, the indices ExB, TGI, and ExG achieved the lowest OI, indicating their superior performance in distinguishing vegetation from other classes. Additionally, when

Table 7 – Weighted Mean of OI obtained for the evaluated color spaces in the RedEdge, Sequoia, and Sugarcane datasets

Color Space	**Background Crop**	**Background Weed**	**Crop Weed**	**Weighted Mean**
YCbCr_Cb	11.73 ± 22.20	28.47 ± 14.48	**42.27 ± 17.29**	**35.62**
YPbPr_Pb	11.73 ± 22.20	28.47 ± 14.48	**42.27 ± 17.29**	**35.62**
YUV_U	11.73 ± 22.20	28.47 ± 14.48	**42.27 ± 17.29**	**35.62**
YIQ_Q	5.08 ± 14.36	5.32 ± 9.82	49.24 ± 11.40	36.03
LUV_V	9.55 ± 17.90	20.36 ± 18.53	46.39 ± 15.07	36.96
LAB_A	5.32 ± 15.04	4.18 ± 10.28	51.31 ± 10.45	37.34
LUV_U	**3.64 ± 10.29**	**3.33 ± 7.22**	53.24 ± 9.04	38.31
LAB_B	13.22 ± 24.86	27.41 ± 25.40	46.74 ± 18.29	38.81
YPbPr_Pr	5.08 ± 14.36	4.24 ± 10.01	55.52 ± 7.95	40.26
YUV_V	5.08 ± 14.36	4.24 ± 10.01	55.52 ± 7.95	40.26
YCbCr_Cr	5.08 ± 14.36	4.24 ± 10.01	55.52 ± 7.95	40.26
HSV_H	8.97 ± 10.31	13.39 ± 9.75	57.34 ± 14.68	43.49
YIQ_I	5.08 ± 14.36	4.24 ± 10.01	61.80 ± 7.47	44.65
RGB_R	7.58 ± 9.55	26.41 ± 24.57	57.69 ± 14.21	45.48
HSV_S	14.28 ± 21.08	26.21 ± 26.47	57.04 ± 14.78	46.00
RGB_G	28.40 ± 16.21	37.31 ± 17.80	52.55 ± 6.72	46.64
HSV_V	24.09 ± 16.34	47.41 ± 30.38	53.29 ± 9.07	48.03
LAB_L	32.31 ± 15.70	48.88 ± 26.35	54.49 ± 8.73	50.32
LUV_L	32.31 ± 15.70	48.88 ± 26.35	54.49 ± 8.73	50.32
XYZ_Y	34.31 ± 14.97	50.42 ± 26.82	55.17 ± 8.47	51.33
RGB_B	39.16 ± 19.60	57.04 ± 31.10	59.78 ± 15.00	56.28
YCbCr_Y	43.68 ± 22.82	59.54 ± 33.58	59.15 ± 12.88	56.89
YIQ_Y	43.68 ± 22.82	59.54 ± 33.58	59.15 ± 12.88	56.89
YPbPr_Y	43.68 ± 22.82	59.54 ± 33.58	59.15 ± 12.88	56.89
YUV_Y	43.68 ± 22.82	59.54 ± 33.58	59.15 ± 12.88	56.89
XYZ_X	49.46 ± 29.25	57.86 ± 33.71	63.63 ± 14.77	60.64
XYZ_Z	47.91 ± 24.09	56.07 ± 29.72	69.89 ± 17.70	64.52

analyzing the OI for crop-weed segmentation, ExB, CI, and ExG were the best performing indices. These results reinforce the effectiveness of these indices in plant segmentation and weed management (TORRES-SÁNCHEZ et al., 2021; FORD; SADGROVE; PAUL, 2023; ANDEREGG et al., 2023).

The ExB and ExG are particularly valuable for vegetation detection using RGB images. ExB was the overall best performing index, demonstrating high accuracy in both OI mean and crop-weed separation. This index emphasizes the blue spectral band, which enhances the contrast between vegetation and the background, particularly in shadowed regions. Its stability under variable lighting conditions makes it highly robust for agricultural applications. ExG, leveraging the strong green reflectance of chlorophyll, also performed well in both general segmentation and crop-weed separation. It remains one of the most widely used indices for early vegetation mapping.

The TGI and CI focus more on vegetation health and chlorophyll content. The TGI ranked second in OI mean, indicating that its ability to balance contributions from the red, green, and blue bands makes it an effective tool for segmenting crops and weeds based on chlorophyll variation. This characteristic is particularly useful for differentiating healthy crops from stressed vegetation or background elements. CI, which was highly effective in the crop-weed segmentation task, likely benefited from its sensitivity to pigment concentration, further improving classification accuracy.

These results highlight the importance of selecting the appropriate vegetation index based on the segmentation task. While ExB and ExG provide strong performance in RGB-based analysis, TGI and CI demonstrate their relevance in cases where spectral variations related to chlorophyll and pigmentation play a key role. Figure 30 shows examples of the vegetation indices ExB and TGI for the RedEdge and Sugarcane datasets. Similarly to the components of color spaces, vegetation indices highlight specific spectral features, facilitating the segmentation process.

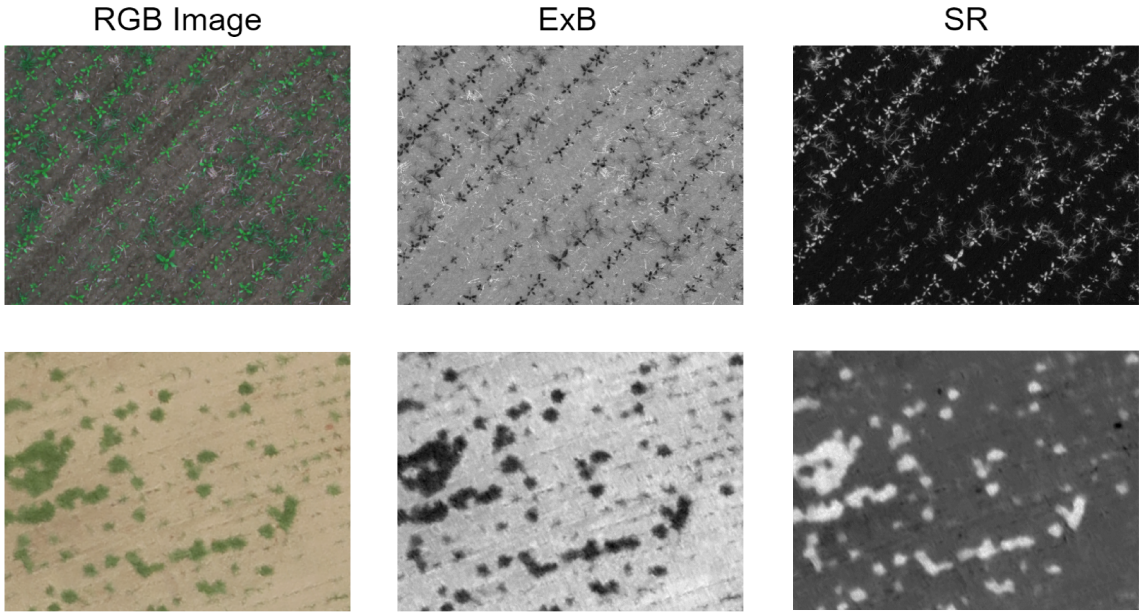


Figure 30 – Example of vegetation index transformation for images from the RedEdge and Sugarcane datasets.

Table 8 – Weighted Mean of OI obtained for the evaluated VI in the RedEdge, Sequoia, and Sugarcane datasets

Vegetation Index	Background Crop	Background Weed	Crop Weed	Weighted Mean
ExB	3.97 ± 9.75	21.48 ± 15.23	**40.13 ± 12.75**	**31.91**
TGI	5.08 ± 13.82	5.39 ± 9.36	49.13 ± 11.53	35.96
ExG	5.08 ± 13.82	**4.94 ± 9.46**	50.00 ± 10.89	36.51
ExR	4.55 ± 12.39	7.34 ± 13.90	51.66 ± 9.38	37.95
NGRDI	3.27 ± 8.89	5.97 ± 14.38	56.27 ± 9.00	40.77
NPCI	3.27 ± 8.89	5.97 ± 14.38	56.27 ± 9.00	40.77
MGRVI	3.73 ± 10.15	6.25 ± 15.42	56.33 ± 9.07	40.93
GNDVI	4.43 ± 5.94	8.56 ± 13.88	55.69 ± 13.14	40.93
VARI	3.27 ± 8.89	5.71 ± 14.38	61.12 ± 9.29	44.13
NDVI	3.06 ± 7.40	5.91 ± 14.29	62.87 ± 16.89	45.35
SAVI	3.24 ± 7.85	6.21 ± 15.02	64.34 ± 16.33	46.46
SR	**1.08 ± 2.62**	5.28 ± 9.34	65.70 ± 18.59	46.94
EVI	22.79 ± 17.50	38.90 ± 25.44	55.50 ± 26.85	48.10
CI	55.41 ± 13.49	34.12 ± 8.53	54.83 ± 11.61	51.81
NDRE	56.52 ± 12.67	34.94 ± 10.47	55.20 ± 11.38	52.36

5.1.3 Multispectral Bands

Additionally, we evaluated the OI for the multispectral bands NIR and RE. As shown in Table 9, the results for NIR and RE were moderate, being particularly useful for background-crop and background-weed segmentation. These bands effectively captured the pronounced differences in reflectance associated with non-vegetated areas (background) versus vegetated regions (crop and weed). The NIR band is highly sensitive to vegetation structure and water content, while the red-edge band is particularly responsive to variations in chlorophyll concentration. These characteristics allowed for clear separation between vegetated and non-vegetated classes, leading to minimal overlap. Figure 31 shows the NIR and RE bands for the RedEdge and Sequoia datasets.

Table 9 – Overlap Index results for multispectral bands.

Multispectral Band	Background Crop	Background Weed	Crop Weed	Weighted Mean
RE	12.27 +/-15.52	16.00 +/-7.96	**58.13 +/-7.67**	**44.93**
NIR	**9.23 +/-14.96**	**11.11 +/-14.20**	66.25 +/-15.77	49.43

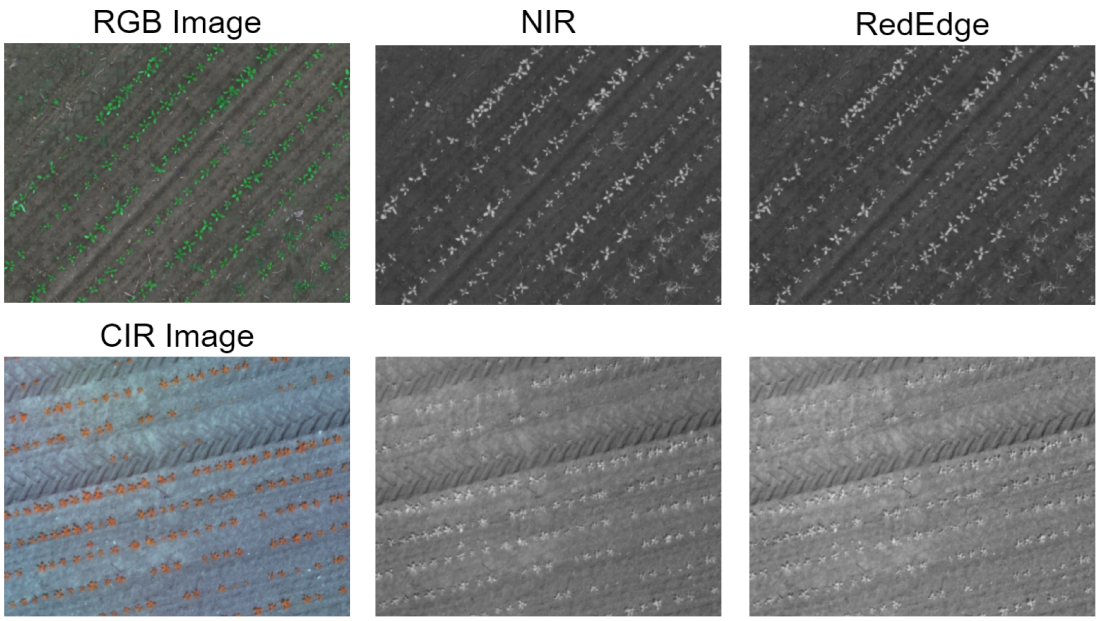


Figure 31 – Example of NIR and RedEdge multispectral bands from the RedEdge and Sequoia datasets.

5.2 Deep learning weed segmentation

5.2.1 Training Process

To address the class imbalance inherent in the weed segmentation problem, the *CrossEntropyLoss* function was employed (MAO; MOHRI; ZHONG, 2023). This loss function is widely used in segmentation tasks as it measures the divergence between the predicted probability distribution (p_{y_i}) and the ground truth labels (y_i):

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N} w_{y_i} \log(p_{y_i}) \tag{7}$$

where N is the number of examples in the batch, w_{y_i} is the weight assigned to class y_i, and p_{y_i} represents the predicted probability for the correct class y_i. This approach penalizes errors in underrepresented classes, ensuring that the model does not disproportionately favor majority classes. We computed the class weights as the inverse of their frequencies in dataset, using the formula:

$$w_c = \frac{1}{f_c} \tag{8}$$

where f_c is the relative frequency of class c. This ensures that less frequent classes, such as weeds, are given greater importance during training. Class frequencies were computed as:

$$f_c = \frac{\text{number of pixels in class } c}{\text{total number of pixels in the dataset}}. \tag{9}$$

The AdamW optimizer was employed due to its ability to dynamically adjust learning rates while applying effective weight regularization, reducing the risk of overfitting. This choice enhances model generalization, especially in scenarios with significant data variability, such as agricultural multispectral imagery. The learning rate and weight decay were set to 10^{-4}, providing a balance between convergence stability and speed (LOSHCHILOV; HUTTER et al., 2017). Each test involved fine-tuning pre-trained models on ImageNet for 100 epochs. All models were trained on a computer equipped with an NVIDIA GeForce RTX 4060 Ti GPU with 8GB of dedicated video memory. During the training of models using early fusion, batch sizes of 20 were used for DeepLabV3+ and 10 for SegFormer. For late fusion, a batch size of 6 was used.

A $70\% - 30\%$ split was used for training and test datasets. Given the significant class imbalance in the dataset, a data balancing strategy was adopted during training. This approach assigns a weight to each sample in the dataset, proportional to its relevance or the need to be sampled more frequently. These weights were used to determine the probability of each sample being selected in each iteration. This ensured that the model learned patterns from all classes more evenly, reducing the bias toward majority classes and improving performance in highly imbalanced tasks.

Augmentations were applied to the input data with a probability of 50% for each, including random resized crops (scaling from 80% to 100%), rotations (ranging from $-90°$ to $90°$), horizontal and vertical flips, and affine transformations (scaling from 90% to 110%, rotation from $-30°$ to $30°$, and shear from -10 to 10). These operations increased the variability of the dataset, mitigating the risk of overfitting and improving the model's robustness in real-world field conditions, as described by (KUMAR et al., 2024).

5.2.2 Early Fusion

Based on the evaluation of the OI obtained for vegetation indices and color spaces, we selected the following features: RGB, YUV_U, LUV_U, LUV_V, LAB_B, and YPbPr_Pr to represent the color spaces, and ExB, TGI, ExG, ExR, SR, and NGRDI for the vegetation indices. We combined features as demonstrated in Table 10.

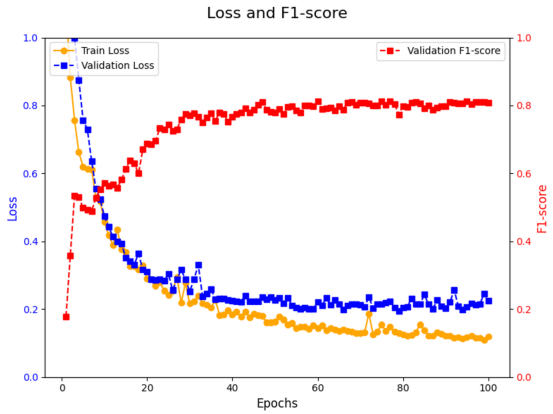
Note that for the Sugarcane dataset, we replaced the vegetation index SR by NGRDI since SR requires multispectral channels for its calculation. Additionally, we didn't test the Multispectral combination on the Sugarcane dataset as it is RGB-only. Another important aspect to note is the use of RGB_B for the Sequoia dataset; this band was estimated as described in the calculation of the color spaces. Each of these combinations was evaluated using the DeepLabV3+ model (with ResNet18 encoder) and the SegFormer model (with MiT-B1 encoder). Table 11 shows the results obtained, and Figure 32 presents the loss and F1-score curves for the best combinations of each model and dataset. As shown in Figure 32, for the Sequoia dataset, the validation loss graph exhibits significant peaks at each epoch, contrasting with the smoothly decreasing curve observed

Table 10 – Combination of inputs for each dataset.

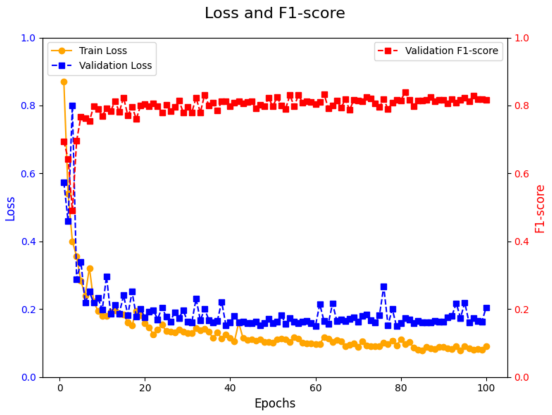
Dataset	Combination Name	Input Features
Sugarcane	RGB	RGB_R, RGB_G, RGB_B
	VI	ExB, TGI, ExG, ExR, NGRDI
	Color Space	YUV_U, LUV_U, LUV_V, LAB_B, YPbPr_Pr
RedEdge	RGB	RGB_R, RGB_G, RGB_B
	Multispectral	RGB_R, RGB_G, RGB_B, NIR, RE
	VI	ExB, TGI, ExG, ExR, SR
	Color Space	YUV_U, LUV_U, LUV_V, LAB_B, YPbPr_Pr
Sequoia	RGB	ExB, TGI, ExG, ExR, SR
	Multispectral	RGB_R, RGB_G, RGB_B, NIR, RE
	VI	YUV_U, LUV_U, LUV_V, LAB_B, YPbPr_Pr
	Color Space	RGB_R, RGB_G, RGB_B

during training. This discrepancy indicates that the validation data exhibit high variability and are not sufficiently representative of the total sample distribution, resulting in pronounced divergences between epochs. This issue aligns with the dataset's distribution, where 99.03% of the pixels belong to the background, leaving only 0.94% as crops and just 0.01% as weeds. In comparison, the RedEdge dataset contains 88.65% background, 5.87% crops, and 5.46% weeds, demonstrating a more balanced class distribution.

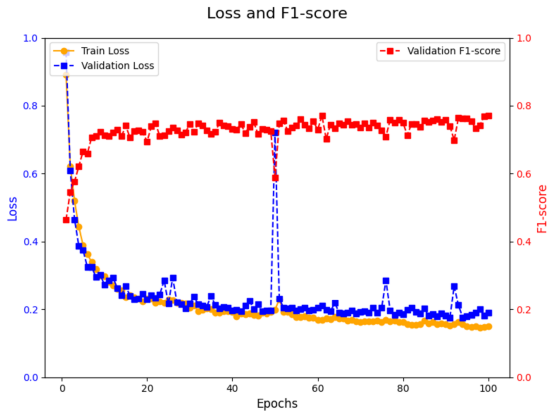
As shown in Table 11, SegFormer achieved superior performance across all evaluated datasets and input combinations. Even when analyzing the results per class, as demonstrated in Table 12, SegFormer outperformed the other models, achieving excellent results even in complex classes such as Weed and Crop. The superior performance of SegFormer over DeepLabv3+ for segmenting background, weeds, and crops can be attributed to its architectural advancements tailored for dense prediction tasks, which are particularly useful in complex agricultural scenarios. SegFormer leverages hierarchical attention mechanisms that enable the extraction of multi-scale contextual information, which is critical for distinguishing fine-grained differences between weeds, crops, and the background. Unlike DeepLabv3+, which relies on pre-defined receptive fields via dilated convolutions, SegFormer dynamically adapts its feature extraction process to capture both local and global patterns across spatial scales. Additionally, the transformer-based architecture of SegFormer excels in modeling long-range dependencies, providing a robust understanding of spatial relationships within the imagery. This capability is especially



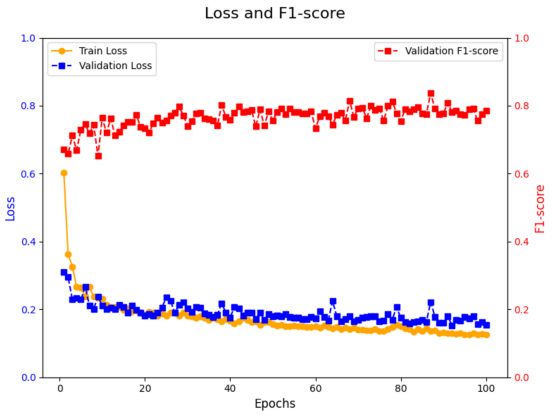
Sugarcane DeepLabV3+ RGB



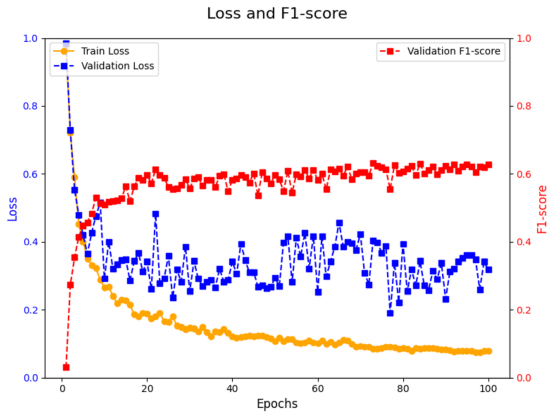
Sugarcane SergFormer RGB



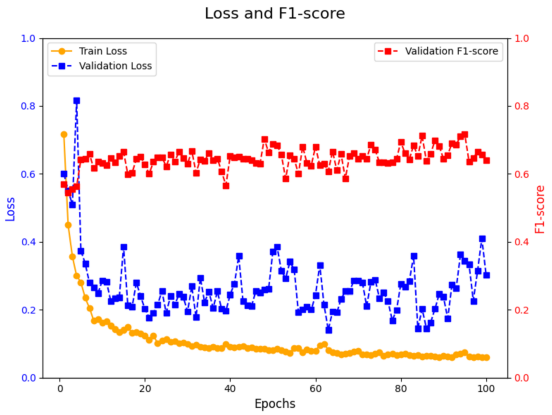
RedEdge DeepLabV3+ VI



RedEdge SegFormer RGB



Sequoia DeepLabV3+ VI



Sequoia SegFormer VI

Figure 32 – Traning loss, validation loss and F1-score curves for the best combinations of each model and dataset.

Table 11 – Results obtained for each combination of dataset, input and model.

Dataset	Model	Combination	Accuracy	Precision	Recall	F1
Sugarcane	**SegFormer**	**RGB**	**0.957**	**0.769**	**0.941**	**0.839**
		VI	0.951	0.750	0.933	0.822
		Color Space	0.953	0.751	0.916	0.818
	DeepLabV3+	**RGB**	0.947	0.739	0.931	0.813
		VI	0.949	0.742	0.922	0.813
		Color Space	0.948	0.742	0.900	0.804
RedEdge	**SegFormer**	**RGB**	**0.965**	**0.781**	0.913	**0.838**
		Multispectral	0.963	0.767	0.927	0.831
		VI	0.961	0.760	0.922	0.827
		Color Space	0.959	0.752	0.929	0.823
	DeepLabV3+	**VI**	0.941	0.685	**0.932**	0.772
		Color Space	0.941	0.685	0.927	0.771
		RGB	0.939	0.679	0.929	0.766
		Multispectral	0.938	0.676	0.929	0.764
Sequoia	**SegFormer**	**VI**	**0.984**	**0.622**	0.909	**0.716**
		Multispectral	0.983	0.613	0.900	0.706
		Color Space	0.981	0.603	0.862	0.689
		RGB	0.981	0.590	0.890	0.682
	DeepLabV3+	**VI**	0.974	0.542	0.882	0.632
		Multispectral	0.975	0.538	**0.937**	0.632
		RGB	0.969	0.531	0.900	0.622
		Color Space	0.974	0.531	0.904	0.621

advantageous in datasets such as RedEdge and Sequoia, where spectral overlaps and small object footprints demand high precision.

Analyzing the SegFormer model, the best combinations were RGB for the Sugarcane and RedEdge datasets, and VI for the Sequoia dataset. Analyzing the strong performance of RGB, we observed that both datasets (Sugarcane and RedEdge) show differences in the growth stages of plants. For example, in the RedEdge dataset, crops are at a more advanced growth stage (15–20 cm) compared to weeds (5–10 cm), and the RGB combination provides clear contrasts in color and morphology, making it sufficient for effective segmentation. Similarly, in the Sugarcane dataset, the weeds are in a more advanced growth stage, further enhancing the effectiveness of RGB for segmentation. However, when analyzing the Sequoia dataset, two main factors contribute to the lower performance of RGB: the absence of the blue spectrum and the early growth stage (crops 8–10 cm and weeds 5–10 cm). These factors render an RGB image, where the blue band is approximated us-

Table 12 – F1 score per class obtained for each combination of dataset, input, and model. BG stands for background.

Dataset	Model	Combination	F1	F1 BG	F1 Crop	F1 Weed
SugarCane	**SegFormer**	**RGB**	**0.839**	**0.977**	**0.786**	**0.753**
		VI	0.822	0.974	0.754	0.737
		Color Space	0.818	0.976	0.767	0.709
	DeepLabV3+	**RGB**	0.813	0.972	0.724	0.744
		VI	0.813	0.973	0.740	0.725
		Color Space	0.804	0.973	0.723	0.716
RedEdge	**SegFormer**	**RGB**	**0.838**	**0.983**	0.798	**0.733**
		Multispectral	0.831	0.981	**0.802**	0.711
		VI	0.827	0.980	0.774	0.726
		Color Space	0.823	0.979	0.770	0.720
	DeepLabV3+	**VI**	0.772	0.969	0.701	0.645
		Color Space	0.771	0.969	0.694	0.649
		RGB	0.766	0.968	0.701	0.628
		Multispectral	0.764	0.968	0.693	0.630
Sequoia	**SegFormer**	**VI**	**0.716**	**0.992**	**0.664**	**0.494**
		Multispectral	0.706	**0.992**	0.663	0.463
		Color Space	0.689	0.990	0.624	0.452
		RGB	0.682	0.990	0.627	0.430
	DeepLabV3+	**VI**	0.632	0.987	0.550	0.360
		Multispectral	0.632	0.987	0.575	0.332
		RGB	0.622	0.984	0.507	0.375
		Color Space	0.621	0.987	0.575	0.300

ing other bands, insufficient for precise segmentation. In this scenario, vegetation indices such as ExB, TGI, ExG, ExR, and SR enhance the distinction by emphasizing subtle differences in reflectance related to plant vigor, leaf structure, and pigmentation. These indices reduce the ambiguity caused by smaller crop and weed footprints by accentuating physiological traits and vegetation health that are not easily discernible through raw bands alone.

5.2.2.1 Qualitative Analysis

Analyzing Figure 33 and Figure 34, we observe differences in segmentation quality between SegFormer and DeepLabv3+ for the RedEdge and Sequoia datasets, respectively. As shown in the images, SegFormer consistently produces finer and more accurate segmentation shapes, while DeepLabv3+ generates coarser edges and exhibits a higher

tendency to misclassify areas, particularly in regions with spectral ambiguity or overlapping features. SegFormer leverages transformer-based attention mechanisms, which capture long-range dependencies across the image, enabling superior differentiation between crops, weeds, and background. This approach is particularly advantageous in agricultural applications, where subtle distinctions in shape, color, and texture are critical for accurate segmentation. The hierarchical multi-scale representations of SegFormer allow it to effectively model objects at different spatial resolutions, ensuring high precision in segmenting small objects such as weeds alongside larger crops. This is especially evident in the Sequoia dataset, where plants are at an earlier growth stage.

In contrast, DeepLabv3+ relies on atrous convolutions to expand its receptive field, which improves its ability to capture global context but compromises the refinement of finer details. The model's dependence on fixed convolutional filters limits its ability to adapt dynamically to complex scenes, leading to coarser boundaries and inaccuracies, particularly at the intersections of crops, weeds, and background.

In the Sequoia dataset, the use of vegetation indices such as ExG, ExB, TGI, and SR amplifies SegFormer's advantages. These indices highlight physiological and spectral differences between crops and weeds, which the attention mechanisms in SegFormer effectively integrate for enhanced segmentation accuracy. DeepLabv3+, however, struggles to fully leverage these indices, as its convolutional architecture does not adapt as well to variations in spectral and spatial properties, leading to less precise segmentation results.

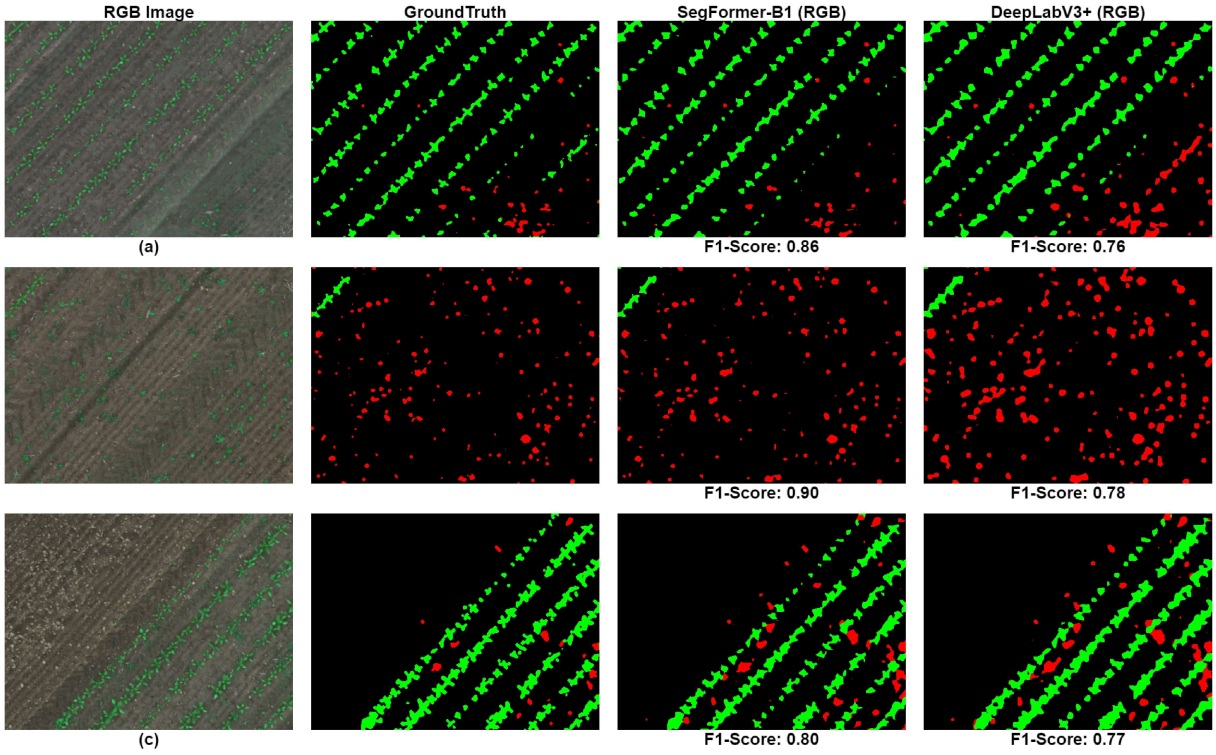


Figure 33 – Results obtained for the RedEdge dataset. Black represents the background, green represents crops, and red represents weeds.

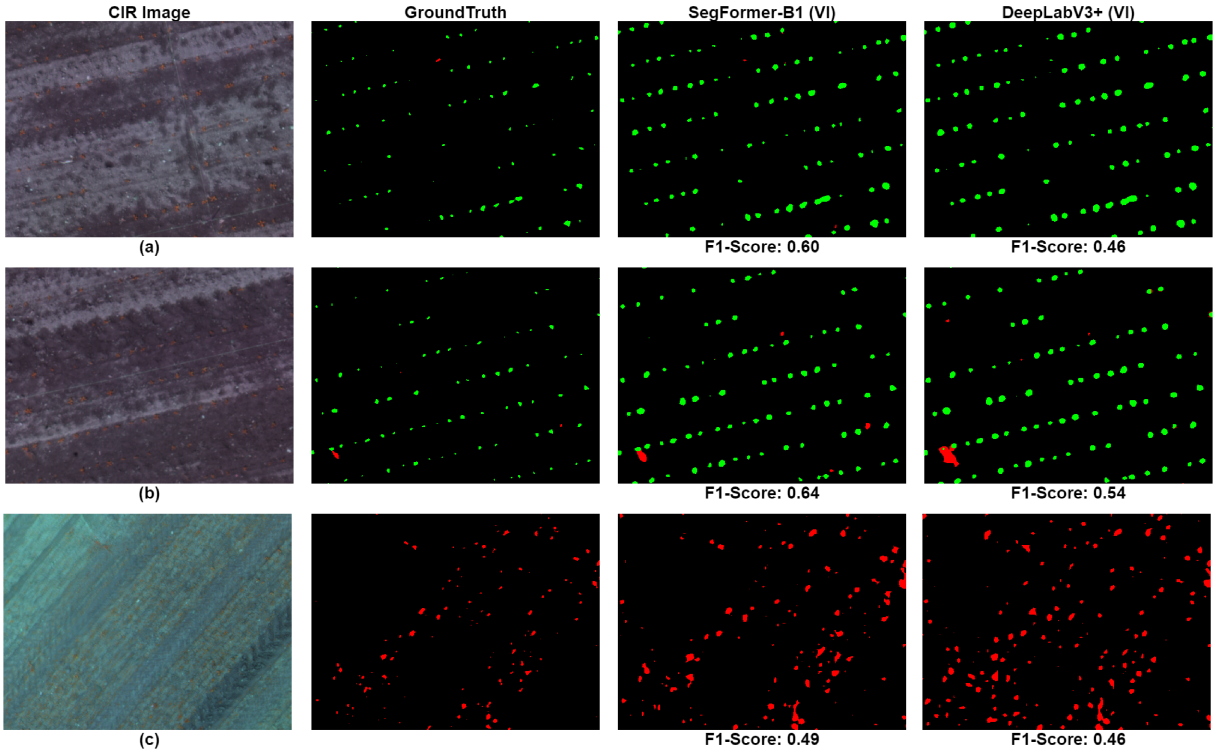


Figure 34 – Results obtained for the Sequoia dataset. Black represents the background, green represents crops, and red represents weeds.

In the case of the Sugarcane dataset, as shown in Figure 35, both SegFormer and DeepLabv3+ exhibited challenges in producing well-defined boundaries during segmentation. This limitation can be attributed to the characteristics of the ground truth annotations provided in the dataset. Unlike datasets with meticulously curated pixel-to-pixel segmentation, the Sugarcane dataset appears to lack precise annotations that align perfectly with object boundaries at a pixel level. Such discrepancies in the ground truth introduce ambiguity during model training, as the models attempt to learn from imprecise or inconsistent labels. This highlights the importance of the quality of ground truth annotations in the performance of segmentation models. Even in this scenario, we observe that SegFormer better approximates the ground truth, achieving superior average results.

5.2.2.2 Analysis of Training Epochs

For the best results shown in Table 11, we investigated whether doubling the number of training epochs, from 100 to 200, would have an impact on the final outcome. As shown in Table 13, training with twice the number of epochs did not significantly improve the results, achieving, at best, an increase of only 0.014 in the F1-score. The additional training was also insufficient to alter the best combination for each dataset, with RGB remaining the optimal combination for the Sugarcane and RedEdge datasets, and VI for the Sequoia dataset.

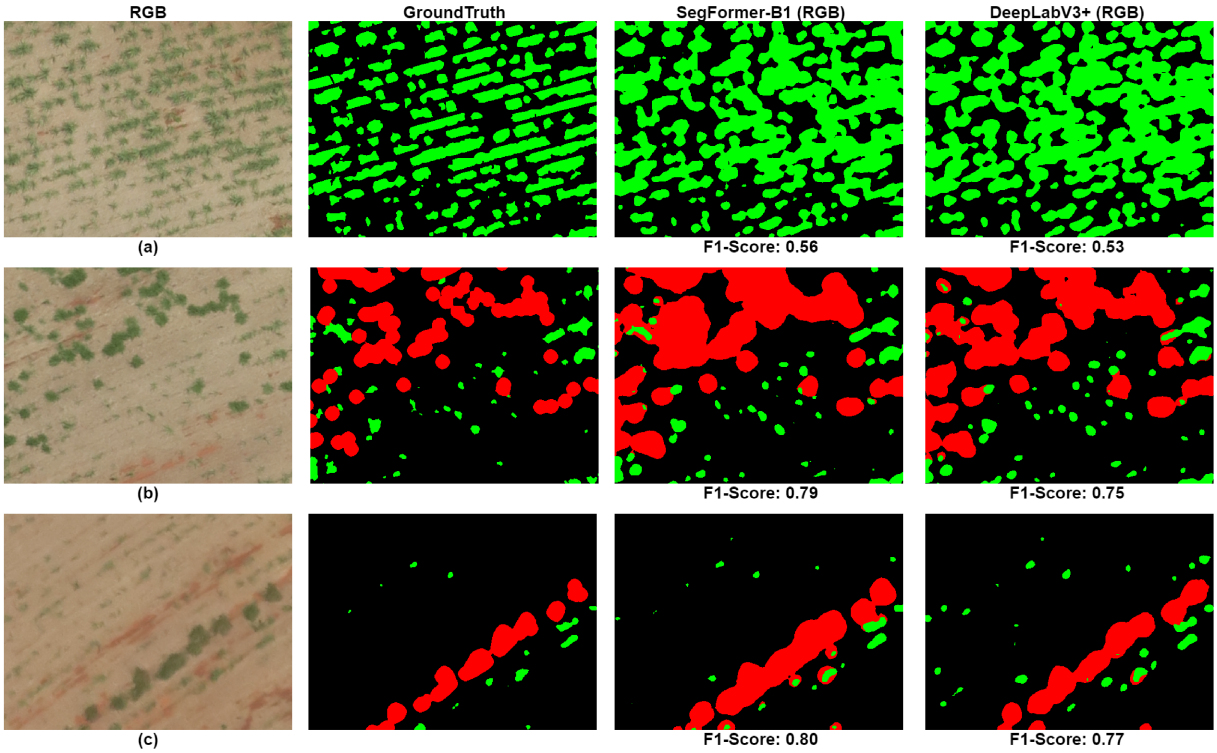


Figure 35 – Results obtained for the Sugarcane dataset. Black represents the background, green represents crops, and red represents weeds.

Table 13 – Results obtained for the best combinations of model, input, and dataset when trained for 200 epochs.

Dataset	Model	Combination	F1-score 100 epochs	F1-score 200 epochs	Delta
Sugarcane	SegFormer	RGB	0.839	**0.842**	0.003
	SegFormer	VI	0.822	0.836	**0.014**
	SegFormer	Color Space	0.818	0.825	0.008
RedEdge	SegFormer	RGB	**0.838**	0.838	-
	SegFormer	Multispectral	0.831	0.831	-
	SegFormer	VI	0.827	0.833	**0.007**
Sequoia	SegFormer	VI	0.716	**0.728**	**0.012**
	SegFormer	Multispectral	0.706	0.708	0.003
	SegFormer	Color Space	0.689	0.689	-

5.2.3 Late Fusion

As described in Section 4.3, we also evaluated late fusion combinations of RGB images, multispectral data, VIs, and color spaces for weed segmentation. Table 14 presents the combinations analyzed in these experiments.

As previously explained in Section 5.2.2, the vegetation index SR was replaced by

Table 14 – Input combination for each encoder in each dataset.

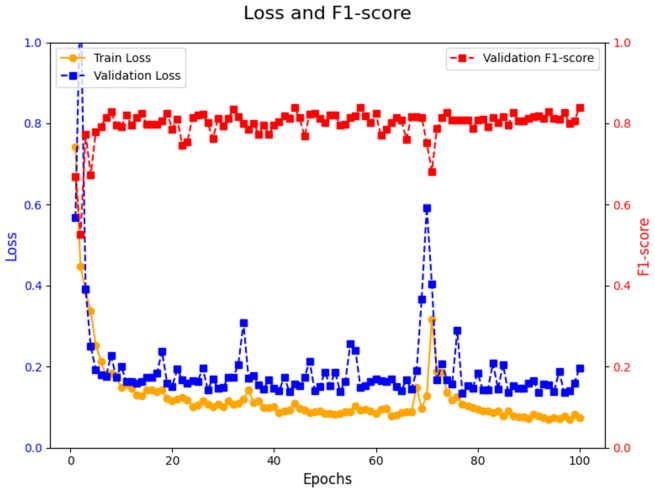
Dataset	Combination Name	Encoder 01 Input Features	Encoder 02 Input Features
Sugarcane	**RGB VI**	RGB_R, RGB_G, RGB_B	ExB, TGI, ExG, ExR, NGRDI
	RGB Color Space	RGB_R, RGB_G, RGB_B	YUV_U, LUV_U, LUV_V, LAB_B, YPbPr_Pr
RedEdge	**RGB VI**	RGB_R, RGB_G, RGB_B	ExB, TGI, ExG, ExR, SR
	RGB Color Space	RGB_R, RGB_G, RGB_B	YUV_U, LUV_U, LUV_V, LAB_B, YPbPr_Pr
	RGB Multispectral	RGB_R, RGB_G, RGB_B	NIR, RE
Sequoia	**RGB VI**	RGB_R, RGB_G, RGB_B	ExB, TGI, ExG, ExR, SR
	RGB Color Space	RGB_R, RGB_G, RGB_B	YUV_U, LUV_U, LUV_V, LAB_B, YPbPr_Pr
	RGB Multispectral	RGB_R, RGB_G, RGB_B	NIR, RE

NGRDI since SR requires multispectral channels for its calculation. The Multispectral combination was not tested on the Sugarcane dataset due to its RGB-only nature, and the RGB_B band was used for the Sequoia dataset. Note that the vegetation indices and color spaces evaluated as inputs for encoder 02 were those that achieved the best OI. Table 15 and Table 16 present the results obtained for this approach, while Figure 36 shows the loss and F1-score curves for the best combinations of each model and dataset. As highlighted in the early fusion approach, a similar training pattern is observed for the Sequoia dataset.

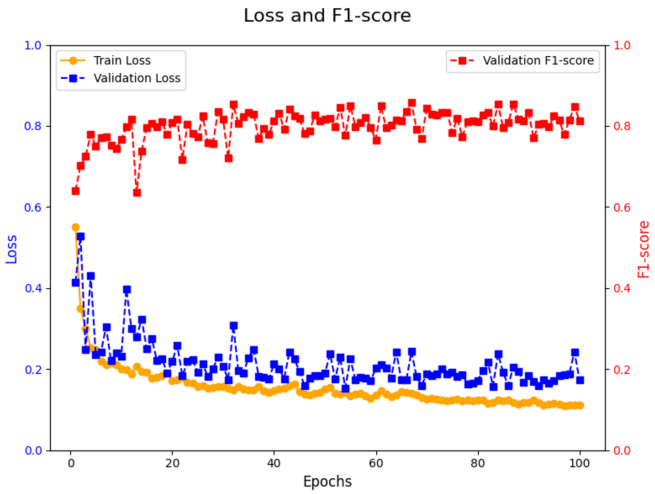
Table 15 – Results obtained in late fusion approach for each combination of dataset, input and model.

Dataset	Model	Combination	Accuracy	Precision	Recall	F1
Sugarcane	**SegFormer**	**RGB VI**	**0.959**	**0.775**	0.935	**0.840**
	SegFormer	**RGB Colorspace**	0.958	0.773	**0.936**	0.839
RedEdge	**SegFormer**	**RGB VI**	**0.971**	**0.818**	0.906	**0.857**
	SegFormer	**RGB Colorspace**	0.970	0.807	**0.922**	0.856
	SegFormer	**RGB Multispectral**	0.969	0.804	0.911	0.850
Sequoia	**SegFormer**	**RGB Colorspace**	0.987	**0.661**	0.859	**0.736**
	SegFormer	**RGB Multispectral**	**0.988**	0.653	0.883	**0.736**
	SegFormer	**RGB VI**	0.986	0.642	**0.901**	0.732

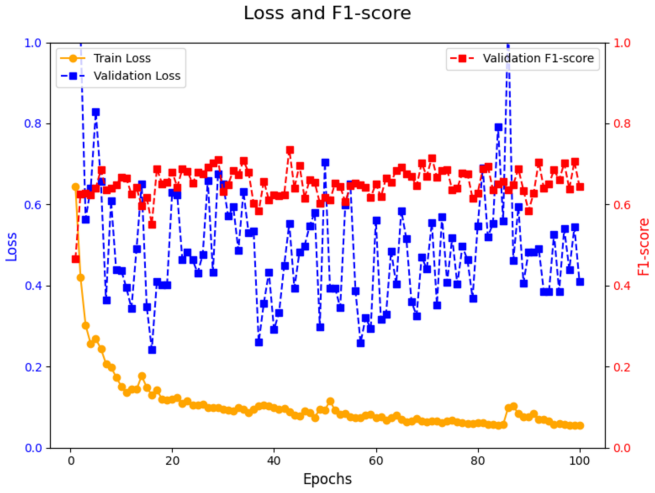
The adoption of late fusion in the SegFormer architecture demonstrated consistent improvements across all datasets and input combinations when compared to the previous approach, which utilized early fusion. By leveraging a structured integration of complementary input sources, late fusion proved to be a robust and efficient strategy for weed



Sugarcane Late SegFormer-B1 (RGB VI)



RedEdge Late SegFormer-B1 (RGB VI)



Sequoia Late SegFormer-B1 (RGB Color Space)

Figure 36 – Traning loss, validation loss and F1-score curves for the best combinations of each model and dataset.

Table 16 – Results per class obtained in the late fusion approach for each combination of dataset, input, and model. BG stands for background.

Dataset	Model	Combination	F1	F1 BG	F1 Crop	F1 Weed
Sugarcane	**SegFormer**	**RGB VI**	**0.840**	**0.979**	**0.776**	0.764
	SegFormer	**RGB Colorspace**	0.839	0.978	0.765	**0.774**
RedEdge	**SegFormer**	**RGB VI**	**0.857**	**0.986**	0.831	**0.755**
	SegFormer	**RGB Colorspace**	0.856	0.985	**0.849**	0.734
	SegFormer	**RGB Multispectral**	0.850	0.985	0.836	0.730
Sequoia	**SegFormer**	**RGB Colorspace**	**0.736**	0.993	0.698	**0.517**
	SegFormer	**RGB Multispectral**	**0.736**	**0.994**	**0.727**	0.487
	SegFormer	**RGB VI**	0.732	0.993	0.702	0.499

segmentation in agricultural scenarios.

For the RedEdge dataset, late fusion achieved notable improvements in segmentation performance. The RGB and VI combination achieved the highest F1-score of 0.857, a significant improvement over the previous best score of 0.838 using RGB-only inputs. This highlights the advantage of combining vegetation indices with RGB data, which enhances the model's ability to distinguish weeds from crops in RedEdge imagery. Additionally, the RGB and Colorspace combination achieved an F1-score of 0.856, underscoring the utility of alternative color spaces in segmentation tasks. The RGB and Multispectral combination also showed strong results, achieving an F1-score of 0.850, compared to the previous 0.831 for multispectral-only inputs. These results confirm that late fusion effectively integrates the spectral and spatial information present in RedEdge imagery, leading to enhanced segmentation performance.

The Sequoia dataset, which traditionally exhibited lower performance compared to other datasets, also benefited from late fusion. The RGB and Colorspace combination, as well as the RGB and Multispectral combination, both achieved an F1-score of 0.736, surpassing the previous best score of 0.716 obtained with VI-only inputs. This result reflects the strength of late fusion in handling challenging spectral data by combining diverse modalities effectively. Furthermore, the RGB and VI combination achieved an F1-score of 0.732, showing moderate improvements over the previous approach. These findings highlight how late fusion enables the network to extract complementary features, even in more complex datasets like Sequoia.

When compared to the early fusion approach, the late fusion strategy consistently outperformed it across all datasets. The results demonstrated higher F1-scores in most scenarios, underscoring the effectiveness of late fusion in integrating complementary data sources such as RGB, vegetation indices, color spaces, and multispectral bands. Unlike early fusion, where data is concatenated at the input level, late fusion allows for independent feature extraction in separate encoders. This approach preserves the unique

characteristics of each modality while enabling their meaningful combination in the final stages of the network.

5.2.3.1 Qualitative Analysis

As shown in Figures 37, 38, and 39, the use of late fusion considerably improved segmentation performance on average. For the Sequoia dataset, the segmentation results were finer, with less noise and greater accuracy in detecting weeds. In Figure 38-b, weeds that were previously misclassified disappear, resulting in superior segmentation.

Analyzing the Sugarcane dataset, the same issue observed in early fusion persists, demonstrating a limitation due to annotations not being at the pixel level. However, a noticeable improvement can be seen in Figure 39-c, with a significant reduction in false positives for weeds, particularly around the crops. This same pattern can also be observed in Figure 39.

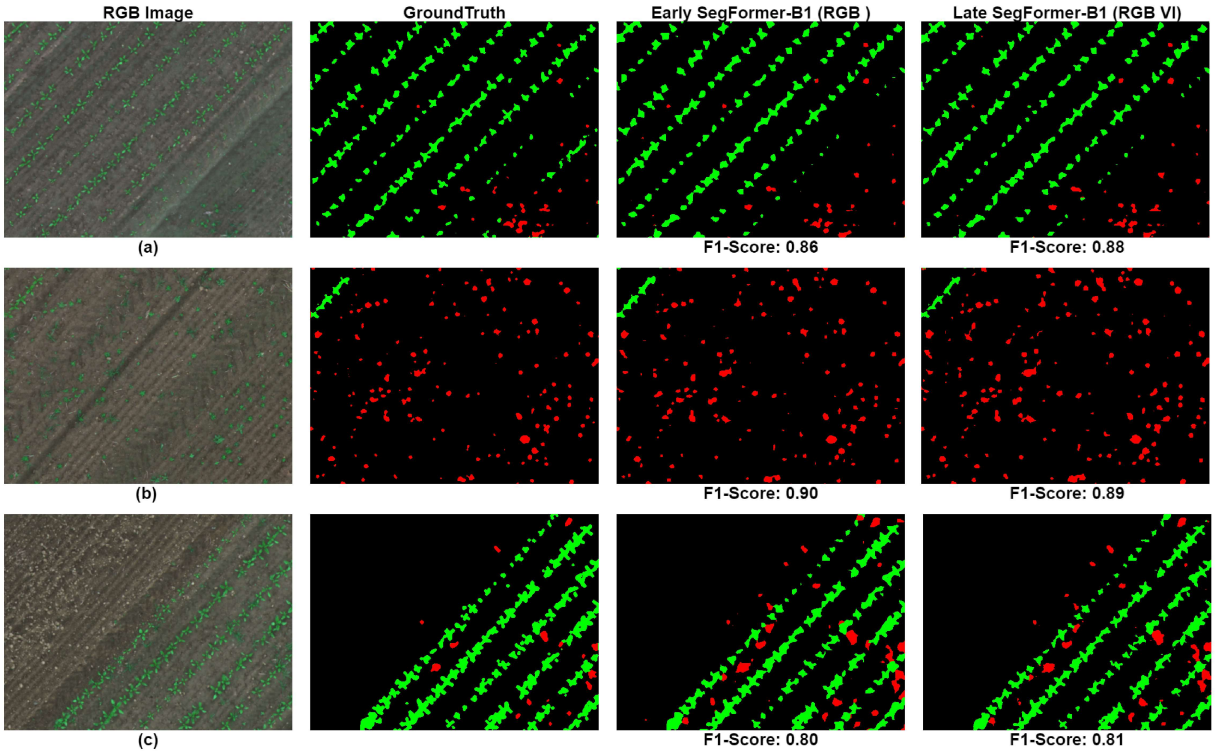


Figure 37 – Results obtained for the RedEdge dataset. Black represents the background, green represents crops, and red represents weeds.

5.2.4 Comparison with the State of the Art

As presented in Table 17, our proposed methodology demonstrates high efficiency, outperforming several state-of-the-art approaches across all datasets. A particularly notable result is observed in the Sequoia dataset, which, as described in Section 4.1, posed the greatest challenge for the segmentation process due to the similarity in growth stages

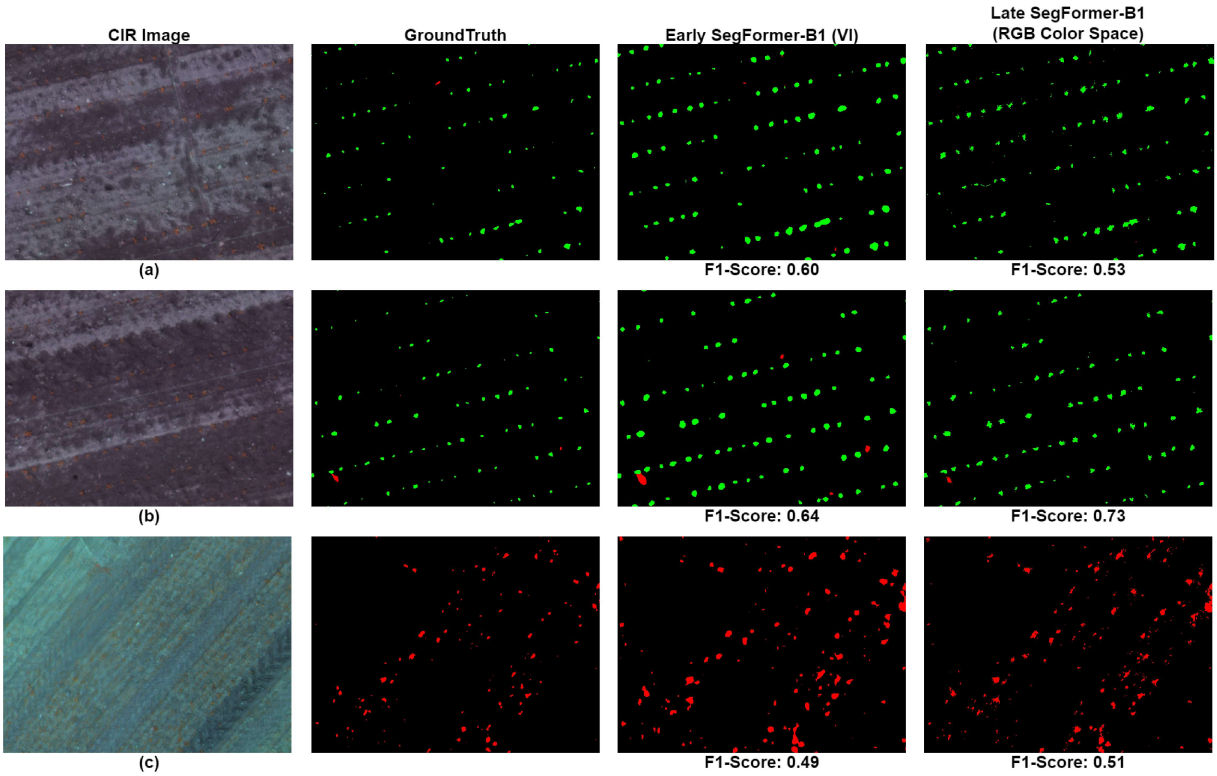


Figure 38 – Results obtained for the Sequoia dataset. Black represents the background, green represents crops, and red represents weeds

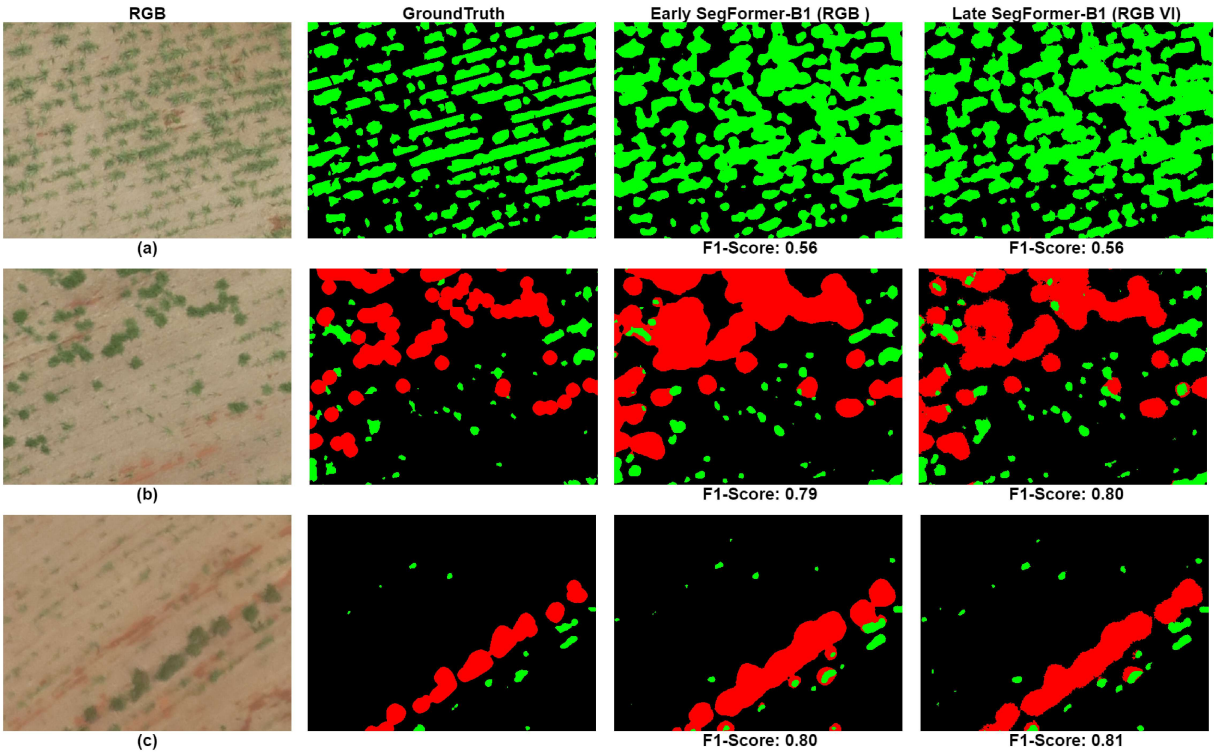


Figure 39 – Results obtained for the Sugarcane dataset. Black represents the background, green represents crops, and red represents weeds

between crops and weeds. Despite these difficulties, our method achieved an F1-score of

0.52 for the weed class, representing an improvement of over 270% compared to other methods. This result underscores the strength of our model in addressing challenging cases with significant spectral and morphological overlaps between vegetation classes.

For the RedEdge dataset, our approach maintained consistent performance, achieving competitive F1-scores across all classes when compared to other methods in the literature. Notably, our method achieved the highest overall F1-score of 0.86 and matched or surpassed state-of-the-art results in background and crop segmentation. This demonstrates the robustness and adaptability of our methodology across datasets with varying levels of complexity, reinforcing its potential applicability in diverse agricultural contexts.

Table 17 – Comparison of our method with the state of the art. BG stands for Background.

Dataset	Method	F1	F1 BG	F1 Crop	F1 Weed
RedEdge	(ours) Late SegFormer VI	**0.86**	**0.99**	0.83	0.76
	(RAMIREZ et al., 2020)	0.84	0.92	0.81	**0.78**
	(CASTELLANO; MARINIS; VESSIO, 2023a)	**0.86**	**0.99**	**0.86**	0.74
	(CASTELLANO; MARINIS; VESSIO, 2023b)	**0.86**			
	(MARINIS; VESSIO; CASTELLANO, 2024)	0.75	0.98	0.71	0.57
Sequoia	(ours) Late SegFormer Color Space	**0.74**	**0.99**	0.70	**0.52**
	(CASTELLANO; MARINIS; VESSIO, 2023a)	0.63	**0.99**	**0.76**	0.14
	(CASTELLANO; MARINIS; VESSIO, 2023b)	0.69			

The results obtained for the Sugarcane dataset, as shown in Table 18, clearly demonstrate the superiority of the proposed method compared to the state of the art. In terms of overall accuracy, our method achieves 0.935, significantly outperforming the 0.631 obtained by FRSEN. This substantial difference highlights the ability of our model to effectively address the inherent complexities of multispectral segmentation in agricultural fields.

The class-specific accuracies further emphasize this advantage. For the background class, Late SegFormer VI achieves an accuracy of 0.961, which is comparable to the 0.992 obtained by FRSEN. While slightly lower in this category, this performance is more than compensated by the significantly better results in the other classes. For crop identification, our model reaches an accuracy of 0.983, greatly surpassing the 0.763 achieved by FRSEN. The weed class also shows remarkably better results, achieving an accuracy of 0.862.

These results underscore the effectiveness of transformer-based models, such as SegFormer, especially when adapted with fusion techniques. The high precision in distinguishing crops from weeds has significant implications for sustainable agriculture, enabling more accurate herbicide applications, reducing environmental impact, and optimizing productivity. The performance achieved by Late SegFormer VI demonstrates its practical applicability and robustness for large-scale segmentation tasks in complex agricultural scenarios, as exemplified by the datasets.

Table 18 – Comparison of our method with the state of the art for the Sugarcane dataset. BG stands for Background.

Dataset	**Method**	**Accuracy**	**Accuracy BG**	**Accuracy Crop**	**Accuracy Weed**	**F1**
Sugarcane	**(ours) Late SegFormer VI**	**0.935**	0.961	**0.983**	**0.862**	**0.840**
Sugarcane	**(HASHEMI-BENI et al., 2022)**	0.631	**0.992**	0.763	0.138	

Chapter 6

Conclusion

Weed segmentation in RGB or multispectral images obtained by UAV is a complex task due to numerous factors influencing plant reflectance and image capture conditions. Variations in plant health cause changes in biochemical composition, directly affecting spectral responses. This thesis addresses these challenges by proposing a segmentation methodology based on transformers, leveraging advanced attention mechanisms and multi-scale contextual features to achieve high segmentation precision across diverse datasets and conditions. The methodology integrates color spaces, vegetation indices, and multispectral information, with early and late fusion strategies, enabling robust weed segmentation.

The experimental results demonstrate the superiority of the proposed approach. Initial analyzes revealed the importance of RGB and alternative color spaces in differentiating crops, weeds, and backgrounds. Components such as YUV_U, LAB_B, and LUV_V were particularly effective, achieving high OI values due to their ability to capture chrominance variations while minimizing sensitivity to lighting changes. Vegetation indices, including ExB, ExG, and SR, further enhanced segmentation by highlighting vegetation-specific spectral features related to chlorophyll content and biomass. Multispectral bands, such as NIR and RE, provided additional precision by capturing reflectance variations in vegetation structure and water content.

Deep learning models, particularly SegFormer, consistently outperformed traditional convolutional architectures such as DeepLabV3+ across all datasets. SegFormer achieved superior F1-scores due to its transformer-based architecture, which excels at capturing long-range dependencies and multi-scale contextual information. Additionally, when considering SegFormer alone, late fusion techniques outperformed early fusion across all datasets. For the Sugarcane and RedEdge datasets, the RGB VI combination proved to be the most effective input combination, with SegFormer achieving an F1-score of 0.840 and 0.857, respectively. The Sequoia dataset, characterized by greater spectral and morphological overlap between crops and weeds, demonstrated the utility of color spaces in late fusion, achieving an F1-score of 0.736. These results underscore the importance of combining spectral, spatial, and contextual information for precise segmentation.

Comparisons with state-of-the-art methods further validate the effectiveness of the proposed approach. For the Sequoia dataset, the methodology achieved an F1-score of 0.52 for the weed class, a 270% improvement over previous approaches, demonstrating the robustness of the proposed framework in handling challenging cases (average F1-score of 0.74). For the RedEdge dataset, our method achieved a result similar to other works with an F1-score of 0.86, highlighting its adaptability to diverse agricultural scenarios. The results of the Sugarcane dataset also exceeded existing methods, with our approach achieving an overall accuracy of 93.5%, significantly outperforming other models.

Qualitative analysis further confirmed the advantages of transformer-based models. SegFormer consistently produced finer segmentation results, with fewer misclassifications and greater boundary precision, even in datasets with imprecise annotations. The integration of complementary data sources through late fusion proved particularly effective, improving model performance by preserving the unique characteristics of each modality while leveraging their synergy.

This thesis presents a robust and adaptable framework for weed segmentation in agricultural fields, demonstrating the efficacy of transformer-based models combined with advanced fusion strategies. The proposed methodology not only achieves state-of-the-art performance but also provides valuable insight into the spectral and spatial features most relevant for weed segmentation. These findings have significant implications for precision agriculture, enabling more accurate weed management, optimizing resource allocation, and reducing environmental impact.

6.1 Future Work

Based on the findings of this study, several directions for future research can be explored to further enhance the effectiveness and applicability of weed segmentation using transformer-based deep learning models.

One promising avenue is the expansion of the dataset to include a wider variety of crop types, weed species, and environmental conditions. Incorporating diverse agricultural landscapes, different soil compositions, and seasonal variations would improve the generalization capability of the model, making it more robust for large-scale deployments. Furthermore, integrating hyperspectral imaging instead of multispectral data could provide richer spectral information, potentially improving the model's ability to distinguish between crops and weeds with similar spectral signatures.

Another important research direction is optimizing transformer-based models for real-time applications. Despite their superior performance in capturing long-range dependencies, transformers are computationally expensive. Future studies should explore efficient model compression techniques, such as quantization, pruning, and knowledge distillation, to reduce inference time while maintaining segmentation accuracy. This would facilitate

the deployment of such models on edge devices or UAVs, enabling real-time detection and management of weeds.

Incorporating temporal analysis through multi-temporal UAV imagery is another promising direction. By analyzing weed growth over different phenological stages, the models could provide more accurate predictions and track the effectiveness of weed control measures over time. This would allow for a more dynamic and adaptive weed management strategy, reducing herbicide usage while maximizing crop yield.

Finally future research should also focus on integrating the proposed segmentation approach with autonomous precision agriculture systems. By linking weed segmentation outputs with robotic weed removal mechanisms or variable-rate herbicide sprayers, fully automated weed control solutions could be developed.

These future research directions aim to refine and extend the current work, pushing the boundaries of transformer-based weed segmentation and contributing to the advancement of precision agriculture technologies.

6.2 Contributions in Bibliographic Production

The publications carried out during the doctorate period are presented below:

- DIAS JUNIOR, J. D.; RIBEIRO, J. B. ; BACKES, ANDRÉ RICARDO . Assessing the impact of JPEG compression on the semantic segmentation of agricultural images. Signal Image and Video Processing, v. 1, p. 1, 2023.

- DIAS JÚNIOR, JOCIVAL; BACKES, ANDRÉ . Unsupervised Segmentation of Leukocytes Images using Particle Swarm. In: 16th International Conference on Computer Vision Theory and Applications, 2021, Online Streaming. Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, 2021. v. 4. p. 439-446.

- RIBEIRO, J. B. ; DA SILVA, RENATO RODRIGUES ; DIAS JUNIOR, J. D. ; ESCARPINATI, MAURICIO CUNHA ; BACKES, ANDRÉ RICARDO. Automated detection of sugarcane crop lines from UAV images using deep learning. INFORMATION PROCESSING IN AGRICULTURE, v. 1, p. 1, 2023.

- DA SILVA, MATHEUS VIEIRA ; SILVA, LEANDRO H.F.P. ; JUNIOR, JOCIVAL DANTAS DIAS ; ESCARPINATI, MAURICIO CUNHA ; BACKES, ANDRÉ RICARDO ; MARI, JOÃO FERNANDO . Generating synthetic multispectral images using neural style transfer: A study with application in channel alignment. COMPUTERS AND ELECTRONICS IN AGRICULTURE, v. 206, p. 107668, 2023.

- FERRAZ, HERICLES ; DIAS JUNIOR, JOCIVAL ; BACKES, ANDRÉ ; ABDALA, DANIEL ; ESCARPINATI, MAURICIO. Multichannel Analysis in Weed

Detection. In: 18th International Conference on Computer Vision Theory and Applications, 2023, Lisbon. Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, 2023. p. 419-426.

- SANTOS, JEAN ; DIAS JUNIOR, JOCIVAL ; BACKES, ANDRÉ ; ESCARPINATI, MAURÍCIO . Segmentation of Agricultural Images using Vegetation Indices. In: 16th International Conference on Computer Vision Theory and Applications, 2021, Online Streaming. Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, 2021. v. 4. p. 506-511.

- SILVA, LEANDRO; DIAS JR, JOCIVAL; ESCARPINATI, MAURÍCIO ; SANTOS, JEAN ; BACKES, ANDRÉ ; MARI, JOÃO . Non-linear Distortion Recognition in UAVs Images using Deep Learning. In: 16th International Conference on Computer Vision Theory and Applications, 2021, Online Streaming. Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, 2021. v. 4. p. 447-454.

Bibliography

AGRICULTURA, P. e. A. Ministério da. **Agropecuária Brasileira em Números**. 2019. Disponível em: <http://www.agricultura.gov.br/assuntos/politica-agricola/agropecuaria-brasileira-em-numeros>.

_____. **Agropecuária Brasileira em Números**. 2023. Disponível em: <https://www.gov.br/agricultura/pt-br/assuntos/politica-agricola/todas-publicacoes-de-politica-agricola/agropecuaria-brasileira-em-numeros/abn-2023-05.pdf/view>.

ALEXEY, D. An image is worth 16x16 words: Transformers for image recognition at scale. **arXiv preprint arXiv: 2010.11929**, 2020. Disponível em: <https://doi.org/10.48550/arXiv.2010.11929>.

ANDEREGG, J. et al. On-farm evaluation of uav-based aerial imagery for season-long weed monitoring under contrasting management and pedoclimatic conditions in wheat. **Computers and Electronics in Agriculture**, Elsevier, v. 204, p. 107558, 2023. Disponível em: <https://doi.org/10.1016/j.compag.2022.107558>.

BRAZ, G. B. P. et al. Sourgrass interference on soybean grown in brazilian cerrado. **Revista Caatinga**, SciELO Brasil, v. 34, p. 350–358, 2021. Disponível em: <https://doi.org/10.1590/1983-21252021v34n211rc>.

BRILHADOR, A. et al. Classification of weeds and crops at the pixel-level using convolutional neural networks and data augmentation. In: IEEE. **2019 IEEE Latin American Conference on Computational Intelligence (LA-CCI)**. 2019. p. 1–6. Disponível em: <https://doi.org/10.1109/LA-CCI47412.2019.9037044>.

CASTELLANO, G.; MARINIS, P. D.; VESSIO, G. Applying knowledge distillation to improve weed mapping with drones. In: IEEE. **2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS)**. 2023. p. 393–400. Disponível em: <https://doi.org/10.15439/2023F960>.

_____. Weed mapping in multispectral drone imagery using lightweight vision transformers. **Neurocomputing**, Elsevier, v. 562, p. 126914, 2023. Disponível em: <https://doi.org/10.1016/j.neucom.2023.126914>.

CELIKKAN, E. et al. Semantic segmentation of crops and weeds with probabilistic modeling and uncertainty quantification. In: **Proceedings of the IEEE/CVF**

International Conference on Computer Vision. [s.n.], 2023. p. 582–592. Disponível em: <https://doi.org/10.1109/ICCVW60793.2023.00065>.

CHEN, L.-C. Rethinking atrous convolution for semantic image segmentation. **arXiv preprint arXiv:1706.05587**, 2017. Disponível em: <https://doi.org/10.48550/arXiv.1706.05587>.

CHEN, L.-C. et al. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: **Proceedings of the European conference on computer vision (ECCV)**. [s.n.], 2018. p. 801–818. Disponível em: <https://doi.org/10.48550/arXiv.1802.02611>.

COX, J. et al. Black-grass (alopecurus myosuroides) in cereal multispectral detection by uav. **Weed Science**, Cambridge University Press, v. 71, n. 5, p. 444–452, 2023. Disponível em: <https://doi.org/10.1017/wsc.2023.41>.

CUNHA, N. et al. Multispectral image segmentation in agriculture: A comprehensive study on fusion approaches. In: SPRINGER. **Iberian Robotics conference**. 2023. p. 311–323. Disponível em: <https://doi.org/10.1007/978-3-031-59167-9_26>.

FERNANDEZ-GALLEGO, J. A. et al. Wheat ear counting in-field conditions: High throughput and low-cost approach using rgb images. **Plant Methods**, BioMed Central, v. 14, n. 1, p. 22, 2018. Disponível em: <https://doi.org/10.1186/s13007-018-0289-4>.

FERRAZ, H. F. et al. Multichannel analysis in weed detection. In: **VISIGRAPP (4: VISAPP)**. [s.n.], 2023. p. 419–426. Disponível em: <https://doi.org/10.5220/0011780900003417>.

FORD, J.; SADGROVE, E.; PAUL, D. Developing an extreme learning machine based approach to weed segmentation in pastures. **Smart Agricultural Technology**, Elsevier, v. 5, p. 100288, 2023. Disponível em: <https://doi.org/10.1016/j.atech.2023.100288>.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [S.l.]: MIT Press, 2016. <http://www.deeplearningbook.org>.

HASAN, A. M. et al. A survey of deep learning techniques for weed detection from images. **Computers and Electronics in Agriculture**, Elsevier, v. 184, p. 106067, 2021. Disponível em: <https://doi.org/10.1016/j.compag.2021.106067>.

HASHEMI-BENI, L. et al. Deep convolutional neural networks for weeds and crops discrimination from uas imagery. **Frontiers in Remote Sensing**, Frontiers Media SA, v. 3, p. 755939, 2022. Disponível em: <https://doi.org/10.3389/frsen.2022.755939>.

HENRICH, V. et al. **Index Database: A database for remote sensing indices**. [S.l.]: WWW document] URL https://www. indexdatabase. de/db/ias. php [accessed 29 . . . , 2017.

HERRMANN, I. et al. Assessment of maize yield and phenology by drone-mounted superspectral camera. **Precision Agriculture**, Springer, v. 21, p. 51–76, 2020. Disponível em: <http://doi.org/10.1007/s11119-019-09659-5>.

HUANG, S.-C. et al. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. **NPJ digital medicine**, Nature Publishing Group UK London, v. 3, n. 1, p. 136, 2020. Disponível em: <https://doi.org/10.1038/s41746-020-00341-z>.

HUNTER, M. C. et al. Agriculture in 2050: Recalibrating Targets for Sustainable Intensification. **BioScience**, v. 67, n. 4, p. 386–391, 02 2017. ISSN 0006-3568. Disponível em: <https://doi.org/10.1093/biosci/bix010>.

JAYANTHI, M.; SHASHIKUMAR, D. Survey on agriculture image segmentation techniques. **Asian Journal of Applied Science and Technology (AJAST)**, v. 1, n. 8, p. 143–146, 2017. Disponível em: <https://doi.org/10.3390/s23010065>.

JIANG, K.; AFZAAL, U.; LEE, J. Transformer-based weed segmentation for grass management. **Sensors**, MDPI, v. 23, n. 1, p. 65, 2022.

KHANAL, S. **Remote Sensing in Precision Agriculture**. 2017. Disponível em: <https://ohioline.osu.edu/factsheet/fabe-5541>.

KHANAL, S.; FULTON, J.; SHEARER, S. An overview of current and potential applications of thermal remote sensing in precision agriculture. **Computers and Electronics in Agriculture**, Elsevier, v. 139, p. 22–32, 2017. Disponível em: <https://doi.org/10.1016/j.compag.2017.05.001>.

KIM, D.-W. et al. Growth monitoring of field-grown onion and garlic by cie l* a* b* color space and region-based crop segmentation of uav rgb images. **Precision Agriculture**, Springer, p. 1–20, 2023. Disponível em: <https://doi.org/10.1007/s11119-023-10026-8>.

KUMAR, T. et al. Image data augmentation approaches: A comprehensive survey and future directions. **IEEE Access**, IEEE, 2024. Disponível em: <https://doi.org/10.1109/ACCESS.2024.3470122>.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **nature**, Nature Publishing Group UK London, v. 521, n. 7553, p. 436–444, 2015. Disponível em: <https://doi.org/10.1038/nature14539>.

LECUN, Y. et al. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, Ieee, v. 86, n. 11, p. 2278–2324, 1998. Disponível em: <https://doi.org/10.1109/5.726791>.

LI, H. et al. An automatic approach for detecting seedlings per hill of machine-transplanted hybrid rice utilizing machine vision. **Computers and Electronics in Agriculture**, Elsevier, v. 185, p. 106178, 2021. Disponível em: <https://doi.org/10.1016/j.compag.2021.106178>.

LIAKOS, K. G. et al. Machine learning in agriculture: A review. **Sensors (Basel, Switzerland)**, MDPI AG, v. 18, n. 8, p. 2674, 2018. Disponível em: <https://doi.org/10.3390/s18082674>.

LIU, B.; BRUCH, R. Weed detection for selective spraying: a review. **Current Robotics Reports**, Springer, v. 1, p. 19–26, 2020. Disponível em: <https://doi.org/10.1007/s43154-020-00001-w>.

LIU, Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows. In: **Proceedings of the IEEE/CVF international conference on computer vision**. [s.n.], 2021. p. 10012–10022. Disponível em: <https://doi.org/10.48550/arXiv.2103.14030>.

LONG, J.; SHELHAMER, E.; DARRELL, T. Fully convolutional networks for semantic segmentation. **Proceedings of the IEEE conference on computer vision and pattern recognition**, p. 3431–3440, 2015. Disponível em: <https://doi.org/10.48550/arXiv.1411.4038>.

LOSHCHILOV, I.; HUTTER, F. et al. Fixing weight decay regularization in adam. **arXiv preprint arXiv:1711.05101**, v. 5, 2017. Disponível em: <https://doi.org/10.48550/arXiv.1711.05101>.

LU, Z. et al. Image segmentation of uav fruit tree canopy in a natural illumination environment. **Agriculture**, MDPI, v. 12, n. 7, p. 1039, 2022. Disponível em: <https://doi.org/10.3390/agriculture12071039>.

MAES, W. H.; STEPPE, K. Perspectives for remote sensing with unmanned aerial vehicles in precision agriculture. **Trends in Plant Science**, v. 24, n. 2, p. 152 – 164, 2019. ISSN 1360-1385. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1360138518302693>.

MAO, A.; MOHRI, M.; ZHONG, Y. Cross-entropy loss functions: Theoretical analysis and applications. In: PMLR. **International conference on Machine learning**. 2023. p. 23803–23828. Disponível em: <https://doi.org/10.48550/arXiv.2304.07288>.

MARINIS, P. D.; VESSIO, G.; CASTELLANO, G. Roweeder: Unsupervised weed mapping through crop-row detection. **arXiv preprint arXiv:2410.04983**, 2024. Disponível em: <https://doi.org/10.48550/arXiv.2410.04983>.

MOAZZAM, S. I. et al. A patch-image based classification approach for detection of weeds in sugar beet crop. **IEEE Access**, IEEE, v. 9, p. 121698–121715, 2021. Disponível em: <https://doi.org/10.1109/ACCESS.2021.3109015>.

MOGILI, U. R.; DEEPAK, B. B. V. L. Review on application of drone systems in precision agriculture. **Procedia Computer Science**, Elsevier, v. 133, p. 502–509, 2018. Disponível em: <https://doi.org/10.1016/j.procs.2018.07.063>.

MONTEIRO, A.; WANGENHEIM, A. von. **Orthomosaic Dataset of RGB aerial Images for Weed Mapping**. [S.l.]: LAPIX/UFSC, 2019. <http://www.lapix.ufsc.br/weed-mapping-sugar-cane>.

MOREIRA, G. et al. Benchmark of deep learning and a proposed hsv colour space models for the detection and classification of greenhouse tomato. **Agronomy**, MDPI, v. 12, n. 2, p. 356, 2022. Disponível em: <https://doi.org/10.3390/agronomy12020356>.

NONG, C.; FAN, X.; WANG, J. Semi-supervised learning for weed and crop segmentation using uav imagery. **Frontiers in Plant Science**, Frontiers, v. 13, p. 927368, 2022. Disponível em: <https://doi.org/10.3389/fpls.2022.927368>.

OSE, K.; CORPETTI, T.; DEMAGISTRI, L. 2 - multispectral satellite image processing. In: BAGHDADI, N.; ZRIBI, M. (Ed.). **Optical Remote Sensing of Land Surface**. Elsevier, 2016. p. 57 – 124. ISBN 978-1-78548-102-4. Disponível em: <http://www.sciencedirect.com/science/article/pii/B9781785481024500028>.

OSORIO, K. et al. A deep learning approach for weed detection in lettuce crops using multispectral images. **AgriEngineering**, MDPI, v. 2, n. 3, p. 471–488, 2020. Disponível em: <https://doi.org/10.3390/agriengineering2030032>.

OTSU, N. A threshold selection method from gray-level histograms. **IEEE transactions on systems, man, and cybernetics**, IEEE, v. 9, n. 1, p. 62–66, 1979. Disponível em: <https://doi.org/10.1109/TSMC.1979.4310076>.

PRASETYO, E. et al. Mango leaf image segmentation on hsv and ycbcr color spaces using otsu thresholding. In: IEEE. **2017 3rd International Conference on Science and Technology-Computer (ICST)**. 2017. p. 99–103. Disponível em: <https://doi.org/10.1109/ICSTC.2017.8011860>.

RADOČAJ, D. et al. State of major vegetation indices in precision agriculture studies indexed in web of science: A review. **Agriculture**, MDPI, v. 13, n. 3, p. 707, 2023. Disponível em: <https://doi.org/10.3390/agriculture13030707>.

RAMADAN, R.; ALY, S. Cu-net: a new improved multi-input color u-net model for skin lesion semantic segmentation. **IEEE Access**, IEEE, v. 10, p. 15539–15564, 2022. Disponível em: <http://dx.doi.org/10.1109/ACCESS.2022.3148402>.

RAMIREZ, W. et al. Deep convolutional neural networks for weed detection in agricultural crops using optical aerial images. In: IEEE. **2020 IEEE Latin American GRSS & ISPRS Remote Sensing Conference (LAGIRS)**. 2020. p. 133–137. Disponível em: <https://doi.org/10.1109/LAGIRS48042.2020.9165562>.

REEDHA, R. et al. Transformer neural network for weed and crop classification of high resolution uav images. **Remote Sensing**, v. 14, n. 3, 2022. ISSN 2072-4292. Disponível em: <https://www.mdpi.com/2072-4292/14/3/592>.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. **International Conference on Medical image computing and computer-assisted intervention**, p. 234–241, 2015. Disponível em: <https://doi.org/10.48550/arXiv.1505.04597>.

ROZNIK, M.; BOYD, M.; PORTH, L. Improving crop yield estimation by applying higher resolution satellite ndvi imagery and high-resolution cropland masks. **Remote Sensing Applications: Society and Environment**, Elsevier, v. 25, p. 100693, 2022. Disponível em: <https://doi.org/10.1016/j.rsase.2022.100693>.

SA, I. et al. Weedmap: A large-scale semantic weed mapping framework using aerial multispectral imaging and deep neural network for precision farming. **Remote Sensing**, MDPI, v. 10, n. 9, p. 1423, 2018. Disponível em: <https://doi.org/10.3390/rs10091423>.

SAHA, S. **A Comprehensive Guide to Convolutional Neural Networks - the ELI5 way**. 2018. Disponível em: <https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53>.

SAHIN, H. M. et al. Segmentation of weeds and crops using multispectral imaging and crf-enhanced u-net. **Computers and Electronics in Agriculture**, Elsevier, v. 211, p. 107956, 2023. Disponível em: <https://doi.org/10.1016/j.compag.2023.107956>.

SANTOS, D. F. D. d. et al. Automated segmentation of tumor regions from oral histological whole slide images using fully convolutional neural networks. Universidade Federal de Uberlândia, 2022. Disponível em: <http://doi.org/10.14393/ufu.te.2022.252>.

SCHRIJVER, R.; POPPE, K.; DAHEIM, C. Precision agriculture and the future of farming in europe. **Teaduslike ja tehnoloogiliste valikute hindamise üksus.[online] http://www. europarl. europa. eu/RegData/etudes/STUD/2016/581892/EPRS_STU (201 6) 581892_EN. pdf (01.05. 2018)**, 2016.

SHIRZI, M. A.; KERMANI, M. R. Adaptive feature-based plant recognition. **IEEE Transactions on AgriFood Electronics**, IEEE, 2024. Disponível em: <https://doi.org/10.1109/TAFE.2024.3444730>.

SINGH, B. M.; KOMAL, C.; VICTOROVICH, K. A. Crop growth monitoring through sentinel and landsat data based ndvi time-series. v. 44, n. 3, p. 409–419, 2020. Disponível em: <http://dx.doi.org/10.18287/2412-6179-CO-635>.

SOZZI, M. et al. Economic comparison of satellite, plane and uav-acquired ndvi images for site-specific nitrogen application: observations from italy. **Agronomy**, MDPI, v. 11, n. 11, p. 2098, 2021. Disponível em: <https://doi.org/10.3390/agronomy11112098>.

SYNGENTA. **Tecnologia é aliada no controle de daninhas, como buva e amargoso**. Portal Syngenta, 2020. *Portal Syngenta*, 2020. Available: https://portalsyngenta.com.br/noticiasdocampo/tecnologia-e-aliada-no-controle-de-daninhas-como-buva-e-amargoso [Last accessed: 27 July 2020]. Disponível em: <https://portalsyngenta.com.br/noticiasdocampo/tecnologia-e-aliada-no-controle-de-daninhas-como-buva-e-amargoso>.

TORRES-SÁNCHEZ, J. et al. Early detection of broad-leaved and grass weeds in wide row crops using artificial neural networks and uav imagery. **Agronomy**, MDPI, v. 11, n. 4, p. 749, 2021. Disponível em: <https://doi.org/10.3390/agronomy11040749>.

TOSCANO, F. et al. Unmanned aerial vehicle for precision agriculture: A review. **IEEE access**, IEEE, 2024. Disponível em: <https://doi.org/10.1109/ACCESS.2024.3401018>.

TOUVRON, H. et al. Training data-efficient image transformers & distillation through attention. In: PMLR. **International conference on machine learning**. 2021. p. 10347–10357. Disponível em: <https://doi.org/10.48550/arXiv.2012.12877>.

TSOUROS, D. C.; BIBI, S.; SARIGIANNIDIS, P. G. A review on uav-based applications for precision agriculture. **Information**, MDPI, v. 10, n. 11, p. 349, 2019. Disponível em: <https://doi.org/10.3390/info10110349>.

TURHAL, U. C.; DAGDELEN, C. Tensor based statistical segmentation of green vegetation canopy images. **Acta Scientiarum. Technology**, Editora da Universidade Estadual de Maringá-EDUEM, v. 44, p. e55708, 2022. Disponível em: <10.4025/actascitechnol.v44i1.55708>.

VASWANI, A. Attention is all you need. **Advances in Neural Information Processing Systems**, 2017. Disponível em: <https://doi.org/10.48550/arXiv.1706.03762>.

WOODS, R. E.; GONZALEZ, R. C. **Digital image processing**. [S.l.]: Pearson Education Ltd., 2008.

XIE, C.; YANG, C. A review on plant high-throughput phenotyping traits using uav-based sensors. **Computers and Electronics in Agriculture**, Elsevier, v. 178, p. 105731, 2020. Disponível em: <https://doi.org/10.1016/j.compag.2020.105731>.

XIE, E. et al. Segformer: Simple and efficient design for semantic segmentation with transformers. **Advances in neural information processing systems**, v. 34, p. 12077–12090, 2021. Disponível em: <https://doi.org/10.48550/arXiv.2105.15203>.

XU, B. et al. Instance segmentation method for weed detection using uav imagery in soybean fields. **Computers and Electronics in Agriculture**, Elsevier, v. 211, p. 107994, 2023. Disponível em: <https://doi.org/10.1016/j.compag.2023.107994>.

_____. Improved weed mapping in corn fields by combining uav-based spectral, textural, structural, and thermal measurements. **Pest Management Science**, Wiley Online Library, v. 79, n. 7, p. 2591–2602, 2023. Disponível em: <https://doi.org/10.1002/ps.7443>.

XUE, J.; SU, B. et al. Significant remote sensing vegetation indices: A review of developments and applications. **Journal of sensors**, Hindawi, v. 2017, 2017. Disponível em: <https://doi.org/10.1155/2017/1353691>.

YAN, H.; ZHANG, C.; WU, M. Lawin transformer: Improving semantic segmentation transformer with multi-scale representations via large window attention. **arXiv preprint arXiv:2201.01615**, 2022. Disponível em: <https://doi.org/10.48550/arXiv.2201.01615>.

ZARCO-TEJADA, P. J. et al. New tools and methods in agronomy. In: _____. **Principles of Agronomy for Sustainable Agriculture**. Cham: Springer International Publishing, 2016. p. 503–514. Disponível em: <https://doi.org/10.1007/978-3-319-46116-8_33>.

ZHANG, R. et al. Sensor fusion for semantic segmentation of urban scenes. In: IEEE. **2015 IEEE international conference on robotics and automation (ICRA)**. 2015. p. 1850–1857. Disponível em: <https://doi.org/10.1109/ICRA.2015.7139439>.

ZHAO, J.; BERGE, T. W.; GEIPEL, J. Transformer in uav image-based weed mapping. **Remote Sensing**, MDPI, v. 15, n. 21, p. 5165, 2023. Disponível em: <https://doi.org/10.3390/rs15215165>.

ZHOU, T.; RUAN, S.; CANU, S. A review: Deep learning for medical image segmentation using multi-modality fusion. **Array**, Elsevier, v. 3, p. 100004, 2019. Disponível em: <https://doi.org/10.1016/j.array.2019.100004>.