TRANSFORMER-BASED HYBRID ARCHITECTURE FOR SEMANTIC SEGMENTATION USING MULTISPECTRAL IMAGERY IN PRECISION AGRICULTURE

By

Zeynep Galymzhankyzy

A THESIS

Submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

In Computer Science

LAWRENCE TECHNOLOGICAL UNIVERSITY

2025

This thesis has been approved in partial fulfillment of the requirements for the Degree of MASTER OF SCIENCE in Computer Science.

Department of Math and Computer Science

Thesis Advisor: *Dr. Eric Martinson*

Committee Member: *Dr. Tao Liu*

Committee Member: *Dr. George Pappas*

Department Chair: *Dr. Eric Martinson*

Contents

Acknowledgments

I want to express my sincere gratitude to my advisor, Dr. Eric Martinson, for his guidance, support, and encouragement throughout this research. His expertise, constructive feedback, and continuous motivation were instrumental in completing this work.

As a multilingual international student with English as my fourth language, I am also grateful for the assistance of language models and AI tools, as they sometimes help me brainstorm and refine ideas. In a world full of opportunities, I believe it is both wise and ethical to use every available resource responsibly to grow, learn, and communicate effectively.

Definitions

Precision Agriculture: A farming management strategy that uses data-driven, site-specific techniques to optimize agricultural inputs and outputs, aiming to improve yield efficiency and sustainability [1].

Site-Specific Weed Management (SSWM): A precision agriculture approach focused on detecting and treating weed-infested areas selectively, thereby reducing herbicide usage, minimizing environmental impact, and improving resource efficiency [2].

Semantic Segmentation: A computer vision task that assigns a categorical label to each pixel in an image, enabling fine-grained classification of different regions such as crops, weeds, and soil in agricultural imagery [3].

Multispectral Imaging: A remote sensing technique that captures image data across multiple spectral bands beyond the visible range (e.g., Near-Infrared, Red-Edge), providing richer information for distinguishing vegetation characteristics [4].

Unmanned Aerial Vehicle (UAV): An aerial platform capable of autonomous or remotely piloted flight, used in agriculture to collect high-resolution imagery for monitoring crop health, weed distribution, and field conditions [5].

Feature Pyramid Network (FPN): A deep learning architecture that combines features at multiple spatial resolutions to improve object detection and segmentation performance [6].

Pyramid Pooling Module (PPM): A component used in deep learning decoders that aggregates global contextual information from multiple scales to refine semantic segmentation outputs [7].

Vision Transformer (ViT): A deep learning model that applies transformer architectures directly to image patches, enabling global attention-based feature modeling across the entire input [8].

List of Abbreviations

This section provides definitions for abbreviations used throughout the thesis.

CNN	Convolutional Neural Network
UAV	Unmanned Aerial Vehicle
SSWM	Site-Specific Weed Management
RGB	Red-Green-Blue (Visible Light Spectrum)
NIR	Near-Infrared
RE	Red-Edge
NDVI	Normalized Difference Vegetation Index
FPN	Feature Pyramid Network
PPM	Pyramid Pooling Module
IoU	Intersection over Union
mIoU	Mean Intersection over Union
ViT	Vision Transformer
MSI	Multispectral Imaging

Abstract

Weed infestation continues to be a significant impediment to sustainable and efficient crop production. Precision agriculture aims to overcome such hurdles through reliable management processes that account for the complexity of site-specific knowledge and data on weeds. UAV-based multispectral imaging can capture bands beyond the visible spectrum including Near-Infrared and Red-Edge, can improve crop–weed differentiation moving from traditional management practices based on traditional RGB images. However, robust semantic segmentation using multispectral data is still challenging due to spectral variation, occlusions, and the heterogeneous nature of field settings.

This thesis proposes a hybrid CNN–Transformer segmentation framework tailored for multispectral crop–weed mapping. The model integrates modality-specific ConvNeXt encoders for spectral feature extraction, Swin Transformer blocks for global contextual reasoning, a gated Feature Pyramid Network (FPN) for adaptive multispectral fusion, and a Pyramid Pooling Module (PPM) for multi-scale decoding.

When evaluated on the WeedsGalore dataset, the proposed model achieved a mean Intersection-over-Union (mIoU) of 90.04%, a considerable improvement over conventional CNN-based and RGB-only baselines. Furthermore, zero-shot and few-shot

fine-tuning studies on carrot and onion field datasets show that the proposed model has promising cross-domain generalization ability while learning from limited labeled examples.

These findings highlight the potential for multispectral fused learning in conjunction with hybrid architectures to drive site-specific weed management, paving the way towards more scalable and sustainable agricultural practices.

Chapter 1

Introduction

1.1 Motivation and Context

Agricultural production is vital to global food security, but it is under increasing pressure from population growth, climate change, soil degradation, and resource constraints. As these issues escalate, the need to implement technology-driven solutions that can effectively increase agricultural productivity while minimizing the environmental footprint becomes increasingly urgent. Precision agriculture represents a paradigm shift enabling the spatially-resolved management of crops, inputs, and field operations based on site-specific data.

Weed management is one of the most significant challenges in precision agriculture.

Weeds absorb sunlight, water, and nutrients and compete with crops, reducing profitability due to yield loss. Traditional weed control approaches employ an overall herbicide application to fields, and this is often inefficient and damaging to the environment. Whether useful or not, this blanket application is excessive, creating high production costs, and aiding future resistant weed evolution [9, 10].

To solve these problems, the idea of Site-Specific Weed Management (SSWM) has become popular. SSWM uses the detection and treatment of just the infested areas of a field in order to maximize use of resources, cutoff unnecessary chemical applications and mitigate environmental impacts of weed management. For SSWM to operate at scale, though, requires accurate, high-resolution mapping of weed distributions under varied and dynamic field conditions, which is challenging due to differences in crop morphology, weed species diversity, soil backgrounds and environmental factors, such as lighting and moisture.

From a new perspective, remote sensing technologies have reshaped agricultural monitoring, providing a low-impact, large-scale, spatially explicit view of farmland and crops. Specifically, Unmanned Aerial Vehicles (UAVs) with imaging sensors have become widely available and low-cost platforms for agricultural data collection. UAVs are able to acquire images at near centimeter-level resolution and provide substantial information concerning crop health, soil health, and weed pressures over three critical development periods of the plant.

Beyond conventional Red-Green-Blue (RGB) imaging, multispectral sensors mounted on UAVs offer access to additional spectral bands, including Near-Infrared (NIR) and Red-Edge (RE). The NIR band provides information related to plant biomass and chlorophyll content, while the Red-Edge band is sensitive to early indicators of vegetation stress [10, 11]. Leveraging these bands allows for the computation of vegetation indices such as the Normalized Difference Vegetation Index (NDVI), enhancing the ability to differentiate crops from weeds even under visually ambiguous conditions. Despite their advantages, multispectral datasets introduce challenges related to sensor calibration variability, illumination inconsistencies, and the complexity of fusing information across different spectral modalities.

1.2 Semantic Segmentation in Precision Agriculture

Semantic segmentation—the process of assigning a categorical label to each pixel in an image—has become a cornerstone of computer vision applications requiring fine-grained scene understanding. [12] In agricultural contexts, semantic segmentation enables detailed mapping of crops, weeds, and background elements at high spatial resolutions, supporting site-specific interventions such as targeted herbicide application or mechanical weeding [13, 14].

Traditional computer vision techniques for segmentation relied on handcrafted features such as color indices, texture descriptors, and geometric properties. While these approaches achieved moderate success under controlled conditions, they often exhibited poor robustness in real-world agricultural environments characterized by significant variability in illumination, soil reflectance, plant morphology, and weed species composition [13].

The advent of deep learning, particularly Convolutional Neural Networks (CNNs), has dramatically advanced semantic segmentation by enabling models to learn hierarchical feature representations directly from raw imagery [12]. CNN-based architectures, including Fully Convolutional Networks (FCNs) [3], U-Net [15], and DeepLabv3+ [16], have demonstrated remarkable performance by extracting fine-grained local textures, edges, and object structures. These models have been widely adopted for agricultural segmentation tasks and have shown promising results in distinguishing crops from weeds under challenging conditions.

Still, CNNs are inherently tied to local receptive fields and capturing long-range dependencies will require stacking more convolutional layers - thus increasing costs and complexity of the model and still not entirely eliminating the global context. In agricultural imagery, where broader spatial patterns—such as planting rows or patch-level distributions—can provide important cues for discrimination, this limitation becomes critical.

Transformers, originally developed for natural language processing, offer a compelling alternative by modeling global relationships through self-attention mechanisms [17]. Vision Transformers (ViTs) and hierarchical variants like the Swin Transformer [18] have shown state-of-the-art performance in various computer vision tasks by capturing both local and global dependencies. Despite their potential, transformer-based architectures remain relatively underexplored in the domain of multispectral crop–weed segmentation [19]. Furthermore, purely transformer-based models often require large training datasets and significant computational resources, making them less practical for many real-world agricultural applications [18, 19, 20].

Thus, a hybrid approach that combines the local feature extraction capabilities of CNNs with the global reasoning strength of Transformers presents a promising design paradigm [21]. Additionally, effectively leveraging multispectral inputs, while accounting for modality-specific characteristics, remains an open research challenge that is crucial for improving segmentation performance in precision agriculture [22].

1.3 Research Gap and Motivation

Despite substantial advances in deep learning for agricultural segmentation, several critical gaps persist. First, the majority of existing models are tailored to RGB imagery and do not fully exploit the rich spectral information available from UAV-based

multispectral sensors. Naive fusion strategies that simply concatenate spectral bands at the input level fail to capture the distinct semantic contributions of modalities like NIR and Red-Edge, limiting their effectiveness [22, 23].

Second, CNN-based architectures, while powerful at extracting local features, are fundamentally constrained in their ability to model long-range spatial relationships—an important consideration in agricultural fields characterized by occlusion, sparse weed distributions, and complex planting patterns [24]. Although Transformer models offer global context modeling, they have not been fully used in multispectral crop–weed segmentation, particularly in a way that balances global reasoning with efficient local feature extraction [25].

Third, cross-domain generalization remains a major bottleneck [14]. Models trained on specific crop types, growth stages, or environmental conditions often fail to transfer effectively to new domains without substantial retraining. Developing architectures that can generalize with minimal labeled data—through few-shot adaptation or lightweight fine-tuning—is critical for practical deployment across diverse agricultural settings [26].

Motivated by these challenges, this thesis proposes a multimodal segmentation framework that integrates modality-specific CNN encoders, Transformer-based refinement modules, adaptive gated fusion strategies, and multi-scale context aggregation. The

goal is to design a system that is robust, spectrally sensitive, and capable of generalizing across different agricultural domains with minimal supervision.

1.4 Research Objectives and Questions

The primary objective of this research is to develop an efficient and scalable multimodal segmentation framework capable of accurately identifying crops, weeds, and background at the pixel level using multispectral imagery gathered by UAVs. In order to answer this aim, the study will address the following research questions:

- **Q1:** How can convolutional and transformer-based modules be fused to take advantage of both fine-grained local features with long-range spatial dependencies for agricultural scenes?

- **Q2:** How can spectral diversity across RGB, NIR, and RE bands be effectively fused to maximize class separability while retaining as much modality-specific information as possible?

- **Q3:** Can hybrid CNN–Transformer model achieve state-of-the-art segmentation performance on a complex multispectral benchmark, WeedsGalore, better than conventional CNN-only models and RGB-only models?

† **Q4:** To what level the proposed framework can generalize to other agricultural domains, such as different crop types or field conditions, through few-shot adaptation with minimal labeled data?

These are critical questions to address for advancing the possibilities of more practical precision agriculture, deployable solutions based on deep learning.

1.5 Contributions

This thesis contributes to the field of precision agriculture and semantic segmentation in the following ways:

First, it introduces a novel hybrid segmentation architecture that combines modality-specific ConvNeXt encoders for spectral feature extraction with Swin Transformer refinement for global context modeling. Second, it proposes a gated Feature Pyramid Network (FPN) for adaptive multispectral fusion that allows the model to learn dynamic weighting of RGB, NIR, and RE modal spatial scales. Third, it demonstrates a benchmark result of 90.04% mean Intersection-over-Union (mIoU) on the Weeds-Galore benchmark suggesting better segmentation performance than previous state-of-the-art results. Finally, consistent generalization of the proposed framework was highlighted through both zero-shot and few-shot trained segmentation experiments

on external carrot and onion field datasets that further established an excellent level of adaptation with minimal supervision.

Altogether, these contributions will further the potential of scalable, data-efficient, and spectrally expanded weed management solutions for sustainable precision agriculture.

Chapter 2

Literature Review

This chapter provides a comprehensive summary of the literature base on semantic segmentation, multispectral imagery, crop–weed classification, and multimodal learning methods. It provides a final account of important developments, notes the limitations of the current practice, and sets the stage for the design choices we made in this thesis.

2.1 Deep Learning for Semantic Segmentation

Semantic segmentation—where a class label is assigned to each pixel in an image—has emerged as an important foundational building block of modern computer vision

that supports applications spanning across autonomous driving, medical imaging, and agricultural monitoring [12]. In agriculture, especially for tasks like crop–weed discrimination, semantic segmentation allows for a fine-grained and spatially coherent understanding of a complex scene for the field.

The traditional way of doing semantic segmentation began with using experts to handcraft features and then later applied classical machine learning classifiers like Support Vector Machines (SVMs) and Random Forests. This approach worked well in structured environments, but these pipelines were brittle to variations in illumination, plant morphology, and soil background; as such, their robustness in actual field conditions was limited.

The emergence of deep learning – and particularly Convolutional Neural Networks (CNNs) – has disrupted semantic segmentation by allowing us to learn end-to-end from raw images. CNNs introduced the capability to automatically extract hierarchical feature representations, capturing local textures, edges, and structures with minimal manual engineering. Early breakthroughs such as Fully Convolutional Networks (FCNs) [3] laid the groundwork by replacing fully connected layers with convolutional layers, allowing variable input sizes and dense output predictions. Subsequent architectures such as U-Net [15] introduced encoder–decoder designs with symmetric skip connections, facilitating the recovery of fine spatial details and improving segmentation accuracy, especially for small or thin structures—an essential requirement

in agricultural mapping. DeepLab series models [16] further advanced the field by introducing atrous (dilated) convolutions and pyramid pooling modules to capture multi-scale contextual information without sacrificing spatial resolution.

Despite these successes, convolutional networks suffer from an inherent weakness: the receptive fields are always local. [3] Of course, the deeper the layers, the larger the receptive field, but capturing the global dependencies over large input areas is a formidable task, particularly in the context of UAV images, where planting patterns and weed clusters may spread over vast regions.

In this case, the ability to aggregate global context without relying solely on local receptive fields presents a compelling advantage for semantic segmentation tasks in agriculture, where understanding broad spatial patterns is critical. In order to address this problem, originally introduced in the natural language processing domain, the Vision Transformers (ViTs) have been introduced [17]. They operate on images as sequences of patches, and in this way, by capturing the relationships between all the patches, they explicitly represent long-range spatial dependencies.

However, pure Vision Transformers come with substantial computational demands and often require extremely large datasets for effective training—an impractical constraint in many agricultural applications. To mitigate these issues, hybrid architectures have been proposed. Models such as the Swin Transformer [18] introduce hierarchical representations with shifted window-based attention, balancing the ability

to model both local and global contexts while maintaining computational efficiency. Hybrid CNN–Transformer frameworks [21] further combine the inductive biases and efficient feature extraction capabilities of CNNs with the global reasoning strengths of transformers, achieving state-of-the-art results across multiple segmentation benchmarks.

Despite these advancements in transfer learning, one area that remains relatively overlooked in the literature is the effective application of transformers and hybrid models to multispectral agriculture images. The majority of the literature remains in a space that revolves around RGB, missing opportunities to take advantage of extended spectral cues such as Near-Infrared (NIR) and Red-Edge (RE) bands for improved vegetation discrimination. This forms a main motivation to address the gap in the literature by integrating multispectral fusion with transformer-based hybrid architectures in the research presented in this thesis.

2.2 Multispectral Imagery for Agricultural Mapping

Remote sensing is an important component of precision agriculture because it enables large-scale, non-destructive monitoring of crop conditions. Multiple remote sensing modalities exist, but multispectral imagery is the most unique of these information

sources because it provides information in the visible spectrum and beyond, allowing for more change detection within vegetation and more information about vegetation health, growth trajectory, or species distinction than with only RGB imagery [27].

Multispectral imaging is the process of recording reflectance data in several spectral regions. These regions, typically, include the visible ones: Red (R), Green (G), and Blue (B), which register and convey color information as the human eye perceives it [28]. In addition to these, multispectral cameras capture Near-Infrared (NIR) and Red-Edge (RE) bands. The NIR band, which is centered at 840 nm, is very sensitive to plant cell structure and biomass, with healthy vegetation showing high reflectance in this region. The Red-Edge band, which is situated at the red edge of the spectrum (around 730 nm), is especially useful for detecting early signs of plant stress and small changes in chlorophyll concentration [10, 11].

Spectral characteristics allow for the derivation of vegetation indices like the Normalized Difference Vegetation Index (NDVI) and Red-Edge NDVI (RENDVI) which amplify the differences between healthy crops, stressed vegetation, and weeds, and these indices have applications in situations from yield predictions to weed detection.

Unmanned Aerial Vehicles (UAVs) which have multispectral cameras have been the predominant method for data collection since these UAVs provide high spatial resolution (centimeter dimensions), operational flexibility, and the opportunity for repeated use at key meteorological growth stages of the crop [29]. In addition to the benefits

mentioned above, UAVs are comparatively cheaper and customizable compared to similar datasets derived from satellite or manned aerial imagers, and they offer rapid spatial and temporal data collection capabilities to address site-specific needs.

Recent large-scale datasets such as WeedMap [11] and WeedsGalore [10] illustrate the trend that a growing amount of UAV multispectral imagery available has been annotated. These datasets have played a crucial role in facilitating the training and benchmarking of deep learning models for crop–weed segmentation at scale.

Despite its advantages, working with multispectral imagery comes with its challenges. Spectral variability that results from sensor specifications, lighting, and soil backgrounds can complicate the task of building models that generalize between fields and seasons [30]. Furthermore, the fusion of multiple spectral bands is non-trivial: naive concatenation often fails to exploit the complementary information embedded in different modalities, while improper integration can introduce redundancy or noise. Additional practical issues, such as motion blur, shadowing, and radiometric inconsistencies in UAV imagery, further degrade the quality of input data [25].

To address these challenges, segmentation architectures will need to extract modality-specific features as well as robustly fuse information across spectral channels. This thesis aims to develop models that are designed specifically for agricultural multispectral datasets.

2.3 Datasets for Crop and Weed Segmentation

The advancement of deep learning models for crop–weed segmentation relies heavily on high-quality, annotated datasets that capture the complexity and variability of real-world agricultural environments. Over the past decade many benchmark datasets have been proposed; each varying in the types of crops, imaging modalities, spatial resolution, and granularity of the annotations that were made. Reviewing datasets can provide necessary context for the dataset choices and evaluation framework used in this thesis.

The **Crop/Weed Field Image Dataset (CWFID)** [31] was among the earliest contributions in this domain, examining early-stage carrot fields utilizing grainy multispectral images obtained through ground robots from the field. CWFID provided Red and Near Infrared (NIR) bands, which allowed for basic vegetation differentiation; however, because it had a limited number of images (only 60) and included only one crop species, it had limited value for training contemporary deep learning architectures that require many different images from many other species in large quantities.

To meet the increasing demand for scale mapping, the **WeedMap** dataset [11] included high-resolution UAV-generated orthomosaics of sugar beet fields. It provided

access to up to 12 spectral bands, including RGB, NIR, and NDVI. WeedMap allowed for large-scale semantic segmentation experiments in realistic situations. However, practical challenges, such as marked class imbalance and the small spatial footprint of many weed instances, often resulted in inferior segmentation performance, especially for minority classes.

Most recently, the **PhenoBench** dataset [32] expanded the scope of the previous datasets by providing dense pixel-wise annotations of crop leaves and weeds over several growing seasons. PhenoBench emphasizes the fine-grained vegetation structure, while the extremely high spatial resolution (1 cm) and dense labeling/annotation introduce significant computation demands when applied to model training and inference pipelines.

The **CropAndWeed** dataset [33] shifted the emphasis toward species-level identification and provides 74 weed classes annotated in RGB images. While the annotated instances present an opportunity for accurate species-level recognition and, ultimately, further species identification in agricultural settings, the complexities of the taxonomy and limited use of sensing modalities create limitations to building generalized multispectral models complementary to agriculture.

Among recent contributions, the **WeedsGalore** dataset [10] is particularly useful for this work. WeedsGalore contains dense, pixel-level annotations of maize fields taken at many dates and seasons with UAV-mounted multispectral cameras. It has

five aligned spectral bands (RGB, NIR, and Red-Edge), a large ground sampling distance (GSD) of 2.5 mm, and encompasses differences in crop growth stages, weed densities, and soil backgrounds. These features make it well suited for investigating multispectral fusion methods and robustness under spectral and temporal variability.

Along with WeedsGalore, two additional datasets were used for cross-domain evaluation: **Carrots 2017** and **Onions 2017**. Both datasets consist of aligned RGB and NDVI images taken using UGVs with manually created weed segmentation masks. Although these two datasets are smaller than WeedsGalore, they provide domain shifts regarding crop morphology, planting structure, and background appearance. These shifts offer a suitable opportunity to evaluate generalization performance in zero-shot and few-shot scenarios.

To provide a clearer comparison of these datasets and highlight their respective strengths and limitations, Table 2.1 summarizes their key characteristics.

Table 2.1
Comparison of Crop–Weed Segmentation Datasets

Dataset	**Year**	**Crop**	**Platform**	**GSD (mm)**	**Modalities**
CWFID [31]	2014	Carrot	UGV	∼9.0	Red, NIR
WeedMap [11]	2018	Sugar Beet	UAV	∼3.8	RGB, NIR, NDVI
PhenoBench [32]	2024	Sugar Beet	UAV	∼1.0	RGB
CropAndWeed [33]	2023	Mixed	UGV	N/A	RGB
WeedsGalore [10]	2025	Maize	UAV	2.5	RGB, RE, NIR
Carrots 2017	2017	Carrot	UGV	∼3.5	RGB, NDVI
Onions 2017	2017	Onion	UGV	∼3.5	RGB, NDVI

Although several datasets have helped advance crop–weed segmentation approaches,

WeedsGalore is this thesis's primary benchmark for training and validation purposes. Carrots 2017 and Onions 2017 are key benchmarks for assessing cross-domain robustness and few-shot adaptations.

2.4 Existing Approaches for Crop–Weed Segmentation

Over the last ten years, crop–weed segmentation has moved from handcrafted feature-engineering approaches to more advanced deep learning architectures. This section covers the varying trends and challenges of the existing approaches and includes significant findings that motivate the architectural design of this work.

The first crop–weed identification methods heavily depended on handcrafted features such as texture features, color indices (e.g., Excess Green, NDVI), and morphological features. Classical machine learning algorithms (Support Vector Machines, Random Forests, or k-nearest Neighbors) were also trained to produce vegetation pixels with these manually extracted features [13, 34]. While traditional methods were effective in a controlled setting, they have inherently low robustness across seasons, fields, or crop types. Variability in illumination, soil backgrounds, plant morphology, and weed species reduced performance and required tedious re-engineering of features to deploy a new crop–weed identification model.

Since the introduction of deep learning and Convolutional Neural Networks (CNN), there has been a clear trend toward performing end-to-end feature learning from imagery [11, 30]. Exemplar CNN-based architectures, such as U-Net [15] and DeepLabv3+ [16], can be used as baseline models for semantic segmentation in precision agriculture. Impressive works like WeedMap [11] show how effective CNNs can be for UAV-based multispectral imagery and significantly improve over prior modalities. Bosilj *et al.* [30] tested transfer learning between cropping types, proving that transfer learning is capable across domains.

Nevertheless, CNN's approaches are inherently limited. CNNs shape their topology to derive local spatial features, like edges and textures. However, they struggle to model more global context, preferencing to distinguish crops and weeds among obscuration or patchy field patterns. Although multilayered CNNs have processed multispectral data, earlier work suffered from naïve fusion methods that concatenated the pixel values across the bands with no respect for the nature of each band's unique semantic latent space.

More recently, researchers began considering transformer-based models to avoid the problem space associated with CNNs. Vision transformers (ViT) or Swin transformers allow for self-attention, which can aid in modeling long-distance dependencies across the entire image [9, 35]. For example, Zhao *et al.* [9] also used a transformer backbone on UAV-based weed mapping, showing enhanced reasoning capabilities compared to

conventional CNNs. Wang *et al.* [35] also used Swin Transformer blocks with transfer learning across two stages to increase generalizability across different agricultural contexts.

Though transformer-based methods show great potential, their use in precision agriculture is still relatively new. Most current applications are exclusively based on RGB images and do not fully take advantage of the range obtained from multispectral imagery. Additionally, transformer usage is typically more computationally intense, making it particularly challenging to deploy on UAVs where memory and processing speed are limited [36].

Using multispectral data—specifically bands from Near Infrared (NIR) and Red-Edge (RE)—is essential for accurate crop–weed classification [10, 11]. Multispectral fusion methods cover this idea quite well. Early fusion methods concatenate the spectral bands into one at the input; while simple to implement, they treat all bands as equal. Thus, valuable modality-specific information content is mainly missed. Late fusion methods, where you have a different feature extractor for each modality and then merge the features later on in the layers [33], discriminate between modalities since each of the feature extractors represent features from each input modality, but they increase complexity. Attention and gating-spectral modalities methods also offer improvements since they are method agnostic in their ability to weight modalities as they merge them. Thus, they can be robust to modality noise and variability

in the field. However, few studies on crop weed segmentation have explored these approaches.

Despite the advances assembled by deep learning, several limitations remain across all existing approaches. First, RGB reliance limits the potential of the spectral information of multispectral images, such as NIR and Red-Edge bands. Second, spectral fusion strategies generally treat all bands the same and provide no means for adaptive, modality-aware fusion strategies. Third, while CNN-based models are biased toward local context, they typically do not have enough long-range reasoning to overcome occlusion biases, dense planting patterns, or nuanced morphological cues. Finally, cross-domain generalization (i.e., transfer to different crops, field conditions, or environmental variances) still requires significant amounts of real labeled data and rewriting.

This proof of concept demonstrates the important ramifications of developing segmentation architectures capable of taking advantage of multispectral diversity, modeling fine-grained and global contexts, and generalizing to agricultural contexts.

The next chapter details the proposed architecture, which explicitly expands on this gap by proposing a hybrid CNN–Transformer model for multispectral crop–weed segmentation from UAV images.

Chapter 3

Methodology

3.1 Problem Formulation

The main methodology of this thesis is to construct a high-accuracy multispectral semantic segmentation model capable of identifying crops, weeds, and backgrounds in imagery taken by UAVs. This chapter will formalize the problem definition, inputs and outputs, and evaluation metrics for assessing model performance, building on the principles of semantic segmentation mentioned in previous chapters.

The model input is a five-channel multispectral image consisting of spatially located bands of Red (R), Green (G), Blue (B), Near-Infrared (NIR), and Red-Edge (RE) with a respective spatial resolution of 600×600 pixels. Each input sample can be

represented as a tensor $\mathbf{X} \in \mathbb{R}^{5\times600\times600}$.

The output from the model is a pixel-by-pixel picture-wide semantic mask $\mathbf{Y} \in \{0, 1, 2\}^{600\times600}$, that classifies the pixels into one of three classes:

† **Class 0:** Background (soil, residues, non-vegetative covers)

† **Class 1:** Crop (the target crop)

† **Class 2:** Weed (the unwanted vegetation)

Formally, the segmentation function f_θ parameterized by θ is defined as:

$$\mathbf{Y} = f_\theta(\mathbf{X})$$

where $\mathbf{X}$ is the multispectral tensor input, and $\mathbf{Y}$ is the predicted class map.

3.2 Dataset Description

This research primarily operated with the WeedsGalore dataset [10], a comprehensive benchmark dataset based on UAV multispectral images for semantic and instance segmentation of crops and weeds in maize fields. The WeedsGalore dataset provides

a realistic and challenging environment for evaluating segmentation models, especially for multispectral learning frameworks in precision agriculture.

3.2.1 Overview of the WeedsGalore Dataset

The WeedsGalore dataset was collected from an agricultural maize field in Marquardt, Potsdam, Germany, about 1840 m^2 in size. The dataset was collected over four campaigns—May 25, May 30, June 6, and June 15, 2023—capturing different crop and weed growing stages. During each campaign, images were acquired using a UAV (DJI Phantom P4 Multispectral) with five spectral bands-RGB (Red, Green, Blue), Red-Edge, and Near-Infrared allocated vertically.

Approximately 1150 raw images were collected for each campaign, and 156 high-quality photos were selected and densely annotated for semantic and instance segmentation. The chosen images were cropped and standardized to be 600×600 pixels, achieving a ground sampling distance (GSD) of 2.5 mm. A GSD of 2.5 mm means each 600×600 image covers a 2.5m×2.5m area, providing granular resolution for crop-weed segmentation. The annotations cover five classes: Maize, Amaranth, Barnyard Grass, Quickweed, and Weed Other. The dataset is split into training (70%), validation (15%), and testing (15%) subsets.

An overview of the main attributes of the dataset is summarized in Table 3.1.

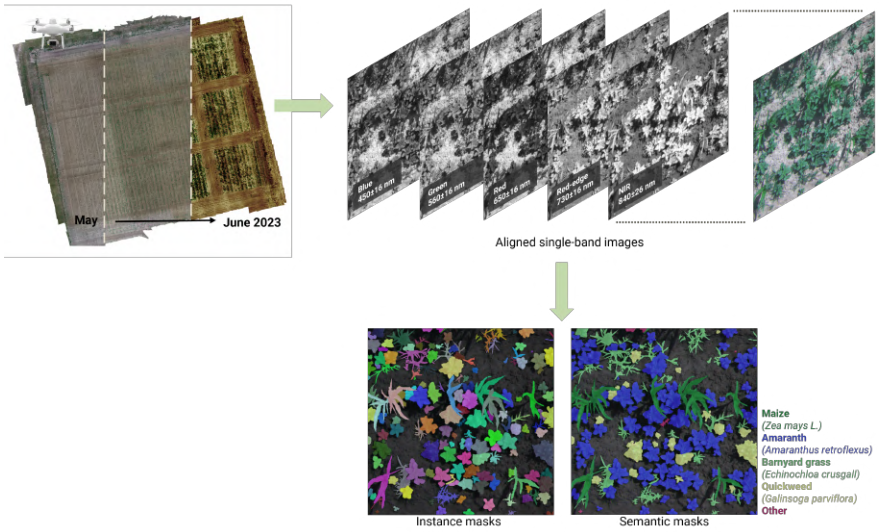

Figure 3.1: Acquisition and annotation workflow for the WeedsGalore dataset. Extracted from the original source [10]

Table 3.1
Key Characteristics of the WeedsGalore Dataset

Attribute	**Details**
Location	Marquardt, Germany
Collection Dates	May 25, May 30, June 6, June 15, 2023
Platform	DJI Phantom P4 Multispectral UAV
Spectral Bands	Red, Green, Blue, Red-Edge, Near-Infrared
Ground Sampling Distance (GSD)	2.5 mm
Image Size	600×600 pixels
Annotated Classes	Maize, Amaranth, Barnyard Grass, Quickweed, Weed Other
Instances per Image	>78 instances on average
Dataset Splits	Train (70%), Validation (15%), Test (15%)

3.2.2 Multispectral Modalities and Data Structure

Every image sample from the WeedsGalore dataset is made up of five spectral bands:

† **RGB Bands (Red, Green, Blue):** Represent the visible portions of the

spectrum that capture the texture, color, and structural patterns of vegetation and soil.

† **Red-Edge (RE):** Sensitive to changes in chlorophyll and early signs of plant stress, thus providing better feature separation for distinguishing healthy vs. stressed vegetation.

† **Near-Infrared (NIR):** measures different biological compositions, such as biomass and canopy structure, to improve the separation of vegetation from bare soil.

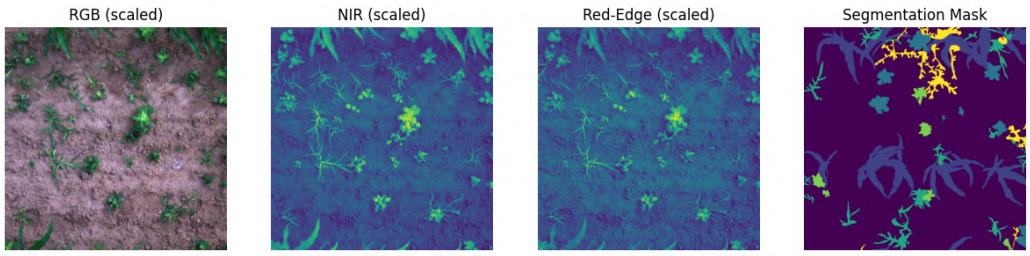

Figure 3.2: Scaled input channels and original segmentation mask

This multispectral information offers complementary views of the field, making it particularly effective for differentiating crops from weeds under varying growth stages and lighting conditions.

3.2.3 Preprocessing Pipeline

To prepare the WeedsGalore dataset for model training and evaluation, the following preprocessing steps were applied:

1. **Spectral Stacking:** Each image's five spectral bands (RGB, RE, NIR) were stacked into a single 5-channel tensor, maintaining perfect pixel-wise registration.

2. **Normalization:** Band-wise normalization was applied to zero mean and unit variance to stabilize optimization dynamics during training.

3. **Semantic Label Remapping:** The original five classes (Maize and four weed types) were remapped into three classes to simplify the segmentation task:

   † **Class 0:** Background (non-vegetation, bare soil)

   † **Class 1:** Crop (Maize)

   † **Class 2:** Weed (all weed species combined)

4. **Data Augmentation:** Data augmentation strategies, including random horizontal flips, vertical flips, rotations, and brightness adjustments, were applied to enhance robustness and mitigate overfitting.

The preprocessing pipeline ensures that the model has experienced a reasonable

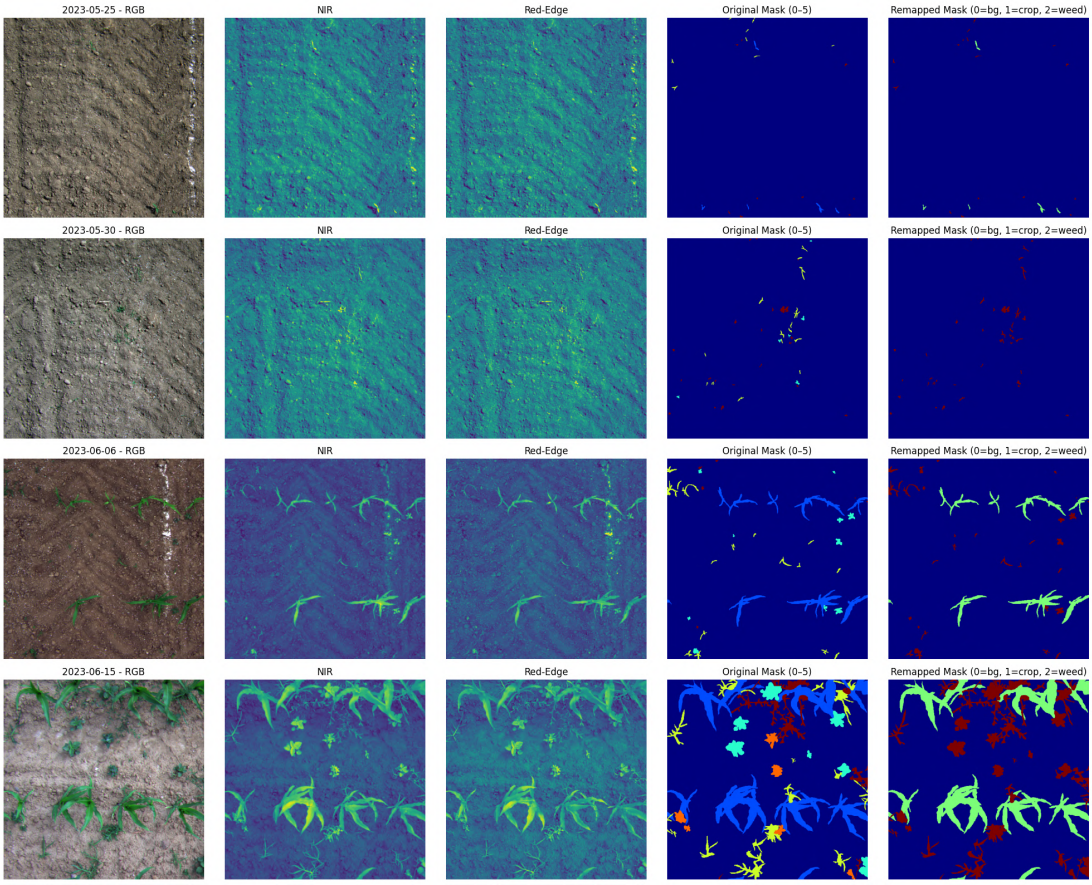

Figure 3.3: Scaled input channels and segmentation mask after preprocessing.

amount of spatial and spectral variability to simulate deployment conditions. Figure 3.3 illustrates the scaled input channels and the resulting segmentation mask after preprocessing.

3.3 Proposed Architecture

The proposed segmentation framework is specifically designed for high-resolution multispectral UAV imagery in precision agriculture. It targets pixel-wise classification into three classes—background, crop, and weed—by efficiently combining modality-specific feature extraction, global context modeling, adaptive fusion, and context-aware decoding.

An overview of the full model architecture is illustrated in Figure 3.4.

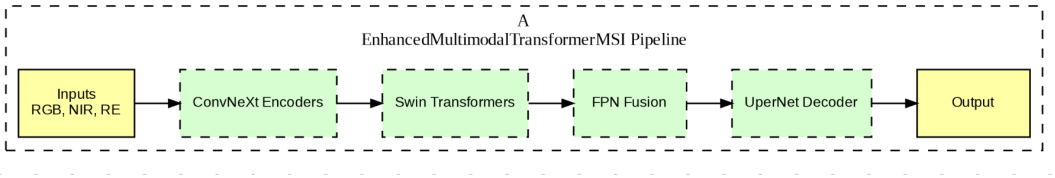

Figure 3.4: Overview of the proposed segmentation model architecture. Separate modality encoders extract multi-scale features, which are refined by Transformer blocks, fused via gated multi-scale integration, and decoded into pixel-wise semantic maps.

The pipeline starts with modality-specific feature extraction. The input image contains five spectral bands: Red, Green, Blue (RGB), Near-Infrared (NIR), and Red-Edge (RE). To fully exploit the unique properties of each modality, the network processes them through separate ConvNeXt-based encoders. The RGB stream uses a pre-trained backbone, while the NIR and RE streams adapt ConvNeXt encoders by

modifying the first convolutional layer to only accept single-channel input. This design ensures that spectral features—such as biomass indicators from NIR or chlorophyll sensitivity from Red-Edge—are captured independently and preserved throughout the early stages of the network.

Following feature extraction, the architecture employs Swin Transformer refinement blocks to improve the highest-level representations. The deepest feature map from each modality is then passed to a Transformer module that models global context with self-attention [37]. This is essential in understanding the high-level patterns in agricultural imagery where crops and weeds are frequently spread over a spatially distributed 'structure.' The Transformer module adds a necessary level of comprehension where local convolutional features are enriched with semantic relationships (contextual relationships), and the model gains a deeper understanding of scene-wide spatial basis.

After transforming features from the three modalities, they are combined using a gated multi-scale Feature Pyramid Network (FPN). The fusion module operates at multiple spatial resolutions, allowing both low-level texture information and high-level semantic context to contribute to the final representation. This also means the contribution of each modality could be adaptively weighed using the learned gating aspects of the model. This means it could learn to weigh features from the more informative modalities more heavily while suppressing noisy features or less

reliable signals, improving robustness under varying illumination, occlusion, or soil background conditions.

Finally, the fused multi-scale feature map is reconstructed back into a dense segmentation map by a decoder with a Pyramid Pooling Module (PPM). The decoder encodes contextual information from multiple scales and projects the fused features to the final three-class output on the exact resolution, 600×600, as the original image. The fusion of fine-level detail recovery with global-scene aggregation is needed to semantically segment small weeds, jagged crop canopies, and mixed background textures accurately [38].

The model processes input samples structured as five-channel images, which are normalized band-wise during preprocessing. The input tensor is divided into three separate streams, RGB, NIR, and RE, but it is not fused together until each stream has passed through the modality-specific branches. This separation ensures that spectral and spatial cues are optimally extracted and fused, rather than naively combined at the input level.

To summarize, the proposed architecture possesses several advantages. By combining modality specialization, global attention refinement, adaptive multi-scale fusion, and efficient decoding, it addresses major challenges in crop–weed segmentation: spectral diversity exploitation, long-range spatial reasoning, robustness to environmental variability, and scalability to high-resolution UAV imagery.

3.4 Modality-Specific Feature Extraction Using ConvNeXt Encoders

The initial stage of the proposed segmentation pipeline extracts rich spatial and spectral features from each input modality— RGB, NIR, and RE—using dedicated ConvNeXt encoders. Processing each modality independently preserves their unique semantic and spectral characteristics before further refinement and fusion, ensuring optimal feature representation for crop–weed segmentation.

ConvNeXt [39] was selected as the backbone for feature extraction due to its modernized convolutional design, which integrates Vision Transformer (ViT) principles while retaining the efficiency and strong inductive biases of convolutional neural networks (CNNs). Compared to traditional CNNs like ResNets or lightweight convolutional backbones, ConvNeXt offers superior modeling of complex spatial relationships through larger kernel sizes, simplified block structures, and deeper hierarchical stages. These attributes make it ideal for high-resolution UAV agricultural imagery, where both fine-grained textures (e.g., leaf edges) and broader contextual patterns (e.g., crop row alignments) are critical for accurate segmentation.

Each modality-specific ConvNeXt encoder follows a four-stage hierarchical structure, based on the ConvNeXt-Tiny variant [39]. The input tensor—three channels for RGB

or one channel for NIR and RE—is first processed by a **stem layer**, which applies a 4×4 convolution with stride 4. This operation reduces the spatial resolution from 600×600 pixels to 150×150 while increasing the channel depth to 96. For the RGB branch, a pretrained ConvNeXt-Tiny model, initialized with ImageNet weights, is used to preserve rich visual priors. For NIR and RE branches, the first convolutional layer is modified to accept single-channel inputs, while deeper layers retain pretrained weights, balancing spectral specificity with faster convergence.

The stem output flows through four sequential stages, each with progressively reduced spatial resolution and increased channel depth. The stages are illustrated in Figure 3.5.

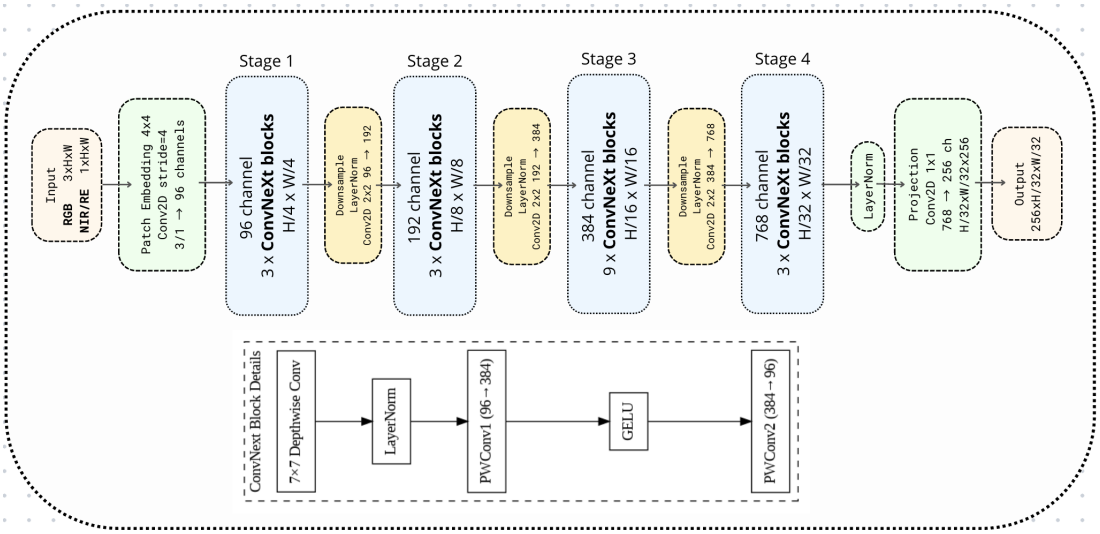

Figure 3.5: Modality-Specific ConvNeXt Encoder Architecture for RGB, NIR, and RE channels. Each modality is processed through a separate ConvNeXt encoder to extract multi-scale features, maintaining spectral specialization before Swin Transformer refinement and FPN fusion.

† **Stage 1 (Shallow Features):** Three ConvNeXt blocks operate at 150×150 resolution with 96 channels. This stage captures fine-grained local details, such as leaf textures, small plant structures, and surface variations, critical for early crop–weed discrimination.

† **Stage 2 (Intermediate Features):** A 2×2 convolution with stride 2 downsamples the feature map to 75×75, increasing channels to 192. This stage encodes mid-level structures, including plant clusters, stems, and soil patterns.

† **Stage 3 (Semantic Features):** Another downsampling reduces the resolution to 37×37, with 384 channels. Nine ConvNeXt blocks model higher-order relationships, such as crop row alignments, weed patches, and canopy coverage patterns.

† **Stage 4 (Global Context Features):** A final downsampling produces feature maps at 18×18 resolution with 768 channels. Three ConvNeXt blocks extract deep semantic features, capturing broad spatial layouts to distinguish crops and weeds under occlusion or complex field conditions.

Each ConvNeXt block employs an inverted bottleneck structure, consisting of:

1. A 7×7 depthwise convolution for spatial mixing, expanding the receptive field compared to standard 3×3 kernels.
2. Layer normalization for stabilizing training.

3. A pointwise 1×1 convolution to expand channels (expansion ratio of 4).

4. GELU activation for non-linearity.

5. A second pointwise 1×1 convolution to reduce channels back to the input dimension.

This design, with large 7×7 kernels, significantly enhances the receptive field, enabling the network to capture contextual relationships across larger spatial extents, which is essential for UAV-based crop mapping where patterns may span tens of meters.

The output feature map from Stage 4 (768 channels, 18×18) is projected to 256 channels using a 1×1 convolution to reduce computational load for subsequent refinement. This projected output is passed to modality-specific Swin Transformer blocks for global context enhancement. Intermediate feature maps from Stages 1, 2, and 3 (at 150×150, 75×75, and 37×37 resolutions, with 96, 192, and 384 channels, respectively) are preserved and, along with the refined Stage 4 output, are later fed to the Feature Pyramid Network (FPN) for multi-scale and cross-modality fusion.

In summary, each modality-specific ConvNeXt encoder produces a pyramid of multi-scale feature maps, with the Stage 4 output (256 channels, 18×18) sent to Swin Transformer blocks for refinement and intermediate features (Stages 1–3) retained for later fusion. The use of ConvNeXt ensures a balance between local detail preservation and global semantic understanding [40], critical for high-accuracy pixel-wise

crop–weed segmentation in heterogeneous agricultural environments. The modular design, with separate encoders for RGB, NIR, and RE, maximizes spectral information exploitation before further processing in the pipeline.

3.5 Transformer-Based Feature Refinement Using Swin Tiny

Following modality-specific feature extraction by the ConvNeXt encoders, the architecture incorporates a transformer-based refinement stage to enhance global context modeling. Agricultural UAV imagery presents challenges such as varying illumination, complex spatial arrangements, and subtle spectral differences between crops and weeds. While ConvNeXt excels at capturing local textures and hierarchical features, it is less effective at modeling long-range spatial dependencies. The Swin Tiny Transformer [18], a lightweight variant of the Swin Transformer, is employed to address this limitation, providing efficient global contextual aggregation.

The Swin Tiny Transformer is designed for hierarchical feature extraction, utilizing a shifted-window multi-head self-attention (W-MSA/SW-MSA) mechanism to balance computational efficiency with global reasoning. Its compact architecture, with approximately 28 million parameters and 4.5G FLOPs for a 224×224 input, makes it

suitable for high-resolution segmentation tasks while maintaining computational feasibility. In this architecture, Swin Tiny is adapted to process the ConvNeXt output, ensuring compatibility with the segmentation pipeline.

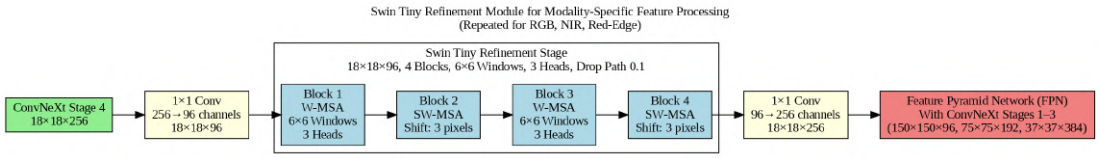

Figure 3.6: Swin Tiny refinement module for modality-specific feature processing. Each module applies four Swin Transformer blocks with window-based (W-MSA) and shifted window (SW-MSA) self-attention at 18×18 resolution to enhance global context.

Each modality-specific ConvNeXt encoder outputs a Stage 4 feature map of size 18×18 with 256 channels. This feature map is projected to 18×18 with 96 channels using a 1×1 convolution, aligning with Swin Tiny's initial embedding dimension. Separate Swin Tiny modules for RGB, Near-Infrared (NIR), and Red-Edge preserve spectral specialization before cross-modality fusion. Given the small input resolution, a single-stage Swin Tiny configuration with four blocks is used to refine features without further downsampling, maintaining the 18×18 resolution.

Swin Tiny Components and Processing

Each Swin Tiny module processes the $18 \times 18 \times 96$ feature map through four Swin Transformer blocks, organized as two pairs alternating between W-MSA and SW-MSA. The blocks are structured as follows:

† **Block 1 (W-MSA):** Applies window-based multi-head self-attention within non-overlapping 6×6 windows, covering approximately three windows per dimension. With three attention heads, W-MSA captures local contextual relationships.

† **Block 2 (SW-MSA):** Uses shifted window multi-head self-attention, applying a cyclic shift of 3 pixels to enable cross-window interactions, propagating information globally.

† **Block 3 (W-MSA):** Repeats W-MSA to further refine local features.

† **Block 4 (SW-MSA):** Repeats SW-MSA to enhance global context.

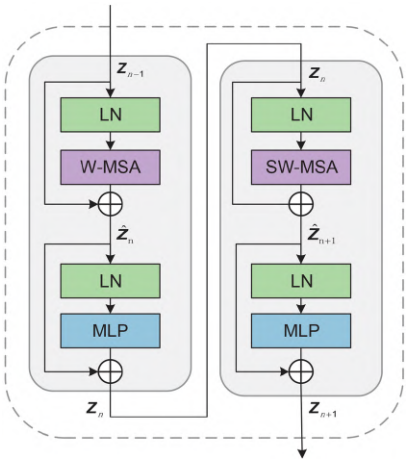

Figure 3.7: Structure of a Swin Transformer block pair, showing the alternating W-MSA and SW-MSA mechanisms with layer normalization, MLP, and residual connections.

Each block includes layer normalization, multi-head self-attention (W-MSA or SW-MSA), a multi-layer perceptron (MLP) with a $4\times$ expansion ratio (96 to 384 intermediate channels), residual connections, and Drop Path regularization (rate 0.1).

The internal structure of a W-MSA/SW-MSA block pair is illustrated in Figure 3.7, showing the alternating attention mechanisms, layer normalization, MLP, and residual connections.

This sequence of blocks expands the effective receptive field, enabling the module to capture large-scale patterns, such as crop row alignments or dispersed weed clusters, which are critical for accurate segmentation in agricultural scenarios.

Output and Integration with FPN

After processing through the four blocks, each Swin Tiny module outputs a refined feature map of size $18 \times 18 \times 96$, enriched with global contextual relationships. This output is projected back to $18 \times 18 \times 256$ using a 1×1 convolution to ensure compatibility with the Feature Pyramid Network (FPN). The projection aligns the channel dimension with other ConvNeXt feature maps for consistent fusion.

The refined Stage 4 feature maps from each modality (RGB, NIR, Red-Edge) are passed to the FPN, along with the intermediate ConvNeXt feature maps from Stages 1–3 ($150 \times 150 \times 96$, $75 \times 75 \times 192$, $37 \times 37 \times 384$). The FPN employs a top-down architecture with lateral connections to integrate these multi-scale and multi-modal features, producing a set of feature maps at resolutions 150×150, 75×75, 37×37, and 18×18, each with 256 channels. The highest-resolution feature map (150×150)

is then fed to the decoder, which generates the final 600×600 segmentation map distinguishing crops, weeds, and background.

The overall architecture of the Swin Tiny refinement module is depicted in Figure 3.6, showing the pipeline from ConvNeXt input to FPN integration.

In summary, the Swin Tiny refinement stage enhances ConvNeXt features by modeling long-range dependencies, critical for recognizing complex agricultural patterns. Its efficient window-based attention and modality-specific design ensure robust performance while preserving spectral information. The refined features, combined with ConvNeXt's multi-scale outputs, enable the FPN to produce accurate semantic segmentation under diverse field conditions.

3.6 Gated Multi-Scale Fusion and Pyramid Pooling Decoder

Following modality-specific feature extraction by the ConvNeXt encoders and global context enhancement by the Swin Tiny Transformer modules, the architecture proceeds with a crucial stage: adaptive fusion and decoding. In precision agriculture, effectively integrating multispectral data is pivotal due to variable field conditions, sensor inconsistencies, and diverse vegetation types. Thus, we introduce a tailored

multi-scale Feature Pyramid Network (FPN) fusion module with an adaptive gating mechanism, followed by a context-aware pyramid pooling decoder.

The multi-scale fusion stage integrates modality-specific feature maps from RGB, Near-Infrared (NIR), and Red-Edge (RE) modalities, extracted by ConvNeXt and refined by Swin Tiny. At each of the four feature scales extracted by ConvNeXt (1/4, 1/8, 1/16, 1/32 of the original 600×600 resolution, corresponding to 150×150, 75×75, 37×37, and 18×18), as well as the high-level transformer-refined maps at 18×18, lateral 1×1 convolutions are first applied to unify the channel dimensions to 256. This ensures consistent channel depths across scales and modalities for subsequent fusion.

To dynamically emphasize the most informative modality at each spatial scale, we employ learned *gating convolutions*. Specifically, at each scale, the modality-specific feature maps are processed by 1×1 convolutions to generate intermediate representations, which are then passed through additional 1×1 convolutions followed by sigmoid activations to produce the gating weights. Each modality's feature map is then element-wise multiplied by its corresponding learned gating weights, enabling the model to adaptively balance modality contributions based on local spectral and spatial context. This gated fusion block, applied at each scale, is illustrated in Figure 3.8.

Formally, for modality-specific feature maps F_{RGB}, F_{NIR}, and F_{RE} at scale s, the gated

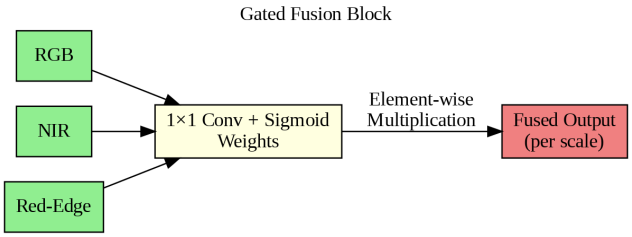

Figure 3.8: Illustration of the gated FPN block, showing the adaptive modality weighting process using 1×1 convolutions and sigmoid activations to generate modality-specific weights for fusion at each scale.

fusion can be expressed as:

$$F_{\text{fused}}^{(s)} = \sigma(G_{\text{RGB}}^{(s)}) \odot F_{\text{RGB}}^{(s)} + \sigma(G_{\text{NIR}}^{(s)}) \odot F_{\text{NIR}}^{(s)} + \sigma(G_{\text{RE}}^{(s)}) \odot F_{\text{RE}}^{(s)} \tag{3.1}$$

where σ denotes the sigmoid activation, $G_{\text{modality}}^{(s)}$ are the gating weights, and $\odot$ represents element-wise multiplication.

After gating-based modality weighting, the resulting fused features at each scale are combined via a hierarchical top-down pathway with lateral connections. Starting from the deepest scale (18×18), features are progressively upsampled to the next scale (e.g., 37×37) using bilinear interpolation and combined with higher-resolution features (37×37, 75×75, 150×150) through element-wise addition, followed by refinement using 3×3 convolutional layers with ReLU activations. This process produces a set of multi-scale feature maps at resolutions 150×150, 75×75, 37×37, and 18×18, each with 256 channels, aggregating both fine-grained and high-level semantic context. Figure 3.9 visually illustrates this gated multi-scale fusion mechanism.

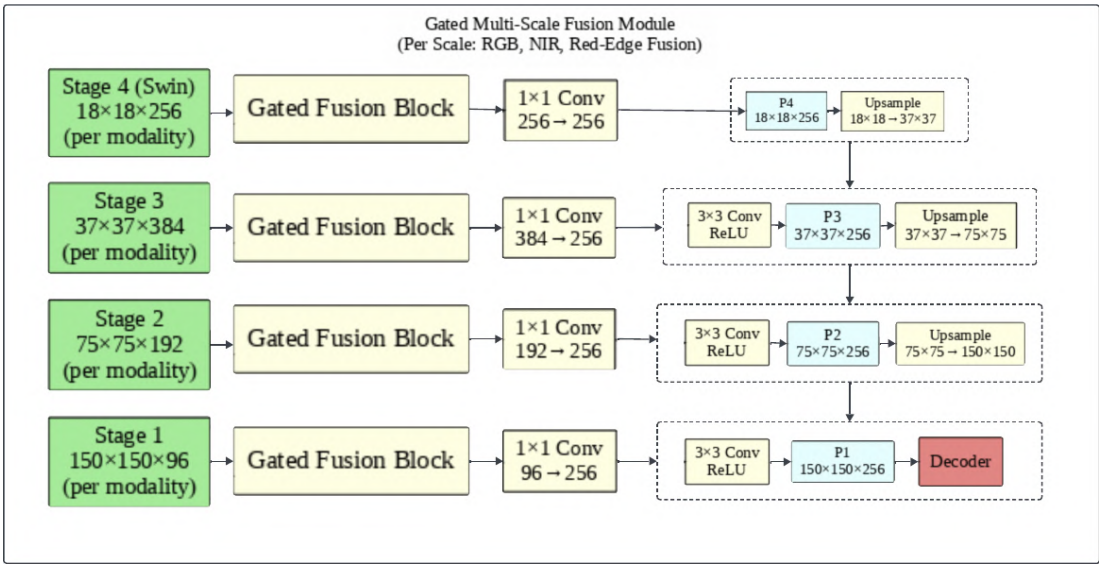

Figure 3.9: Illustration of the gated multi-scale fusion module, employing adaptive modality weighting via gating convolutions and hierarchical multi-scale aggregation through an FPN-style structure.

To decode the fused multi-scale representation into dense, pixel-level segmentation maps, we employ a context-aware decoder inspired by the UPerNet architecture. The decoder takes the highest-resolution fused feature map from the FPN at 150×150 and utilizes a Pyramid Pooling Module (PPM) to capture spatial context at multiple scales effectively. Specifically, the fused features undergo spatial pooling at four predefined scales (1×1, 2×2, 3×3, and 6×6), each followed by a 1×1 convolution to compress feature dimensions. These multi-scale pooled representations are then upsampled to 150×150 using bilinear interpolation and concatenated with the original fused feature map, enriching the representation with comprehensive global context alongside local details.

Subsequently, this combined feature map passes through two additional 3×3 convolutional layers, each with Batch Normalization and ReLU activation, to refine spatial consistency and enhance semantic accuracy. These layers enable the network to smooth out inconsistencies introduced by multi-scale concatenation and learn complex feature interactions, improving boundary delineation and semantic precision for distinguishing crops and weeds. A final 1×1 convolution generates pixel-wise class logits for the three semantic classes: background, crop, and weed. Finally, the class logits are upsampled to the original 600×600 resolution using bilinear interpolation for pixel-wise classification. This decoding approach significantly improves boundary localization and segmentation robustness, particularly crucial for accurate delineation of complex vegetation structures in agricultural imagery. Figure 3.10 illustrates the pyramid pooling decoding structure employed.

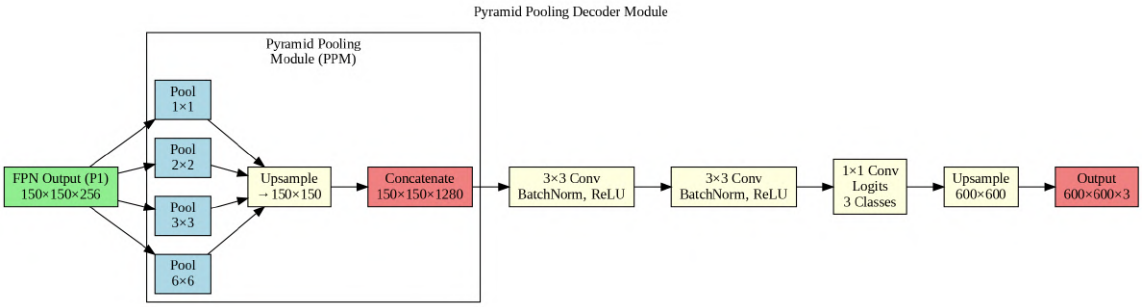

Figure 3.10: Context-aware Pyramid Pooling Decoder module. Multi-scale context pooling and fusion, followed by refinement convolutions, enhance boundary precision and semantic consistency, crucial for agricultural segmentation.

Overall, the integration of gated multi-scale fusion and pyramid pooling decoding is essential for producing robust segmentation results. The adaptive fusion strategy

dynamically emphasizes the most informative spectral modalities and scales, while the context-rich decoding ensures accurate, fine-grained delineation of crops and weeds. Collectively, these modules significantly contribute to the proposed architecture's effectiveness and reliability in multispectral semantic segmentation tasks for precision agriculture.

Chapter 4

Evaluation

This chapter shows the comprehensive evaluation of the proposed architecture. It includes the complete training pipeline, loss functions, optimization methods and experimental setup, with an analysis of both quantitative and qualitative results. Finally, a conclusion with a performance comparison against state-of-the-art baselines from the WeedsGalore benchmark.

4.1 Training Strategy and Implementation Details

4.1.1 Loss Function and Class Imbalance Handling

In this thesis, due to the extreme class imbalance in the WeedsGalore dataset, containing a significant number of background pixels and far fewer crop and weed pixels, a composite loss function consisting of Dice Loss and Class-Balanced Focal Loss (CB-Focal) was used.

† **Dice Loss** encourages overlap between prediction and ground truth masks, especially effective for imbalanced segmentation and boundary-sensitive regions [41]:

$$\mathcal{L}_{Dice} = 1 - \frac{2\sum_{i=1}^{N} p_i y_i + \epsilon}{\sum_{i=1}^{N}(p_i^2 + y_i^2) + \epsilon}$$

† **Class-Balanced Focal Loss** scales the classic focal loss by the inverse effective number of samples for each class, reducing bias toward dominant classes and emphasizing hard examples [42]:

$$\mathcal{L}_{CB-Focal} = -\frac{1-\beta}{1-\beta^{n_c}}(1-p_t)^{\gamma}\log(p_t)$$

The total training loss is defined as:

$$\mathcal{L}_{\text{total}} = \lambda_1 \cdot \mathcal{L}_{Dice} + \lambda_2 \cdot \mathcal{L}_{CB-Focal}, \quad \text{with} \quad \lambda_1 = \lambda_2 = 1.0$$

4.1.2 Optimization and Scheduling

Training is conducted using the **AdamW** optimizer with the following hyperparameters:

† Learning rate: 1×10^{-4}

† Weight decay: 1×10^{-2}

† Scheduler: Cosine Annealing with 10 warm-up epochs, total training = 200 epochs

† Mixed Precision: Enabled using PyTorch Automatic Mixed Precision (AMP)

4.1.3 Data Augmentation and Normalization

Real-time data augmentation to improve model generalization was applied:

† Horizontal and vertical flips

† Random rotations (0°, 90°, 180°, 270°)

† Brightness and contrast jittering

† Horizontal scaling and aspect ratio warping

All five spectral channels are normalized to zero mean and unit variance using statistics computed from the training set.

4.1.4 Training Infrastructure

† **Hardware:** NVIDIA A100 GPU (80GB VRAM)

† **Software Stack:** PyTorch 2.0, CUDA 12.4, cuDNN 8.9

† **Environment:** Ubuntu 20.04, Python 3.10

† **Libraries:** TIMM 0.9.10, HuggingFace Transformers 4.37.2

† **Memory Optimization:** `PYTORCH_CUDA_ALLOC_CONF=expandable_segments:True`

4.2 Quantitative Results

Two input configurations were studied to assess the performance of the proposed architecture, and detailed experimentation was performed on the WeedsGalore test

set:

1. **Multispectral (MSI) input:** 5-channel input comprising RGB, Near-Infrared (NIR), and Red-Edge (RE) bands.

2. **RGB-only input:** Standard 3-channel RGB input for baseline comparison.

To evaluate segmentation quality, the following metrics are used:

† **Mean Intersection-over-Union (mIoU):** Measures the average overlap between predicted and ground truth regions across all classes. For class c, IoU is calculated as:

$$\text{IoU}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c + \text{FN}_c}$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively.

† **Pixel Accuracy (Acc):** Computes the ratio of correctly predicted pixels to the total number of pixels:

$$\text{Acc} = \frac{\sum_{c=0}^{2} \text{TP}_c}{\text{Total Pixels}}$$

† **F1 Score (F1):** Represents the harmonic mean of precision and recall for each class:

$$\mathrm{F1}_c = 2 \times \frac{\mathrm{Precision}_c \times \mathrm{Recall}_c}{\mathrm{Precision}_c + \mathrm{Recall}_c}$$

where:

$$\mathrm{Precision}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c} \quad \text{and} \quad \mathrm{Recall}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c}$$

These metrics are selected to address class imbalance and ensure per-class performance, critical for agricultural applications where weeds may be sparse compared to crops and background.

Performance Metrics

The test set evaluation demonstrates a significant performance improvement when leveraging multispectral inputs over RGB alone.

The MSI configuration significantly outperforms RGB-only across all metrics, especially for the crop and weed classes. The improvements in crop IoU (+47.95%) and weed IoU (+31.87%) highlight the strong benefits of integrating spectral modalities

Table 4.1
Performance Comparison: Multispectral vs RGB-only Inputs on WeedsGalore Test Set

Class	Input	IoU	F1 Score	Precision	Recall
Background	MSI	0.9900	0.9950	0.9950	0.9949
	RGB	0.9794	0.9896	0.9863	0.9930
Crop	MSI	0.8700	0.9305	0.9316	0.9294
	RGB	0.3905	0.5616	0.6477	0.4957
Weed	MSI	0.8411	0.9137	0.9116	0.9158
	RGB	0.5224	0.6863	0.7028	0.6705
Mean IoU		**90.04%** (MSI)	63.08% (RGB)		
Overall Accuracy		**99.03%** (MSI)	97.12% (RGB)		

such as NIR and Red-Edge.

Confusion Matrices

The following confusion matrices summarize the raw pixel classification results for both models:

Multispectral (MSI) Model:

$$\begin{bmatrix} \mathbf{15791643} & 21114 & 60107 \\ 20382 & \mathbf{310255} & 3202 \\ 58389 & 1677 & \mathbf{653231} \end{bmatrix}$$

RGB-only Model:

$$\begin{bmatrix} \mathbf{8615339} & 19656 & 41184 \\ 40879 & \mathbf{92496} & 53210 \\ 79047 & 30647 & \mathbf{223266} \end{bmatrix}$$

In both cases, background pixels dominated in the predictions. However, the MSI model shows much better classification consistency for both crop and weed pixels, reducing misclassification errors across minority classes. The substantial drop in false negatives and improved diagonal dominance validate the model's improved precision and recall for each class.

The proposed model achieved a **mean IoU value of 90.04%** which is significantly higher than the models with RGB only data. This highlights the importance of spectral diversity in pixel-level classification and confirms the value of attention-based multimodal fusion and multi-scale decoding in real-world agricultural segmentation tasks.

4.3 Qualitative Results

In addition to quantitative measurements, qualitative assessment provides further visual evidence that the proposed model works well to segment complex scenes from the field. This section presents side-by-side visual comparisons of the RGB-only and

multispectral (MSI) input configurations using a few representative samples from the WeedsGalore dataset test set.

Each visualization includes the original input image, ground truth annotation, and predicted segmentation mask. Additionally, the individual spectral bands—RGB, Near-Infrared (NIR), and Red-Edge (RE) were visualized for the MSI setup to highlight their contribution to model performance.

Example 1: Row-Structured Maize with Sparse Weeds

In Figure 4.1, the RGB-only model correctly identifies most background and some crop regions, but fails to delineate small weeds and produces noisy predictions along row boundaries. Thin crop structures are often missed or misclassified as weeds due to limited spectral contrast in the RGB channels.

Example 2: Multispectral Input with Enhanced Delineation

Figure 4.2 and Figure 4.3 shows the segmentation output using MSI input. Compared to the RGB-only result, the proposed model generates cleaner boundaries, accurately distinguishes maize crops from surrounding weeds, and correctly identifies sparse and occluded weed patches. The inclusion of NIR improves vegetation-background

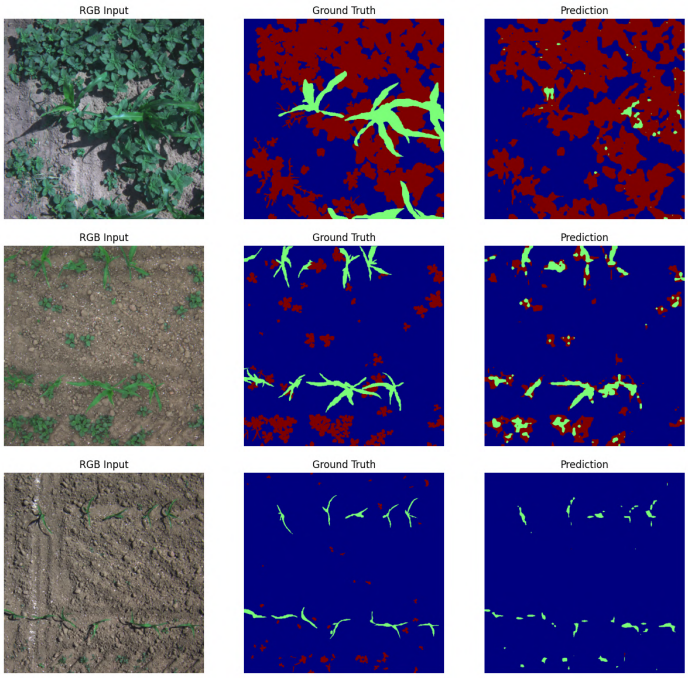

Figure 4.1: RGB-only configuration: input image (RGB), ground truth mask, and prediction segmentation. There are blurred crop boundaries and false-positive weed predictions.

separation, while Red-Edge contributes to detecting subtle differences in chlorophyll concentration between crops and weeds.

4.4 Comparison with Baseline Methods

A further comparative analysis was conducted against baseline models reported in the WeedsGalore benchmark [10] such as DeepLabv3+ and MaskFormer, respectively, two different semantic segmentation paradigms: a CNN and a Transformer. All models were evaluated on the identical 3-class background, crop, and weed segmentation, using RGB only and multispectral (MSI: RGB + NIR + RE) input modalities under

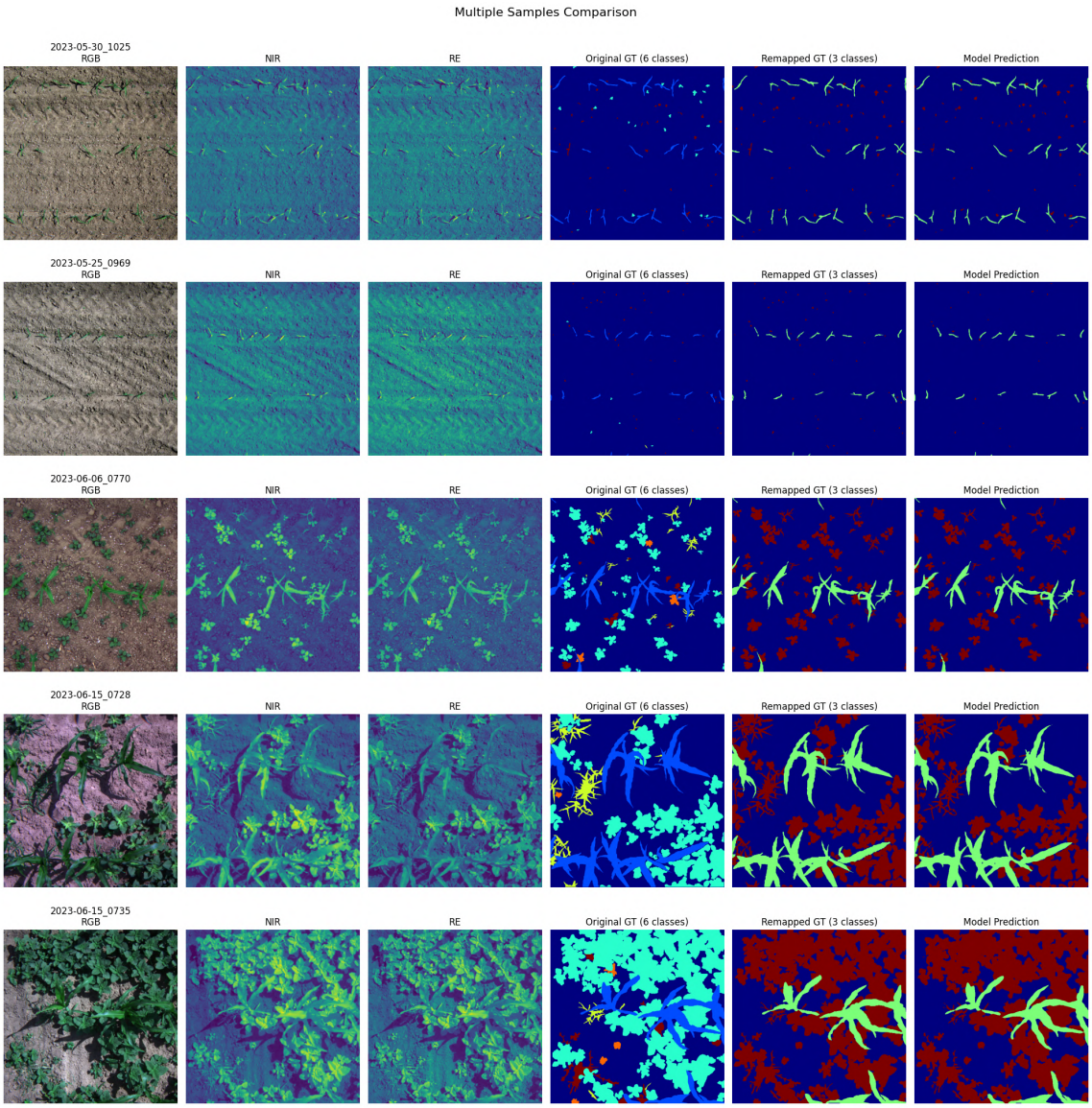

Figure 4.2: Multispectral configuration: input image (RGB, NIR, and RE bands), ground truth, and predicted segmentation based on MSI input. The output produces a sharp and accurate boundary for all classes.

the same testing conditions.

As presented in Table 4.2, the proposed model demonstrates superior performance across both vegetation classes. It achieves a mean IoU (mIoU) of **90.04%**, outperforming the best-performing baseline (DeepLabv3+ with MSI input) by a margin of

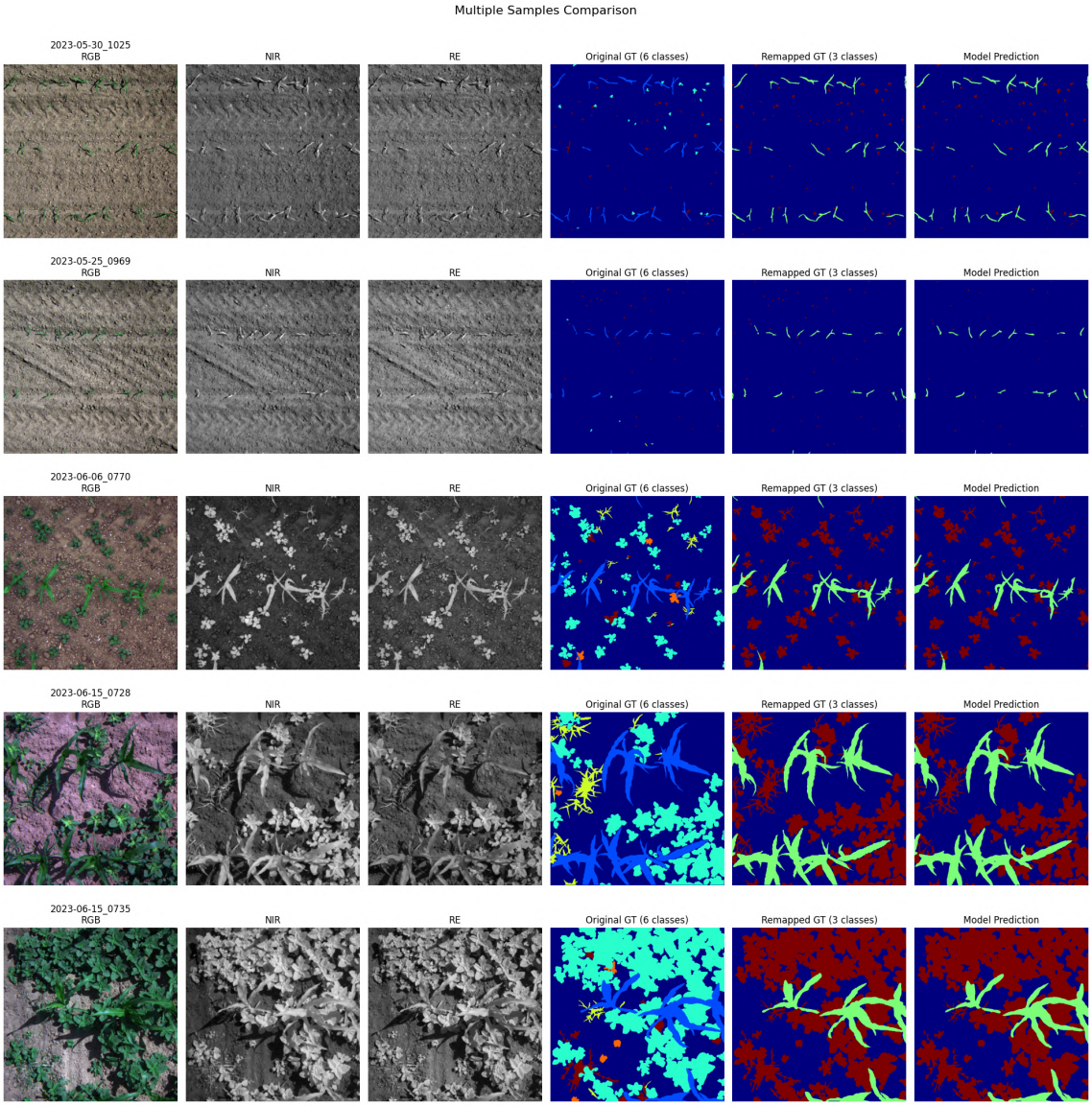

Figure 4.3: Multispectral configuration: (RGB, NIR (grey scaled), and RE (grey scaled) bands), followed by ground truth and MSI-based predicted segmentation.

+7.14 percentage points. Relative to MaskFormer, improvements of +9.77% (MSI) and +10.78% (RGB) are also observed.

Table 4.2
Semantic segmentation performance (%) on the WeedsGalore test set (3-class configuration).

Model	**Input Modality**	**IoU$_{\text{Crop}}$**	**IoU$_{\text{Weed}}$**	**Mean IoU**
DeepLabv3+ [10]	RGB	67.93	72.08	79.33
DeepLabv3+ [10]	MSI	72.93	77.31	82.90
MaskFormer [10]	RGB	70.18	69.85	79.26
MaskFormer [10]	MSI	69.49	73.33	80.27
Proposed Model	RGB	39.05	52.24	63.08
Proposed Model	MSI	**87.00**	**84.11**	**90.04**

The improvements in performance - and a comparison to existing architectures - confirm the complementary development of architectural combinations made throughout this design. Specifically, the use of modality-specialized ConvNeXt encoders allows independent extraction of spectral features, while Swin Transformer modules introduce global context modeling through efficient self-attention. The gated multi-scale fusion mechanism further enhances feature integration across modalities and spatial resolutions, enabling robust boundary localization and semantic discrimination.

The RGB-only variant of the proposed model, while achieving an mIoU of 63.08%, performs less competitively when compared to baseline models optimized for single-modality settings. This can be attributed to several design-related factors:

† The architecture is tailored for multispectral inputs and includes modality-specific encoders and fusion blocks. When restricted to RGB-only input, two out of three encoder branches are inactive, resulting in reduced model utilization.

† Unlike the baseline models, which are fully optimized for RGB, the proposed model's components—particularly the gated fusion and Swin Transformer modules—are under-used in the absence of multispectral diversity.

† The training configuration was primarily optimized for MSI, and no separate tuning or architecture pruning was performed for RGB-only inference.

Despite this, the large gap between the RGB-only (63.08%) and MSI configuration (90.04%) underscores the critical role of spectral diversity in achieving high segmentation accuracy.

The proposed model achieved the highest segmentation performance in the WeedsGalore benchmark compared to previously reported studies. Its performance margin across both crop and weed classes demonstrates its potential for use in precision agriculture, particularly when collected in multispectral UAV conditions.

Chapter 5

Cross-Domain Generalization Experiments

For real-world precision agriculture applications, semantic segmentation models must be able to perform universally across various field conditions, crop types, and sensor arrangements. However, most state-of-the-art approaches are trained and evaluated only in a single dataset domain, reducing their robustness in domain shift.

This chapter evaluates the proposed segmentation model's generalization performance beyond the source domain (WeedsGalore) using two external datasets: *Carrots 2017* and *Onions 2017*. These datasets pose notable challenges due to their different crop structures, imaging setups, and spatial configurations. Both zero-shot inference and

few-shot adaptation experiments are conducted to assess model robustness and adaptability.

5.1 Target Datasets: Carrots 2017 and Onions 2017

The Carrots 2017 (CA17) and Onions 2017 (ON17) datasets were collected using a ground-based robotic platform equipped with dual RGB and NIR cameras. NDVI images were derived from the raw RGB-NIR data to facilitate spectral vegetation analysis. Pixel-wise ground truth masks are available for three classes: background, crop, and weed.

Table 5.1
Summary of Carrots 2017 and Onions 2017 Datasets

Attribute	**Carrots 2017**	**Onions 2017**
Collection Date	June 2017	April 2017
Location	North Scarle, UK	South Scarle, UK
Number of Images	20	20
Image Resolution	2428×1985	2419×1986
Modalities	RGB, NIR, NDVI	RGB, NIR, NDVI
Ground Truth Labels	Background, Crop, Weed	Background, Crop, Weed
Vegetation Density	High	Sparse
Crop Stage	Late-stage carrots	Early-stage onions
Crop Instances/Image	88	52
Weed Instances/Image	86	22

Compared to the structured maize fields of WeedsGalore, CA17 and ON17 introduce

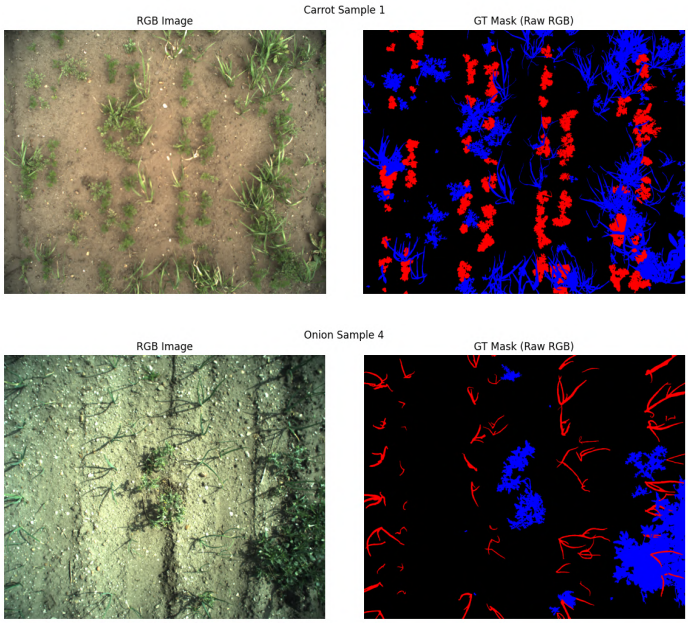

Figure 5.1: Top: Sample from Carrots 2017. Bottom: Sample from Onions 2017. Left: RGB input; Right: Ground truth mask (Red = Weed, Blue = Crop, Black = Background).

domain shifts in canopy geometry, crop spacing, sensor angles, and spectral coverage (absence of Red-Edge). These differences make them suitable benchmarks for testing the generalization capacity of the proposed architecture.

5.2 Zero-Shot Evaluation

To evaluate out-of-domain generalization, the pretrained model was directly tested on CA17 and ON17 without any fine-tuning. The model was trained on WeedsGalore with five-channel inputs (RGB+NIR+RE), while CA17 and ON17 inputs were aligned using RGB+NDVI and a zero-filled RE channel to maintain input dimensionality.

Carrots 2017

The model showed strong performance on background pixels but failed to generalize crop and weed classes due to occlusion, dense foliage, and structural variance.

† **mIoU:** 0.3596 **Accuracy:** 87.08%

† **Crop IoU:** 0.0141 **Weed IoU:** 0.1874

† **F1 Scores:** Crop = 0.0271, Weed = 0.3110

Onions 2017

Despite accurate background predictions, segmentation of thin, sparse onion crops remained poor.

† **mIoU:** 0.3473 **Accuracy:** 91.89%

† **Crop IoU:** 0.0899 **Weed IoU:** 0.0210

† **F1 Scores:** Crop = 0.1509, Weed = 0.0411

Zero-shot transfer exhibited limited generalization. While background segmentation

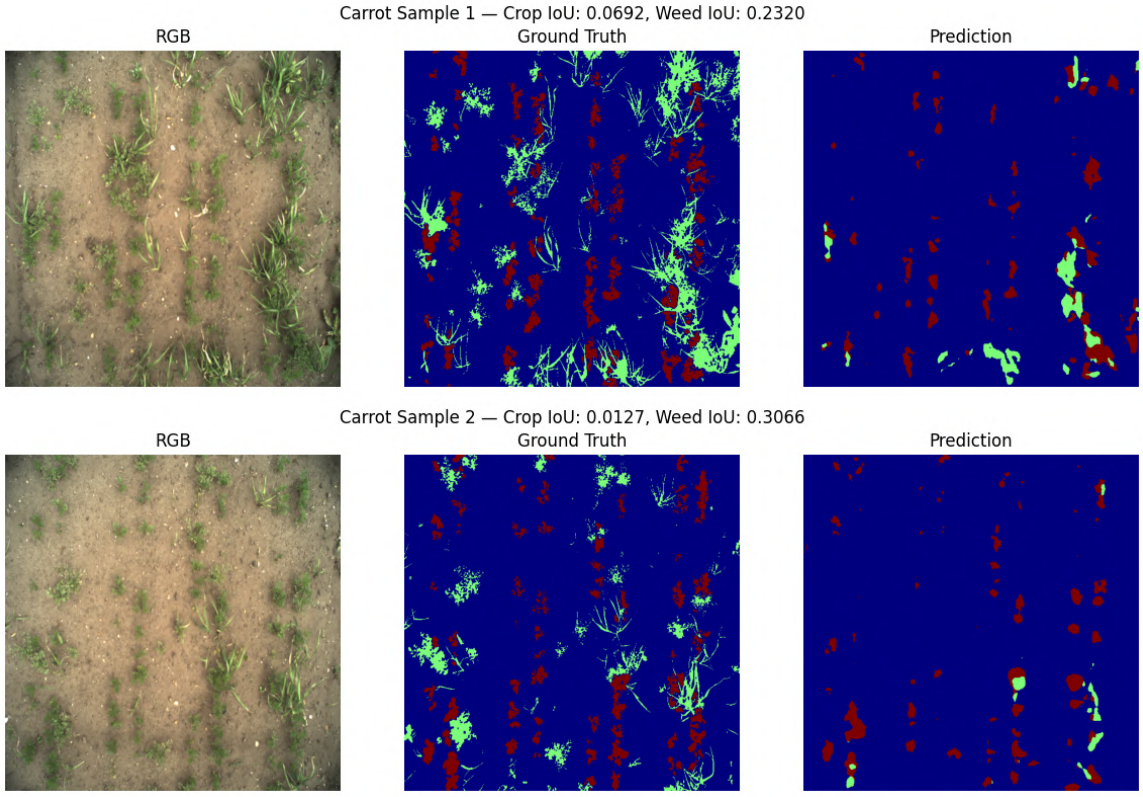

Figure 5.2: Carrots 2017 – Zero-shot prediction vs. ground truth. Crops are under-segmented and misclassified as background.

remained effective, crop and weed classes were poorly segmented, motivating the need for minimal adaptation.

5.3 Few-Shot Adaptation

To improve performance under domain shift, few-shot adaptation was explored. The model was fine-tuned on $N = \{5, 10, 15\}$ samples from each dataset using three random splits per shot size. Only the decoder and fusion modules were updated

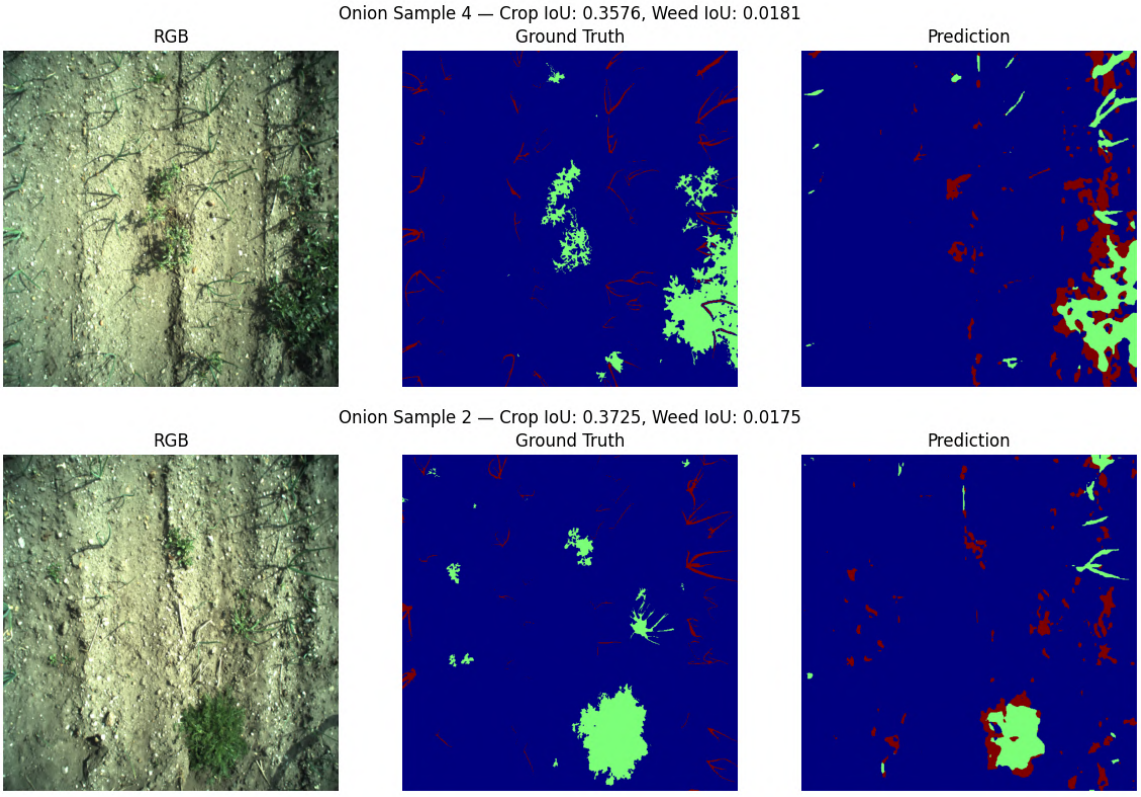

Figure 5.3: Onions 2017 – Zero-shot prediction vs. ground truth. Sparse vegetation leads to under-segmentation of foreground classes.

while the pretrained ConvNeXt and Swin Transformer backbones remained frozen.

Table 5.2
Few-shot adaptation results on Carrots17 and Onions17 (mean ± std over 3 random splits).

Dataset	Shots	mIoU	F1 Score	Accuracy
Carrots17	5	0.6415 ± 0.0084	0.7610 ± 0.0087	0.9221 ± 0.0012
	10	0.6764 ± 0.0029	0.7922 ± 0.0024	0.9283 ± 0.0003
	15	0.6923 ± 0.0081	0.8051 ± 0.0066	0.9317 ± 0.0011
Onions17	5	0.5762 ± 0.0145	0.6825 ± 0.0152	0.9706 ± 0.0016
	10	0.6154 ± 0.0052	0.7210 ± 0.0051	0.9743 ± 0.0004
	15	0.6299 ± 0.0033	0.7341 ± 0.0037	0.9757 ± 0.0003

Adaptation led to substantial recovery in segmentation quality. The Carrots17 dataset benefited the most, with mIoU reaching 69.2% under 15-shot supervision.

Onions17 exhibited more modest improvements, likely due to its simpler structure and limited visual variation.

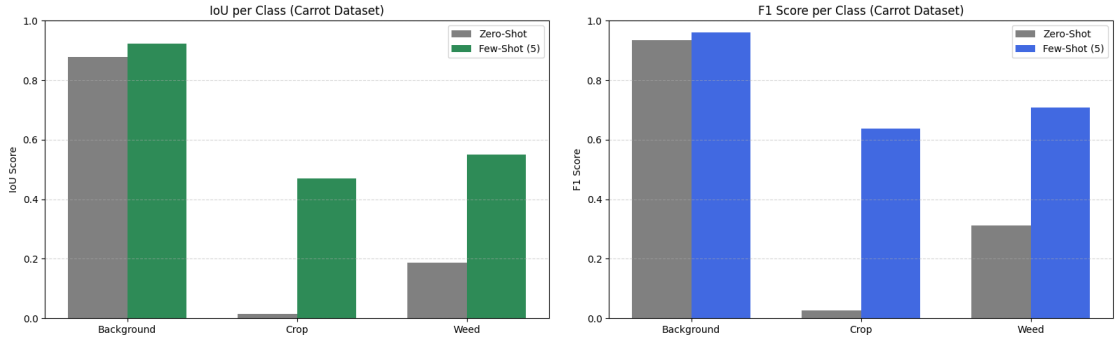

Figure 5.4: Carrots17 – Per-class IoU and F1 score before and after few-shot adaptation.

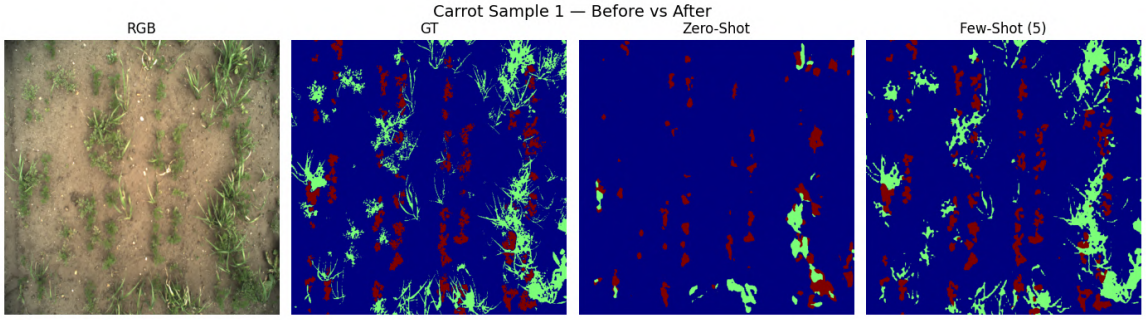

Figure 5.5: Qualitative results on Carrots17. Left to right: RGB input, ground truth, zero-shot prediction, few-shot prediction (5-shot).

Interestingly, performance gains plateaued between 10 and 15 shots in both datasets. This suggests decreasing returns from additional supervision beyond a small sample threshold. A potential explanation lies in scale mismatches between the source (UAV-based) and target (ground robot) imagery—differences in viewpoint, spacing, and object scale may introduce inconsistencies that limit transfer. To address this, future adaptation efforts may benefit from scale-aware augmentations or larger support sets (e.g., 50–100 samples) to cover broader intra-domain variability.

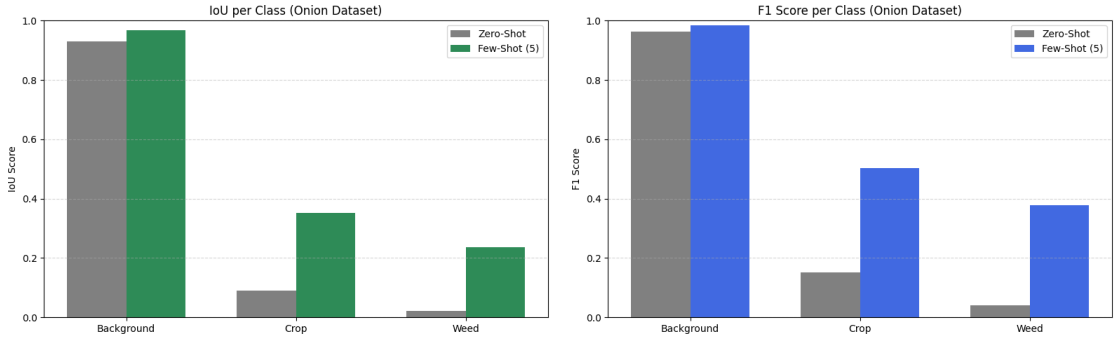

Figure 5.6: Onions17 – Per-class IoU and F1 score before and after few-shot adaptation.

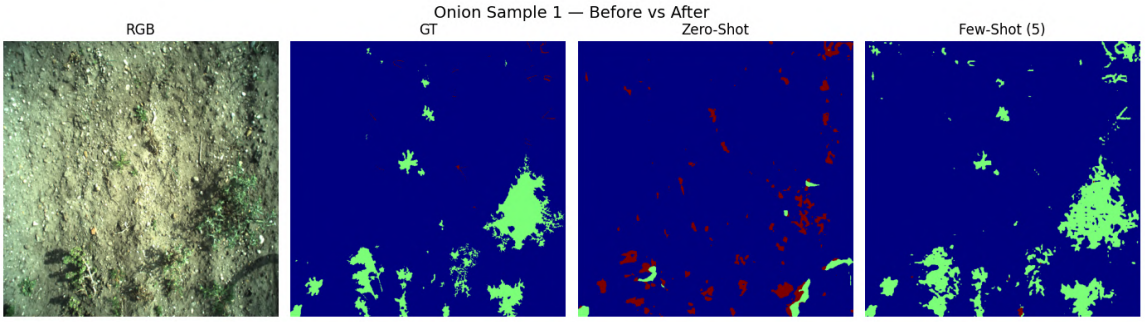

Figure 5.7: Qualitative results on Onions17. Left to right: RGB input, ground truth, zero-shot prediction, few-shot prediction (5-shot).

The proposed segmentation model exhibited limited zero-shot generalization, but strong adaptability with minimal supervision. Few-shot fine-tuning with as few as 5–10 labeled images enabled rapid recovery of segmentation accuracy, validating the model's suitability for deployment in new field environments where annotations are scarce.

Chapter 6

Conclusion

The thesis presented a novel multimodal deep learning architecture for the semantic segmentation of crops and weeds to enhance precision agriculture. Drawing on the spectral richness of multispectral UAV imagery to look beyond simple RGB images, the architecture contained modality-specific ConvNeXt encoders, Swin Transformer refinement modules, a gated multi-scale fusion approach, and a context-aware decoder. The architecture was modular in its design (modular, scalable, etc.) and could accommodate the spatial and spectral variability observed within agricultural fields.

The model showed state-of-the-art segmentation performance on the WeedsGalore benchmark, which consists of RGB, NIR, and Red-Edge (RE) bands, achieving a mean

Intersection-over-Union (mIoU) of **90.04%**—surpassing existing baselines including DeepLabv3+ and MaskFormer by a significant margin. In addition to being statistically accurate, the model also showed robustness across class boundaries—particularly for underrepresented or occluded crop and weed structures, due to its attention-based fusion and spectral-aware design.

To evaluate the ability to generalize, a comprehensive range of cross-domain tests was done using the Carrots 2017 and Onions 2017 datasets, which set new and difficult challenges of crop geometry, planting density, and acquisition conditions. With zero-shot evaluation, the deficiencies of the direct transfer approach were disclosed by the fact that the model was not capable of correctly segmenting non-seen crop types with just the pre-trained weights. In contrast, few-shot adaptation experiments, where 5 to 15 labeled samples were used, demonstrated the model's good capability to gain back strong performance only with a little supervision, i.e. the mIoU scores were over 69% on Carrots 2017 and 63% on Onions 2017. So these results can be taken as evidence of a practical application of the model in the real-world situations where there are marked limitations on labeled data.

The study has limitations despite its contributions. The model was trained on a single source domain, which may limit its exposure to broader crop and field variations. Additionally, the lack of Red-Edge channels in the external datasets required

zero-padding during domain adaptation, slightly reducing spectral effectiveness. Furthermore, the observed performance plateau between 10 and 15 samples suggests a possible bottleneck related to scale mismatches between datasets. Variations in plant spacing, sensor height, and object size may restrict the model's capacity to generalize spatial patterns beyond those learned in the source domain.

Future work can build upon this foundation in several ways. Incorporating scale-aware representations or positional encodings may address cross-domain resolution issues. Integrating domain adaptation techniques—such as adversarial learning or meta-learning—could further improve generalization to new crop types without relying on labeled data. Additionally, extending the model to support fine-grained weed species segmentation would provide more actionable insights for agricultural robotics. From a deployment perspective, efforts toward pruning, optimization, quantization, and distillation can make the model feasible for real-time inference on edge devices such as UAVs or embedded platforms.

In conclusion, this work demonstrates that deep multimodal architectures, when guided by spectral understanding and modular design, can effectively advance the state of semantic segmentation for precision agriculture. The model's adaptability with minimal supervision highlights its potential for rapid deployment across diverse field conditions, contributing toward scalable, data-efficient solutions in sustainable crop management.

References

[1] Government Accountability Office, "Precision agriculture: Benefits and challenges for technology adoption and use," U.S. Government Accountability Office, Tech. Rep. GAO-24-105962, January 2024, published: January 31, 2024. Publicly Released: January 31, 2024. [Online]. Available: https://www.gao.gov/products/gao-24-105962

[2] R. Gerhards, D. Andujar Sanchez, P. Hamouz, G. G. Peteinatos, S. Christensen, and C. Fernandez-Quintanilla, "Advances in site-specific weed management in agriculture—a review," *Weed Research*, vol. 62, no. 2, pp. 123–133, 2022.

[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[4] S. Candiago, F. Remondino, M. De Giglio, M. Dubbini, and M. Gattelli, "Evaluating multispectral images and vegetation indices for precision farming applications from uav images," *Remote sensing*, vol. 7, no. 4, pp. 4026–4047, 2015.

[5] J. L. E. Honrado, D. B. Solpico, C. M. Favila, E. Tongson, G. L. Tangonan, and N. J. Libatique, "Uav imaging with low-cost multispectral imaging system for precision agriculture applications," in *2017 IEEE Global Humanitarian Technology Conference (GHTC)*. IEEE, 2017, pp. 1–7.

[6] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[7] B. Yu, L. Yang, and F. Chen, "Semantic segmentation for high spatial resolution remote sensing images based on convolution neural network and pyramid pooling module," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 9, pp. 3252–3261, 2018.

[8] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 558–567.

[9] J. Zhao, T. W. Berge, and J. Geipel, "Transformer in uav image-based weed mapping," *Remote Sensing*, vol. 15, no. 21, p. 5165, 2023.

[10] E. Celikkan, T. Kunzmann, Y. Yeskaliyev, S. Itzerott, N. Klein, and M. Herold, "Weedsgalore: A multispectral and multitemporal uav-based dataset for crop and weed segmentation in agricultural maize fields," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 4767–4777.

[11] I. Sa, M. Popović, R. Khanna, Z. Chen, P. Lottes, F. Liebisch, J. Nieto, C. Stachniss, A. Walter, and R. Siegwart, "Weedmap: A large-scale semantic weed mapping framework using aerial multispectral imaging and deep neural network for precision farming," *Remote Sensing*, vol. 10, no. 9, p. 1423, 2018.

[12] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," 2017. [Online]. Available: https://arxiv.org/abs/1704.06857

[13] A. Milioto, P. Lottes, and C. Stachniss, "Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in cnns," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 2229–2235.

[14] F. Magistri, J. Weyler, D. Gogoll, P. Lottes, J. Behley, N. Petrinic, and C. Stachniss, "From one field to another—unsupervised domain adaptation for semantic segmentation in agricultural robotics," *Computers and Electronics in Agriculture*, vol. 212, p. 108114, 2023.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[16] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[18] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[19] R. Reedha, E. Dericquebourg, R. Canals, and A. Hafiane, "Transformer neural network for weed and crop classification of high resolution uav images," *Remote sensing*, vol. 14, no. 3, p. 592, 2022.

[20] G. Castellano, P. De Marinis, and G. Vessio, "Weed mapping in multispectral drone imagery using lightweight vision transformers," *Neurocomputing*, vol. 562, p. 126914, 2023.

[21] J. Fang, H. Lin, X. Chen, and K. Zeng, "A hybrid network of cnn and transformer for lightweight image super-resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1103–1112.

[22] J. Su, D. Yi, M. Coombes, C. Liu, X. Zhai, K. McDonald-Maier, and W.-H. Chen, "Spectral analysis and mapping of blackgrass weed by leveraging machine learning and uav multispectral imagery," *Computers and Electronics in Agriculture*, vol. 192, p. 106621, 2022.

[23] M. Dalla Mura, S. Prasad, F. Pacifici, P. Gamba, J. Chanussot, and J. A. Benediktsson, "Challenges and opportunities of multimodality and data fusion in remote sensing," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1585–1601, 2015.

[24] B. Kayalibay, G. Jensen, and P. van der Smagt, "Cnn-based segmentation of medical imaging data," *arXiv preprint arXiv:1701.03056*, 2017.

[25] N. Genze, R. Ajekwe, Z. Güreli, F. Haselbeck, M. Grieb, and D. G. Grimm, "Deep learning-based early weed segmentation using motion blurred uav images of sorghum fields," *Computers and Electronics in Agriculture*, vol. 202, p. 107388, 2022.

[26] M. Halstead, P. Zimmer, and C. McCool, "A cross-domain challenge with panoptic segmentation in agriculture," *The International Journal of Robotics Research*, vol. 43, no. 8, pp. 1151–1174, 2024.

[27] L. Deng, Z. Mao, X. Li, Z. Hu, F. Duan, and Y. Yan, "Uav-based multispectral remote sensing for precision agriculture: A comparison between different cameras," *ISPRS journal of photogrammetry and remote sensing*, vol. 146, pp. 124–136, 2018.

[28] B. T. W. Putra and P. Soni, "Evaluating nir-red and nir-red edge external filters with digital cameras for assessing vegetation indices under different illumination," *Infrared Physics & Technology*, vol. 81, pp. 148–156, 2017.

[29] P. Velusamy, S. Rajendran, R. K. Mahendran, S. Naseer, M. Shafiq, and J.-G. Choi, "Unmanned aerial vehicles (uav) in precision agriculture: Applications and challenges," *Energies*, vol. 15, no. 1, p. 217, 2021.

[30] P. Bosilj, E. Aptoula, T. Duckett, and G. Cielniak, "Transfer learning between crop types for semantic segmentation of crops versus weeds in precision agriculture," *Journal of Field Robotics*, vol. to be determined (published online), 2019.

[31] S. Haug and J. Ostermann, "A crop/weed field image dataset for the evaluation of computer vision based precision agriculture tasks," in *Computer Vision-ECCV*

2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part IV 13. Springer, 2015, pp. 105–116.

[32] J. Weyler, F. Magistri, E. Marks, Y. L. Chong, M. Sodano, G. Roggiolani, N. Chebrolu, C. Stachniss, and J. Behley, "Phenobench: A large dataset and benchmarks for semantic image interpretation in the agricultural domain," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[33] D. Steininger, A. Trondl, G. Croonen, J. Simon, and V. Widhalm, "The cropandweed dataset: A multi-modal learning approach for efficient crop and weed manipulation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 3729–3738.

[34] G. W. Pereira, D. S. M. Valente, D. M. de Queiroz, N. T. Santos, and E. I. Fernandes-Filho, "Soil mapping for precision agriculture using support vector machines combined with inverse distance weighting," *Precision Agriculture*, vol. 23, no. 4, pp. 1189–1204, 2022.

[35] Y. Wang, S. Zhang, B. Dai, S. Yang, and H. Song, "Fine-grained weed recognition using swin transformer and two-stage transfer learning," *Frontiers in Plant Science*, vol. 14, p. 1134932, 2023.

[36] B. Jankovic, S. Jangirova, W. Ullah, L. U. Khan, and M. Guizani, "Uav-assisted real-time disaster detection using optimized transformer model," *arXiv preprint arXiv:2501.12087*, 2025.

[37] B. Madhavi, M. Mahanty, C.-C. Lin, B. O. Lakshmi Jagan, H. M. Rai, S. Agarwal, and N. Agarwal, "Swinconvnext: a fused deep learning architecture for real-time garbage image classification," *Scientific Reports*, vol. 15, no. 1, p. 7995, 2025.

[38] Q. Wang, X. Dong, R. Wang, and H. Sun, "Swin transformer based pyramid pooling network for food segmentation," in *2022 IEEE 2nd International Conference on Software Engineering and Artificial Intelligence (SEAI)*. IEEE, 2022, pp. 64–68.

[39] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 976–11 986.

[40] X. Tang, B. Li, J. Guo, W. Chen, D. Zhang, and F. Huang, "A cross-modal feature fusion model based on convnext for rgb-d semantic segmentation," *Mathematics*, vol. 11, no. 8, p. 1828, 2023.

[41] S. Jadon, "A survey of loss functions for semantic segmentation," in *2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB)*. IEEE, 2020, pp. 1–7.

[42] A. Golnari and M. Diba, "Adaptive real-time multi-loss function optimization using dynamic memory fusion framework: A case study on breast cancer segmentation," *arXiv preprint arXiv:2410.19745*, 2024.

[43] J. You, W. Liu, and J. Lee, "A dnn-based semantic segmentation for detecting weed and crop," *Computers and Electronics in Agriculture*, vol. 178, p. 105750, 2020.

[44] A. Olsen, D. A. Konovalov, B. Philippa, P. Ridd, J. C. Wood, J. Johns, W. Banks, B. Girgenti, O. Kenny, J. Whinney *et al.*, "Deepweeds: A multiclass weed species image dataset for deep learning," *Scientific reports*, vol. 9, no. 1, p. 2058, 2019.

[45] Y.-Y. Zheng, J.-L. Kong, X.-B. Jin, X.-Y. Wang, T.-L. Su, and M. Zuo, "Cropdeep: The crop vision dataset for deep-learning-based classification and detection in precision agriculture," *Sensors*, vol. 19, no. 5, p. 1058, 2019.

[46] X. Wu, S. Aravecchia, P. Lottes, C. Stachniss, and C. Pradalier, "Robotic weed control using automated weed and crop classification," *Journal of Field Robotics*, vol. 37, no. 2, pp. 322–340, 2020.

[47] M. Fawakherji, A. Youssef, D. D. Bloisi, A. Pretto, D. Nardi *et al.*, "Crop and weed classification using pixel-wise segmentation on ground and aerial images," *International journal of robotic computing*, vol. 2, no. 1, pp. 39–57, 2020.

[48] N. Yokoya, C. Grohnfeldt, and J. Chanussot, "Hyperspectral and multispectral data fusion: A comparative review of the recent literature," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 2, pp. 29–56, 2017.