UNET-SA: A Spatial Attention Enhanced UNET for Crop–Weed Segmentation in UAV-Based Pigeon Pea Fields

Original Scientific Paper

Vaibhav Dhore*
Veermata Jijabai Technological Institute
Department of Computer Engineering & Information Technology
Mumbai, Maharashtra, India
vddhore@ce.vjti.ac.in

Seema Shrawne
Veermata Jijabai Technological Institute
Department of Computer Engineering & Information Technology
Mumbai, Maharashtra, India
scshrawne@ce.vjti.ac.in

Mohan Khedkar
Veermata Jijabai Technological Institute
Department of Computer Engineering & Information Technology
Mumbai, Maharashtra, India
mskhedkar_p22@ce.vjti.ac.in

Vijay Sambhe
Veermata Jijabai Technological Institute,
Department of Computer Engineering & Information Technology
Mumbai, Maharashtra, India
vksambhe@it.vjti.ac.in

*Corresponding author

Abstract *– As the global population continues to expand and the effects of climate change become increasingly evident, the demand for sustainable agricultural practices has grown more urgent. A persistent challenge in crop cultivation lies in the intense competition between crops and weeds for essential resources such as water, nutrients, and sunlight—often leading to substantial yield losses. Conventional approaches that rely heavily on herbicides and pesticides, while effective in the short term, can degrade soil health and harm the surrounding ecosystem. Hence, developing environmentally friendly and efficient weed management strategies has become a priority in precision agriculture. In this study, we introduce UNET-SA, an improved semantic segmentation framework that integrates a spatial attention mechanism into the traditional UNet architecture. The addition of spatial attention enables the model to better identify small or scattered weeds by concentrating computational focus on key regions within the image—areas that standard segmentation networks often overlook. The proposed model was trained and evaluated using a dataset of 1,727 annotated images collected from pigeon pea fields in the Vidarbha region of India. To correct manual annotation inconsistencies, HSV color space transformation was applied during preprocessing. Experimental findings demonstrate that UNET-SA delivers notable performance gains over the baseline UNet, achieving a mean Intersection over Union (IoU) of 94.44% and an overall accuracy of 98.64%, reflecting improvements of +1.74% and +1.04%, respectively. Additional testing on the larger CropAndWeed dataset further validated the model's generalization capability, where UNET-SA achieved 98.81% accuracy and a 55.79% mean IoU, outperforming the baseline UNet (98.49% accuracy, 51.81% mean IoU). The disparity between high accuracy and moderate IoU highlights the impact of class imbalance—large background regions can inflate accuracy without reflecting true segmentation precision. Consequently, mean IoU serves as a more reliable indicator of model effectiveness. Overall, UNET-SA surpasses leading architectures such as DeepLabv3+, SegFormer, PSPNet, and LinkNet, demonstrating strong potential for practical, long-term deployment in crop–weed segmentation tasks under real agricultural conditions.*

Keywords*: Crop and Weed Detection, Pigeon Pea, Spatial Attention, Deep Learning, UNET*

Received: October 6, 2025; Received in revised form: November 28, 2025; Accepted: November 28, 2025

1. INTRODUCTION

By 2050, the world's population is expected to reach nine billion, which means that food production will need to increase by about 70% to meet human needs [1]. But this growth is in danger because there is less arable land, water is becoming scarcer, and climate change is hurting crop yields and soil health. To make sure that there is enough food for everyone and to protect the environment, it is important to sustainably increase agricultural productivity in the face of rising global demand. Precision agriculture is a promising so-

lution that uses advanced sensing, automation, and artificial intelligence (AI) technologies to make better use of resources and manage crops more efficiently [1-3].

Weeds are still one of the biggest problems that keep us from being more productive. Weeds compete with crops for important resources like nutrients, water, sunlight, and space, which lowers both the yield and the quality [4, 5]. Preventive, cultural, mechanical, biological, and chemical weed control methods [6] are often expensive, time-consuming, and bad for the environment. Overuse of herbicides, especially, harms the soil, makes weeds resistant to chemicals, and puts people and animals at risk [6, 7].

The need for weed management that is both sustainable and accurate has led to a lot of interest in systems that can automatically find and kill weeds. These kinds of systems use computer vision and machine learning to find, sort, and pinpoint weeds at the plant level. These technologies cut down on the use of herbicides and labor costs while also protecting the environment by allowing for selective spraying or mechanical removal [8, 9]. However, it is still hard to tell the difference between crops and weeds in natural field conditions because of things like changes in light, blockage, and the fact that crops and weeds look very similar in color, texture, and shape.

Deep learning-based semantic segmentation accurately identifies plants, allowing for targeted actions in weed control and crop management. Fully Convolutional Networks (FCNs), UNet, DeepLab, and SegFormer are examples of architectures that have shown promise in segmenting agricultural images. But new ways to add data or transfer learning methods can help reduce the need for large, well-annotated datasets in agricultural image segmentation. For example, there is no public dataset for pigeon pea (Cajanus cajan), which is a major legume crop in India and other tropical areas. To fill this gap, this study provides a high-resolution UAV-based pigeon pea dataset with three annotated semantic classes: crop, weed, and background. The data was collected in real field conditions. The dataset was obtained from three separate pigeon pea fields utilizing both UAV and handheld camera systems over a span of six consecutive days to guarantee variability in lighting, soil background, and stages of plant growth.

Another big problem is manual annotation, which takes a long time and is easy to make mistakes, which makes it harder for the model to generalize. To fix this, we use HSV (Hue–Saturation–Value) color space transformation to help make the annotations more accurate. This preprocessing method makes the color contrast between areas of vegetation and soil stronger, which lowers the chance of boundary errors and makes the labels more consistent.

This study presents the UNET-SA model to enhance segmentation accuracy by utilizing a spatial attention mechanism that emphasizes crucial spatial areas and sharpens segmentation boundaries within the UNet architecture. The original UNet does a good job of capturing multiscale contextual features, but it doesn't always do a good job of highlighting spatially important areas in complicated field scenes. With spatial attention, the network can focus on important spatial features, like the fine lines that separate crop and weed areas, while ignoring background noise. This results in enhanced feature representation and segmentation accuracy, especially in diverse agricultural settings. Additionally, performing comparative experiments on various agricultural datasets or integrating user feedback for model enhancement can further substantiate the efficacy of the UNET-SA model. Performance is evaluated based on accuracy, precision, recall, F1-score, and Intersection over Union (IoU), demonstrating the effectiveness of the spatial attention mechanism in enhancing discriminative ability and minimizing false classifications. In summary, the major contributions of this work are as follows:

- Development of a high-resolution UAV-based pigeon pea dataset with three annotated classes—crop, weed, and background—captured under realistic field conditions.
- Enhancement of annotation quality using HSV color space transformation to minimize manual labeling errors and improve dataset consistency.
- Proposal of the UNET-SA model, an enhanced UNet architecture that integrates a spatial attention mechanism to focus on relevant spatial regions and suppress irrelevant background information.
- Comprehensive comparative analysis of the proposed model against baseline UNet and other state-of-the-art deep segmentation networks to validate performance improvements.

The remainder of this paper is structured as follows. Section 2 reviews related work in the field of crop–weed detection and attention-based segmentation models. Section 3 presents the dataset, preprocessing strategies, and proposed methodology. Section 4 discusses the experimental setup and performance analysis. Finally, Section 5 concludes the study and outlines directions for future research.

2. LITERATURE REVIEW

There are a few publicly available datasets for crop–weed segmentation, like the Sugar Beet dataset from Bonn University [10], the Crop and Weed Field Image Dataset (CWFID) [11], and the UAV-based CWFID dataset [12]. However, most of them are only concerned with crops like sugar beet, spinach, maize, and beans. These datasets are useful, but they don't have a lot of different crops, growth stages, or geographic contexts. Pigeon pea (Cajanus cajan) is a major legume crop grown a lot in tropical and subtropical areas, but it is still not well represented in open-access collections, even though it is very important for farming, especially in India and sub-Saharan Africa. Its diverse canopy structure, intercropping arrangements, and fluctuating weed density pose unique obstacles for

semantic segmentation and model generalization. To fill this gap in research, we created a high-resolution UAV-based pigeon pea dataset with 1,727 annotated images taken in natural light from different fields in the Vidarbha region of India. Each image has pixel-level labels for three groups: crop, weed, and background. This makes them a useful standard for testing deep learning models in real-world farming situations. This dataset is meant to help researchers study how to tell the difference between crops and weeds, how to use transfer learning, and how to automate precision weeding, especially for legume species that don't get a lot of attention. For strong segmentation models in UAV imagery, high-quality annotation is a must because changes in lighting, soil background, and vegetation that is very close together make labels less certain. Color-space preprocessing, especially HSV transformation, is often used to make it easier to separate vegetation and soil and to create consistent initial masks before manual refinement. When combined with morphological filtering and annotator verification, HSV-based pipelines lower the amount of variation between annotators and speed up the process of curating large orthomosaics [13, 14]. Recent studies have integrated color indices (such as ExG and NDVI, where multispectral sensors are accessible) with superpixel and clustering methodologies (for instance, LAB-ab K-means) to generate dependable pseudo-masks that annotators subsequently refine, enhancing overall annotation quality and facilitating improved model generalization. [15, 17].

The U-Net family remains the preeminent framework for pixel-wise segmentation in agricultural applications, owing to its encoder-decoder architecture and skip connections that maintain intricate spatial details [18], [19]. But baseline U-Net can have trouble with long-range context and fine boundary delineation when crops and weeds have similar spectral signatures or when there are small weeds. Two distinct trends have manifested in research from 2023 to 2025:

(1) Adding attention and multi-scale context modules (SE, ECA, CBAM, ASPP, MGA) to U-Net variants has always improved mean IoU and boundary F1. Some examples are Coordinate Attention UNet, U-MGA, MSF-CA-Net, and Dilated Multi-Scale Attention UNet [20-23].

(2) Hybrid and transformer-based encoder designs, like Visual Mamba UNet, SSMR-Net, and Dual-Task Enhanced UNet, get the big picture while keeping the small details, which makes them work better on hard UAV datasets [24-27].

Additionally, lightweight U-Net variants and orthomosaic-aware pipelines have been suggested to preserve real-time inference capabilities on embedded hardware like NVIDIA Jetson and Xavier platforms, guaranteeing their appropriateness for precision agriculture applications [19, 25].

Spatial attention, which explicitly models the "where" to look, has been very useful for UAV agricultural imagery, where standard convolutions often get confused by background noise and changes in lighting. Recent agricultural segmentation studies [25, 26, 28-31] have shown that spatial attention modules (either on their own or as part of hybrid attention like CBAM) make it easier to find small objects and separate classes in dense canopies. Based on these ideas, the proposed UNET-SA adds an efficient spatial attention (SA) block to the U-Net decoder to selectively boost crop/weed discriminative regions while lowering irrelevant background activations. This design uses HSV-assisted annotation for cleaner supervision and keeps the model small for practical UAV/edge deployment.

Finally, strict benchmarking is necessary to show real progress. Modern comparative studies usually use mIoU, mPA (mean pixel accuracy), boundary F1, inference speed (FPS or ms/image), and parameter/FLOP budgets to compare models on UAV and field datasets. To validate UNET-SA, we employ the same methodology and conduct comparisons against baseline U-Net, UNet++, DeepLabv3+, SegFormer, Swin-UNet, and specific lightweight U-Net variants on the new pigeon pea dataset and on standard public datasets where relevant. This comparative evaluation measures the benefits of spatial attention and high-quality HSV-enhanced annotations, situating UNET-SA within the latest developments in attention-driven crop–weed segmentation.

3. MATERIALS AND METHODS

3.1. DATASET DESCRIPTION

The Pigeon Pea dataset utilized in this research comprises 1,727 RGB images of crops and weeds, gathered from three separate agricultural fields in the Vidarbha region of India, encompassing areas of 1.62 ha, 1.24 ha, and 1.46 ha, respectively. We surveyed each field on two consecutive days to make sure that the light, weed density, and crop growth stages were all different. The dataset includes pigeon pea plants at different stages of their life cycle, which gives it a lot of variety in terms of appearance and canopy coverage.

We took pictures with both a smartphone camera (1,634 pictures) and a DJI Mavic Air 2S drone (93 pictures). The drone took pictures from 20 cm to 1 m above the ground, and the handheld pictures were taken from 10 to 30 cm above the ground. GPS coordinates were used to mark the edges of the fields, and there was always at least a 2 m gap between each frame to make sure that no one plant instance was recorded more than once. This gets rid of redundancy and makes sure that each session has its own sample.

To keep data from leaking, the dataset was split up by field. Seventy percent of the images were used for training, ten percent for validation, and twenty percent for testing. This made sure that each subset was fairly represented. This strategy makes sure that images that are close in time or space don't show up in more than one subset, which makes the generalization test fair.

We used polygonal masks to manually label each image into three groups: crop, weed, and background. The average class distribution was 8.68% background, 7.39% crop, and 83.93% weed. This is what you would expect to see in a real field with a moderate imbalance. The dataset has enough variation to test how well the proposed segmentation models work in different natural situations. Table 1 shows a summary of the day-by-day image collection details, and Fig. 1 shows some sample images of crops and weeds.

Fig. 1. Crop and Weed Images

Table 1. Dataset Description

Date	Capturing Device	Resolution	Number of Images	Training	Validation	Testing
16th July 2023	Mobile Camera	2016×4480 and 4480×2016	137	96	14	27
17th July 2023	Mobile Camera	2016×4480 and 4480×2016	422	295	43	84
18th July 2023	Mobile Camera	2016×4480 and 4480×2016	440	308	44	88
19th July 2023	Drone	5472×3648	93	65	9	19
20th July 2023	Mobile Camera	2016×4480 and 4480×2016	256	179	26	51
22nd July 2023	Mobile Camera	2016×4480 and 4480×2016	379	265	28	76
Total			**1727**	**1208**	**174**	**345**

3.2. DATASET PREPROCESSING

The raw images have different resolutions, such as 2016x4480, 4480x2016, and 5472x3648. Noise in images is when pixel values change randomly, which makes the image look worse. It can happen when an image is taken or sent. Noise in pictures can make edges, textures, and object boundaries look different, which makes it harder to find useful patterns. Deep learning models can get more useful information from clean images. When using deep learning to analyze images, steps like noise reduction, normalization, and resizing are very important for getting reliable and consistent results. Below are the steps taken to prepare the samples that were collected.

- Noise removal: An important way to process images that can be used alone or with other methods. There are many ways to get rid of noise in an image. One way to get rid of noise is to find it with other information and then use the best filtering algorithms that don't hurt the picture quality and make it smoother for analysis. This work used a Gaussian filter to get rid of noise.
- Image normalization: Image normalization changes the range of pixel intensities, which speeds up execution. There is one channel and 0–255 pixels of intensity in grayscale images. Normalization changes intensity from 0 to 1. Normalization makes the intensity go from 0 to 1. There are three channels in an RGB image, and the pixel intensities range from 0 to 255. This changes the range of pixel intensities for all three channels from 0 to 1.
- Changing the size of an image: A mobile camera and a drone took pictures at 2016 x 4480, 4480 x 2016, and 5472 x 3648 pixels. Before this stage, all of the pictures are made smaller to 640 by 640.

3.3. DATASET ANNOTATION AND MASK GENERATION

Deep learning architecture uses supervised learning for training, where a labeled dataset is needed to learn the probability distribution. The images after preprocessing are required to be labeled at the pixel level, where each pixel will be classified into one of the three categories (background, crops, and weeds). For the task of annotation, we have used Roboflow. The dataset is uploaded to a server, and images are annotated using the smart polygon tool. After annotation, the image annotation file is downloaded in YOLOv8 format for further use.

As the Roboflow online platform provides many tools for annotation, it is selected for the annotation. Among

different tools of annotation, the polygon tool is selected due to the shape of size of the weed. Each image is annotated and contains three classes, i.e., Background, Crop, and Weed. After the completion of annotation of all the images, the annotation file is downloaded in JSON format. But semantic segmentation models require mask images for training. The JSON file is read, and different mask images are generated. The steps followed for the generation of mask files are given in Algorithm 1.

Algorithm 1. Generation of Mask images from JSON Annotations file

Input: JSON annotation file, Image dimensions

Output: Mask images corresponding to annotated images

1. Load the JSON annotation file.
2. For each annotated image in the JSON file:
   2.1. Extract the image filename and dimensions.

2.2. Create a blank mask image with all pixel values set to 0 (background).

2.3. For each annotated object in the image:

a. Extract polygon coordinates and class label.

b. Assign a unique integer value to the class (e.g., 0: Background, 1: Crop, 2: Weed).

c. Draw the polygon on the mask using the assigned class value.

2.4. Save the generated mask image with the same filename as the original image.

3.4. MASK IMPROVEMENT USING HSV COLORSPACE TRANSFORMATION

Algorithm 2 offers a methodical strategy for enhancing manually annotated segmentation masks through the utilization of the Hue–Saturation–Value (HSV) color space. The main purpose of this algorithm is to make the annotated masks more accurate by highlighting a range of green color components. These usually stand for plants like crops and green weeds in pictures of farms. During manual annotation, small weeds or areas that are only slightly green are often incorrectly labeled as background. This automated refinement process is needed to fix this.

The algorithm starts by reading the RGB image and the mask that was manually marked up. The OpenCV function cv2.cvtColor then changes the RGB image into the HSV color space. The HSV representation is better because it separates color information (hue) from brightness (value), which makes it more stable when the lighting changes or when there are shadows, which is common in field conditions.

Then, using the predefined green color range with lower bounds (30, 40, 40) and upper bounds (90, 255, 255), the HSV image is thresholded to make a binary mask. The Mgreen mask separates the areas of vegetation that correspond to crops and weeds. It does a great job of capturing the different shades of green that can be found in natural agricultural scenes.

Then, a copy of the green mask is set up as the better mask. Setting the pixel values of non-green areas that were wrongly marked as vegetation in the manual mask to background (0) fixes the problem. On the other hand, areas marked as green are kept or improved based on the original annotated mask, making sure that both crop and weed areas are still shown correctly. This step also helps bring back small or subtle weed patches that were previously thought to be background.

Finally, the improved mask output saves the refined mask, which makes it easier to tell where vegetation ends and starts and cuts down on mistakes in the annotations. This improvement makes the training data cleaner and more accurate, which makes the segmentation model work better during training and evaluation.

Algorithm 2. Mask Improvement Using HSV Color Space Transformation

Input: RGB image I_rgb, manually annotated mask M_annotated, green color thresholds: lower green (30, 40, 40), upper green (90, 255, 255)

Output: Improved mask M_improved

1. Read the input RGB image I_rgb and manually annotated mask M_annotated.
2. Transform the RGB image to HSV color space: cv2.cvtColor(I_rgb, cv2.COLOR_BGR2HSV)
3. Create a binary mask for the green region: cv2.inRange (I_hsv, lower_green, upper_green)
4. Initialize M_improved as a copy of M_green.
5. Modify the background in M_improved:
   M_improved[(M_green == 0) & (M_annotated != 0)] = 0
6. Preserve non-background regions from M_annotated in M_improved:
   M_improved[(M_green != 0)] = M_annotated[(M_green != 0)]
7. Save the improved mask.

3.3. UNET ARCHITECTURE

Deep learning architecture especially intended for semantic segmentation problems is the U-NET model, particularly in biomedical image segmentation, but it has proven effective in many domains, including agricultural applications like crop and weed detection. For semantic segmentation, the UNET model is trained to distinguish between different classes in images, such as crops, weeds, and background. The architecture consists of two primary parts: the encoder and the decoder. The encoder takes the input image, stores its context, and then uses that information to extract features by progressively reducing the spatial dimensions, which helps the model learn higher-level representations. In contrast, the decoder restores the spatial resolution by up-sampling the feature maps, ultimately generating pixel-wise class predictions for the input image. The UNET architecture features skip connections that link corresponding layers between the encoder and decoder components. The implementation of skip connections allows the model to preserve detailed spatial information from preceding layers, which is essential for accurately segmenting objects at a pixel level. For crop-weed semantic segmentation, this means the model can precisely differentiate between crops, weeds, and the background, even when these objects are nearby or share similar textures. The architecture of basic UNET is provided in Fig. 2.

3.3 SPATIAL ATTENTION MECHANISM

The Spatial Attention (SA) module focuses on the most important spatial regions in the feature maps. It generates an attention map that highlights important

spatial regions by suppressing irrelevant or background information. This is accomplished by applying both average pooling and Max Pooling along the channel axis. Both the pooling operations reduce the size of input feature maps into two 2-D maps, which detect complementary contextual information. Max pooling detects the most important regions, and average pooling focuses on the overall distribution of feature maps across the channels. The output of two pooling operations is concatenated. Then the concatenated map is given input to a convolutional layer having a kernel size of 7x7 to combine pooled information and find spatial dependencies. Next, the output of the convolution is passed through the sigmoid function, which produces values in the range of 0 to 1, which is the spatial attention mask. The value 1 represents more important regions, and the value 0 corresponds to less important regions.

Lastly, the spatial attention mask is multiplied by input feature maps, which helps to highlight more important spatial regions and to reduce the effect of less important spatial regions. In this way, spatial attention improves the feature representation and increases the model's accuracy of predictions. The SA module is represented in Fig. 3.

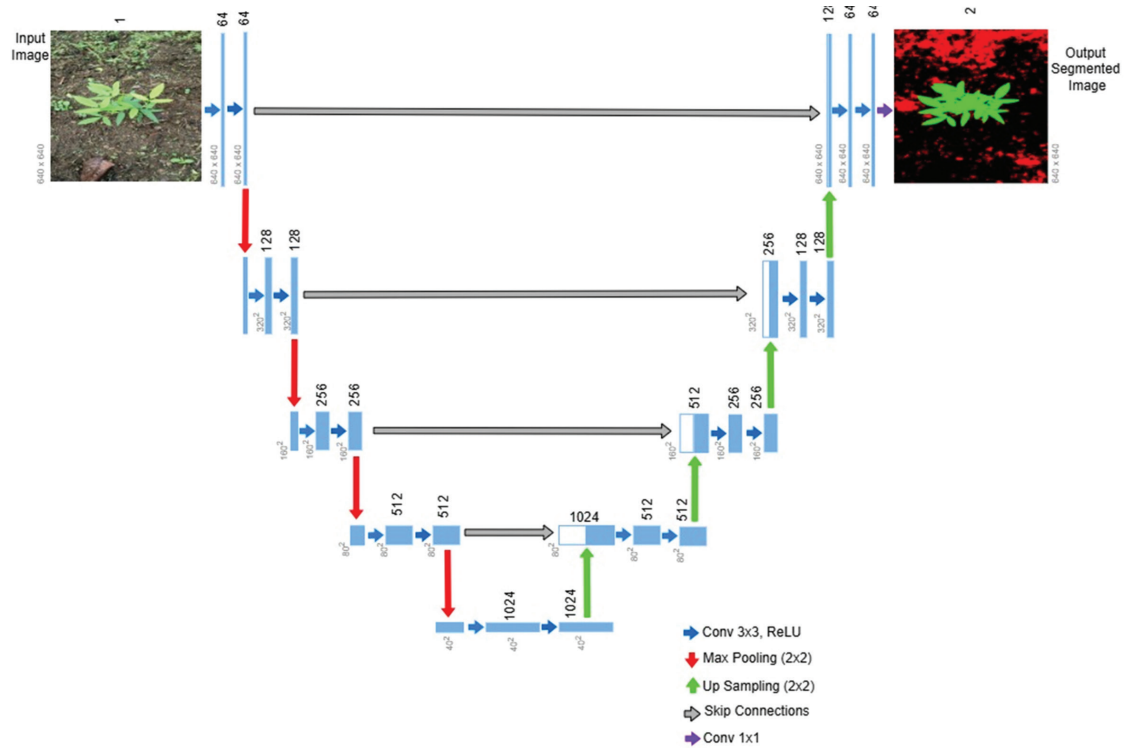


Fig. 2. Basic UNET Architecture

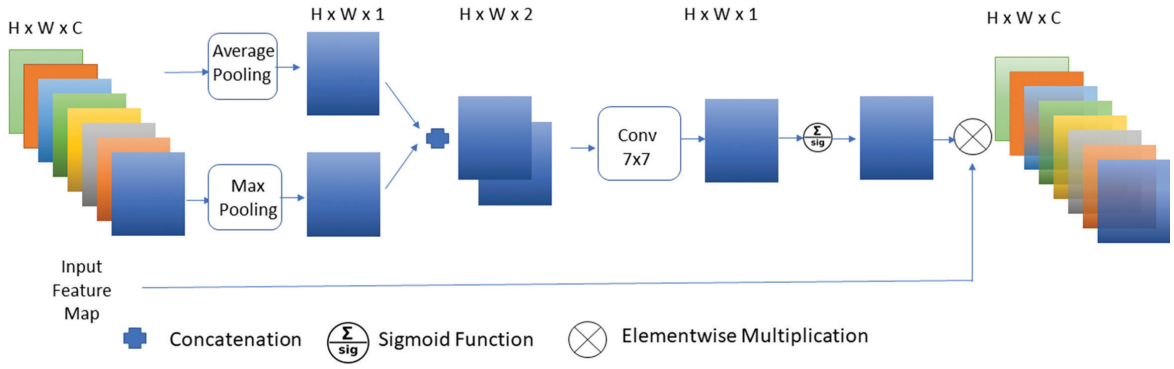


Fig. 3. Architecture of Spatial Attention Module

3.4. PROPOSED METHODOLOGY

In this section, we discuss the proposed methodology named UNET with Spatial Attention (UNET-SA). The UNET-SA is an enhanced version of the traditional UNET architecture, incorporating spatial attention mechanisms to improve semantic segmentation performance. The steps in the proposed methodology is given the Fig. 4. The images are acquired from the fields using a drone and a smartphone camera. Then the images are annotated using the polygon tool of the Roboflow platform. Then the JSON annotations are ex-

tracted. As UNET requires mask images, they are generated using the JSON annotations. For mask generation, algorithm 1 is used.

The generated mask has human annotation errors. To refine the mask further, the image is converted to HSV colorspace. Then the green part(usual crop and weed) is extracted from the image using a range of green color(lower green (30, 40, 40), upper green (90, 255, 255) in HSV color space. Using the earlier-generated mask and image with only the green part, a new refined mask is generated. The detailed steps are provided in Algorithm II.

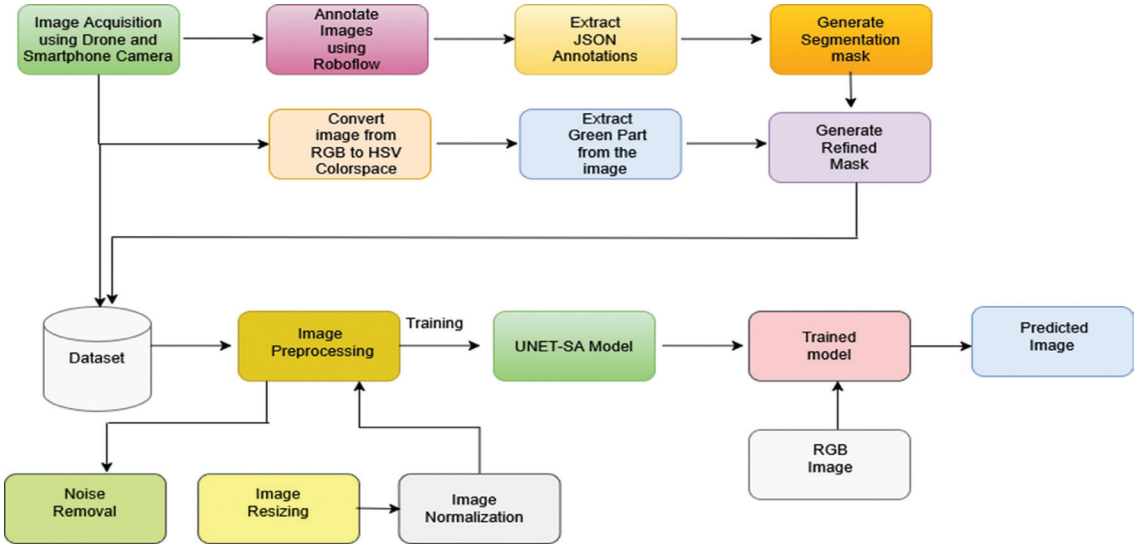


Fig. 4. Proposed methodology

In Fig. 5, it can be observed that the annotation of images is improved after the HSV color transformation. The first image is the sample input image, the second image is a mask generated using Algorithm I, and the third image is a mask generated using Algorithm II. It can be observed that Fig. 5(c) is more refined and accurate than Fig. 5(b) in terms of object detection and segmentation.

(a) (b) (c)

Fig. 5. Mask improvement, (a) Input Image, (b) Mask generated using JSON file after applying Algorithm-I, (c) Improved mask generated after applying Algorithm-II

He UNET model follows the encoder-decoder structure with skip connections, making it well-suited for multi-class segmentation. It has an encoder containing four convolutional blocks, each followed by a max-pooling layer to progressively reduce spatial dimensions while increasing feature richness. The spatial attention modules are applied after each encoder block, helping the network focus on discriminative regions in the image, which is particularly useful for complex segmentation tasks such as distinguishing between crops and weeds in dense field environments. The bridge layer at the bottleneck stage also incorporates a spatial attention module, ensuring that the most critical high-level features are retained before upsampling begins. The decoder progressively reconstructs the spatial resolution using transposed convolutions, with skip connections reintroducing fine-grained details from the encoder. Since the encoder's feature maps are already enhanced through spatial attention, the skip connections further improve boundary precision and segmentation accuracy.

A. UNET-SA model

The SA module helps the model focus on relevant regions, making it particularly effective in agricultural segmentation tasks where objects may be overlapping, occluded, or highly similar in appearance. This results in better object delineation, improved segmentation accuracy, and robustness to background noise. By integrating spatial attention within the UNET framework, UNET-SA achieves superior performance in tasks like automated weed detection and crop segmentation, enhancing precision agriculture applications.

4. RESULTS AND DISCUSSION

For all the experiments performed the hyperparameters used are learning rate is 0.001, epochs are 100, optimizer adam, loss function is cross entropy, batch size is 2, and image height and width is 640.

4.1. PERFORMANCE EVALUATION METRICS

To evaluate the performance of the proposed model, testing is required. Various metrics are used to evalu-

ate the model. The selection of evaluation metrics is dependent on the specific criteria. If the dataset has a uniform distribution of samples among all the classes, then accuracy is used for evaluation. For a non-uniform distribution of samples among all the classes, recall, precision, and F1-score are useful. Precision and recall are used to detect the false positive rate and false negative rate, respectively. To balance both the rates, the harmonic mean is calculated, and it is known as the F1-score. Intersection over union (IoU) and mean IoU are used in the case of segmentation tasks. IoU measures how much the prediction matches the ground truth mask. It is calculated per-class. Mean IoU is nothing but the mean of IoU of all the classes.

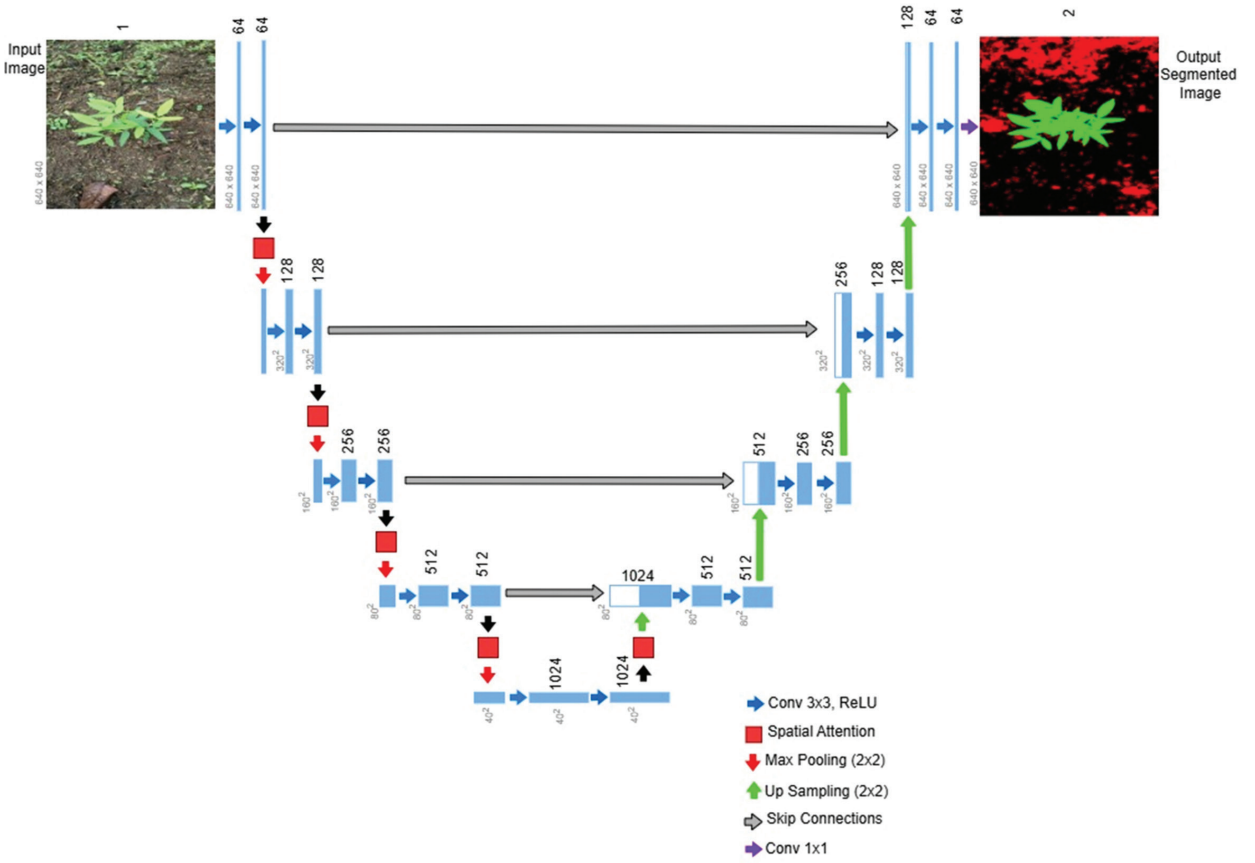


Fig. 6. UNET architecture with Spatial Attention module integrated into the encoder and bridge layer

4.2. EXPERIMENTAL RESULTS OF SOTA MODELS WITHOUT MASK ENHANCEMENT

The images collected using a drone and mobile camera are annotated using Robolflow. Additionally, the dataset is used to select and train various SOTA models used for semantic segmentation. In terms of accuracy, precision, recall, F1-score, and mean IoU, the UNET model outperforms other state-of-the-art models, demonstrating its effectiveness in semantic segmentation tasks. It attains an accuracy of 88.26%, a precision of 88.29%, a recall of 88.26%, an F1-score of 88.23%, and a mean IoU of 76.35. It can be observed from Table 4 that UNET performs better for all the performance metrics.

The per-class performance of all the SOTA models is analyzed, which is given in Table 5. It can be observed that UNET results are much better than those of other models. The highest F1-score, IoU, is 89.09 %, 82.46 % respectively, for the crop class, which is attained by the UNET. For the weed class, the highest F1-score and IoU are attained by the UNet++ model, which are 81.56% and 71.28%, respectively. For other performance metrics, the values attained by the UNET model are much closer to the highest value. The UNet++ is the variant of UNET. After going through all the results, it can be inferred that the UNET and its variant UNET perform better for most of the performance metrics.

Table 2. Overall Performance of SOTA models without mask enhancement

Model Name	Accuracy	Precision	Recall	F1 Score	Mean IoU
Unet	0.8826	0.8829	0.8826	0.8823	0.7635
Unet++	0.8804	0.8809	0.8804	0.8804	0.7603
MAnet	0.8793	0.8794	0.8793	0.8792	0.7582
Linknet	0.8765	0.8781	0.8765	0.8767	0.7466
FPN	0.8760	0.8788	0.8760	0.8763	0.7556
PSPNet	0.8754	0.8777	0.8754	0.8757	0.7460
PAN	0.8185	0.8249	0.8185	0.8156	0.6088
DeepLabV3	0.8736	0.8760	0.8736	0.8738	0.7455
DeepLabV3+	0.8798	0.8802	0.8798	0.8799	0.7613
UPerNet	0.8745	0.8775	0.8745	0.8749	0.7488

Table 3. Per-class Performance of SOTA models without mask enhancement

Model Name	Class	Accuracy	Precision	Recall	F1-Score	IoU
Unet	Background	0.8871	0.8510	0.8871	0.8597	0.7690
	Crop	0.8909	0.9045	0.8909	0.8909	0.8246
	Weed	0.7692	0.8619	0.7692	0.8027	0.6967
Unet++	Background	0.8408	0.8761	0.8408	0.8481	0.7539
	Crop	0.8928	0.8844	0.8928	0.8825	0.8142
	Weed	0.8202	0.8286	0.8202	0.8156	0.7128
MAnet	Background	0.8761	0.8539	0.8761	0.8548	0.7623
	Crop	0.8622	0.9171	0.8622	0.8786	0.8079
	Weed	0.7996	0.8390	0.7996	0.8097	0.7044
Linknet	Background	0.8296	0.8829	0.8296	0.8454	0.7492
	Crop	0.8729	0.8762	0.8729	0.8635	0.7842
	Weed	0.8361	0.8017	0.8361	0.8110	0.7064
FPN	Background	0.8206	0.8898	0.8206	0.8443	0.7469
	Crop	0.8864	0.8872	0.8864	0.8783	0.8098
	Weed	0.8499	0.7957	0.8499	0.8133	0.7101
PSPNet	Background	0.8134	0.8880	0.8134	0.8354	0.7376
	Crop	0.8567	0.9130	0.8567	0.8753	0.8015
	Weed	0.8239	0.8058	0.8239	0.8047	0.6989
PAN	Background	0.8992	0.7956	0.8992	0.8314	0.7297
	Crop	0.7997	0.6746	0.7997	0.6825	0.5624
	Weed	0.5975	0.8270	0.5975	0.6659	0.5342
DeepLabV3	Background	0.8113	0.8758	0.8113	0.8324	0.7342
	Crop	0.8852	0.8682	0.8852	0.8687	0.7942
	Weed	0.8485	0.7953	0.8485	0.8127	0.7080
DeepLabV3+	Background	0.8458	0.8734	0.8458	0.8496	0.7547
	Crop	0.8869	0.8979	0.8869	0.8865	0.8213
	Weed	0.8255	0.8162	0.8255	0.8120	0.7079
UPerNet	Background	0.8191	0.8882	0.8191	0.8419	0.7441
	Crop	0.8551	0.9044	0.8551	0.8660	0.7980
	Weed	0.8476	0.7915	0.8476	0.8097	0.7044

4.3. EXPERIMENTAL RESULTS ON SOTA MODELS AFTER MASK ENHANCEMENT

The images collected using a drone and mobile camera are annotated using Robolflow. After annotation, HSV color transformation is used to remove the manual annotation errors. Also, different SOTA models for semantic segmentation are selected and trained on the newly generated dataset. Among all the SOTA models, UNet performs better. It attains an accuracy of 97.51%, a precision of 97.53%, a recall of 97.51%, an F1-score of 97.51%, and a mean IoU of 92.87%. It can be observed from Table 6 that UNET performs better for all the performance metrics.

The per-class performance of all the SOTA models is analysed, which is given in Table 7. It can be observed that UNET results are much better than those of other models. The highest accuracy of 95.25%, recall of 95.25%, F1-score of 94.77%, and IoU of 91.07% for the crop class are attained by the UNET.

The highest Accuracy of 95.72 %, Recall of 95.72%, F1-score of 94.88%, and IoU of 90.51 % for the weed class, which is attained by the UNET. For other performance metrics, values attained by the UNET model are much closer to the highest value. After going through all the results, it can be inferred that the UNET performs better for all the performance metrics except precision.

Table 4. Overall Performance of SOTA models after mask enhancement

Model Name	Accuracy	Precision	Recall	F1 Score	Mean IoU
Unet	0.9751	0.9753	0.9751	0.9752	0.9287
Unet++	0.9740	0.9740	0.9740	0.9740	0.9268
MAnet	0.9728	0.9727	0.9728	0.9727	0.9195
Linknet	0.9722	0.9723	0.9722	0.9723	0.9228
FPN	0.9474	0.9474	0.9474	0.9474	0.8811
PSPNet	0.9299	0.9303	0.9299	0.9301	0.8501
PAN	0.9092	0.9092	0.9092	0.9081	0.7793
DeepLabV3	0.9295	0.9292	0.9295	0.9292	0.8469
DeepLabV3+	0.9502	0.9502	0.9502	0.9502	0.8855
UPerNet	0.9520	0.9525	0.9520	0.9522	0.8914

Table 5. Per-class performance of SOTA models after mask enhancement

Model Name	Class	Accuracy	Precision	Recall	F1-Score	IoU
Unet	Background	0.9809	0.9889	0.9809	0.9849	0.9703
	Crop	0.9525	0.9450	0.9525	0.9477	0.9107
	Weed	0.9572	0.9410	0.9572	0.9488	0.9051
Unet++	Background	0.9847	0.9825	0.9847	0.9835	0.9677
	Crop	0.9369	0.9598	0.9369	0.9472	0.9090
	Weed	0.9466	0.9503	0.9466	0.9483	0.9036
MAnet	Background	0.9859	0.9818	0.9859	0.9838	0.9683
	Crop	0.9338	0.9462	0.9338	0.9381	0.8943
	Weed	0.9347	0.9538	0.9347	0.9435	0.8959
Linknet	Background	0.9796	0.9855	0.9796	0.9825	0.9658
	Crop	0.9489	0.9444	0.9489	0.9456	0.9065
	Weed	0.9519	0.9362	0.9519	0.9436	0.8961
FPN	Background	0.9595	0.9606	0.9595	0.9600	0.9234
	Crop	0.9435	0.9398	0.9435	0.9408	0.8983
	Weed	0.8978	0.9013	0.8978	0.8992	0.8215
PSPNet	Background	0.9362	0.9510	0.9362	0.9434	0.8936
	Crop	0.9351	0.9267	0.9351	0.9297	0.8784
	Weed	0.8751	0.8703	0.8751	0.8721	0.7784
PAN	Background	0.9591	0.9212	0.9591	0.9393	0.8864
	Crop	0.8782	0.8206	0.8782	0.8264	0.7384
	Weed	0.7823	0.8823	0.7823	0.8276	0.7132
DeepLabV3	Background	0.9479	0.9397	0.9479	0.9434	0.8936
	Crop	0.9502	0.9160	0.9502	0.9317	0.8814
	Weed	0.8332	0.9012	0.8332	0.8628	0.7658
DeepLabV3+	Background	0.9629	0.9619	0.9629	0.9623	0.9277
	Crop	0.9498	0.9375	0.9498	0.9427	0.9014
	Weed	0.8928	0.9138	0.8928	0.9026	0.8273
UPerNet	Background	0.8191	0.8882	0.8191	0.8419	0.7441
	Crop	0.8551	0.9044	0.8551	0.8660	0.7980
	Weed	0.8476	0.7915	0.8476	0.8097	0.7044

4.4. PERFORMANCE IMPROVEMENT AFTER MASK ENHANCEMENT

The HSV color transformations are used to remove the manual annotation errors. Both crops and weeds mostly possess a green color. Also, the variance of

the size of weeds is very high. Due to this, annotating weeds is a very challenging task. HSV transformation addresses this issue effectively.

This is how the two datasets, Dataset-1 and Dataset-2, are made. Images and mask images made with Algorithm-I are in Dataset-1. Dataset-2 has images and mask images that were made with Algorithm-II.

The performance comparison in Table 6 shows that the mask enhancement technique made a big difference in the performance of many state-of-the-art (SOTA) semantic segmentation models. Among the architectures tested, PAN had the best overall performance, with an accuracy of 11.08%, a precision of 10.22%, a recall of 11.08%, an F1 score of 11.33%, and a mean IoU of 28.01%. This shows that it is better at using refined boundary information from enhanced masks. The LinkNet and U-Net++ models also showed big improvements, with mean IoU values of 23.60% and 21.90%, respectively. This shows that encoder-decoder architectures with efficient skip connections benefit a lot from better mask delineation.

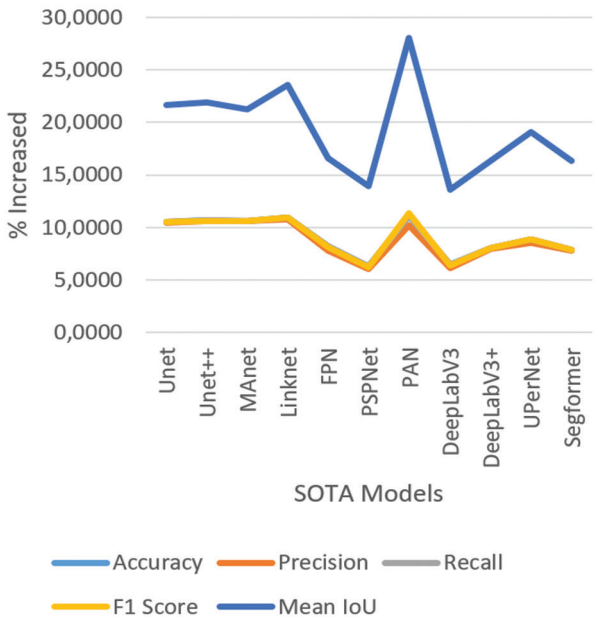


Fig. 7. Percentage improvement in performance of SOTA models after using mask images generated using Algorithm-II

The improvements seen in all models show that mask enhancement helps with better feature localization and boundary refinement, especially when the original segmentation masks had rough or noisy annotations. Traditional models like U-Net and MAnet showed steady but moderate improvements in performance. This suggests that these architectures do a good job of capturing contextual information, but their decoders may not be able to take full advantage of improved edge precision because they are not very deep. On the other hand, models like DeepLabV3 and PSPNet, which use dilated convolutions and pyramid pooling to combine context, saw smaller improvements.

This may be because they rely on large receptive fields instead of fine-grained spatial accuracy, which makes them less sensitive to mask-level changes.

In general, the results show that mask enhancement techniques can greatly improve segmentation quality, especially for architectures that rely on multi-scale spatial correspondence between the encoder and decoder stages. This finding emphasizes the significance of high-quality annotation refinement as an adjunct to model optimization, which can directly impact the learning of object boundaries and improve the discriminative ability of deep segmentation networks in intricate agricultural imagery.

Table 6 Performance improvement of SOTA models after Mask Enhancement

Model Name	Accuracy	Precision	Recall	F1 Score	Mean IoU
Unet	10.4790	10.4633	10.4790	10.5247	21.6467
Unet++	10.6373	10.5689	10.6373	10.6255	21.9041
MAnet	10.6330	10.6087	10.6330	10.6350	21.2740
Linknet	10.9201	10.7346	10.9201	10.8957	23.6009
FPN	8.1502	7.8027	8.1502	8.1175	16.6078
PSPNet	6.2280	5.9926	6.2280	6.2060	13.9532
PAN	11.0813	10.2198	11.0813	11.3373	28.0104
DeepLabV3	6.4047	6.0793	6.4047	6.3428	13.6096
DeepLabV3+	8.0101	7.9415	8.0101	7.9887	16.3103
UPerNet	8.8652	8.5419	8.8652	8.8315	19.0327

4.5. PERFORMANCE COMPARISON OF UNET AND UNET-SA MODEL ON PIGEON PEA DATASET

The spatial attention module improves the performance of UNET. It can be observed from Table 7 that the performance of the UNET-SA model is superior to the UNET model for all the performance metrics. The SA module focuses on the most important part, which helps in improving performance. As per Table 7, the UNET-SA model attains an accuracy of 98.53%, a precision of 98.54%, a recall of 98.53%, an F1-score of 98.53, and a mean IoU of 94.49%. The performance of the UNET-SA model is improved for all the metrics. The accuracy, precision, recall, F1-score, and mean IoU are increased by 1.04%, 1.04%, 1.04%, 1.04%, and 1.74%, respectively.

Table 7. Performance comparison of UNET and UNET-SA models on Pigeon pea Dataset

Model Name	Accuracy	Precision	Recall	F1 Score	Mean IoU
Unet	0.9751	0.9753	0.9751	0.9752	0.9287
UNET-SA	0.9864	0.9854	0.9853	0.9853	0.9449

The confusion matrix and training and validation loss curves are given in Figs. 8 and 9. The training and validation accuracy are above 95% and the training and validation IoU are above 90%. Also, all the losses are below 0.1.

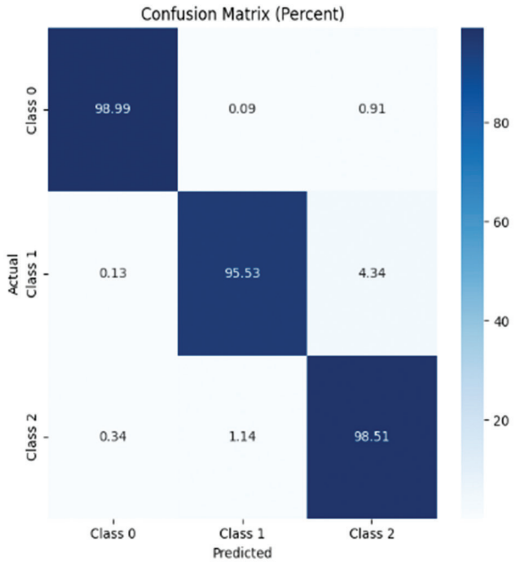
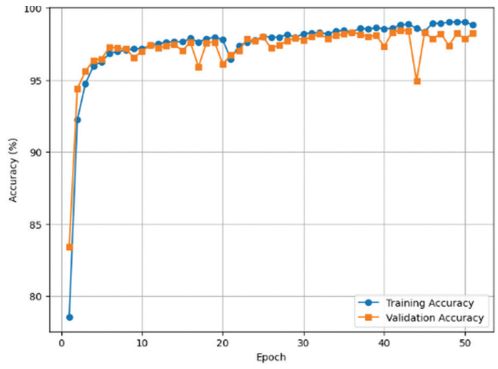
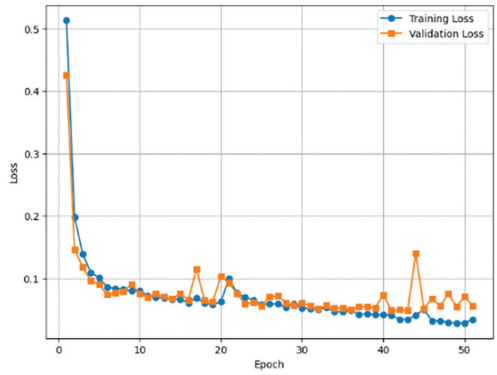


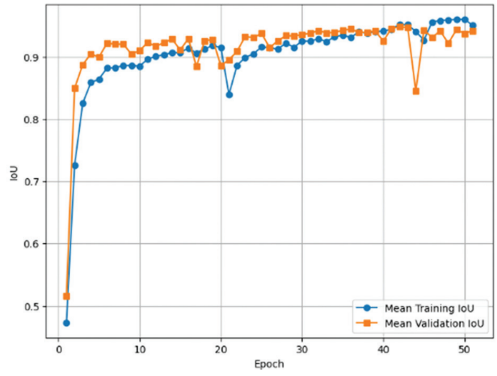
Fig. 8. Confusion Matrix of the UNet-SA model(proposed)



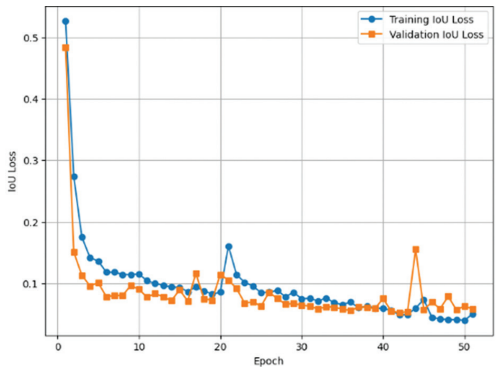
(a)



(b)



(c)



(d)

Fig. 9. Training and Validation curves for the UNet-SA model(proposed). (a) Training vs Validation Accuracy, (b) Training vs Validation loss, (c) Mean Training IoU vs Mean Validation IoU, 9(d)Training vs Validation IoU Loss

4.6. PERFORMANCE COMPARISON OF UNET AND UNET-SA MODEL ON THE CROPANDWEED DATASET [32]

The CropAndWeed dataset [32] has 111,953 images in all. There are 74 classes in all. 16 of them are for the types of crops, like sugar beets, soybeans, sunflowers, potatoes, peas, beans, pumpkins, maize, and so on. The other 58 classes are types of weeds like grasses, knotweed, goosefoot, thistle, and others. The dataset includes a wide range of soil types, plant types, weed types, and more because it was collected from many different places. This benchmark dataset is used to compare the original UNet and the proposed model in terms of MeanIoU, accuracy, precision, and other metrics.

Table 8. Performance comparison of UNET and UNET-SA models on the CropAndWeed Dataset

Model Name	Accuracy	Precision	Recall	F1 Score	Mean IoU
Unet	98.49	98.40	98.49	98.41	51.81
UNET-SA	98.81	98.79	98.81	98.80	55.79

The CropAndWeed dataset has a class imbalance and a pixel distribution that makes the mean IoU low but the accuracy high. The model can get a high accuracy by correctly classifying these big areas because most of the pixels are in the background or the main crop areas. But mean IoU treats all classes the same and is more likely to make mistakes in smaller or less common weed classes. IoU also punishes mistakes on the boundaries and partial mis-segmentations, which are common in fine plant structures. So, even though the model's pixel-level accuracy is very high, its segmentation consistency across all classes is still lower than it could be.

4.7. DISCUSSION

The Pigeon Pea dataset is used in this work. The dataset is collected from three different fields, which have different growth stages of the crop of Pigeon pea, soil con-

ditions, and types of weeds. Also, the dataset is collected on six different days. The lighting conditions during data collection were different. These factors add variability to the dataset. The model training is carried out on a high-end workstation. Various SOTA models used for semantic segmentation are employed. These SOTA models are efficient and resource-intensive, and are selected due to their advantages for attaining the set objectives.

The images captured through a smartphone and a drone are annotated using Roboflow. The annotations have errors. To reduce these errors, RGB images are converted to the HSV color space for efficient detection of crops and weeds, which are green in color. The two datasets are created, one without and one with HSV color transformation, named Dataset-1 and Dataset-2. The SOTA models are trained on both datasets. The performance on the dataset with HSV transformation is improved. Also, on both datasets, UNET has better performance compared to all the SOTA models.

On Dataset-1, UNET achieves better performance compared to all the SOTA models. It attains an accuracy of 88.26%, a precision of 88.29%, a recall of 88.26%, an F1-score of 88.23%, and a mean IoU of 76.35%. On Dataset-2, it attains an accuracy of 97.51%, a precision of 97.53%, a recall of 97.51%, an F1-score of 97.52%, and a mean IoU of 92.87. It can be concluded that due to HSV color transformation, the masks' images are improved, and annotation errors are reduced. With reduced errors, models get trained better and perform better.

As UNET is performing better than all the SOTA models, it is selected for further improvement. To enhance its performance, the spatial attention mechanism is integrated, which helps in focusing on the most important part of the image. The spatial attention mechanism is integrated in the encoder part. After each down-sampling operation, the image is passed through the spatial attention. Also, it is particularly effective in segmentation tasks where objects may be overlapping, occluded, or highly similar in appearance. The UNET-SA model attains an accuracy of 98.64%, a precision of 98.54%, a recall of 98.53%, an F1-score of 98.53, and a mean IoU of 94.49%. The performance of the UNET-SA model is improved for all the metrics. The accuracy, precision, recall, F1-score, and Mean IoU are increased by 1.04%, 1.04%, 1.04%, 1.04%, and 1.74%, respectively.

Table 9 shows how different U-Net architectures with different attention mechanisms compare in terms of performance. The evaluation metrics—accuracy, precision, recall, F1 score, and mean Intersection over Union (mIoU)—show how well each model can accurately separate crop and weed areas in field images.

U-Net-SA (U-Net with Spatial Attention) is the best of all the variants. It has an accuracy of 98.64%, a precision of 98.54%, a recall of 98.53%, an F1 score of 98.53%, and a mean IoU of 94.49%. These numbers are higher than those of other U-Net models that use attention, like U-Net-DA (Dual Attention) and U-Net-AT (Attention Gate), which have Mean IoUs of 92.68% and 89%91, respectively.

The Spatial Attention (SA) mechanism, which focuses on finding the most important spatial regions in the feature maps by highlighting areas with a lot of texture and spatial changes, is what makes U-Net-SA work so well. This is especially helpful for separating crops from weeds, since changes in the background soil, shadows, and lighting can make it hard to see the edges of plants. U-Net-SA effectively reduces irrelevant background noise and improves fine-grained boundary details by learning how to highlight discriminative spatial locations. This leads to more accurate mask generation.

Table 9. Performance comparison of the proposed model with other UNET models with attention modules

Model Name	Accuracy	Precision	Recall	F1 Score	Mean IoU
UNET-CBAM	0.9417	0.9399	0.9417	0.9405	0.7848
UNET-DA	0.9802	0.9803	0.9802	0.9802	0.9268
UNET-ST	0.9642	0.9645	0.9642	0.9643	0.8558
UNET-RA	0.9666	0.9668	0.9666	0.9667	0.8599
UNET-AT	0.9751	0.9752	0.9751	0.9752	0.8991
UNET-SE	0.9592	0.9595	0.9592	0.9593	0.8348
UNET-SCSE	0.8954	0.9039	0.8954	0.8881	0.6571
UNET-SA	0.9864	0.9854	0.9853	0.9853	0.9449

Other attention mechanisms, like SE (Squeeze-and-Excitation) and CBAM (Convolutional Block Attention Module), on the other hand, focus mostly on channel-wise dependencies. This means they might miss small spatial cues that are needed to tell the difference between crops and weeds that are overlapping or close together. The same goes for U-Net-SCSE, which combines spatial and channel attention by concatenating them. It doesn't work as well (mean IoU = 65.71%) because it might have too many parameters and doesn't do a good job of feature fusion.

Even though U-Net-DA and U-Net-AT also work well, they use more complicated attention aggregation, which can make it harder to accurately locate small objects in space. The U-Net-SA variant, on the other hand, strikes the perfect balance between model complexity and feature enhancement. It focuses on vegetation pixels without slowing down the computer.

We also tested the UNET-SA model on the large CropAndWeed dataset, which has 111,953 images of 74 different types of crops and weeds. This was done to see how well it could generalize. The proposed model got 98.81% accuracy and a mean IoU of 55.79% on this benchmark, which was better than the baseline UNET (accuracy = 98.49%, Mean IoU = 51.81%). The difference between high accuracy and a lower mean IoU is mostly due to class imbalance, where large background ar-

eas have more pixels than smaller weed classes, which makes accuracy higher and IoU lower. Still, the fact that both metrics keep getting better shows that spatial attention helps the model generalize better to different types of fields, soils, and vegetation structures.

The results show that adding the Spatial Attention mechanism to the UNET architecture greatly improves the quality of semantic segmentation by making spatial feature representation and boundary precision stronger. The model is efficient and flexible, which makes it a good choice for real-time crop-weed identification on portable IoT devices like the Raspberry Pi or Jetson Nano. But spatial attention also adds more parameters, which can make it easier to overfit on smaller datasets and slightly increase the amount of computation needed during inference. Future research may investigate Vision Transformer (ViT)–based encoders to capture more intricate spatial dependencies and multimodal data fusion (RGB + NIR) to improve the differentiation between crops and weeds. Also, making fake samples with generative adversarial networks (GANs) could help with the lack of data and make the model even better at generalizing.

4.8. LIMITATIONS AND SCALABILITY CONSIDERATIONS

The proposed U-Net-SA model performs exceptionally well on the Pigeon Pea dataset; however, its scalability to various agricultural domains and unfamiliar environments poses significant challenges. The dataset, despite being gathered from three separate domains with differing growth stages, soil types, and lighting conditions, may not comprehensively represent the diversity of actual agricultural ecosystems. As a result, the model may not work as well in areas with very different canopy structures, weed densities, or background textures. Changes in the color of the soil, the shape of the weeds, and the maturity of the crops can affect the spectral properties of plants. This can make it harder to segment them correctly when tested outside of the current domain. The HSV color space transformation does help with annotation consistency and segmentation performance, but it only works well for crops and weeds that are in the green color range. This method might not work as well when there are a lot of non-green crops, dead leaves, or dry weeds in the field. This makes the model less flexible in other agricultural situations. Another issue is that U-Net-SA's computational scalability is limited because adding spatial attention parameters makes training longer and inference latency higher. This makes it hard to use U-Net-SA on low-power IoT devices for large-scale or real-time deployments.

Synthetic data augmentation can help get around these problems. Using Generative Adversarial Networks (GANs) or diffusion-based models to create realistic images of crops and weeds can add a lot of different field textures, lighting conditions, and plant shapes to the dataset. This would make the model more robust to changes across fields and help prevent overfitting. Also, domain adaptation methods and transfer learning can make scalability even better by making small changes to the trained model for new areas or crop types with little effort to annotate. Consequently, forthcoming research ought to concentrate on utilizing synthetic augmentation and domain adaptation to enhance the applicability of the U-Net-SA model for extensive precision agriculture in diverse field conditions.

5. CONCLUSION

A novel approach for detecting crops and weeds in Pigeon Pea fields by integrating a spatial attention module into the UNET model is proposed in this work. The proposed model attains an accuracy of 98.64%, precision of 98.54%, recall of 98.53%, F1-score of 98.53%, and mean IoU of 94.49%. The annotation errors are minimized using the HSV color space. The performance of SOTA models is improved on the dataset in which annotations are transformed using the HSV colorspace. The performance of all the SOTA models is improved after HSV color transformation in all the performance metrics. As UNET is performing better, it is selected, and all the attention mechanisms are integrated into it. After performance analysis, it is found that UNET with a spatial attention mechanism is performing better.

The proposed UNET-SA model is also tested on the large CropAndWeed dataset [32], which has 74 different types of crops and weeds, to make sure it is even more reliable. The model does better than the baseline UNET (accuracy 98.49%, mean IoU 51.81%) with an accuracy of 98.81% and a mean IoU of 55.79%. The high accuracy and low mean IoU show that there is class imbalance. This is because dominant background pixels raise overall accuracy, while the mean IoU shows how well the segmentation works across all classes. This shows that the proposed model works well with a wide range of agricultural datasets that are both diverse and complex.

We used images from three fields and six days to diversify the dataset and reduce overfitting to demonstrate that the algorithm can handle new data. The model scored well on all assessment criteria on a test set not used during training. While the UNET-SA model offers advantages, it also presents certain limitations that need to be addressed. These drawbacks impact the model's generalizability and computational efficiency. Although spatial attention enhances predictions, it introduces characteristics that can exacerbate overfitting on datasets with limited samples. This aspect of the model's behavior needs to be carefully addressed to ensure robust performance. This increase in processing overhead and inference latency due to spatial attention can affect the model's real-time performance and computational efficiency, necessitating optimization strategies for deployment. The integration of vision transformers could enhance the model's capacity to capture intricate details of different crops and weeds, potentially improving

segmentation accuracy and feature representation in agricultural imagery. Future research endeavors could involve testing the proposed model with RGB+NIR data to explore the synergistic benefits of combining visible and near-infrared spectral information for improved crop and weed segmentation accuracy. Also, we can add synthetic data to the dataset using a generative adversarial network (GAN) to alleviate data scarcity.

6. REFERENCES:

[1] P. Radoglou-Grammatikis, P. Sarigiannidis, T. Lagkas, I. Moscholios, "A compilation of UAV applications for precision agriculture", Computer Networks, Vol. 172, 2020.

[2] R. Lal, "Soil structure and sustainability", Journal of Sustainable Agriculture, Vol. 1, No. 4, 1991.

[3] S. K. Seelan, S. Laguette, G. M. Casady, G. A. Seielstad, "Remote sensing applications for precision agriculture: A learning community approach", Remote Sensing of Environment, Vol. 88, No. 1-2, 2003.

[4] D. D. Patel, B. A. Kumbhar, "Weed and its management: A major threats to crop economy", Journal of Pharmaceutical Sciences and Bioscientific Research, Vol. 6, No. 6, 2016.

[5] N. Iqbal, S. Manalil, B. S. Chauhan, S. W. Adkins, "Investigation of alternate herbicides for effective weed management in glyphosate-tolerant cotton", Archives of Agronomy and Soil Science, Vol. 65, No. 13, 2019.

[6] Lead Team, "Five general categories of weed control methods", https://greenrootltd.com/2019/02/19/five-general-categories-of-weed-control-methods/ (accessed: 2025)

[7] J. S. Holt, "Principles of weed management in agroecosystems and wildlands", Weed Technology, Vol. 18, No. sp1, 2004.

[8] B. Liu, R. Bruch, "Weed detection for selective spraying: A review", Current Robotics Reports, Vol. 1, No. 1, 2020.

[9] P. Lameski, E. Zdravevski, A. Kulakov, "Review of automated weed control approaches: An environmental impact perspective", Proceedings of the International Conference on Telecommunications, Ohrid, Macedonia, 17-19 September 2018, pp. 132-147.

[10] P. Lottes, J. Behley, N. Chebrolu, A. Milioto, C Stachniss, "Joint stem detection and crop-weed classification for plant-specific treatment in precision farming", Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Madrid, Spain, 1-5 October 2018, pp. 8233-8238.

[11] S. Haug, J. Ostermann, "A crop/weed field image dataset for the evaluation of computer vision based precision agriculture tasks", Proceedings of the Computer Vision - ECCV 2014 Workshops, Zurich, Switzerland, 6-7 September 2015, pp. 105-116

[12] N. Chebrolu, P. Lottes, A Schaefer, W. Winterhalter, W. Burgard, and C Stachniss, "Agricultural robot dataset for plant classification, localization and mapping on sugar beet fields", The International Journal of Robotics Research, Vol. 36 No. 10, 2017, pp. 1045-1052.

[13] T. Liu, Y. Zhao, H. Wang, W. Wu, T. Yang, W. Zhang, S. Zhu, C. Sun, Z. Yao, "Harnessing UAVs and deep learning for accurate grass weed detection in wheat fields: a study on biomass and yield implications", Plant Methods, Vol. 20, No. 1, 2024, p. 144.

[14] P. Kumar, S. J. Miklavcic, "Analytical study of colour spaces for plant pixel detection", Journal of Imaging, Vol. 4, No. 2, 2018, p. 42.

[15] J. Kailun, H. Wenjiang, P. Huang, "Dual-task segmentation of oilseed rape and weeds in agricultural fields: A hybrid approach combining enhanced UNet and unsupervised clustering", Computers and Electronics in Agriculture, Vol. 238, 2025, p. 110827.

[16] H. Xu, Y. Lan, S. Zhang, B. Tian, H. Yu, X. Wang, S. Zhao, Z. Wang, D. Yang, J. Zhao, "Research on vegetation cover extraction method of summer maize based on UAV visible light image", International Journal of Precision Agricultural Aviation, Vol. 6, No. 1, 2023.

[17] A. Nițu, C. Florea, M. Ivanovici, A. Racoviteanu, "NDVI and Beyond: Vegetation Indices as Features for Crop Recognition and Segmentation in Hyperspectral Data", Sensors, Vol. 25, No. 12, 2025, p. 3817.

[18] J. Cui, F. Tan, N. Bai, Y. Fu, "Improving U-Net network for semantic segmentation of corn seed-

lings and weeds during the corn seedling stage in field", Frontiers in Plant Science, Vol. 14, 2024, p. 1344958.

[19] Y. Cai, L. Wang, "Attention-aided semantic segmentation network for weed distribution mapping in UAV platforms", Computers and Electronics in Agriculture, Vol. 211, 2023, p. 108403.

[20] X. Yi, J. Wang, G. Chen, H. Zhang, "U-Net with Coordinate Attention and VGGNet backbone for grape leaf disease segmentation", Agronomy, Vol. 14, No. 5, 2024, p. 925.

[21] Y. Chen, Y. Xie, W. Yao, Y. Zhang, X. Wang, Y. Yang, L. Tang, "U-MGA: A Multi-Module U-Net Optimized with Multi-Scale Global Attention Mechanisms for Fine-Grained Segmentation of Cultivated Areas", Remote Sensing, Vol. 17, No. 5, 2025, p. 760.

[22] Q. Yang, Y. Ye, L. Gu, Y. Wu, "MSFCA-net: A multi-scale feature convolutional attention network for segmenting crops and weeds in the field", Agriculture, Vol. 13, No. 6, 2023, p. 1176.

[23] X. Wang, S. Zhang, T. Zhang, "Crop insect pest detection based on dilated multi-scale attention U-Net", Plant Methods, Vol. 20, 2024, p. 34.

[24] K. Zhao, Q. Zhang, C. Wan, Q. Pan, and Y. Qin, "Visual Mamba UNet fusion multi-scale attention and detail infusion for unsound corn kernels segmentation", Scientific Reports, Vol. 15, No. 1, 2025, pp. 1-20.

[25] Y. Cai, L. Wang, "Attention-aided semantic segmentation network for weed distribution mapping in UAV platforms", Computers and Electronics in Agriculture, Vol. 211, 2023, p. 108403.

[26] X. Mei, Y. Sun, M. Zhou, L. Liu, H. Yang, "SSMR-Net and across-feature-mapping attention combined with U-Net for improved weed segmentation", Computers and Electronics in Agriculture, Vol. 260, 2025, p. 110827.

[27] J. Kailun, H. Wenjiang, H. Ping, "Dual-task segmentation of oilseed rape and weeds in agricultural fields: A hybrid approach combining enhanced UNet and unsupervised clustering", Computers and Electronics in Agriculture, Vol. 238, 2025, p. 110827.

[28] Y. Li, R. Guo, R. Li, R. Ji, M. Wu, D. Chen, C. Han, R. Han, Y. Liu, Y. Ruan, J. Yang, "An improved U-Net and attention mechanism-based model for sugar beet and weed segmentation", Frontiers in Plant Science, Vol. 15, 2025, p. 1449514.

[29] S. D. Khan, S. M. Khan, M. Ejaz, Z. He, "Weed-Crop Segmentation in Drone Images with a Novel Encoder-Decoder Framework Enhanced via Attention Modules", Remote Sensing, Vol. 15, No. 23, 2023, p. 5615.

[30] Y. Lu, H. Li, C. Zhang, S. Zhang, "Object-Based Semi-Supervised Spatial Attention Residual U-Net for Urban High-Resolution Remote Sensing Image Classification", Remote Sensing, Vol. 16, No. 8, 2024, p. 1444.

[31] A. Syed, D. K. Sharma, "MSEA-Net: Multi-Scale and Edge-Aware Network for Weed Segmentation via Spatial-Channel Attention", Agriculture, Vol. 7, No. 4, 2025, p. 103.

[32] D. Steininger, A. Trondl, G. Croonen, J. Simon, V. Widhalm, "The CropAndWeed Dataset: a Multi-Modal Learning Approach for Efficient Crop and Weed Manipulation", Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2-7 January 2023.