Semantic Weed Segmentation using U-Net with ResNet101

Sagnik Barik
School of Computing Science and Engineering
VIT Bhopal University
Madhya Pradesh, India
sagnikbarik2022@vitbhopal.ac.in

***Abstract*—Precision agriculture requires accurate, automated weed detection to enable targeted herbicide application and reduce environmental impact. This project implements a deep learning pipeline for the semantic segmentation of agricultural weeds using the CoFly-WeedDB aerial image dataset. A U-Net architecture with a pre-trained ResNet101 backbone was developed to classify pixels into four categories: background and three distinct weed types. The methodology involved patch-based training, extensive data augmentation, and a composite loss function combining weighted Dice and Focal Loss to address class imbalance. The model was trained and evaluated on a held-out test set, achieving a Mean Intersection over Union (mIoU) of 0.54 and a weighted average precision of 0.84. The results demonstrate the model's strong capability for accurately delineating weeds from crops, providing a robust foundation for real-world automated weed management systems.**

I. Introduction

Effective weed management is a cornerstone of modern precision agriculture, essential for safeguarding crop yields and promoting sustainable farming practices. Weeds compete directly with crops for vital resources like water, nutrients, and sunlight, leading to significant economic losses. While traditional methods rely on broad-acre herbicide application, this approach is often inefficient and carries a substantial environmental footprint. The paradigm of Site-Specific Weed Management (SSWM) has emerged as a superior alternative, advocating for targeted interventions that minimize chemical usage and operational costs. The success of SSWM, however, is fundamentally dependent on the accurate and automated detection of weeds.

Deep learning, particularly semantic segmentation, has become the state-of-the-art for agricultural image analysis. Unlike object detection, which provides bounding boxes, semantic segmentation offers a pixel-level classification, enabling the precise delineation required for targeted spraying. As demonstrated in numerous studies, architectures like U-Net have proven exceptionally effective for this task, producing detailed segmentation masks even in complex canopies. The performance of these models is further enhanced by leveraging powerful, pre-trained backbones like ResNet, which apply features learned from vast datasets to the specific domain of weed identification.

This project implements and evaluates a complete pipeline for weed segmentation using the publicly available CoFly-WeedDB aerial dataset. We employ a U-Net architecture with a pre-trained ResNet101 backbone to accurately classify each pixel into one of four categories: a background class and three distinct weed types. This combination of a proven segmentation model with a robust feature extractor is chosen to achieve high accuracy and generalizability, addressing the challenge of identifying visually similar plant species from an aerial perspective.

To build a robust model, key methodological steps were implemented, including patch generation to handle high-resolution imagery, extensive data augmentation to increase dataset variance, and a composite loss function to manage class imbalance. The following sections will describe this methodology in detail, present the training and evaluation results, and discuss the model's performance. Ultimately, this work contributes to the development of automated, data-driven systems for advancing precision agriculture.

II. Related Work

The application of computer vision in precision agriculture has rapidly evolved from traditional image processing to sophisticated deep learning methodologies. Early approaches required manual feature engineering, which proved brittle and unable to generalize across varying field conditions. The advent of deep learning, particularly Convolutional Neural Networks (CNNs), has revolutionized weed detection by enabling automatic and robust feature extraction directly from raw image data (Hasan et al., 2021). This has led to the widespread adoption of semantic segmentation, which provides the pixel-level precision necessary for tasks like targeted herbicide spraying.

Among the various deep learning architectures, encoder-decoder models have become the standard for semantic segmentation in agriculture. The U-Net architecture, originally developed for biomedical imaging, has been exceptionally successful due to its use of skip connections, which help preserve high-resolution spatial information crucial for precise localization (Ronneberger et al., 2015). Numerous comparative studies have confirmed the superior performance of U-Net and its variants, such as U-Net++, especially when working with limited or challenging UAV datasets (Fathipoor et al., 2023; Shahi et al., 2023). The effectiveness of these architectures is often amplified by using pre-trained backbones like ResNet, VGG, or EfficientNet, which leverage transfer

learning to improve model performance and reduce training time (Alirezazadeh et al., 2024).

A critical challenge in weed segmentation is the inherent class imbalance present in most agricultural datasets, where weed pixels are significantly outnumbered by crop and soil pixels. This imbalance can bias a model towards the majority classes, resulting in poor detection of the target weeds. To mitigate this, researchers have moved beyond standard cross-entropy to adopt more advanced loss functions. A survey by Jadon (2020) categorizes and reviews various loss functions, highlighting the efficacy of region-based methods like Dice Loss and distribution-based methods like Focal Loss. Combining these into a composite loss function has become a common and effective strategy to force the model to focus on the under-represented weed classes and handle hard-to-classify examples.

This project builds directly upon these established findings. By implementing a U-Net architecture with a pre-trained ResNet101 backbone, we adopt a proven model structure for agricultural segmentation. Furthermore, our use of a composite Dice and Focal Loss function follows the best practices identified in the literature for addressing class imbalance. By applying this robust pipeline to the CoFly-WeedDB, a public UAV dataset, our work aligns with the current research trajectory aimed at developing practical, high-performance models for real-world automated weed management.

III. Methodology

The project was executed through a systematic pipeline encompassing dataset acquisition, comprehensive data preparation, and a carefully configured model training and evaluation process. Each stage was designed to address the specific challenges of segmenting high-resolution aerial imagery in an agricultural context.

A. Dataset

The study utilized the CoFly-WeedDB dataset, a publicly available collection of high-resolution RGB aerial images acquired by a UAV. This dataset is specifically designed for semantic segmentation tasks in precision agriculture. It contains images of crops alongside corresponding pixel-level annotations that delineate four distinct classes: background (including soil and crop), 'Weed Type 1', 'Weed Type 2', and 'Weed Type 3'. The detailed, multi-class annotations make this dataset well-suited for developing and evaluating models capable of differentiating between various types of weeds, which is a critical requirement for selective herbicide application.

B. Data Preparation

A multi-step data preparation pipeline was implemented to transform the raw dataset into an optimal format for training the deep learning model. To address the computational constraints imposed by the high-resolution UAV imagery, a patch-based strategy was adopted. The original images and their corresponding masks were systematically partitioned into smaller, 256x256 pixel patches. To ensure the model trained on agriculturally relevant information and to begin mitigating the dataset's inherent class imbalance, a filtering step was implemented. This process discarded any patch containing less than 3

Following the initial patch generation, the training dataset was significantly enhanced to improve model robustness and prevent overfitting. Leveraging the albumentations library, a series of random geometric transformations including horizontal and vertical flips, 90-degree rotations, and grid distortions were applied to each training patch. This procedure created two new, unique samples from each original, effectively tripling the size of the training set and exposing the model to a wider variety of visual scenarios. The augmented dataset was then partitioned into training and testing sets using an 80/20 ratio, establishing a clear separation for model development and subsequent evaluation.

The final stage of preparation involved formatting the data to meet the specific input requirements of the pre-trained model and the loss function. The training and testing images were normalized using the dedicated preprocessing function associated with the ResNet101 backbone. This critical step ensures that the input data's statistical distribution matches that of the ImageNet dataset on which the backbone was originally trained, which is essential for successful transfer learning. Concurrently, the single-channel integer-based masks were converted into a one-hot encoded format. This transformation produced a four-channel binary mask for each patch, where each channel corresponds to one of the four classes, aligning the ground truth labels with the requirements of the categorical loss function used during training.

C. Model Training

The training process was centered on a U-Net architecture with a ResNet101 backbone, chosen for its proven effectiveness in segmentation tasks.

The U-Net is a fully convolutional neural network designed for precise image segmentation. Its architecture consists of a contracting path (encoder) to capture context and a symmetric expanding path (decoder) that enables precise localization. A key innovation of U-Net is the use of "skip connections," which concatenate feature maps from the encoder directly to the decoder. This allows the network to combine high-level contextual information with fine-grained spatial details, making it exceptionally effective at delineating object boundaries (Ronneberger et al., 2015).

The ResNet101 (Residual Network) is a state-of-the-art deep CNN used as the encoder, or backbone, for our U-Net. ResNet architectures introduced "residual blocks" that allow the network to learn an identity function, which effectively mitigates the vanishing gradient problem that plagues very deep networks. This enables the training of models with over 100 layers, resulting in highly powerful feature extraction capabilities.

This specific combination was chosen to leverage the benefits of transfer learning. By using a ResNet101 pre-trained on

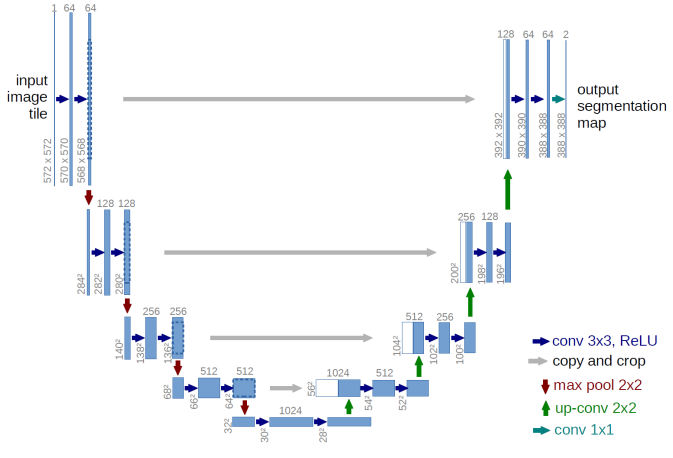

Fig. 1. U-Net architecture (Ronneberger et al., 2015).

the massive ImageNet dataset, the model's encoder is already a proficient feature extractor. This powerful encoder is then combined with the U-Net's superior localization structure, creating a model that excels at both identifying what is in the image and pinpointing where it is. The ResNet backbone was unfrozen to allow its weights to be fine-tuned on the weed dataset, adapting its learned features for this specific agricultural task.

The model was compiled using a composite loss function that combines a class-weighted Dice Loss with Categorical Focal Loss. This hybrid loss is particularly effective for imbalanced datasets, as Dice Loss optimizes for region overlap while Focal Loss focuses the model on hard-to-classify examples (Jadon, 2020). The AdamW optimizer was used with an initial learning rate of 0.0001. Training was conducted for up to 50 epochs with a batch size of 8. Callbacks such as ModelCheckpoint and EarlyStopping were employed to save the best-performing model and prevent overfitting.

IV. Results

The performance of the trained U-Net model with a ResNet101 backbone was quantitatively evaluated on the held-out test set, which comprised 20% of the patched dataset and was not used during training or augmentation. The evaluation focused on standard semantic segmentation metrics, including per-class and overall Intersection over Union (IoU) and F1-Score.

A. Quantitative Evaluation

The model effectively segmented the aerial imagery into four classes. The Mean Intersection over Union (mIoU) metric, a primary semantic segmentation metric, was 0.54, indicating a 54% overlap between the model's predictions and ground truth labels. The weighted average F1-Score was 0.84, reflecting a good balance between precision and recall. Table 1 provides a detailed breakdown of the model's performance for each class. The model excelled in identifying the 'Background' class, achieving an IoU of 0.8212 and an F1-Score of 0.90, as expected given its visual distinctiveness and prevalence in the dataset.

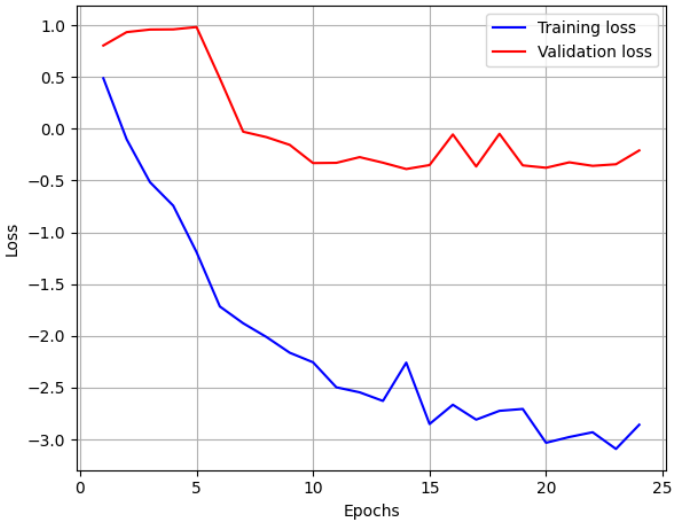

Fig. 2. IoU vs. Step plot.

Performance on the individual weed classes varied. 'Weed Type 1' was the most successfully identified weed, with an IoU of 0.5306 and an F1-Score of 0.69. 'Weed Type 3' also showed reasonable performance with an IoU of 0.4402. The model found 'Weed Type 2' to be the most challenging to segment, resulting in a lower IoU of 0.2255. This variation in performance is likely attributable to the differing visual characteristics of the weeds and their representation in the training dataset.

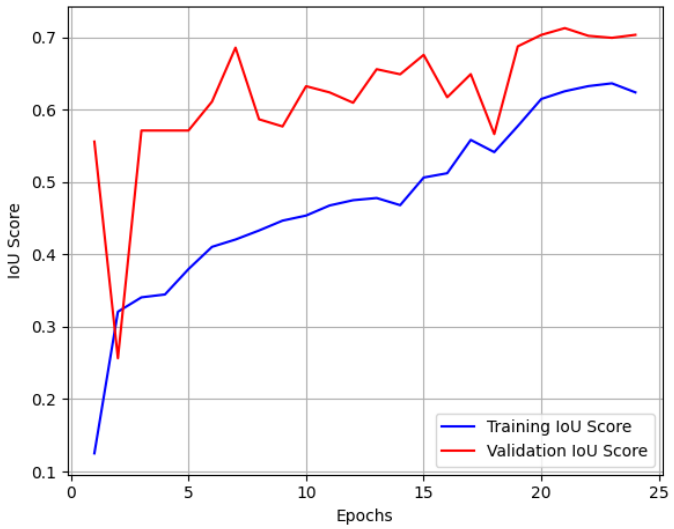

Fig. 3. Learning Rate vs. Step plot.

B. Visual Results

Visual inspection of the model's predictions on test images confirms the quantitative findings. The segmentation masks generated by the model generally align well with the ground

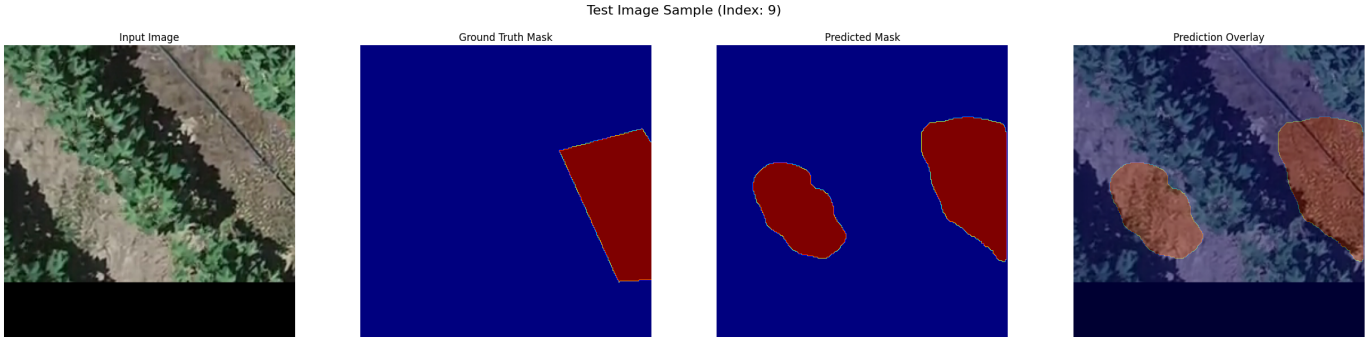

Fig. 4. Sample Interaction with the Model

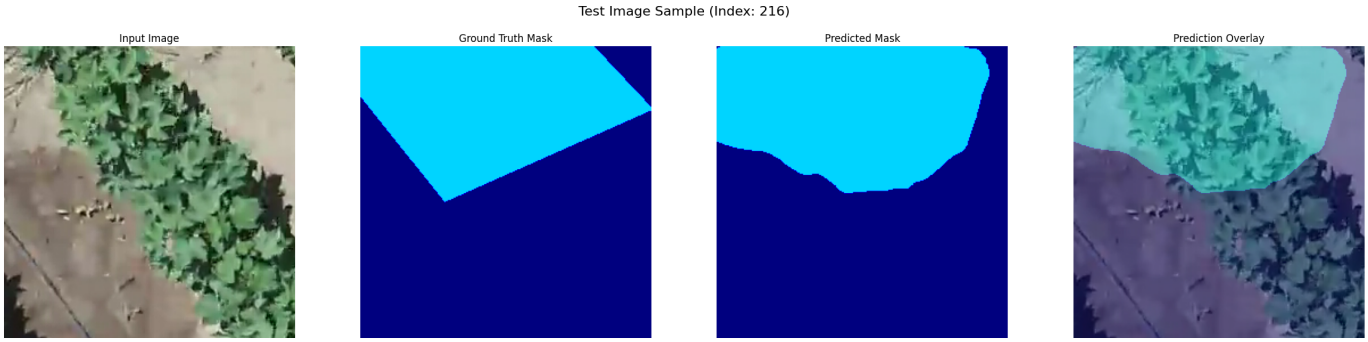

Fig. 5. Sample Interaction with the Model

TABLE I
PER-CLASS PERFORMANCE METRICS ON THE TEST SET

Class Name	Performance Metrics			
	IoU	Precision	Recall	F1-Score
Background	0.8212	0.93	0.87	0.90
Weed Type 1	0.5306	0.66	0.73	0.69
Weed Type 2	0.2255	0.32	0.43	0.37
Weed Type 3	0.4402	0.54	0.70	0.61
Macro Avg.	**0.5044**	**0.61**	**0.68**	**0.64**
Weighted Avg.	—	**0.86**	**0.84**	**0.85**

truth, accurately delineating large, contiguous areas of background and weeds. The model is particularly effective at identifying the dominant weed types within an image. However, the visual results also highlight areas for potential improvement. The model sometimes struggles with accurately segmenting fine details, such as thin weed foliage or instances where different weed types are closely intertwined. These challenging scenarios are likely the primary contributors to the lower IoU scores observed for the less-represented weed classes. Overall, the visual inference demonstrates that the model has successfully learned the key visual features necessary to distinguish between weeds and background from an aerial perspective.

CONCLUSION

This project successfully developed and evaluated a deep learning pipeline for weed segmentation, applying a U-Net architecture with a ResNet101 backbone to the CoFly-WeedDB aerial dataset. The model achieved a Mean Intersection over Union of 0.5044 and a weighter average F1-score of 0.85, demonstrating a robust capability to generate precise weed maps. The results indicate strong performance in identifying the background class, with varied accuracy across the three distinct weed types, which is likely attributable to differences in their visual complexity and representation within the training data.

Future work should focus on enhancing performance for the more challenging weed classes, potentially through targeted data augmentation or the integration of attention mechanisms. Validating the model across more diverse agricultural datasets and exploring the benefits of multi-spectral imagery would further strengthen its real-world applicability.

In summary, the implemented pipeline provides an effective solution for automated weed detection. By enabling the creation of accurate, pixel-level maps, this work is a critical step toward advancing Site-Specific Weed Management, supporting the deployment of more sustainable and efficient agricultural technologies like robotic weeding and precision spraying.