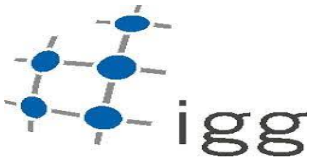

University of Bonn
Institute of Geodesy and Geo-information
Master in Geodetic Engineering
Major in [MSR]

Enhancing Cross-Domain Performance of Panoptic One-Click Segmentation For Sugar Beet Fields: Towards an Efficient Annotation Tool

Master Thesis

In partial fulfilment of the requirements for the Master of Science
in Geodetic Engineering (MSc GE)

Submitted by:
Saad Naseer

Bonn, [December 27, 2023]

Supervisor:
Prof. Dr. Chris McCool
Institut für Landtechnik (ILT)

Supervisor:
Dr. Michael Halstead
Institut für Landtechnik (ILT)

Advisor:
Ph.D. Patrick Zimmer
Institut für Landtechnik (ILT)

UNIVERSITÄT BONN

RHEINISCHE
FRIEDRICH-WILHELMS-
UNIVERSITÄT

igg

Institut für Geodäsie
und Geoinformation

Professur für Geodäsie

Nußallee 17
53115 Bonn
Tel. 0228/73-2622
Fax 0228/73-2988

www.igg.uni-bonn.de

Description: The objective of this master's thesis is to enhance the cross-domain performance of panoptic one-click segmentation Zimmer et al. (2023) specifically tailored for sugar beet fields, with the ultimate goal of developing an efficient one-click annotation tool. This tool will streamline the annotation process, automate the segmentation of sugar beet fields, and enable more accurate and efficient agricultural practices.

The research will be conducted in two main tasks. Firstly, an extensive dataset will be collected from 11 different fields at various growth stages and under diverse environmental conditions. This dataset will undergo manual annotation and proper separation into training, testing, and evaluation sets. It will serve as the foundation for training the model and assessing the performance of the developed methods. The inclusion of data from different fields and growth stages ensures the robustness and generalizability of the model, accounting for the inherent variability in real-world sugar beet cultivation scenarios. The second task involves the implementation of advanced techniques to improve the model's cross-domain generalizability. Specifically, vegetation assessment approach such as vegetation indices calculated from rgb color space will be incorporated into the training dataset, enabling the model to learn more invariant features Milioto et al. (2018). This integration of vegetation indices enhances the model's ability to accurately segment sugar beet fields across various domains and different environmental conditions. Additionally, a combination of image enhancement techniques, such as contrast adjustment Wang et al. (2020), will be applied to optimize the input imagery. By enhancing the quality and visibility of the images, the model can extract more precise features, leading to improved segmentation results.

By enhancing the cross-domain performance of panoptic one-click segmentation and developing an efficient annotation tool for sugar beet fields, this research aims to significantly reduce the manual effort and time required for segmentation tasks in agricultural applications. The outcomes of this study have the potential to revolutionize agricultural practices, optimizing resource allocation, and contributing to more sustainable and productive farming systems.

Through rigorous experimentation, evaluation, and analysis, this thesis will advance computer vision techniques in the field of precision agriculture, offering practical solutions for efficient data annotation and analysis specifically tailored to sugar beet fields.

In summary, this master's thesis pushes the boundaries of panoptic one-click segmentation, highlighting its applicability in the agricultural domain. The results hold the promise of transformative impact in precision farming, providing farmers with valuable tools to make data-driven decisions and drive efficiency in the management of sugar beet fields.

References

Milioto, A., Lottes, P., and Stachniss, C. (2018). 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)*, pages 2229–2235.

Wang, A., Xu, Y., Wei, X., and Cui, B. (2020). Semantic segmentation of crop and weed using an encoder-decoder network and image enhancement method under uncontrolled outdoor illumination. *IEEE Access*, 8:81724–81734.

Zimmer, P., Halstead, M., and McCool, C. (2023). Panoptic one-click segmentation: Applied to agricultural data. *IEEE Robotics and Automation Letters*, 8(5):2478–2485.

Abstract

Deep learning methods have revolutionized agricultural practices, yet their dependence on extensive annotated data poses challenges. Panoptic One-Click segmentation offers a streamlined approach that generates high-quality pseudo labels with just a single click on an object. However, its adaptability across diverse domains remains underexplored. This thesis investigates the generalization capabilities of this method across various sugar beet domains.

To bridge this gap, we gathered diverse sugar beet data from 11 fields and experimented with different field splits, dividing them into two distinct groups. One group was utilized for training, validation, and evaluation to assess in-domain performance, while the other group was exclusively reserved for evaluation to gauge cross-domain performance. Our findings highlight that employing a 2/9 split between these groups presents the most challenging cross-domain scenario. To improve the model's generalization capabilities, we implemented a multichannel input approach, data augmentation, and a dilated panoptic one-click system. Notably, the multichannel approach, incorporating the Excess Green Minus Excess Red (ExGR) index as an additional channel, enhanced cross-domain performance by 5.6 %, marking the highest improvement among all methods. This enhancement is attributed to ExGR's provision of pre-segmentation information and its robustness in outdoor conditions. Conversely, the data augmentations such as random rotation and color jitter suffered from redundant information and misalignment problems in the case of multichannel input having ExGR. While the dilated panoptic one-click system encountered noise amplification issues with multichannel input having ExGR.

Our work's primary contributions are the introduction of a unique dataset for the cross-domain assessment of crop/weed object segmentation models in diverse sugar beet fields and the presentation of an improved panoptic one-click technique that outperforms the baseline in a variety of fields.

Acknowledgements

I would like to express my sincere gratitude to my advisor, Patrick Zimmer, a Ph.D. student at the University of Bonn, for his invaluable guidance and support throughout the thesis. Additionally, I extend my thanks to my supervisor, Professor Chris McCool from the University of Bonn, and Dr. Michael Halstead, also from the University of Bonn, for providing both the platform and valuable feedback.

Contents

List of Figures

List of Tables

Acronyms

B Blue.

BiLSTM Bidirectional Long Short-Term Memory.

C Click.

CFD Confidence Failure Diversity.

CieLab CIE 1976 L*, a*, b*.

CIVE Color Index of Vegetation Extraction.

CNN Convolutional Neural Network.

COM1 Combined Indices 1.

COM2 Combined Indices 2.

ExG Excess Green.

ExGR Excess Green minus Excess Red.

ExR Excess Red.

FCN Fully Convolutional Network.

FFNN Feedforward Neural Networks.

G Green.

GANs Generative Adversarial Networks.

GLI Green Leaf Index.

GRU Gated Recurrent Unit.

HSI Hue, Saturation, Intensity.

HSV Hue, Saturation, Value.

LSTM Long Short-Term Memory.

MExG Modified Excess Green.

MGRVI Modified Green Red Vegetation Index.

MPRI Modified Photochemical Reflectance Index.

NDI Normalized Difference Index.

NIR Near-Infrared.

R Red.

RGB Red, Green, Blue.

RGBD Red, Green, Blue, Depth.

RGBVI Red Green Blue Vegetation Index.

VEG Vegetative Index.

VI Vegetation Index.

1 Introduction

Agriculture serves a crucial role in ensuring food security, environmental sustainability, and economic stability. Sugar beet, accounting for 14% of the world's sugar production, holds significant prominence. Europe cultivates approximately 50% of the global sugar beet. In 2019, Germany reported a cultivation area of 408 thousand hectares, contributing to Europe's overall cultivation area of 1640 thousand hectares (Vladu et al., 2021). However, substantial losses in production and quality arise due to inefficient weed control and chemical spraying methods (Bhadra et al., 2020). To boost production and quality, various deep learning methods have emerged, aiding farmers in decision-making and precise agricultural operations. For detailed crop inspection, panoptic segmentation stands out as one of the recognized methods. It generates pixel-level labels in images, providing in-depth insights into individual objects (Kirillov et al., 2018). This labeled imagery aids automated systems and farmers in informed decision-making processes. Implementing such methods demands extensive, high-quality annotated data for training. These annotations serve as ground truth labels that can be used to compare with predictions from the model and measure the error. Based on these errors the model tunes its internal parameters until the desired level of performance is achieved. The high-quality annotations ensure that the model learns accurate and reliable patterns and achieves optimal performance on the given task. The most accurate method of doing annotations is by pixel-wise labeling for each object present in the image which requires a great amount of time and effort. To address this problem, various weak learning-based approaches (Xu et al., 2016), (Ramadan et al., 2020), (Lin et al., 2020), (Zimmer et al., 2023) have been developed to generate high-quality pseudo-labels for the target objects through minimal user interactions. Among these methods, panoptic one-click segmentation (Zimmer et al., 2023) stands out as the most advanced technique in the agricultural domain that is capable of generating pseudo-labels for target objects with just a single click on the object. The main advantage of this approach is that it requires less training time compared to other approaches because of its ability to simultaneously estimate the location of multiple objects present in a scene, unlike other weak learning approaches where the location of multiple objects present in a scene are estimated independently.

It's important to note that the panoptic one-click segmentation model has indeed simplified the annotation process. However, its applicability to new domains has not been thoroughly examined, representing a significant gap in this approach. This gap serves as the

primary focus of my thesis. Understanding the cross-domain applicability of panoptic one-click segmentation stands as a pivotal step towards revolutionizing agricultural annotation practices. Bridging this gap not only enhances the model's capabilities but also holds the promise of significantly advancing precision agriculture practices.

To address this gap, we initially collected a diverse and challenging sugar beet dataset. Then, we determined the best composition of the dataset for training, validation, and evaluation sets that present a challenge for cross-domain performance. This was essential to establish a proper evaluation system aimed at testing the cross-domain limits of panoptic one-click segmentation. Next, to expand its cross-domain capabilities we have applied various domain generalization approaches that have proven effective in other systems and also studied their effects on panoptic one-click segmentation.

The methods we have employed include a multichannel approach, data augmentation, and the dilated panoptic one-click segmentation model. In the multichannel approach, we calculated pre-segmentation information using vegetation indices and added it as an extra channel in our input image. For data augmentation, we applied color jitter and random rotation to increase diversity in our training data by adjusting the colors and varying the angles of the samples. Lastly, in the dilated panoptic one-click model, we implemented a dilation stack comprising layers of dilated convolutions with different dilation rates in our model to capture global and fine-grained information for a better understanding of the scene.

The main contributions of our work are introducing a novel dataset for cross-domain evaluation of crop/weed object segmentation models across various sugar beet fields, and presenting an enhanced panoptic one-click method that can perform better than the baseline across different fields.

1.1 Thesis Structure

This thesis covers different parts of the research journey, beginning with Chapter 2, the Literature Review, where we explore foundational knowledge and analyze relevant literature in detail. In Chapter 3, Methodologies, we present the strategies to tackle identified challenges and outline the procedures for data processing and experimental setup for conducting the experiments. In Chapter 4, Results and Discussion, we present the results obtained from the performed experiments along with comprehensive discussions. In Chapter 5, Conclusion, we summarize the key findings, ensuring a concise wrap-up of the research outcomes. In our

final Chapter 6, Future Work, we discuss future directions to further improve the results and their applicability to other domains.

2 Literature Review

Our primary objective is to enhance the cross-domain performance of panoptic one-click segmentation to simplify the process of annotating datasets. In this literature review, we will delve into several crucial areas. Firstly, we will explore panoptic segmentation and assess its significance in the field of computer vision. Next, we will examine the concept of interactive segmentation and its role in the data annotation process. We will also investigate the advantages of integrating the idea of interactive segmentation into panoptic segmentation. Furthermore, we will explore the concept of domain generalization, along with its various techniques, to understand its impact on expanding the cross-domain capabilities of deep learning models. Lastly, we will analyze the existing research and identify gaps within these areas, paving the way for a discussion of potential solutions in the following chapter.

2.1 Panoptic Segmentation

Panoptic segmentation explained in (Kirillov et al., 2018) is a powerful technique used to simultaneously segment 'stuff' and 'things' within an image. 'Stuff' refers to the elements in an image that are uncountable, such as the sky, land, or water. On the other hand, 'things' encompass elements that are countable, including people, plants, and animals. This method assigns labels and unique instance IDs to each pixel in the image, effectively distinguishing and categorizing both types of information present in the scene. Building upon this foundational concept, Cheng et al. (2020)introduced a simple and faster method for panoptic segmentation. Their approach utilizes a three-head decoder structure to predict semantic segmentation (labels for each pixel), instance center (center of mass of each object), and instance center regression (offset of each foreground pixel to their corresponding center of mass) maps. which are later combined using post-processing techniques to generate a panoptic map.

In general, there are different schemes available to perform panoptic segmentation which are sharing backbone, explicit connections, one-shot model, and cascade model (Elharrouss et al., 2021). In Figure 1, various panoptic segmentation models share a common backbone for segmentation, but their methods of combining instance (things) and semantic (stuff) segmentation differ. The shared backbone computes them separately, which may lead to confusion when objects are both stuff and things. While explicit connections enhance reli-

ability by sharing information between instance and semantic blocks, this approach can be complex and time-consuming. The Cascade model employs a sequential scheme, computing semantic information first, followed by instance results, which helps maintain a clear distinction between stuff and things and reduces confusion. However, it can be complex and computationally intensive. The One-shot model offers a simplified solution for real-world applications by combining both instance and semantic components in a single block.

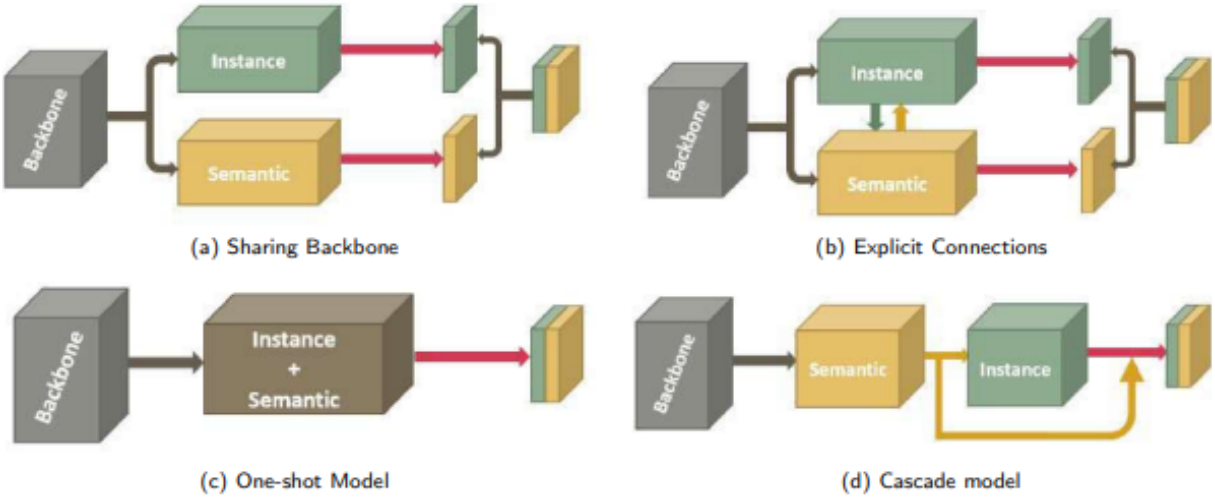

Figure 1: Networks methodologies for panoptic segmentation methods (Elharrouss et al., 2021)

Panoptic segmentation can be used for various applications, which include remote sensing, medical image analysis, data augmentation, and precision farming (Elharrouss et al., 2021).

2.2 Interactive Segmentation

Interactive segmentation is a method that uses human-provided cues in the form of scribbles, bounding boxes, and points to segment the region of interest or target object present in an image with better accuracy (Ramadan et al., 2020). Among these cues, point-based methods are often considered the fastest and most efficient due to their minimal user interaction. In contrast, bounding boxes require more user input and may introduce errors due to their loose boundaries around the target object. Scribbles, on the other hand, can produce highly accurate segments but typically demand more user interaction, such as dragging to delineate object boundaries (Lin et al., 2020). Traditionally, point-based methods require multiple positive and negative clicks, where positive clicks correspond to the target object and negative clicks correspond to the background. These clicks are used to generate Euclidean maps,

which are then combined with RGB images as additional channels to produce accurate segmentation (Xu et al., 2016). Recent advances in deep learning have made the transition to single-click methods possible, improving accuracy and efficiency (Majumder et al., 2020). The models used in these approaches primarily focus on single object segmentation and require multiple click maps for multiple instances within a scene. They typically process each object or instance independently, necessitating multiple passes through the image to segment each one. Consequently, this leads to long inference and training times. To address this challenge, Zimmer et al. (2023) introduced a novel approach known as panoptic one-click segmentation. A discussion of this method will be presented in the upcoming section.

2.3 Panoptic One Click Segmentation

In an agricultural domain, for the first time the concept of a panoptic segmentation to solve a one-click system was introduced by Zimmer et al. (2023) called panoptic one-click segmentation. The implementation of this method follows the same structure as Panoptic DeepLab (Cheng et al., 2020) to generate the panoptic map. The main difference is that Panoptic DeepLab predicts the center location of objects, while in Panoptic One-Click Segmentation the annotator provides the center location of the objects as a click map. This click map is then used as an additional channel alongside RGB input to solve the one-click system. This difference in approach can also be seen in the architecture diagram of the panoptic one-click system presented in Figure 2.

This innovative approach not only addresses the challenges of manual annotation but also possesses the capability to segment all objects present in a scene in a single forward pass, because of its unique input click map strategy. Instead of using multiple input click maps for multiple objects present in a scene, they utilized a single click map that contains the center locations of all n objects. Furthermore, instead of processing each click separately, they process all the clicks together to segment all the objects simultaneously. This difference in approach significantly reduces both training and inference times when compared to other interactive annotation methods (Xu et al., 2016), (Lin et al., 2020), (Majumder et al., 2020).

This method has shown good performance in terms of mIOU on challenging sugar beet and corn datasets. The paper does not investigate how well the method adapts to different crops or environments, leaving a gap in understanding its generalization capability, which I aim to address in my thesis.

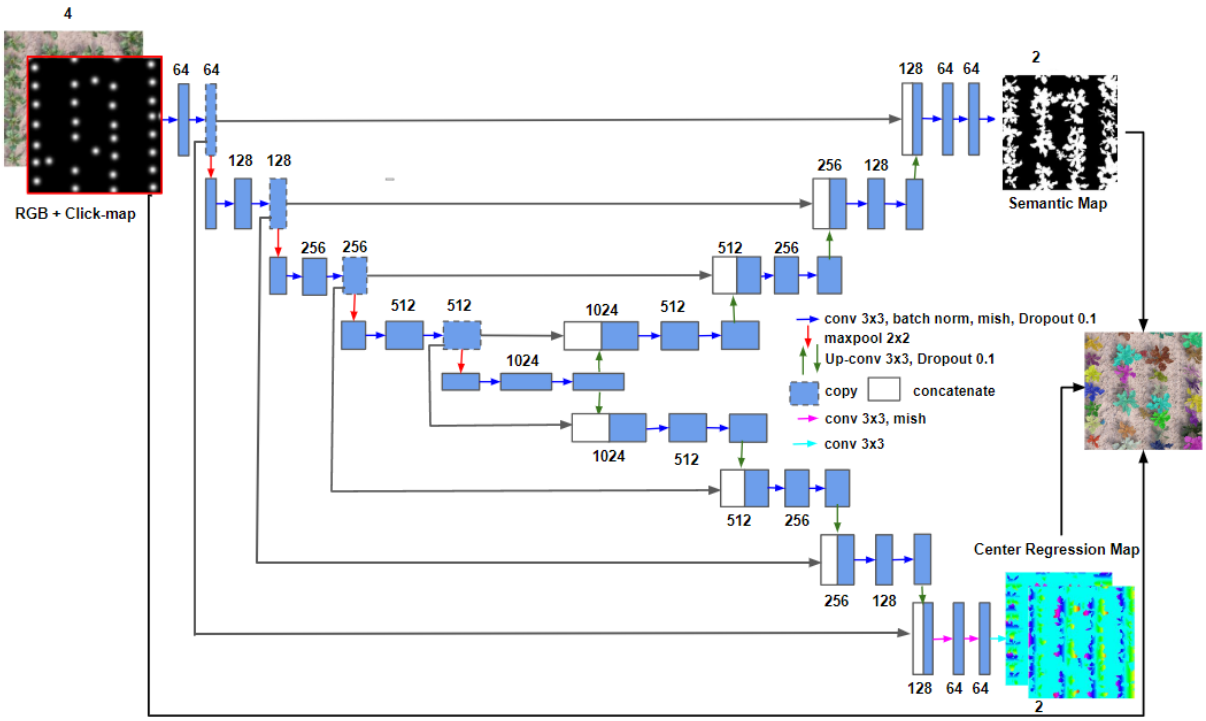

Figure 2: Panoptic One-Click Segmentation Architecture

2.4 Domain Generalization

In the field of computer vision, domain generalization pertains to a method's ability to perform effectively on data it has not seen before or on data from different domains, which may exhibit various types of differences. These differences can encompass variations in lighting conditions, environmental settings, camera angles, data collection methods, and several other factors. The key challenge in domain generalization is that it's impractical to collect and train a model for every possible variation or domain (Zhou et al., 2022). To address this challenge in the context of the agricultural domain, there exist various methods, including image processing techniques, data manipulation approaches, and a variety of learning approaches (Wang et al., 2022). In the forthcoming sections, we will delve into these methods in more detail, drawing insights from relevant literature.

2.4.1 Image Processing Methods

In the agricultural context, before the advancement of deep learning methods, plants were located, recognized, and segmented in images using image processing techniques, enabling automated plant extraction and accurate differentiation from the surrounding soil. A fundamental strategy for plant extraction involved leveraging vegetation indices or color spaces to

highlight plant pixels within images based on their color, reflectance, or absorption properties. This was followed by the application of thresholding techniques for effective separation from the background as discussed in a survey conducted by Luo et al. (2023).

In (Jafari et al., 2006), the authors investigated the extraction of plants from soil using various color spaces, including chromatic, HSI (Hue, Saturation, Intensity), and Greenness. A key advantage of employing these color spaces is their resilience to changes in illumination, making them suitable for real-world conditions. Notably, the study demonstrated that 'greenness' exhibits exceptional robustness across different illumination conditions when tested on a diverse dataset comprising 300 sugar beet images. In another study conducted by Aureliano Netto et al. (2018), three automatic gray-level histogram-based threshold schemes, namely Otsu, Ridler, and Triangle, were evaluated on grayscale vegetation indices of maize fields captured under varying illumination and soil conditions (normal, no-tillage, and straw residue). The study found that the combination of NDI (Normalized Difference Index) with Triangle and ExGR (Excess Green minus Excess Red) with Triangle produced the best results when compared to ExG with all the thresholding methods. This is because ExG (Excess Green) is more sensitive to illumination conditions compared to other indices. The study also revealed that the presence of soil with straw residues can lead to confusion in vegetation indices, causing misclassification of non-vegetation areas as vegetation in an image.

To investigate the impact of varying illumination conditions on vegetation indices, an experiment detailed in (Upendar et al., 2021) applied the ExG, ExR (Excess Red), and ExGR indices to six color palettes representing vegetation and soil under controlled varying illumination. The study found that as illumination increases, the difference in vegetation index values between plants and soil begins to decrease. Notably, among all tested illumination conditions, ExGR exhibited the largest difference in values among all vegetation indices. Additionally, the study conducted tests on actual plant and soil samples, revealing that ExG with a threshold value of 10 and ExGR with a threshold value of 0 yielded effective segmentation results. In contrast, ExR performed poorly due to minimal differentiation between plant and soil values. The study emphasizes the importance of selecting the right index and threshold values for accurate segmentation in varying lighting conditions.

To evaluate the effectiveness of various vegetation indices, Meyer and Neto (2008) conducted a study comparing binary images generated from ExG + Otsu, NDI + Otsu, and ExGR with a positive threshold. This analysis was conducted across diverse backgrounds (dry, wet, wheat

straw) and lighting conditions (greenhouse and natural). The findings demonstrated that ExGR consistently yielded superior outcomes without necessitating distinct threshold values for different images. Additionally, a broader survey byHamuda et al. (2016) encompassed a comparative analysis of distinct vegetation indices (ExG, ExR, ExGR, CIVE (Color Index of Vegetation Extraction), MExG (Modified Excess Green), VEG (Vegetative Index), COM1 (Combined Indices 1), COM2 (Combined Indices 2)) across varying lighting and soil conditions. The outcomes from multiple studies presented in this survey underscored the variability in the performance of these vegetation indices in response to differing lighting conditions. Specifically, suboptimal outcomes were noted under both high-intensity and low-light situations. The results further revealed that distinct vegetation indices exhibited strong performance for specific types of datasets. Notably, the excess green vegetation index (ExG) demonstrated greater stability and reduced sensitivity to variations in illumination and soil conditions.

In another research study mentioned by Santos et al. (2021), a bunch of different plant indices (MGRVI (Modified Green Red Vegetation Index), GLI (Green Leaf Index), RGBVI (Red Green Blue Vegetation Index), MPRI (Modified Photochemical Reflectance Index), ExG, VEG) and methods for setting thresholds (Otsu and k-means) were tested using various sets of data that covered different types of lighting and soil conditions. The study particularly points out how the MGRVI performed well across these different data sets. It highlights that these indices are sensitive to things like how much light there is, the weather, and the state of the soil. Additionally, the study shows that relying on just one vegetation index might not be enough, and it suggests that combining multiple indices could be a better way to make the results more dependable and effective, especially when dealing with different kinds of crops.

Considering the impact of lightning, weather, and soil conditions, the cited work (Riehle et al., 2020)introduces a color transformation approach. This method is grounded in the concept that color schemes like HSV (Hue, Saturation, Value) and CieLab (CIE 1976 L*, a*, b*) incorporate distinct channels that convey color brightness and intensity. This division facilitates segmentation, effectively mitigating the influence of brightness variations. The experiment specifically examines the 'H' and 'a' channels, revealing that the 'a' channel yields improved results when compared not only to the 'H' channel but also to vegetation indices like ExG and ExGR.

In another paper (Seong-Heon Kim, 2015), it was observed that existing vegetation indices, such as ExR and ExG, tend to misclassify non-vegetation objects like gravel, white reflectors,

and rubber boots. To address this issue, the paper proposes a new vegetation index based on the 'a' and 'b' channels from the LAB color space, which closely aligns with human perception and offers improved differentiation between green and non-green objects. The study also tested the proposed index in conjunction with a combined thresholding approach that combines Otsu and Triangle methods for segmentation. The results indicate that the combined thresholding approach is more effective than using Otsu and Triangle methods individually.

The conclusion I have drawn from these studies is that vegetation indices and color schemes are powerful tools for plant analysis as they are easy to implement and save a lot of time and memory. However, their practical application is restricted due to their reliance on outdoor lighting conditions. Additionally, determining new thresholds for various conditions or plants poses a significant challenge. Unfortunately, these methods struggle to handle overlapping cases, which are common in agricultural settings. To address these challenges, upcoming sections will explore various deep learning-based methods.

2.4.2 Data Augmentation Methods

Deep learning methods require a large amount of diverse data for training to perform well in various scenarios. However, acquiring and processing a significant amount of data can be a laborious task, which can be mitigated by using data augmentation techniques. Data augmentation is defined as a strategy that is used to artificially expand the training samples while preserving the original labels (Taylor and Nitschke, 2018).

In their comprehensive survey on data augmentation methods in deep learning, referenced as Khosla and Saini (2020), the authors explore a range of techniques to facilitate data generation and mitigate the risk of overfitting. These methods encompass geometric transformations and color adjustments, which modify data's geometric and color attributes, addressing issues like position, orientation, scale, brightness, contrast, and sharpness. Geometric and color transformations enhance the model's robustness across varying conditions and locations. Random Erasing is particularly effective for handling occlusions, as it introduces random alterations through rectangular patches. For artistic data, Neural Style Transfer is a favored choice, creating new data by merging artistic styles with input images. Dealing with complex scenes is facilitated through the mixing of images, where two dataset images are randomly combined. Feature Space Augmentation is employed to enhance robustness with noisy or incomplete data, involving feature vector interpolation or extrapolation. On the other

hand, Adversarial Training and GANs (Generative Adversarial Networks) represent advanced approaches, generating synthetic images that closely resemble real ones instead of altering input images. These synthesized data are used to challenge and bolster the model's performance on diverse target distributions by introducing artificial elements. However, it's important to note that data augmentation approaches may come with drawbacks such as increased training time and higher memory requirements, with geometric and color transformations generally being easier to implement compared to more advanced techniques like GANs. In another survey from Yang et al. (2023), various data augmentation techniques, including geometric methods, color methods, image erasing, and image mixing, were applied to the PASCAL VOC dataset. The results of the study, which included testing with different deep learning models, concluded that data augmentation led to improved performance compared to scenarios with no augmentation.

In a study conducted by Taylor and Nitschke (2018), the authors evaluated various augmentation methods based on geometric transformations (flipping, rotating, cropping) and photometric techniques (color jitter, edge enhancement, Fancy PCA). They found that all augmentation techniques helped the CNN (Convolutional Neural Network) model avoid overfitting and improved generalization performance compared to no augmentation. Furthermore, they observed that among these methods, geometric transformations, especially cropping, outperformed all the other techniques as it generate more samples than others. In another paper by Brilhador et al. (2019), the authors introduced 'patch augmentation' and tested it on Crop/Weed Field Image Dataset (CWFID) taken from Haug and Ostermann (2015). This approach divides an input image into n×n small patches, treating each patch as a separate image. Patches are padded to match the input resolution, preventing information loss seen in other augmentation techniques that involves resizing. The study found that when tested separately, both patch augmentation and traditional techniques (horizontal and vertical flips, rotation, shifts, shear, zoom) consistently outperformed no augmentation, even for images of varying resolutions. Additionally, the study tested the combined application of all traditional approaches and found that this ensemble approach significantly increases diversity, particularly in complex environments. Higher resolutions were found to be optimal for augmentations, increasing the number of samples and enhancing model performance. The effectiveness of patch augmentation depends on the number of patches, making it particularly advantageous for high-resolution images.

In (Fawakherji et al., 2020), an instance generative GAN approach was introduced and tested on the open-source sugar beet 2016 dataset. This approach leverages the mask information of the target object to generate instances rather than generating full images, as seen in other GAN approaches. This approach also allows for the generation of more samples for weak classes or objects of interest. The mixture of original and GAN generative images demonstrated performance improvements when tested with different deep learning models (Unet, Bonnet, Segnet, Unet-Resnet) compared to using only original or only GAN generative images.

In (Blok et al., 2021), a series of experiments were conducted on different broccoli cultivars to assess the effectiveness of data augmentation on both seen and unseen data. Data augmentation methods were categorized into three groups: geometric (involving cropping, partitioning, rotation, and scaling), photometric (including texture enhancement, texture blur, light transformations, and color transformations), and the combined geometric and photometric group. Notably, all augmentation approaches consistently outperformed scenarios with no augmentation when Mask R-CNN was employed for training and testing, whether on the same broccoli cultivar or different ones. The study also recommended the inclusion of a few samples from the target set to further enhance model performance on unseen data and also highlighted that transformations bearing a closer resemblance to the target set have a more significant impact on the model's performance.

The conclusion I have drawn from these findings is that data augmentation remains an important tool for diversifying a limited dataset. Yet, it's impractical to encompass all real-world variations through augmentation alone. Especially within a constrained target domain, selecting the appropriate augmentation method is crucial, and incorrect choices can lead to diminished performance. Implementing these approaches demands careful handling to prevent label alterations.

2.4.3 Multichannel input methods

Several studies have demonstrated the benefits of incorporating additional channels alongside the standard RGB (Red, Green, Blue) channels to enhance the model's ability to learn invariant features and improve generalization (Milioto et al., 2018), (Yang et al., 2020).

In the past, segmentation tasks predominantly relied on RGB images to differentiate between crops, weeds, and soil. The study by Milioto et al. (2018) presents an encoder-

decoder CNN network to solve the plant/weed segmentation task. The network was trained with a 14-channel input comprising RGB and vegetation indices. The results showcased improved cross-domain performance compared to traditional RGB and RGB + NIR (Near-Infrared) inputs. The findings suggest the value of incorporating task-relevant vegetation indices, providing a cost-effective alternative while achieving high segmentation accuracy. The transferability of the model to new domains is also highlighted.

In the research conducted by Kerkech et al. (2018), the analysis of different color spaces and combinations with vegetation indices demonstrates the significance of these factors for improved segmentation. By incorporating task-relevant vegetation indices with color spaces, the model learned more invariant vegetation features, resulting in enhanced classification performance. The study emphasizes the importance of considering complementary sources of information for accurate segmentation.

Another research by Yang et al. (2020) advances semantic segmentation using deep networks for rice lodging identification by incorporating additional vegetation indices alongside RGB data. Notably, the integration of ExGR and ExG with RGB (RGB + ExGR and RGB + ExG) enhances segmentation accuracy and domain generalization across datasets from different years. However, the inclusion of three indices (RGB + ExG + ExGR) results in reduced accuracy, suggesting a risk of feature redundancy. This highlights the intricate balance between enriching information and avoiding confusion. In a separate study, (Yang et al., 2021) introduced a lightweight model for wheat lodging extraction utilizing 4-channel input data. The combination schemes (RGB + DSM) and (RGB + ExG) demonstrated improved performance compared to RGB alone when trained and tested on a dataset spanning three growth stages. Their lightweight model, based on separable convolutions, not only exhibited enhanced performance but also reduced training time due to fewer parameters compared to traditional architectures such as Unet and FCN (Fully Convolutional Network).

In the study by Kitzler et al. (2023), the development of a low-cost RGBD (Red, Green, Blue, Depth) camera and its application for segmentation tasks demonstrates innovation. By capturing data in various conditions and growth stages, the researchers aimed to improve classification performance. The incorporation of a depth map as an input channel enhanced spatial information and addressed lighting and color variations. However, the additional cost and increased processing power required may limit practicality and scalability.

The literature emphasizes the importance of integrating task-specific information in cross-

domain applications. However, it's crucial to note that an excess of irrelevant or extraneous information can detrimentally impact the model's performance. These methods demand small additional memory and time during computation.

2.4.4 Ensemble Learning

Ensemble learning is another domain generalization technique that utilizes the strength of diversity to enhance model performance across various scenarios. This involves combining predictions from different models trained on distinct subsets of the training data, employing different parameters, and utilizing diverse learning schemes. By doing so, ensemble learning utilizes the strength of different models to produce more stable output and increase overall robustness (Sagi and Rokach, 2018).

In (Mesbah et al., 2021), the author employed four different ensemble schemes to combine the outputs of five distinct deep learning models. The ensemble methods included averaging outputs, weighted average outputs using single-layer perceptrons, and weighted average outputs with multi-layer perceptrons. The results revealed enhanced generalization performance when the models were trained and tested on different single-source datasets, as compared to the performance of individual models. However, in some cases, no improvement was observed due to significant gaps between the source and target domains. In another paper (Guo and Gould, 2015), the authors experimented with ensemble networks comprising different numbers of networks on varying sample sizes for the task of object detection on the challenging PASCAL VOC dataset. Their findings suggested that increasing the number of networks indeed elevates performance in both in-domain and cross-domain scenarios, and it reduces overfitting compared to individual networks. However, they observed that increasing the number of samples resulted in a reduction in performance. The new stacking ensemble model, composed of three base learners named LSTM (Long Short-Term Memory), BiLSTM (Bidirectional Long Short-Term Memory), GRU (Gated Recurrent Unit), and one meta learner named GRU, was introduced in Muhammad et al. (2023) for the detection of face presentation attacks. The purpose of the meta-learner is to process the outputs from the different base learners and generate the final predictions. To test the model's generalization performance, it was trained on three source domains and tested on a fourth, unseen domain. The obtained results showed an improvement in cross-domain performance compared to the individual models. The paper also found that decreasing the number of source domains

negatively impacts the ensemble model's performance, and biases or limitations in certain models also propagate to the meta-model predictions.

In the study (Bian and Wang, 2007), the authors introduced a mechanism for selecting base learner models for ensembles based on diversity measures. The findings explored 10 different diversity measures and suggested that CFD (Confidence Failure Diversity) is the most appropriate measure for choosing base learners. CFD was preferred because it is independent of the number of base learners and the accuracy of individual learners. The paper suggested that higher diversity among the base learners leads to better ensemble performance. However, the optimal choice when selecting base learners involves a balanced tradeoff between diversity and the accuracy of models. In another study (Ortega et al., 2022), the authors explored the relationship between diversity and the generalization performance of ensemble methods. The study claims that a combination of high diversity and accuracy among base learner models is optimal for reducing generalization errors in ensembles.

The literature highlights that the ensemble methods are useful for reducing overfitting and getting more stable outputs by combining the different base learner models. However, the implementation of many base learners is a time-consuming task and requires additional computational resources and memory and also increases the complexity of the model (Baba et al., 2015).

2.4.5 Dropout Methods

Dropout is a regularization technique that leverages the concept of diversity to address overfitting issues in deep neural networks and improve generalization performance. Dropout methods enhance model diversity by deactivating elements within the network during training, preventing the model from relying too heavily on specific features, neurons, channels, or structures. This encourages the development of more generalized and adaptable representations (Srivastava et al., 2014).

The standard dropout scheme, involving the random deactivation of neurons in a network to disrupt coadaptation and prevent overfitting, is explored in (Srivastava et al., 2014). Testing on various challenging datasets demonstrated a significant decrease in errors compared to scenarios without dropout. However, dropout's effectiveness depends on the dataset's size and diversity. For small datasets, dropout may struggle to address overfitting, as the model tends to memorize the limited data excessively. Conversely, with large and diverse datasets, models

showed resistance to overfitting. Therefore, the paper suggests determining the dropout rate by assessing the extent of overfitting, offering a more tailored regularization strategy based on dataset characteristics. The inclusion of standard dropout to a CNN increases training time by generating matrices with random zeros at various locations. The computations involving these non-contributing zeros lead to extra and unnecessary processing. To address this issue in (Ko et al., 2017), the authors implemented a technique of dropping matrix elements either row-wise or column-wise to suppress matrices by eliminating zeros. The method was tested on FFNN (Feedforward Neural Networks) and CNN across various datasets, including MINST, CIFAR-10, and SVHN. This approach led to reduced training time while maintaining comparable accuracy compared to standard dropout.

In a study conducted by Park and Kwak (2017), it was observed that incorporating dropouts in convolutional layers activates previously dead neurons, enhancing the model's ability to learn informative features. However, randomly selecting neurons with high probabilities can result in the loss of valuable information. To address this, the authors implemented two dropout methods: feature-wise and channel-wise dropout. In these methods, neurons with the maximum value in feature maps or across feature maps are dropped at specified locations with a given probability. To boost model robustness, the authors introduced stochastic dropout, injecting noisy variations by changing the dropout value according to a Gaussian distribution at each iteration. The results indicated that all dropout schemes had a more significant impact in the absence of data augmentation and on smaller datasets. Notably, incorporating dropouts in higher-level convolutions, which contain more specific information, proved more beneficial. The findings consistently showed that all dropout schemes outperformed the scenario with no dropout.

In contrast to deactivating individual neurons,Hou and Wang (2019) deactivates entire channels within convolution layers based on assigned weights. The decision stems from the observation that only a few channels in higher convolution layers exhibit significant activity, while neurons in other channels tend to remain inactive. Each channel in the current convolutional layer receives a score, and a binary mask is generated accordingly. To address the likelihood of repeated masks in successive iterations, a random number introduces diversity by filtering out channels. This method demonstrates adaptability across various CNNs, exhibiting improved performance compared to standard and no dropout on diverse datasets. It proves particularly effective for small datasets prone to overfitting. In a different

study (Guo et al., 2023), researchers investigated a hybrid dropout approach that combined layer-wise and channel-wise dropout techniques, incorporating a progressive dropout rate. In this method, a variable number of channels were randomly dropped based on a progressive ratio within a randomly selected layer. The noteworthy discovery was a reduction in the domain gap between the source and unseen data across various layers. This decrease was attributed to the compensation for the introduced noise from dropped channels by other layers, ultimately enhancing the overall model robustness. To address overfitting, the study introduced a progressive dropout rate that dynamically increased during training steps. This adaptive strategy effectively mitigated the growing risk of overfitting in the later stages of training. Significantly, this approach demonstrated superior performance compared to subsequent dropout techniques when evaluated on three challenging datasets. Importantly, this improved performance was achieved without incurring additional computing costs.

The literature emphasizes the utility of dropout in improving generalization by creating subsets of the network without inflating the parameters. However, selecting an appropriate dropout rate is vital, considering the level of overfitting, excessive dropout rates might result in losing valuable information. Dropout methods also showcase varied performances across different architectures (Lim, 2021).

2.4.6 Early Stopping

Early stopping is a simple and effective technique for minimizing overfitting and underfitting in machine and deep learning models. The core concept involves terminating the training process when the difference between the training and validation loss is minimized. The goal is to halt training at the point where the model achieves a balance between fitting the training data well and maintaining good generalization to unseen data (Ying, 2019), (Bisong, 2019). The main advantage of early stopping is that it is parameter-free and don't require retraining like other regularization methods which require tunning by adjusting hyperparameters and retraining to test the effect of these hyperparameters on the model. However, early stopping relies on a validation set, which is not necessary for other regularization methods (Allamy, 2014). Sometimes validation error fluctuates up and down due to varying levels of noise captured by the model during training, making it challenging to determine the right stopping point. Stopping too early can result in underfitting, while stopping too late can lead to overfitting, both of which can damage the model's generalization ability. In the analysis

presented by Syakrani (2022), the authors observed the validation error over a period based on the 'patience' hyperparameter, even after discovering the first local minima. To assess its impact, they employed the YOLOv4 object detection model, training it on four distinct datasets with varying classes. Their findings revealed that early stopping with a lower patience value proved more effective, while higher values sometimes failed to trigger early stopping. This technique demonstrated effectiveness, particularly on smaller datasets known for their susceptibility to overfitting, thus exhibiting a higher likelihood of early stopping activation. Surprisingly, the number of classes didn't seem to affect the effectiveness of early stopping. In terms of accuracy, models with and without early stopping showcased similar results, but the early stopping approach notably reduced the number of iterations. This reduction in iterations translates to considerable savings in computation power and cost.

The literature emphasizes the advantages of the early stopping approach as it doesn't require retraining, is easy to implement, reduces training time, and saves power costs. However, it is dependent on the validation set, so choosing the right data for the validation set is key to maximizing its benefits.

2.4.7 L1 and L2 Regularization

These regularization methods are effective in reducing overfitting by incorporating a penalty term into the cost function (Gupta et al., 2017). L1 regularization, also known as Lasso Regression, introduces the sum of the absolute values of model coefficients in the cost function, thereby diminishing the influence of irrelevant features and simplifying the complex function that might cause overfitting. Conversely, L2 regularization, or Ridge Regression, adds the sum of the squares of model coefficients to the cost function, aiming to diminish the impact of high correlations between coefficients that could lead to increased model sensitivity and complexity, contributing to overfitting. The penalty term's strength is controlled by a hyperparameter (Sherzodjon, 2023).

In a referenced study by Gupta et al. (2018), suggested smaller hyperparameter values closer to 0.001 were deemed effective. The study evaluated L1, L2, and dropout regularization on the Airfoil and MNIST datasets, concluding that, in this context, L2 was the most successful approach for reducing overfitting compared to L1 and dropout. Specifically, L2 proved to be the most effective for models with fewer features, such as the MNIST dataset with its 784 features, whereas dropout regularization appeared more beneficial for models with a higher

number of features. Similarly in another study (Mehdi[1] et al., 2023), it was observed that L2 regularization tends to exhibit better performance than dropout when applied to models containing fewer than 300 hidden neurons. However, beyond this threshold, dropout surpasses L2 regularization in effectiveness. Additionally, the rate of accuracy improvement in both cases tends to decrease as the number of hidden neurons increases. In the study conducted by Kamalov and Leung (2020), the researchers explored regularization techniques on an imbalanced Reuters dataset. Their findings suggested that, in the context of imbalanced datasets, L1 regularization outperformed L2 regularization and dropout methods in enhancing the model's generalization performance. Notably, they observed that using L1 regularization often required lower hyperparameter values to achieve optimal performance. In a different study, Kim and Kang (2020) observed that increasing the amount of data and using deeper architectures were more effective than employing techniques like L1, L2 regularization or dropout for reducing overfitting.

The literature underscores the effectiveness of L1 and L2 regularization, particularly in scenarios with datasets featuring fewer features or imbalanced data, where dropout might not be the optimal choice. These methods' performance is regulated by hyperparameters and doesn't necessarily hinge on a validation set for tuning. Importantly, they are easy to implement and do not substantially increase memory requirements or the number of parameters in the model.

2.4.8 Architectural Modifications

To mitigate overfitting and enhance model performance, several intelligent approaches have been developed to modify components within existing deep networks without significantly increasing complexity.

In (Zhou et al., 2019), the authors addressed overfitting in deep neural networks like AlexNet, which perform well on large datasets such as ImageNet but exhibit reduced accuracy when applied to smaller datasets like CIFAR-10. They initially introduced the Basic network, consisting of simple convolutions, max pooling, and fully connected layers. Subsequently, they replaced the simple convolutions with dilated convolutions to capture features at different scales without increasing the number of parameters. They replaced the fully connected layers with global average pooling to reduce parameter count and added batch normalization after each convolutional layer to fasten the training process. The results demonstrated im-

proved performance compared to both AlexNet and the BaseNet. Moreover, by increasing the depth of the modified network, they observed a reduction in error rate from 14.06 % to 12.5 %. In a study from Lei et al. (2019), the use of dilated convolutions over normal convolutions enhanced training efficiency and accuracy. Normal convolutions require larger kernel sizes to cover expanded receptive fields, increasing parameters, and computational costs. In contrast, dilated convolutions introduce "holes" in the kernel, expanding the receptive field without enlarging its size or parameters. Initially replacing all convolution layers with fixed-rate dilated convolutions reduced training time and errors but caused a slight drop in testing accuracy due to information loss across scales. To address this, researchers stacked dilated convolutions with varying dilation rates, enabling comprehensive information capture at different scales, surpassing both the fixed-rate dilated convolution and traditional CNNs in terms of training and testing performance while maintaining parameter efficiency.

The literature highlights that instead of creating entirely new architectures, modifying existing ones can maximize performance across various tasks. This involves adapting specific components carefully within the architecture to suit different requirements.

2.5 Summary

Manual annotation demands substantial time and effort. The panoptic one-click system, introduced in the paper by Zimmer et al. (2023), stands out as an efficient annotation tool that generates pseudo-labels with just one click on the target objects. The main advantage of this approach, as highlighted in the same paper, is that it is faster to train because it utilizes one click map for all the objects in a scene and estimates the location of all objects simultaneously, as compared to other click approaches (Xu et al., 2016), (Majumder et al., 2020) which require a separate click map for each object and process each object independently. However, the applicability of this method is limited across different domains. To enhance its cross-domain limitations, we have studied various generalization methods in the reviewed literature. These methods encompass image processing, data augmentation, multichannel integration, ensemble techniques, dropout mechanisms, early stopping, regularization, and architectural modifications. Image processing methods, such as threshold-based separation of foreground from background based on color detection, are straightforward but susceptible to complex scenes and illumination variations, demanding threshold adjustments for different conditions, as observed in a survey by Hamuda et al. (2016). Data augmentation,

used to diversify training data, has proven effective in reducing overfitting, particularly in small datasets. However, it requires careful selection of augmentation methods specific to the target domain and involves hyperparameter tuning, as observed in (Zhou et al., 2022). Multichannel methods incorporate supplementary information as an additional channel to mitigate overfitting. They are straightforward to implement and do not necessitate retraining or parameter tuning as observed in (Milioto et al., 2018). However, they can lead to an increase in model parameters. Ensemble methods, aggregating multiple model outputs for enhanced accuracy. They are difficult to implement, time-consuming, memory-intensive, and increase model parameters (Baba et al., 2015). Dropout, by deactivating neurons or channels to prevent over-specific learning, requires parameter tuning, and retraining. At times, it can lead to the loss of crucial information. Early stopping halts training at an optimal point without retraining but relies on a validation set to identify the optimal point and has a high chance of failure in case of fluctuations in validation error (Allamy, 2014). Regularization methods impose additional costs on the cost function to simplify the model and prevent overfitting, necessitating retraining and fine-tuning. Architectural modification methods have emphasized the advantages of dilated convolutions over standard convolutions, particularly in their ability to capture multiscale information without significantly increasing the model's complexity and parameters (Zhou et al., 2019).

All the studied methods have proven effective in enhancing generalization in other systems. To address gaps in the panoptic one-click system, we have selected specific approaches based on their simplicity and effectiveness. Our primary choice involves the multichannel input method utilizing vegetation indices due to its ease of implementation and resilience in various environmental conditions, as well as its proven effectiveness in numerous agricultural scenarios. Next, we've chosen data augmentation because we're working with a small training dataset more prone to overfitting. These methods increase diversity in the training set, making the model robust against various variations, and have also proven very effective in many systems. We have applied basic random rotation and color jitter to capture angular and color variations encountered in real-world scenarios. Thirdly, we have integrated a dilation stack into our model to prevent information loss and capture more detailed features from our small training dataset.

3 Research Method

The main goal of our work is to explore and enhance the generalization ability of the Panoptic One-Click annotation tool to reduce manual annotation efforts. To achieve this goal, our research is conducted in two main parts. The first part involves the collection and manual annotation of diverse sugar beet data gathered from 11 different farms across Germany and finding a split of these farms into training, validation, and evaluation sets that create a challenging cross-domain scenario. In the second part, we have employed various cross-domain generalization techniques to enhance its cross-domain performance. These include the multichannel approach, data augmentation, and the dilated Panoptic One-Click model. In the multichannel approach, we incorporated additional task-relevant information such as vegetation indices as extra inputs to the Panoptic One-Click model. For data augmentation, we employed methods such as random rotation and color jitter to increase diversity in our small training dataset. With the dilated Panoptic One-Click model, we employed the dilation stack to capture more detailed information about the task. The detailed procedure for these parts is discussed in the following subsections.

3.1 Data Collection and Preprocessing

To ensure the model's effectiveness across various situations, we assembled a diverse dataset by collecting data from 11 distinct sugar beet fields located in Bonn, Germany. This data collection effort spanned three different days for every field and during each of these days, the data collection happened at different times. We employed a DJI Matrice 300 drone, flying at an altitude of 12 meters, equipped with a non-integrated 45-megapixel DJI Zenmuse P1 RGB camera for image capture. The images were captured at a resolution of 8192 x 5460 pixels. The example image from each of the farms along with their names can be seen in Figure 3.

For 10 of the fields (ve, wy, bo, br, di, do, ja, ka, oy, wa), which covered smaller areas, we chose to capture images in a raw format to preserve the highest image quality. Afterward, we processed these raw images using RawTherapee (Raw, 2005) software to convert them into jpg format. However, for the 11th field known as "ck," which encompassed a larger ground area, we captured images directly in jpg format due to storage constraints. Our dataset was intentionally designed to include examples from various growth stages, diverse soil compositions, different environmental conditions, and different weed species. This diversity

was essential to create a solid testing set to ensure that our panoptic one-click model could effectively handle a wide range of real-world scenarios commonly encountered in agricultural settings. Following the data collection phase, we conducted a meticulous manual annotation

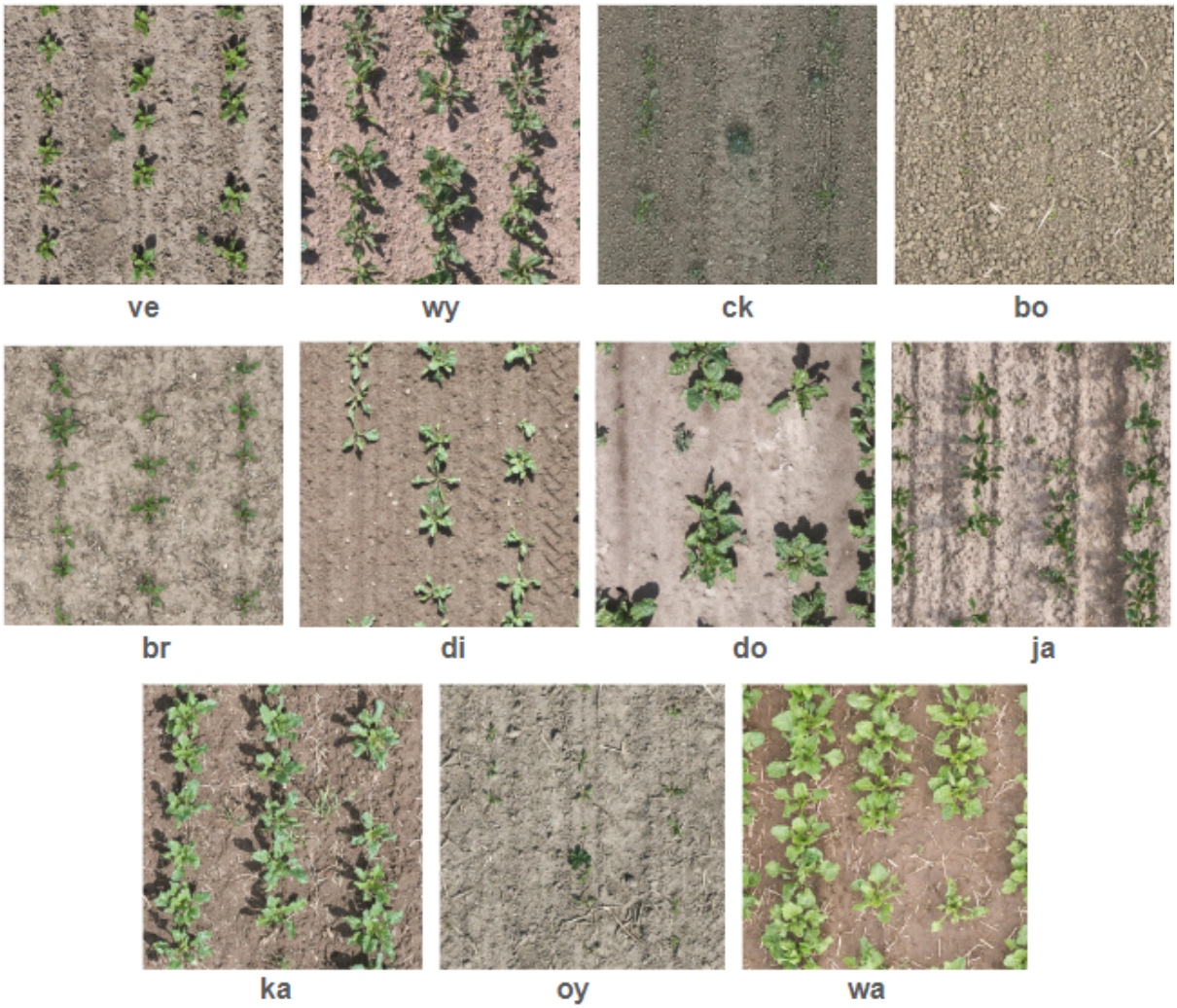

Figure 3: Image example from each farm along with their names

process, utilizing a COCO annotator tool (Brooks, 2019). For annotations, we chose not to use the original resolution of 8192 x 5640 pixels. Instead, we took smaller crops sized at 1024 x 1024 due to our limited GPU resources. In total, we annotated 143 images, a process that involved creating masks for each plant using a polygon tool. Additionally, we marked the central points (nodes) of each annotated plant which are later used as input clicks in the panoptic one-click system. This rigorous and attentive annotation process was pivotal in establishing a dependable and accurate ground truth dataset. This dataset serves as a ground turth, ensuring that our panoptic one-click model is trained and tested on precise and reliable data. Lastly, we exported our dataset in a COCO format for further processing.

Following the annotation procedure, we conducted a few experiments (see section 3.5) to find the proper split of the 11 farms into training, validation, and evaluation sets with the aim of establishing a challenging cross-domain scenario. It is necessary to establish a proper evaluation system for measuring the model's cross-domain performance.

In summary, collecting and preprocessing this dataset provides a strong foundation for our research. It enables us to conduct experiments and assess the generalization performance of the panoptic one-click model in real-world agricultural scenarios with confidence.

3.2 Integration of Vegetation Indices as Additional Inputs

The enhancement of the generalization capability of the Panoptic One-Click model can be achieved by enriching the model with domain-invariant features that facilitate the segmentation process. Different vegetation indices, based on their unique properties, extract various useful invariant features like chlorophyll content, greenness, and overall vegetation health, which aid in the segmentation of plants and soil (Hamuda et al., 2016). To leverage this advantage, we have integrated several robust vegetation indices from Hamuda et al. (2016) into our methodology, namely Excess Green (ExG), Excess Red (ExR), Excess Green Minus Excess Red (ExGR), Modified Excess Green (MExG), Color Index of Vegetation Extraction (CIVE), and Combined Index (COM). Each of these indices brings its own set of strengths to the table, contributing to a comprehensive understanding of the vegetation in the scene and enhancing segmentation performance both within and across different domains. In the equations provided below, 'r,' 'g,' and 'b' represent the 3 channels of the normalized images computed by subtracting the mean of the dataset from the RGB values and dividing by the standard deviation of the dataset.

- Excess Green (ExG):

$$ExG = 2 * g - r - b$$

  ExG quantifies the excess of greenness in the image with reduced susceptibility to background noise and outdoor environmental conditions. High ExG values represent regions with an excess of green.

- Excess Red (ExR):

$$ExR = 1.4 * r - g$$

ExR measures redness with less influence from non-vegetation elements, particularly those in green. High ExR values suggest chlorophyll-rich vegetation, aiding in capturing plant health variations.

- Excess Green Minus Excess Red (ExGR):

$$ExGR = ExG - ExR$$

ExGR enhances vegetation analysis by assessing the green-red color balance. Positive ExGR values represent dominant green, signifying vegetation. While negative values indicate red dominance, often associated with non-vegetation elements. It combines the properties of Exg and ExR which results in better separation and reduced noise.

- Modified Excess Green (MExG):

$$MExG = 1.262 * g - 0.884 * r - 0.311 * b$$

MExG is an adapted version of ExG designed with adjusted weights to emphasize vegetation while reducing sensitivity to non-vegetation elements. Similarly to ExG, high values of MExG represent the abundance of greenness.

- Color Index of Vegetation Extraction (CIVE):

$$CIVE = 0.441 * r - 0.811 * g + 0.385 * b + 18.78745$$

CIVE leverages weighted differences between color channels to highlight the presence of healthy green vegetation in images. Its formula is designed to emphasize vegetation features while minimizing the influence of non-vegetation elements. High CIVE represents the region with dense and healthy vegetation.

- Combined Index (COM):

$$COM = ExG + ExR + ExGR + MExG$$

COM aggregates multiple vegetation indices to provide a holistic representation of vegetation features. These combined features help in a better understanding of plants

and other regions in a scene. High values of COM represent the dense vegetation.

- Sobel Filter (Edge Detection):

  The Sobel filter is an edge detection filter that is used to effectively highlight edges and transitions in pixel intensities (Vincent and Folorunso, 2009). The Sobel operator consists of two convolution kernels: one for horizontal gradients (often denoted as $Sobel_x$) and the other for vertical gradients ($Sobel_y$). These kernels are defined as follows:

$$Sobel_x \text{ Kernel}: \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \quad Sobel_y \text{ Kernel}: \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}$$

  In our methodology, we apply these Sobel kernels from Vincent and Folorunso (2009) to single-channel plant images obtained from vegetation indices using convolution operations. This process allows us to compute gradient magnitude and direction at each pixel. The resulting gradient images are then used to extract edge information, which provides crucial prior knowledge about the plant's structure and boundaries.

To offer a more tangible understanding of how these indices distinguish the plants from the background, we have included visual aids (see Figure 4) demonstrating their impact.

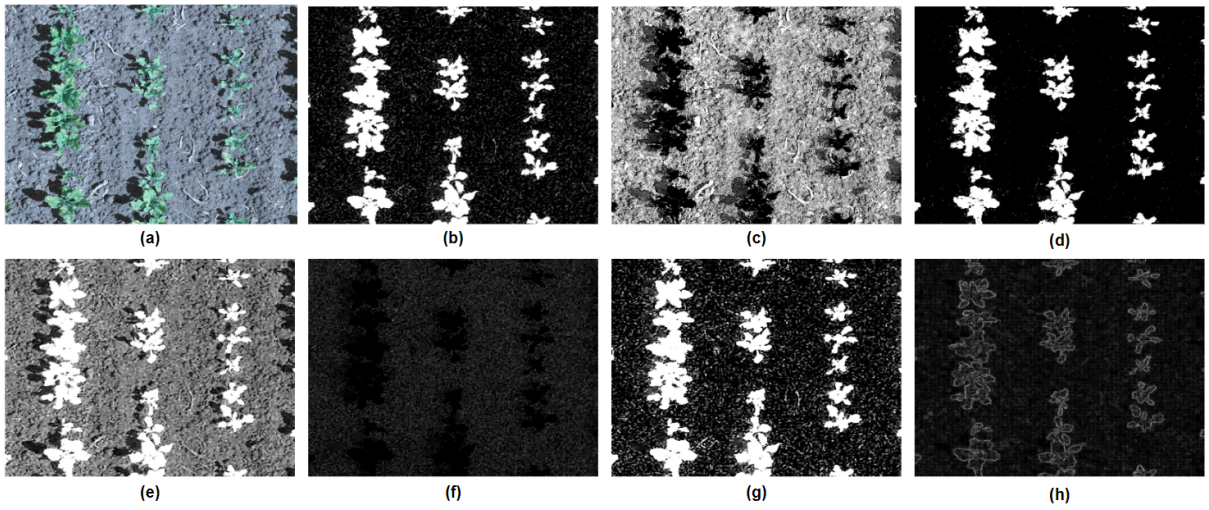

Figure 4: Vegetation indices (*a*) RGB (*b*) ExG (*c*) ExR (*d*) ExGR (*e*) MExG (*f*) CIVE (*g*) COM (*h*) Sobel

Next, The extracted features from these indices can be incorporated both individually (R (Red), G (Green), B (Blue), C (Click), VI (Vegetation Index)) and in combinations (R, G, B, C, VI(i), ..., VI(n)) (see Section 3.5) as additional channels on top of the 4 channel (R, G,

B, C) input representations used in the Panoptic One-Click segmentation model, which is depicted in Figure 5. This incorporation aimed to support the model in plant segmentation. Later, this input is fed into the Panoptic One-Click model, which is a simple U-Net model (Ronneberger et al., 2015) with two heads where one head is used to predict a semantic map, and the other head is used to predict the center regression map. Subsequently, these predicted maps are combined with the input click map using post-processing methods to generate the final panoptic map.

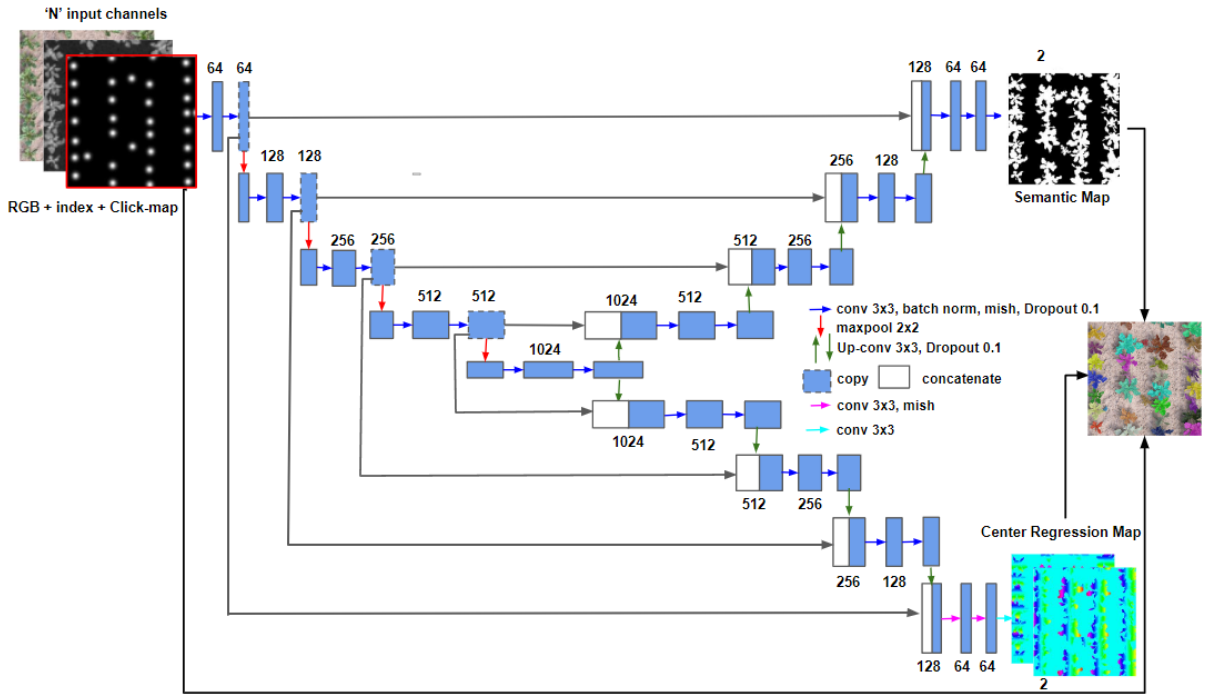

Figure 5: Panoptic one-click model with additional vegetation index input

3.3 Data Augmentation Methods

To enhance the model's ability to generalize effectively, we have employed various data augmentation techniques, including random rotation and color jitter (Khosla and Saini, 2020). In our approach, random rotation is simultaneously applied to all multichannel data (R, G, B, C, VI(1), VI(2), ..., VI(n)) with a rotation degree of 1 and a rotation probability of 0.1, determined empirically through iterative experimentation. This augmentation technique helps the model to learn angular variations.

Additionally, color jitter is applied specifically to the R, G, and B channels with a range and probability of 0.3 and 0.1. These values were determined through an empirical process of iterative experimentation. The resulting image is then combined with the remaining channels

(C, VI(1), VI(2), ..., VI(n)) to ensure that color alterations do not affect critical components like the click maps and vegetation indices. This augmentation technique helps the model to learn different color representations encountered in real-world scenarios.

We have included these methods in our system, as they have been proven effective in enhancing generalization in other systems, and we expect similar positive effects in our system.

3.4 Dilated Panoptic one-click model

To further improve the generalization performance some modifications are done in the architecture of the panoptic one-click model. We modified the standard convolution layers at the bottleneck of the model with the dilated convolution stack (Piao and Liu, 2019) (see Figure 6).

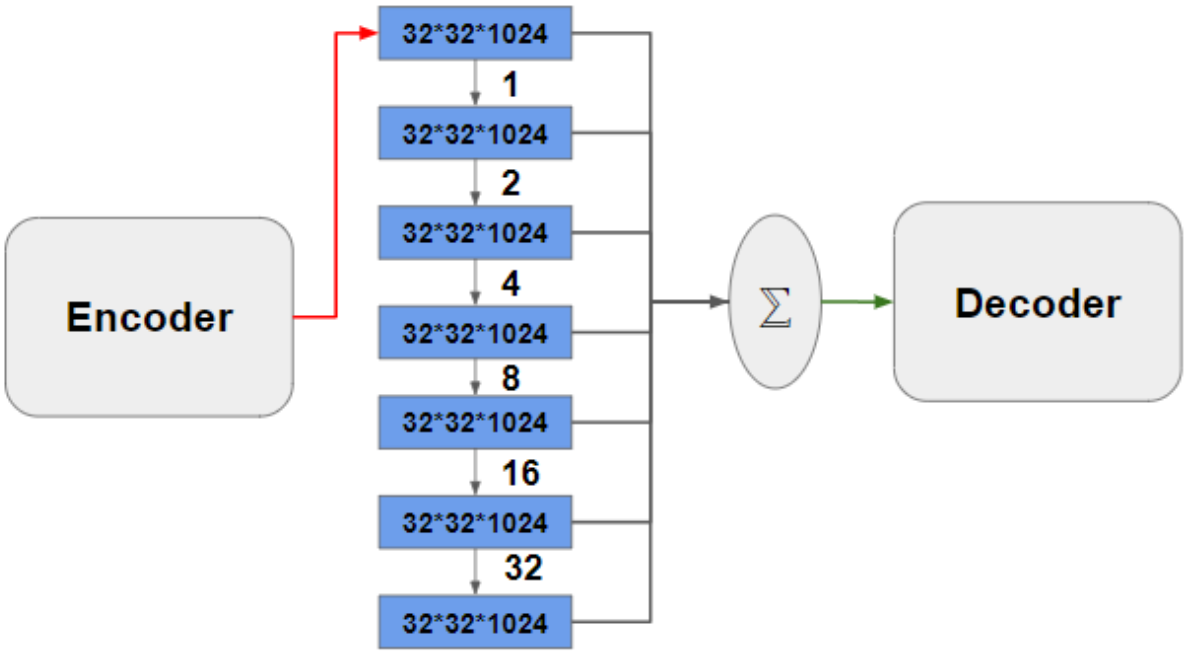

Figure 6: Panoptic one-click model with replaced bottleneck by dilated convolution stack

When we go into deeper layers of the deep neural networks, the semantic information increases while the spatial information decreases due to a series of pooling and convolution operations (Piao and Liu, 2019). This dilated convolution stack plays a crucial role in enabling the model to capture information at different scales which helps in enriching the semantic understanding while preserving spatial information. Layers with low dilation rates focus on fine-grained details, while layers with high dilation rates gather broader contextual

information.

In our approach, we have utilized the 7 dilation convolution operations with different dilation rates ranging from 1 to 32 to increase the receptive fields of our 32x32 image. We have maintained the resolution of the image through different dilation layers by the operation of the padding. In the end, we have fused the multi-scale information from all layers so we have a better understanding of the scene which ,may results in improved performance and reduced overfitting.

3.5 Experimental System

To comprehensively assess the model's performance, we divided the 11 fields taken from different farms into two distinct subsets set A and set B. In this partition, Set A served as the designated dataset for all stages of model development, including training, validation, and evaluation to access the model's in-domain performance, while Set B was exclusively reserved for evaluating the model's performance on data from unseen fields. To find a split of the number of fields into set A and set B that gives the challenging cross-domain scenario we have conducted two different experiments.

First Experiment:

Set A: 3 fields (ck, ve, wy)

Train: 9 images

Validation: 24 images

Evaluation: 25 images

Set B: 8 fields (bo, br, di, do, ja, ka, oy, wa)

Evaluation: 72 images

Second Experiment:

Set A: 2 fields (ve, wy)

Train: 6 images

Validation: 19 images

Evaluation: 21 images

Set B: 9 fields (ck, bo, br, di, do, ja, ka, oy, wa)

Evaluation: 78 images

In the first experiment, we randomly split the data into 3/8 and found out that having 3 diverse fields in set A already led to good generalization performance. So in the second

experiment, we deliberately decided to move one more field to the cross-domain testing set B to make it more challenging.

Subsequently, we selected the second experiment as our baseline. We then conducted several additional experiments to investigate how including vegetation indices as additional channels affects the generalization performance of the panoptic one-click model. The following list outlines these experiments:

1. Integration of ExG with (R, G, B, C) channels
2. Integration of ExR with (R, G, B, C) channels
3. Integration of ExGR with (R, G, B, C) channels
4. Integration of MExG with (R, G, B, C) channels
5. Integration of CIVE with (R, G, B, C) channels
6. Integration of COM with (R, G, B, C) channel
7. Integration of Sobel Filter (ExG) with (R, G, B, C) channel

Following this, we selected some of the vegetation indices that improved the generalization performance from the above experiments, and their different combinations were tested to further enhance the general capabilities of the panoptic one-click model. The experiments conducted for this purpose are listed below:

1. Integration of ExG, ExGR with (R, G, B, C) channels
2. Integration of ExR, ExGR with (R, G, B, C) channels
3. Integration of MExG, ExGR with (R, G, B, C) channels
4. Integration of COM, ExGR with (R, G, B, C) channels
5. Integration of Sobel Filter, ExGR with (R, G, B, C) channels

In an effort to further improve the generalization capabilities of the model five other experiments were performed where data augmentation methods were applied on the baseline (R, G, B, C) and the best-performing combination of multichannel (R, G, B, C, ExGR) input data. The performed experiments are listed below:

1. Random Rotation on baseline (R, G, B, C)

2. Random Rotation on multichannel input (R, G, B, C, ExGR)

3. Color Jitter on R, G, B channels of baseline (R, G, B, C)

4. Color Jitter on R, G, B channels of multichannel input (R, G, B, C, ExGR)

5. Random Rotation on (R, G, B, C) channels and Color Jitter on (R, G, B) channels combined on baseline (R, G, B, C)

Following this, to further enhance the generalization of the model two more experiments were performed with the modified network on the baseline (R, G, B, C) and the best-performing combination of multichannel (R, G, B, C, ExGR) input data. The performed experiments are listed below:

1. Dilated Panoptic one-click model on 4-channel input data (R, G, B, C)

2. Dilated Panoptic one-click model on 5-channel input data (R, G, B, C, ExGR)

All of these experiments are conducted on a GeForce RTX 2080 Ti using Python and PyTorch Lightning.

3.6 Implementation

To train the Panoptic One-Click model using multichannel data, we normalize the RGB image by subtracting the mean and dividing it by the standard deviation calculated from the training dataset for each channel. This ensures a consistent range of values. Additionally, we resize all the channels present in our input images to 512 x 512 pixels due to limited GPU constraints. A batch size of 1 is employed because of the small training dataset. This smaller batch size introduces noise into the weight updates, which serves as a regularization technique (Keskar et al., 2017). It helps prevent overfitting, encourages the model to avoid sharp minima, and ultimately enhances its generalization to new, unseen data.

Throughout the training, we maintain a fixed learning rate of 0.001. The network is trained for a total of 600 epochs which takes around 60 minutes. The choice of these parameters is made by monitoring the model's performance on the validation set from set A, and then these settings are retained for all subsequent experiments.

To address the class imbalance in our dataset, we use weighted cross-entropy loss (Smitt et al., 2022) which assigns lower loss to a class having more samples. The formula used to calculate the weights for the classes (plant and background), taken from Zimmer et al. (2023) is as follows:

$$w_c = 1/log(Area_{class}/Area_{backgorund} + 1.02) \tag{1}$$

In this formula, $Area_{class}$ and $Area_{background}$ represent the average areas covered by the chosen class and soil in pixels within a training set. The ratio $Area_{class}/Area_{background}$ indicates whether the chosen class or soil exhibits greater visibility within the training set. Based on this ratio, we assign weight to the classes. In our case, plants are assigned higher weights as they are on average less visible than the ground.

3.7 Performance Evaluation Metric

To evaluate the performance of the model we measured the mean quality of each individual mask by a metric called mean object intersection over union (mIOU), which is a metric of choice for such systems (Xu et al., 2016), (Zimmer et al., 2023). This metric quantifies how well the model aligns its predicted object masks with the corresponding ground truth object masks. The foundation of mIOU is the object-level IoU (Intersection over Union) formula, which computes the ratio of the overlapping area between predicted and ground truth objects to their combined area.

$$IoU = (A \cap B)/(A \cup B)$$

In this representation:

- "$(A \cap B)$" represents the region where the model's prediction for a specific object matches the ground truth mask for the same object.
- "$(A \cup B)$" encompasses the combined region covered by both the predicted object mask and the ground truth mask for that object.

Object-level IoU is calculated individually for each object within an image. This allows for a precise evaluation of how accurately the model delineates each object.

To derive an overall assessment of the model's performance across all objects and classes in a dataset, the object-level IoU scores are summed and then divided by the total number of

instances. This calculation results in mIOU.

$$mIOU = (IoU1 + IoU2 + ... + IoU_n)/n$$

mIOU provides a comprehensive evaluation of how effectively the model distinguishes and outlines individual objects in images from the entire dataset. Higher mIOU scores indicate superior instance-based segmentation accuracy and alignment between predictions and ground truth for all classes and instances combined.

4 Results and Discussion

In this chapter, we will explore and discuss the outcomes of experiments conducted to enhance the overall performance of the Panoptic One-Click model. In the first section, we explored the outcomes of different field splits within our Set A and Set B to identify the split that offers a challenging cross-domain scenario. Moving to the second part, we evaluated the additional channel approach, where we added vegetation indices as extra input channels in the panoptic one-click model. Following this, we assessed data augmentation approaches such as random rotation and color jitter, comparing their performance against both our baseline model (R, G, B, C) and the best-performing model (R, G, B, C, ExGR). Lastly, we evaluated our modified network and compared it against the baseline model (R, G, B, C) and the best-performing model (R, G, B, C, ExGR).

4.1 Evaluation of Different Compositions of Fields

We will begin by presenting the results of our initial experiments when introducing the new dataset, which includes the evaluation of different compositions of our sets A and B. The findings from this experiment are listed in the table below:

Table 1: Performance of Panoptic one-click model on different compositions of our sets A and B

In Domain Evaluation			
Training Dataset	Evaluation Dataset	mIOU Plant[%]	
		3 farms (ck,ve,wy)	2 farms (ve,wy)
Set A	Set A	66.1	75.3
Cross Domain Evaluation			
Training Dataset	Evaluation Dataset	mIOU Plant[%]	
		8 farms	9 farms
Set A	Set B	72.9	66.6

In Table 1, we can discern considerable differences in the in-domain and cross-domain performance when the panoptic one-click model is trained with Set A having 3 farms as compared to when it's trained with Set A having 2 farms. Notably, we observed that its in-domain accuracy is lower by 9.2 % when trained with three farms. while in the cross-domain case when the trained model with 3 farms is tested on Set B having 8 different farms, it leads to leads to a 6.3 % increase compared to when the trained model with 2 farms is tested on Set B having 9 different farms. This discrepancy can be attributed to the data collection process.

Additionally, in the 'ck' field, the plants were hoed instead of sprayed, and we observed potential mechanical damage to the plants. This mechanical damage, combined with the higher compression levels used for capturing data in .jpg format, resulted in a reduction in image quality. The higher compression introduced significant data loss, including crucial color information and finer image details like edges. In contrast, the other fields were captured in raw format, ensuring higher quality and preserving a greater amount of image information. Furthermore, the 'ck' field exhibited a specific type of weed not present in the other fields. This unique weed species introduced variability in the 'ck' field's images, making them distinct from those of the other fields. Furthermore, it was underrepresented in the small training set. The model's limited exposure to this specific weed type during training hindered its ability to generalize effectively to images with this weed presence during in-domain tasks.

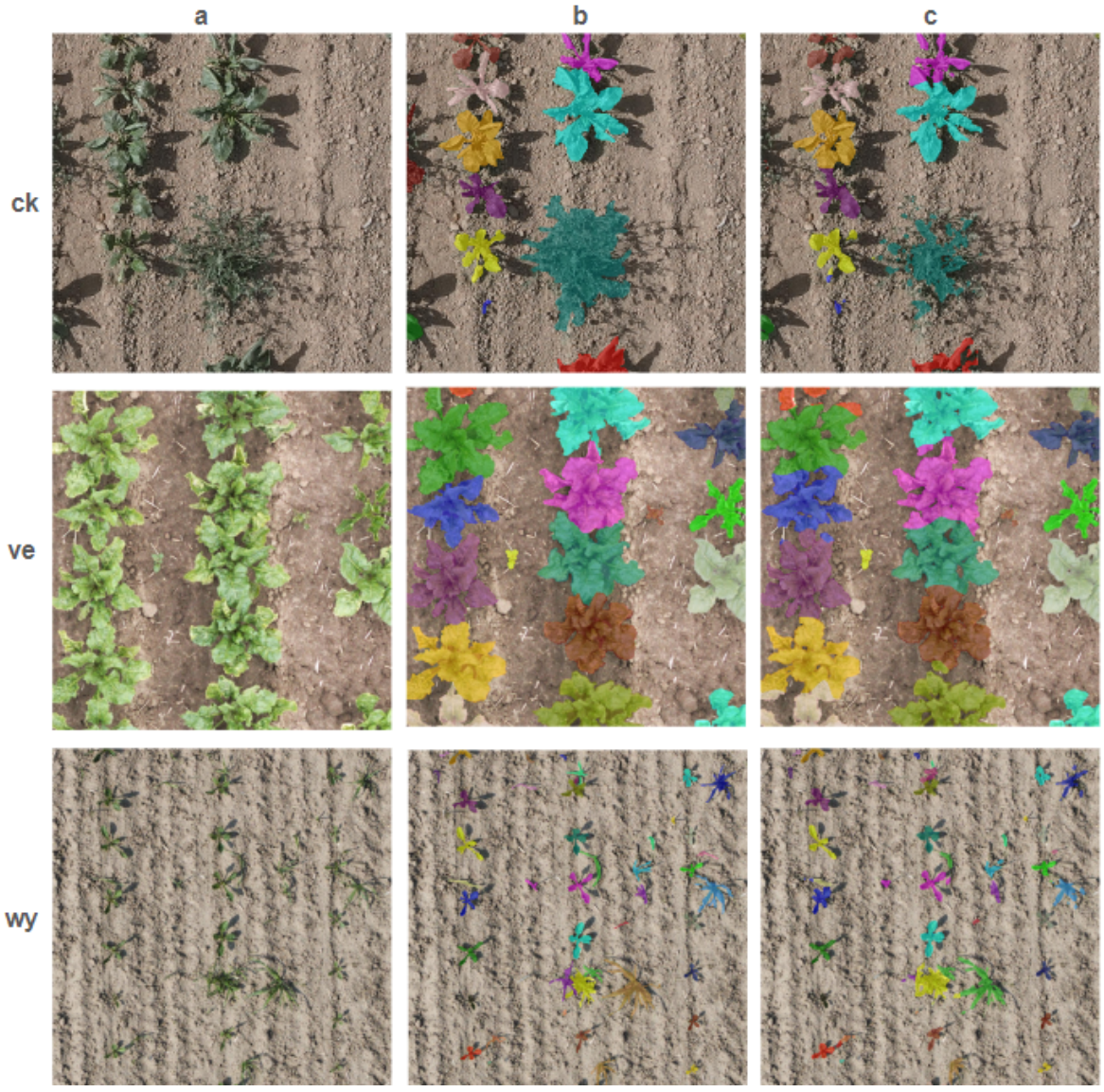

Figure 7: Evaluation of Set A (a) RGB (b) Ground Truth (c) Panoptic One-click (R, G, B, C)

To provide a visual representation of these differences, Figure 7 presents a comparison of images from the 'ck' field and the other two fields used for training, along with their corresponding ground truth and segmentation results. This visual comparison further illustrates how the combination of JPEG compression, mechanical plant damage, and the presence of a distinct weed species impacted the model's segmentation performance on in-domain tasks. Additionally, it can also be seen that the panoptic one-click segmentation performed well at non-overlapping regions but some misclassifications were observed in the overlapping regions.

In the second composition, when 'ck' is shifted from Set A to Set B, resulting in a drop in cross-domain performance while increasing in-domain performance due to the aforementioned reasons. Consequently, we selected this second composition as our baseline, as it presents a more challenging scenario for generalization.

4.2 Evaluation of Vegetation Indices as Extra Channels

In this section, we evaluate the impact of incorporating various vegetation indices as an additional channel in the model's input data, alongside the R, G, B, and click channels. The results of these experiments, where each vegetation index is added individually as a fifth channel to the input data, are presented in the table below:

Table 2: performance of Panoptic one-click model on 5-channel input

In Domain Evaluation									
Training Dataset	Evaluation Dataset	mIOU Plant[%]							
		Baseline	ExG	ExR	MExG	Sobel Filter (ExG)	ExGR	CIVE	COM
Set A	Set A	75.3	78.5	79.1	78.4	78.9	77.5	79.0	76.9
Cross Domain Evaluation									
Training Dataset	Evaluation Dataset	mIOU Plant[%]							
		Baseline	ExG	ExR	MExG	Sobel Filter (ExG)	ExGR	CIVE	COM
Set A	Set B	66.6	69.3	68.7	69.5	70.1	72.2	54.3	71.3

Incorporating the ExG channel, as evidenced by Table 2, significantly boosts model accuracy in both in-domain and cross-domain by 3.2 % and 2.7% compared to baseline. This enhancement is attributed to the ExG channel's ability to emphasize green elements in images and its capacity to handle environmental and lighting variations. What makes the ExG channel particularly valuable is its consistent 'green clue' that remains stable across diverse datasets and environmental conditions, akin to a universal feature shared by various plant

types. This green clue not only aids the model in better distinguishing plants from the soil but also enhances its ability to generalize effectively across different datasets and environments, as shown in an example in Figure 9(c).

In Table 2, we can see that when we included the ExR channel as an extra channel in our model, both the in-domain and cross-domain performances improved by 3.8 % and 2.1 % respectively, compared to the baseline. This enhancement is attributed to the ExR channel's ability to differentiate between plants and soil by examining the red color in objects. As healthy plants have high chlorophyll content, they reflect less red light compared to the background, which reflects high red light. Providing this additional color clue to the model helps in better differentiation between plants and soil. In comparison with ExG, the cross-domain performance experiences a slight decrease of about 0.6%. This decline can be attributed to ExR's limitations in accurately detecting plants under specific lighting conditions. For instance, in scenarios with high illumination intensity, ExR may struggle with plant detection. As the illumination increases, the ExR value gap between plants and soil becomes smaller as discussed by Upendar et al. (2021). In contrast, ExG performs more reliably under these conditions, leading to better segmentation results. This decline in performance can also be visually observed in Figure 9(d).

The inclusion of the MExG vegetation index as a fifth channel in the model also resulted in a significant improvement in both in-domain and cross-domain performances by 3.6 % and 3.5 % compared to the baseline results, as listed in Table 2. This improvement can be attributed to the unique characteristics of MExG, which differentiates it from ExG. MExG, with its optimized weighting scheme, excels in extracting plants from the background with higher accuracy compared to ExG in different environment conditions as discussed by Guo et al. (2013). The weighting scheme of MExG enhances its ability to capture color features that are more robust to different illumination conditions. As a result, the model's performance benefits from these robust features, leading to the observed 0.4 % improvement in in-domain performance and 0.8 % improvement in cross-domain performance compared to ExG, as listed in Table 2. The segmentation result can also be visually seen in Figure 9(e).

The incorporation of the Sobel filter (ExG) as an extra channel also leads to an improvement in both in-domain and cross-domain performance by 3.1 % and 2.9 % respectively compared to the baseline, as documented in Table 2. This enhancement can be attributed to the Sobel filter's distinctive ability to extract edges in the scene, which facilitates the creation

of clear boundaries between plant and non-plant regions. This pre-segmentation information provides the model with valuable clues, enabling it to better discern between plant and non-plant regions within a scene. It's worth noting that the effectiveness of the Sobel filter is dependent on the choice of the vegetation index to which it is applied. In our case, we applied this edge extraction to the 'EXG' image. This strategic choice further amplifies the benefits of the Sobel filter, resulting in a slight 0.2% improvement in cross-domain performance compared to 'EXG' see Figure 9 (f).

The inclusion of ExGR as an additional channel in our model also boosted the model's performance in both in-domain and cross-domain scenarios by 2.2 % and 5.6 % respectively compared to the baseline, as demonstrated in Table 2. This enhancement can be attributed to ExGR's ability to measure the relative difference between greenness and redness within an image. By combining these two characteristics, ExGR effectively extracts green vegetation while reducing background noise from non-plant green elements (Hamuda et al., 2016), (Aureliano Netto et al., 2018). This dual-feature extraction approach offers more invariant features across domains, allowing the model to distinguish plants from soil more effectively. Furthermore, the ExGR is more robust to environmental and illumination variations as compared to ExG (Meyer and Neto, 2008), (Aureliano Netto et al., 2018), (Upendar et al., 2021), ExR (Upendar et al., 2021), and MExG (Guo et al., 2013). Learning from this robust information, the model exhibits improved generalization compared to using ExG, ExR, and MExG as seen in Table 2. This enhanced performance can also be visually seen in Figure 9(g).

Incorporating CIVE as an additional channel to our model significantly improved in-domain performance by 3.7 %, but a considerable decline of 12.3% was observed in cross-domain performance compared to the baseline. This drop in cross-domain performance can be attributed to the poor adaptability of CIVE to varying lighting conditions, especially in the presence of shadows and high illumination, as discussed by Zheng et al. (2009), Yu et al. (2013). The impact of these conditions on the segmentation performance with their corresponding ground truths can also be visually seen in some of the example images taken from set B, see Figure 8. It's noteworthy that the cross-domain results with CIVE also fall below the performance achieved with other tested vegetation indices and can be visually seen in Figure 9(h).

The inclusion of the Composite (COM) index as an additional channel also leads to an improvement in both in-domain and cross-domain performance by 1.6 % and 4.7 % respect-

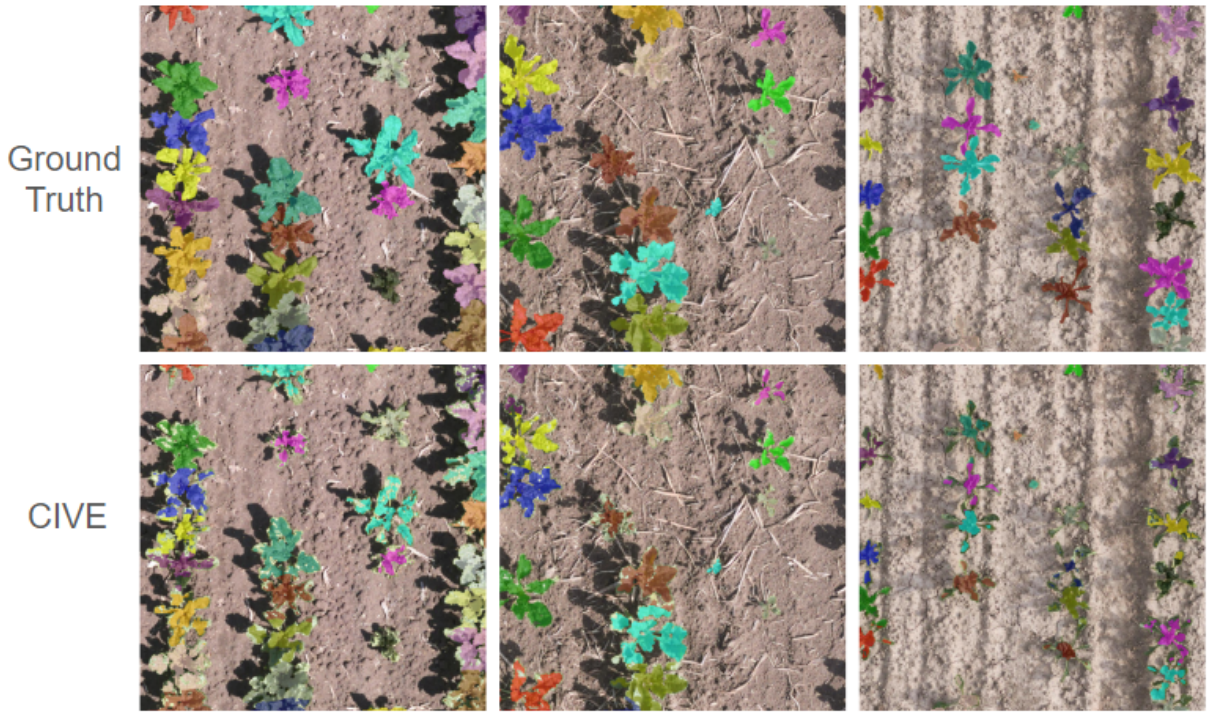

Figure 8: Evaluation of set B: Impact of varying illumination conditions on the performance of panoptic one-click (R, G, B, C, CIVE) when CIVE is added as additional channel

ively compared to the baseline, as documented in Table 2. The primary contributing factor to this enhancement is the fusion of multiple vegetation indices into a single, composite index. This composite index incorporates the diverse properties extracted from different individual indices, enhancing the model's robustness in handling variations in lighting, growth stages, and soil conditions. The result is a significant improvement in generalization performance across a range of scenarios. Conversely, a minor drop in in-domain performance is observed. This decline could be attributed to the reduction in overfitting achieved through the use of combined indices. The combination of multiple indices helps prevent the model from focusing excessively on domain-specific features, resulting in a more balanced and adaptable model see Figure 9(i).

Next, instead of passing one vegetation index at a time, we have incorporated 2 vegetation indices as an additional channel to see the effect on the generalization performance of the panoptic one-click model. In our approach, we have decided to test the combination of ExGR with other indices to further push the high performance achieved. As each index offers unique characteristics, combining the indices may more enrich the input representation. This enrichment may help the model to have a better understanding of vegetation characteristics. The obtained results are listed in 3.

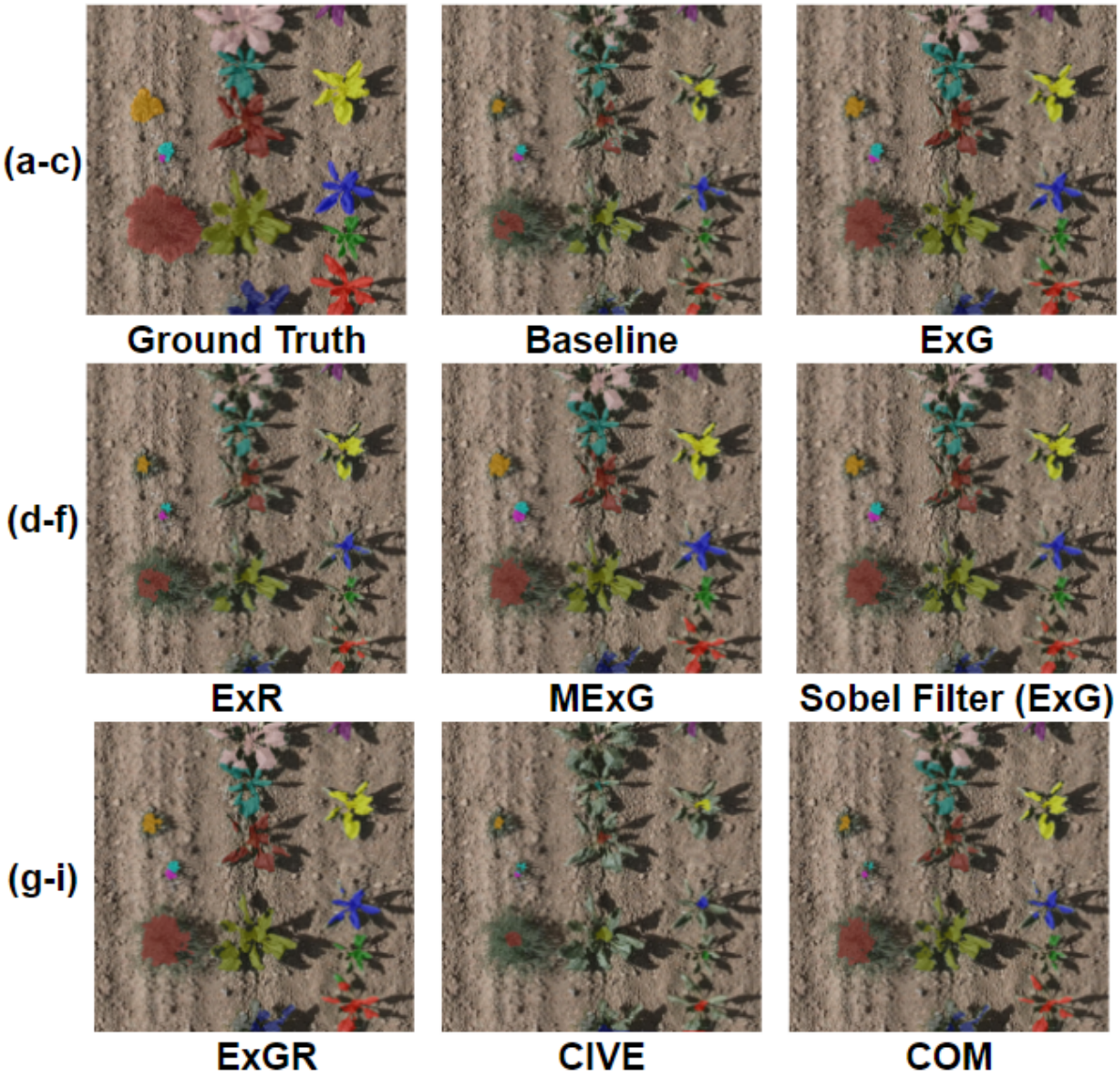

Figure 9: Evaluation of Set B (a) Ground Truth (b) Baseline (R, G, B, C) (c) Panoptic One-click (R, G, B, C, ExG) (d) Panoptic One-click (R, G, B, C, ExR) (e) Panoptic One-click (R, G, B, C, MExG) (f) Panoptic One-click (R, G, B, C, Sobel Filter (ExG)) (g) Panoptic One-click (R, G, B, C, ExGR) (h) Panoptic One-click (R, G, B, C, CIVE) (i) Panoptic One-click (R, G, B, C, COM)

The 6-channel input (R, G, B, C, ExR, ExGR), achieved in-domain and cross-domain accuracies of 79.2 % and 71.9 %, respectively. These were the highest among other 6-channel combinations and baseline. However, its in-domain performance was higher by 1.7 % while the cross-domain performance was slightly lower by 0.3 % compared to the maximum accuracy achieved with the 5-channel input (R, G, B, C, ExGR). This discrepancy in performance is due to redundancy and similarity among the features, as discussed by Yang et al. (2020). This overlapping information added complexity without providing new useful details, potentially causing overfitting and hindering the model's ability to generalize to new data. This pattern

remained consistent across all other examined combinations, consistently impacting the model's generalization ability.

Table 3: performance of Panoptic one-click model on 6-channel input

In Domain Evaluation							
Training Dataset	Evaluation Dataset	mIOU Plant[%]					
		Baseline	ExG, ExGR	ExR, ExGR	MExG, ExGR	COM, ExGR	Sobel Filter (Exg), ExGR
Set A	Set A	75.3	76.7	79.2	78.2	79.1	77.5
Cross Domain Evaluation							
Training Dataset	Evaluation Dataset	mIOU Plant[%]					
		Baseline	ExG, ExGR	ExR, ExGR	MExG, ExGR	COM, ExGR	Sobel Filter (Exg), ExGR
Set A	Set B	66.6	71.4	71.9	68.8	70.7	69.4

In summary, the findings in this section highlight that the inclusion of invariant vegetation-related features in the input helps the model to become more robust and accurate. The inclusion of ExGR as an extra channel has produced the highest cross-domain result of 72.2 % mIOU because of its robustness against environmental and illumination conditions as compared to other indices.

4.3 Evaluation of Data Augmentation Methods

In this section, we will evaluate the performance of the presented data augmentation approaches against our baseline and the best result obtained so far, which is a 5-channel image containing ExGR. The results obtained are listed in 4.

Table 4: Performance of data augmentation methods (random rotation, color jitter, and combined) on the panoptic one-click model(R, G, B, C) and panoptic one-click model(R, G, B, C, ExGR)

In Domain Evaluation								
Training Dataset	Evaluation Dataset	mIOU Plant[%]						
		No Augmentation		Color Jitter		Random Rotation		Combined
		Baseline	ExGR	Baseline	ExGR	Baseline	ExGR	Baseline
Set A	Set A	75.3	77.5	78.8	77.2	78.6	75.6	77.6
Cross Domain Evaluation								
Training Dataset	Evaluation Dataset	mIOU Plant[%]						
		No Augentation		Color Jitter		Random Rotation		Combined
		Baseline	ExGR	Baseline	ExGR	Baseline	ExGR	Baseline
Set A	Set B	66.6	72.2	68.6	72.2	70.1	70.9	65.7

The incorporation of color jitter into the baseline model notably improved in-domain performance by 3.5 % and enhanced cross-domain performance by 2 %. Color jitter effectively captures diverse illumination variations, thereby expanding the training sample's effectiveness and mitigating risks of overfitting within the model.

In the case of ExGR, with or without color jitter, the cross-domain performance remained consistent at 72.2 %. This consistency suggests that ExGR has a more pronounced impact on model performance compared to color jitter. ExGR demonstrates robustness against illumination variations and offers pre-segmentation information, aiding the model in robust segmentation even in the face of changes in illumination conditions. Consequently, applying color jitter augmentation in the presence of ExGR does not contribute to a further performance increase due to redundant information.

The implementation of random rotation significantly enhances baseline performance, improving both in-domain and cross-domain accuracy by 3.3 % and 3.5 %, respectively. Random rotation aids the model in learning location and angle variations by introducing transformed training samples, reducing overfitting, and boosting performance. However, in the case of random rotation applied to the 5-channel input (R, G, B, C, ExGR), there is a notable performance drop of 1.9 % in the in-domain and 1.3 % in the cross-domain as compared to no random rotation. These results indicate that ExGR is sensitive to random rotation, causing a misalignment between RGB and ExGR information. This misalignment may affects the model's ability to interpret the relationship between different types of information, leading to decreased performance. Increasing the random rotation value might enhance this issue, resulting in a further drop in model performance.

To investigate the impact of random rotation on feature representation, we isolated a random feature map from the model's bottleneck and reduced its dimensionality to 2-D using t-SNE visualization method (Van der Maaten and Hinton, 2008). When the baseline model is applied without random rotation, the features representing plants and background lack clear distinguishability, resulting in ambiguous boundaries and higher misclassification errors see Figure 10(a). However, upon applying random rotation to the baseline model, the boundaries between plant and background features become somewhat clearer, and the feature spread reduces, leading to a decrease in misclassification errors compared to the previous scenario see Figure 10(b). In the case of a 5-channel input (utilizing R, G, B, C, and ExGR), as illustrated in Figure 10(c), the features exhibit a small spread, tightly compacted with clear distinguishable boundaries. This characteristic results in very low misclassification errors, explaining why this model outperforms other approaches. However, when random rotation is applied to the 5-channel input, although clear boundaries persist, the spread increases see Figure 10(d). This suggests that certain plant data points might share similarities

or characteristics with soil data, leading to a slight performance drop. Moreover, as the amount of random rotation on the 5-channel input increases further, the boundaries become unclear, causing all points to scatter see Figure 10(e). This scenario results in higher misclassification errors compared to other scenarios.

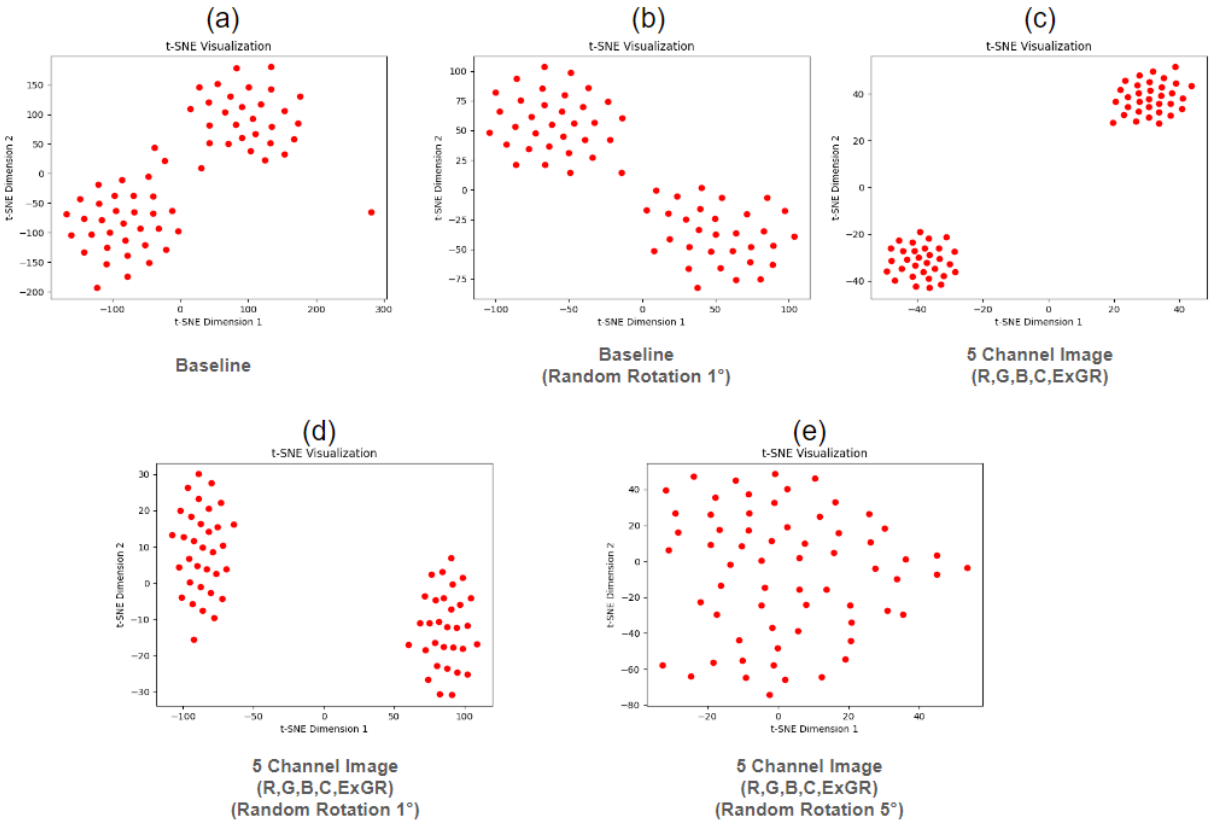

Figure 10: t-SNE Visualization: Panoptic One-Click Bottleneck Feature Map in 2D

To test the limitations of data augmentation, we combined color jitter and random rotation on the baseline model (R, G, B, C). The outcomes showed an in-domain performance of 77.6 % and a cross-domain performance of 65.7 %. This decline in performance suggests that excessive augmentation led the model to adapt specifically to the augmented data, resulting in overfitting rather than learning generalizable features. The combined augmentation approach was not tested on the model (R, G, B, C, ExGR) due to observed overfitting. Additionally, when applying color jitter and random rotation individually in the presence of ExGR, we noticed no noticeable additional benefits.

In summary, these findings emphasize that the ExGR channel has a greater impact on model performance compared to the applied augmentation approaches. The implementation of color jitter resulted in redundant information in the presence of the ExGR channel, while random rotation may cause misalignment effects between the ExGR information and other

channels. Additionally, excessive use of augmentation approaches led to massive overfitting.

4.4 Evaluation of Architectural Modification

In this section, we will evaluate the performance of the dilated panoptic one-click model against our Panoptic one-click model in case of 4-channel and 5-channel input data. The results obtained are presented in Table 5.

Table 5: Performance comparison of Dilated panoptic one-click against Panoptic one-click model

In Domain Evaluation					
Training Dataset	Evaluation Dataset	mIOU Plant[%]			
		Panoptic one-click		Dilated panoptic one-click	
		(R, G, B, C)	(R, G, B, C, ExGR)	(R, G, B, C)	(R, G, B, C, ExGR))
Set A	Set A	75.3	77.5	77.9	77.3
Cross Domain Evaluation					
Training Dataset	Evaluation Dataset	mIOU Plant[%]			
		Panoptic one-click		Dilated panoptic one-click	
		(R, G, B, C)	(R, G, B, C, ExGR)	(R, G, B, C)	(R, G, B, C, ExGR))
Set A	Set B	66.6	72.2	70.7	71.2

The modified network exhibits an enhanced performance of 2.6 % and 4.1 % on 4-channel input data in in-domain and cross-domain scenarios compared to the panoptic one-click model (R, G, B, C) which is our baseline. These findings suggest that the incorporation of stacked dilated convolutions assists the model in capturing larger spatial information, enabling it to comprehend complex patterns. Moreover, the expanded receptive field facilitates learning features across multiple scales, aiding in understanding relationships within the data. This approach allows the model to effectively capture more invariant features, consequently improving its overall performance.

While in the case of 5-channel input data, the modified network exhibits a slight performance drop of 0.2 % and 1 % in in-domain and cross-domain scenarios compared to panoptic one-click model (R, G, B, C, ExGR) which is our best performing model so far. This drop suggests that the ExGR channel, aimed at separating plants from the background, could potentially contain noise stemming from varied environmental conditions. As a result, this pre-segmentation data from ExGR might occasionally possess inaccuracies or noise as shown in Figure 11(a). When this noisy ExGR data is processed through the dilation stack, both the noise amplifies and the fine details in plant structures are lost, particularly with the increase in dilation rates, as shown in Figure 11(b-d) This amplification of noise and loss of information

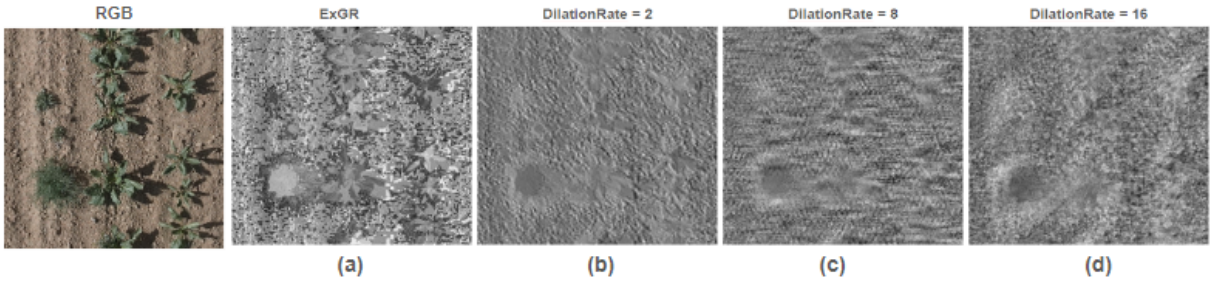

Figure 11: Impact of noise on dilated convolution

by the dilation stack might lead to flawed feature learning or misinterpretation of the data. Consequently, it could affect the model's capacity to accurately segment the plants, resulting in an overall reduction in segmentation performance.

In summary, these findings highlight that the introduction of dilation layers in the panoptic one-click model causes the amplification of noise present in the data which hurts its performance.

5 Conclusion

In our research, we explored various domain generalization approaches to enhance the capabilities of the panoptic one-click model, intending to leverage it as an annotation tool for agricultural data and minimize manual annotation efforts. Our initial focus was on the multichannel approach, and the inclusion of the ExGR index as an additional channel yielded the most promising results in cross-domain performance, achieving 72.2 % mIOU, surpassing all other approaches. ExGR demonstrated superior performance due to its reduced sensitivity to noise and environmental conditions. Additionally, the inclusion of multiple vegetation indices as extra channels did not provide added benefits due to the redundancy of information. These findings highlight the significance of selecting the appropriate vegetation index or indices, while considering factors such as noise levels and information redundancy, especially within multichannel methodologies.

Secondly, the implementation of augmentation techniques, such as color jitter and random rotation, failed to further extend the highest achieved performance. This limitation resulted from the dominance of pre-segmentation information provided by ExGR, which exerted a more substantial influence on the model compared to augmentation approaches. The implementation of color jitter provided redundant information in the presence of ExGR, while random rotation may have had misalignment effects on the pre-segmentation information from ExGR. Moreover, a combined application of augmentation techniques could potentially increase overfitting issues. These findings suggest the potential importance of considering data augmentation strategies based on the variations present in the target domain. In multichannel scenarios, augmentation should complement rather than impede or duplicate the extra information provided Finally, the implementation of the dilated panoptic one-click model also failed to further improve the maximum performance achieved. This occurred because, in the case of multichannel input (R, G, B, C, ExGR), the dilation stack amplified the noise present in the ExGR, which had a negative impact on the model's performance.

Overall, all the methods exhibited an improvement in performance compared to our baseline. Among these, the multichannel approach utilizing the ExGR vegetation index as an additional input demonstrated superior cross-domain performance. This suggests that incorporating invariant vegetation-related features has a greater impact on the robustness of the model.

6 Future Work

In the future, the performance of the panoptic one-click model could be further enhanced by expanding the training set with a more diverse range of data. Another approach could involve the use of zero-shot learning methods that leverage the model's existing knowledge learned from one domain to generalize and perform tasks in entirely new or unseen domains without the need for explicit training on those specific tasks (Romera-Paredes and Torr, 2015).

Moreover, it's crucial to explore how well the panoptic one-click model can adapt to different agricultural datasets. Understanding its flexibility in handling various types of agricultural data will be essential for its broader use in the field.

References

(2005). Rawtherapee. Computer software.

Allamy, H. (2014). Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study).

Aureliano Netto, A., Martins, R., de Souza, G., De Moura Araújo, G., Almeida, S., and Capelini, V. (2018). Segmentation of rgb images using different vegetation indices and thresholding methods. *Nativa*, 6:389.

Baba, N. M., Makhtar, M., Fadzli, S. A., and Awang, M. K. (2015). Current issues in ensemble methods and its applications. *Journal of Theoretical & Applied Information Technology*, 81(2).

Bhadra, T., Mahapatra, C. K., and Paul, S. K. (2020). Weed management in sugar beet: A review. *Fundamental and Applied Agriculture*, 5(2):147–156.

Bian, S. and Wang, W. (2007). On diversity and accuracy of homogeneous and heterogeneous ensembles. *Int. J. Hybrid Intell. Syst.*, 4:103–128.

Bisong, E. (2019). *Regularization for Deep Learning*, pages 415–421. Apress, Berkeley, CA.

Blok, P. M., van Evert, F. K., Tielen, A. P. M., van Henten, E. J., and Kootstra, G. (2021). The effect of data augmentation and network simplification on the image-based detection of broccoli heads with mask r-cnn. *Journal of Field Robotics*, 38(1):85–104.

Brilhador, A., Gutoski, M., Hattori, L., de Souza Inácio, A., Lazzaretti, A., and Lopes, H. (2019). Classification of weeds and crops at the pixel-level using convolutional neural networks and data augmentation. pages 1–6.

Brooks, J. (2019). COCO Annotator. `https://github.com/jsbroks/coco-annotator/`.

Cheng, B., Collins, M. D., Zhu, Y., Liu, T., Huang, T. S., Adam, H., and Chen, L.-C. (2020). Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation.

Elharrouss, O., Al-Maadeed, S., Subramanian, N., Ottakath, N., Almaadeed, N., and Himeur, Y. (2021). Panoptic segmentation: A review.

Fawakherji, M., Potena, C., Prevedello, I., Pretto, A., Bloisi, D. D., and Nardi, D. (2020). Data augmentation using gans for crop/weed segmentation in precision farming. In *2020 IEEE Conference on Control Technology and Applications (CCTA)*, pages 279–284.

Guo, J. and Gould, S. (2015). Deep cnn ensemble with data augmentation for object detection.

Guo, J., Qi, L., Shi, Y., and Gao, Y. (2023). PLACE dropout: A progressive layer-wise and channel-wise dropout for domain generalization. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 20(3):1–23.

Guo, W., Rage, U., and Ninomiya, S. (2013). Illumination invariant segmentation of vegetation for time series wheat images based on decision tree model. *Computers and Electronics in Agriculture*, 96:58–66.

Gupta, A., Sharma, A., and Goel, A. (2017). Review of regression analysis models. *Int. J. Eng. Res. Technol*, 6(08):58–61.

Gupta, S., Gupta, R., Ojha, M., and Singh, K. (2018). A comparative analysis of various regularization techniques to solve overfitting problem in artificial neural network. In *Data Science and Analytics: 4th International Conference on Recent Developments in Science, Engineering and Technology, REDSET 2017, Gurgaon, India, October 13-14, 2017, Revised Selected Papers 4*, pages 363–371. Springer.

Hamuda, E., Glavin, M., and Jones, E. (2016). A survey of image processing techniques for plant extraction and segmentation in the field. *Computers and Electronics in Agriculture*, 125:184–199.

Haug, S. and Ostermann, J. (2015). A crop/weed field image dataset for the evaluation of computer vision based precision agriculture tasks. In *Computer Vision - ECCV 2014 Workshops*, pages 105–116.

Hou, S. and Wang, Z. (2019). Weighted channel dropout for regularization of deep convolutional neural network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:8425–8432.

Jafari, A., Eghbali Jahromi, H., Mohtasebi, S., and Omid, M. (2006). Color segmentation scheme for classifying weeds from sugar beet using machine vision. *International Journal of Information Science and Management (IJISM)*, 4(1):1–12.

Kamalov, F. and Leung, H. H. (2020). Deep learning regularization in imbalanced data. In *2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*, pages 1–5.

Kerkech, M., Hafiane, A., and Canals, R. (2018). Deep leaning approach with colorimetric spaces and vegetation indices for vine diseases detection in uav images. *Computers and Electronics in Agriculture*, 155:237–243.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. (2017). On large-batch training for deep learning: Generalization gap and sharp minima.

Khosla, C. and Saini, B. S. (2020). Enhancing performance of deep learning models with different data augmentation techniques: A survey. In *2020 International Conference on Intelligent Engineering and Management (ICIEM)*, pages 79–85.

Kim, H.-C. and Kang, M.-J. (2020). A comparison of methods to reduce overfitting in neural networks. *International journal of advanced smart convergence*, 9(2):173–178.

Kirillov, A., He, K., Girshick, R. B., Rother, C., and Dollár, P. (2018). Panoptic segmentation. *CoRR*, abs/1801.00868.

Kitzler, F., Barta, N., Neugschwandtner, R. W., Gronauer, A., and Motsch, V. (2023). We3ds: An rgb-d image dataset for semantic segmentation in agriculture. *Sensors*, 23(5).

Ko, B., Kim, H.-G., and Choi, H. (2017). Controlled dropout: A different dropout for improving training speed on deep neural network. *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 972–977.

Lei, X., Pan, H., and Huang, X. (2019). A dilated cnn model for image classification. *IEEE Access*, 7:124087–124095.

Lim, H.-i. (2021). A study on dropout techniques to reduce overfitting in deep neural networks. In Park, J. J., Loia, V., Pan, Y., and Sung, Y., editors, *Advanced Multimedia and Ubiquitous Engineering*, pages 133–139, Singapore. Springer Singapore.

Lin, Z., Zhang, Z., Chen, L.-Z., Cheng, M.-M., and Lu, S.-P. (2020). Interactive image segmentation with first click attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Luo, Z., Yang, W., Yuan, Y., Gou, R., and Li, X. (2023). Semantic segmentation of agricultural images: A survey. *Information Processing in Agriculture*.

Majumder, S., Khurana, A., Rai, A., and Yao, A. (2020). Multi-stage fusion for one-click segmentation.

Mehdi[1], C. A., Nour-Eddine, J., and Mohamed[1], E. (2023). Check for updates regularization in cnn: A mathematical study for l1, l2 and dropout regularizers. In *International Conference on Advanced Intelligent Systems for Sustainable Development: Volume 1-Advanced Intelligent Systems on Artificial Intelligence, Software, and Data Science*, volume 637, page 442. Springer Nature.

Mesbah, Y., Ibrahim, Y. Y., and Khan, A. M. (2021). Domain generalization using ensemble learning.

Meyer, G. E. and Neto, J. C. (2008). Verification of color vegetation indices for automated crop imaging applications. *Computers and Electronics in Agriculture*, 63(2):282–293.

Milioto, A., Lottes, P., and Stachniss, C. (2018). 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)*, pages 2229–2235.

Muhammad, U., Laaksonen, J., Beddiar, D. R., and Oussalah, M. (2023). Domain generalization via ensemble stacking for face presentation attack detection.

Ortega, L. A., Cabañas, R., and Masegosa, A. (2022). Diversity and generalization in neural network ensembles. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I., editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 11720–11743. PMLR.

Park, S. and Kwak, N. (2017). Analysis on the dropout effect in convolutional neural networks. pages 189–204.

Piao, S. and Liu, J. (2019). Accuracy improvement of unet based on dilated convolution. *Journal of Physics: Conference Series*, 1345(5):052066.

Ramadan, H., Lachqar, C., and Tairi, H. (2020). A survey of recent interactive image segmentation methods. *Computational Visual Media*, 6:355 – 384.

Riehle, D., Reiser, D., and Griepentrog, H. W. (2020). Robust index-based semantic plant/background segmentation for rgb- images. *Computers and Electronics in Agriculture*, 169:1–12.

Romera-Paredes, B. and Torr, P. (2015). An embarrassingly simple approach to zero-shot learning. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2152–2161, Lille, France. PMLR.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597.

Sagi, O. and Rokach, L. (2018). Ensemble learning: A survey. *WIREs Data Mining and Knowledge Discovery*, 8(4):e1249.

Santos, J., Dantas Dias Junior, J., Backes, A., and Escarpinati, M. (2021). Segmentation of agricultural images using vegetation indices. pages 506–511.

Seong-Heon Kim, Chan-Seok Ryu, Y.-S. K. Y.-B. M. (2015). Improved Plant Image Segmentation Method using Vegetation Indices and Automatic Thresholds. *Journal of Agriculture Life Science*, 49:333–341.

Sherzodjon, Y. (2023). The problem of overfitting in machine learning and its solutions. *International Journal of Contemporary Scientific and Technical Research*, page 144–147.

Smitt, C., Halstead, M., Ahmadi, A., and McCool, C. (2022). Explicitly incorporating spatial information to recurrent networks for agriculture.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

Syakrani, A. R. M. H. P. U. P. H. N. (2022). Early stopping effectiveness for yolov4. *Universitas Airlangga*, Vol. 8 No. 1 (2022): April.

Taylor, L. and Nitschke, G. (2018). Improving deep learning with generic data augmentation. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1542–1547.

Upendar, K., Agrawal, K., Chandel, N., and Singh, K. (2021). Greenness identification using visible spectral colour indices for site specific weed management. *Plant Physiology Reports*, 26.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Vincent, O. and Folorunso, O. (2009). A descriptive algorithm for sobel image edge detection.

Vladu, M., Tudor, V. C., Mărcuță, L., Mihai, D., and Tudor, A. D. (2021). Study on the production and valorization of sugar beet in the european union. *Romanian Agricultural Research*, 38:447–455.

Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., Chen, Y., Zeng, W., and Yu, P. S. (2022). Generalizing to unseen domains: A survey on domain generalization.

Xu, N., Price, B., Cohen, S., Yang, J., and Huang, T. (2016). Deep interactive object selection.

Yang, B., Zhu, Y., and Zhou, S. (2021). Accurate wheat lodging extraction from multi-channel uav images using a lightweight network model. *Sensors (Basel, Switzerland)*, 21.

Yang, M.-D., Tseng, H.-H., Hsu, Y.-C., and Tsai, H. P. (2020). Semantic segmentation using deep learning with vegetation indices for rice lodging identification in multi-date uav visible images. *Remote Sensing*, 12:633.

Yang, S., Xiao, W., Zhang, M., Guo, S., Zhao, J., and Shen, F. (2023). Image data augmentation for deep learning: A survey.

Ying, X. (2019). An overview of overfitting and its solutions. *Journal of Physics: Conference Series*, 1168(2):022022.

Yu, Z., Cao, Z., Wu, X., Bai, X., Qin, Y., Zhuo, W., Xiao, Y., Zhang, X., and Xue, H. (2013). Automatic image-based detection technology for two critical growth stages of maize: Emergence and three-leaf stage. *Agricultural and forest meteorology*, 174:65–84.

Zheng, L., Zhang, J., and Wang, Q. (2009). Mean-shift-based color segmentation of images containing green vegetation. *Computers and Electronics in Agriculture*, 65(1):93–98.

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., and Loy, C. C. (2022). Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20.

Zhou, S., Chen, C., Han, G., and Hou, X. (2019). Deep convolutional neural network with dilated convolution using small size dataset. In *2019 Chinese Control Conference (CCC)*, pages 8568–8572.

Zimmer, P., Halstead, M., and McCool, C. (2023). Panoptic one-click segmentation: Applied to agricultural data. *IEEE Robotics and Automation Letters*, 8(5):2478–2485.

Statement of Authorship

I hereby certify that this master thesis has been composed by myself. I have not made use of the work of others or presented it here unless it is otherwise acknowledged in the text. All references and verbatim extracts have been quoted, and all sources of information have been specifically acknowledged.

Date : _____ Signature : _____