Semantic Segmentation of Plant Structures with Deep Learning and Channel-wise Attention Mechanism

Mukund Kumar Surehli[1], Naveen Aggarwal[2], Garima Joshi[2], and Harsh Nayyar[2]

[1]*VIT-Bhopal University, Madhya Pradesh, India,*
[2]*Panjab University, Chandigarh, India*

https://doi.org/10.26636/jtit.2025.1.1853

Abstract — Semantic segmentation of plant images is crucial for various agricultural applications and creates the need to develop more demanding models that are capable of handling images in a diverse range of conditions. This paper introduces an extended DeepLabV3+ model with a channel-wise attention mechanism, designed to provide precise semantic segmentation while emphasizing crucial features. It leverages semantic information with global context and is capable of handling object scale variations within the image. The proposed approach aims to provide a well generalized model that may be adapted to various field conditions by training and tests performed on multiple datasets, including Eschikon wheat segmentation (EWS), humans in the loop (HIL), computer vision problems in plant phenotyping (CVPPP), and a custom "botanic mixed set" dataset. Incorporating an ensemble training paradigm, the proposed architecture achieved an intersection over union (IoU) score of 0.846, 0.665 and 0.975 on EWS, HIL plant segmentation, and CVPPP datasets, respectively. The trained model exhibited robustness to variations in lighting, backgrounds, and subject angles, showcasing its adaptability to real-world applications.

Keywords — channel-wise attention, computer vision, DeepLabV3+, deep learning, plant segmentation, semantic segmentation

1. Introduction

Computer vision-based segmentation methods have been used to address some data-rich agriculture problems. Segmentation of plant structures enables precise crop monitoring, disease detection, and weed management. It facilitates targeted interventions, optimizes usage of resources and contributes to sustainable production.

Semantic segmentation of plant structures involves classifying image pixels into distinct categories, such as leaves, stems, and fruits. Conventionally, thresholding for semantic segmentation of plant images involves differentiating between the foreground (green color) and background. Such an approach was employed in works [1] and [2]. This involved using such indices as excess green index (ExG) [3], normalized green-red difference index (NGRDI) [4], and color index of vegetation extraction (CIVE) [5] to enhance the green color of plants in the images.

The reliance of such methods on the green color limits their effectiveness and applicability to cases where the plant color differs significantly from the background. Additionally, variations in lighting conditions and reflections can impact segmentation accuracy, posing challenges for generalization across different scenarios [6]. Therefore, classic methods used for semantic segmentation of plant images pay attention to the features in an image and compare the differences between and/or gradients of pixels. These methods employ mathematical models and algorithms to identify regions of interest within an image. To identify these regions, common characteristics such as color, texture, and intensity are used.

These segmentation techniques, though simple, fast and memory-efficient, are more applicable to simple segmentation tasks. They require fine tuning for the specific use case and provide limited accuracy for complex scenes, which makes them considerably unsuitable for dealing with plant images.

On the other hand, deep learning-based methods perform segmentation-related tasks by employing neural networks to identify the vital features in an image [7]. These developments have resulted in some decent image segmentation models, boasting remarkable performance improvements over their classic predecessors.

2. Related Work

Research and development related to semantic segmentation focused on deep learning approaches and has resulted in the creation of various models relying on a wide array of architectures. The authors of [8] proposed a two-stage deep learning approach for plant disease detection. In the first stage, semantic segmentation models (U-Net [9], SegNet [10], and DeepLabV3+ [11]) were employed to extract plants from the input, with U-Net achieving the highest mean weighted intersection over union (mwIOU) of 0.9422. In the second stage, DeepLabV3+ outperformed the previous approach achieving mwIOU of 0.7379 for disease localization. Such an integrated model combining U-Net and DeepLabV3+ demonstrated robust performance. Unfortunately, the paper lacks a discussion on the generalization methods making it suitable to various crop species or diseases and omits

insights into computational costs – a crucial aspect for real-time segmentation tasks. These limitations should be taken into account while considering broader applicability of the proposed two-stage model.

Paper [12] introduces a semantic segmentation framework leveraging both real and synthetic data. The proposed approach employes a mask region-based convolutional neural network (R-CNN) model with a ResNet101 backbone [13] and a feature pyramid network (FPN) [14]. Synthetic images were generated from a dataset focusing on computer vision problems in plant phenotyping (CVPPP) encountered in the leaf segmentation challenge (LSC) [15]. Training the model on a dataset comprising both real and synthetic images, the authors achieved a leaf segmentation score of 90% on the A1 subset of the CVPPP dataset, with a mean score of 81% across the entire dataset.

Article [16] utilizes U-Net for semantic segmentation of leaf structures in plants. Leveraging the architecture's lightweight structure, as well as its computationally less intensive nature and fast inference, the authors trained it on the CVPPP-LSC dataset and achieved an intersection over union (IoU) of 90.56% and 98.69% on training and testing sets, respectively. Utilizing the U-Net architecture, the proposed model was resistant to varying input image dimensions. However, it is crucial to take note of the fact that the proposed approach is overly reliant on the training dataset.

The dataset contains images of *Arabidopsis thaliana* and *Nicotiana tabacum* (Tobacco) plants, with the structures being highly similar and all green. This affects the model's ability to generalize over a diverse range of plant structures and varying colors. Additionally, the images were captured in indoor conditions, in ideal lighting environments and have no structural overlap. This creates reservations concerning the segmentation quality of the model when used outdoors, where lighting, shadows, occlusions and position of the subject all could vary. In order to be used in real-world applications, the model would require rigorous tuning and diverse datasets.

The authors of [17] performed semantic segmentation on tall fig shrubs under real-world, open-field cultivation conditions. The proposed methodology made use of a convolutional neural networks (CNN) architecture inspired by SegNet with fewer trainable parameters. It was trained on a custom dataset comprising fig shrub images, captured from a drone, at a relatively high altitude and achieved an impressive accuracy of 93.84%. The introduced model was robust enough to handle varying outdoor visual conditions, such as shadows, occlusions, plant overlap, sunlight illumination. Additionally, owing to its smaller size, it was relatively computationally less intensive for inference purposes.

The literature review indicates that complexity of plant structures is an important factor, as plants contain structures of varying scale – from fine vein-like elements to the shape of the entire plant – making it increasingly difficult to have one model to detect them all. Illumination is another key factor, as variations in lighting conditions, especially in outdoor settings, affect the plants" appearance and visibility. The background is the next factor that needs to be taken into consideration, as images may often contain cluttered backdrops such as weeds, parts of other plants, soiling caused by wetting or drying, moss, etc. The availability of high-quality annotated data is another challenge. Annotation of plant images for segmentation purposes may be highly be labor-intensive and time-consuming Therefore, the process of creating large and diverse datasets poses a demanding challenge [18].

The culmination of these factors, from variations in outdoor conditions to the availability of data, creates a challenging scenario while developing the segmentation model. However, taking account of the dataset's inherent nature, it would be increasingly difficult to adapt the proposed model to segmenting plants that exhibit different structures, or to plant images that have been captured from a closer distance. The authors noted poor accuracy in scenes in which miscellaneous structures were visible along the fields or plants. Furthermore, the model's smaller size creates uncertainty regarding its ability to capture increasingly complex scenes for effective segmentation.

To overcome the "less-than-ideal" condition and over-reliance on training datasets, we introduce a DeepLabV3+ model coupled with a channel-wise attention mechanism referred to as "squeeze & excitation" (SE). The aforementioned model has been tested using our custom dataset, i.e. "botanic mixed set" [19], to evaluate the its generalization capabilities and applicability to data that is comprehensively unseen and different.

3. Methodology

The work presented in this paper introduces a robust semantic segmentation model with the ability to generalize over a wide array of plant species and handle various issues, as discussed in Section 1, by exploiting modern deep learning methods. The encoder's shallow layers represent the image as a low-level feature map presenting basic, simple, and local characteristics of an object in the image, such as edges, textures and corner points. The deeper layers output a high-level representation of the image, focusing on complex shapes and a deeper understanding of the global context. High-level features are often composites of multiple low-level features.

The DeepLabV3+ segmentation model was chosen based on its encoder-decoder architecture. The model makes use of the atrous spatial pyramid pooling (ASPP) module (fed with high-level features from the encoder), which internally makes use of atrous (or dilated) convolutions. Atrous convolutions differ from normal convolutions in the way that they introduce gaps in the kernel with a parameter called dilation rate. When these dilated kernels stride over image pixels, they may capture a wider field of view, thus producing a feature map that has a certain spatial context.

In the ASPP module, several atrous convolutions at different dilation rates are performed in parallel in order to obtain their corresponding feature maps. These maps, along with a global average pooling map, are concatenated to form the ASPP output. This output is rich in spatial context at different

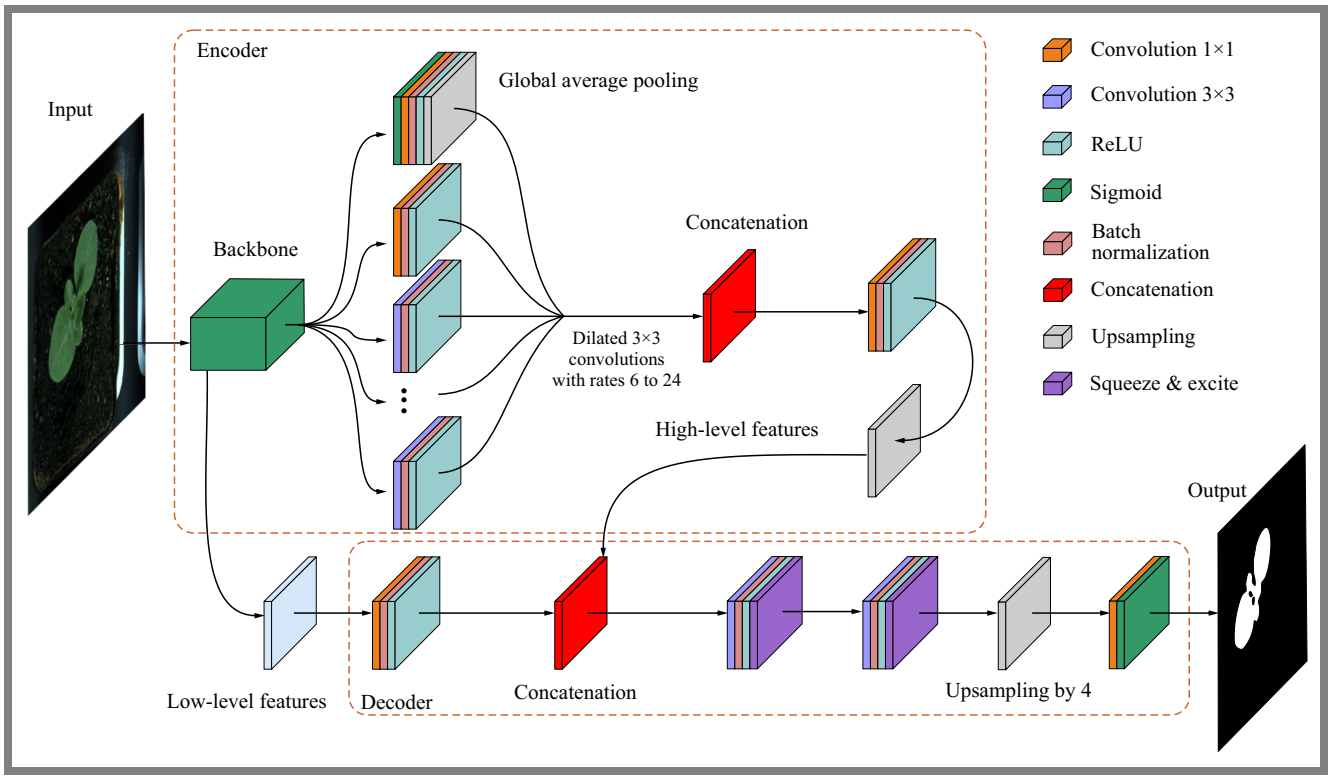

Fig. 1. Architecture of the implemented model.

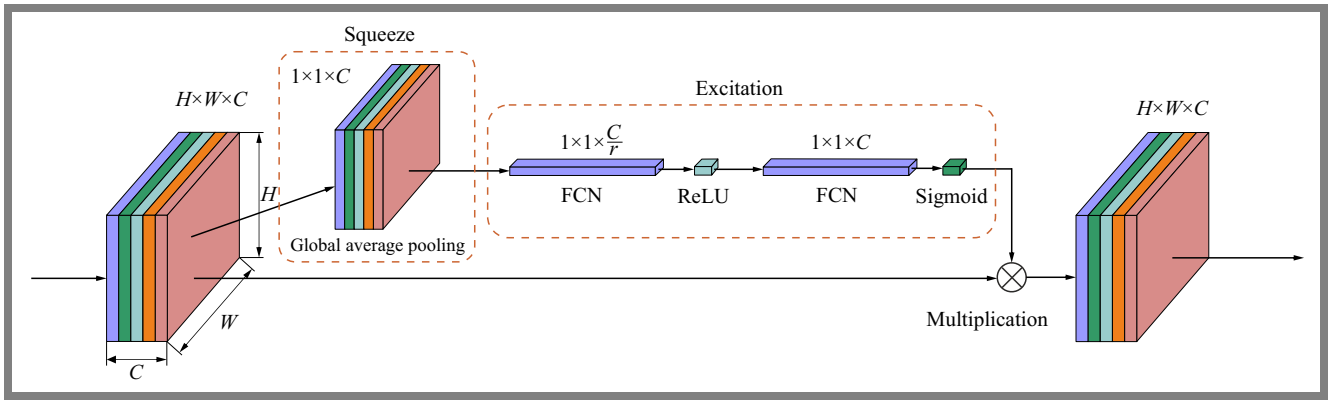

Fig. 2. Architecture of the squeeze and excitation (SE) module.

scales. High-level features from ASPP and low-level features from the encoder are concatenated together in the decoder to combine semantic information with increased spatial context.

Pixel-wise classifications are performed and the original input resolution is achieved with consequent 3×3 convolutions and bilinear up-sampling by 4. A simpler way to look at the architecture is to observe that at the model is initially aware of the position of a given object in the image (spatial information) but is not exactly aware of what that object is (lacking semantic information). As the input propagates forward, the model becomes aware of what the object is, thus gaining semantic information, but because of the repeated convolutions, it lacks global spatial information.

Segmentation models, with their goal being to assign class labels to every pixel in the image, require spatial information to delineate object boundaries and semantic information to differentiate object categories within the image. Thus, skip-connections are utilized to transport both spatial and semantic information (through ASPP) to the decoder. The decoder combines both and gets a concatenated feature map which not only knows what the object is, but also where exactly it is.

Figure 1 provides an overview of the model described in this paper. CNNs at each layer output feature maps that assist the network in extracting hierarchical information from the input. These feature maps are represented as tensors of $B \times C \times H \times W$ dimensionality, where B denotes the batch size, C denotes the number of channels, H denotes the height, and W denotes the width. Feature maps are representations of different features in an image, and directly affect the quality of output. The aim is to recalibrate these feature maps in such a way that, for a given task, they capture and highlight only those properties of an object that benefit the output.

A way of achieving feature recalibration is channel-wise attention. When CNN outputs feature maps, the channels within it are equally weighted. Since channels represent the derivation of different features from an input (convolutional filters), it is given that not all the channels within a feature map hold equal representational importance for a specific task. Paper [20] proposed a method of applying channel-wise weights to describe their representativeness. Thus, more important features are amplified the less useful ones are suppressed. This method involved using a SE module responsible for the squeeze and excite operation.

Figure 2 illustrates the SE operation that is employed in this study. For the squeeze operation, consider the feature map as a tensor of dimensionality $C \times H \times W$. A channel descriptor is formed here by aggregating the input feature map across its spatial dimensions $H \times W$, forming a single numerical value. The goal of this aggregation is to capture global information about the feature map in a channel-wise manner. Global average pooling is a way to perform these aggregations. The output of the squeeze operation is in the $C \times 1 \times 1$ dimensions and is forwarded to the excite operation.

The excite operation consists of two fully convolutional network (FCN) blocks with a ReLU and sigmoid activation, respectively. The first block performs dimensionality reduction on C channels by a reduction factor r, with the goal of this operation being to decrease computational complexity and maintain global information at smaller scales. The input is now reduced to dimensions of $\frac{C}{r} \times 1 \times 1$ and is passed onto the next FCN block for scaling the reduced map to the original dimension of $C \times 1 \times 1$. The output of the excitation operation is a set of scaling factors for each channel, represented by a weighted tensor of dimensions $C \times 1 \times 1$.

The weighted tensor can now be multiplied with the original feature map with dimensions $C \times H \times W$ to obtain the output of the SE module, i.e. a re-calibrated feature map. Within this study, SE modules with a reduction factor r of 8 have been incorporated in the final 3×3 convolutions, following the fusion of spatial and semantic information.

3.1. Implementation Details

The segmentation model was implemented with TensorFlow and Keras libraries. It was trained using the Nvidia Tesla T4 unit with 16 GB GPU memory. The model follows a DeepLabV3+ architecture with a choice of backbone between ResNet50 [13] and Xception [21], pre-trained on ImageNet weights [22]. For training datasets data augmentations applied as detailed in Section 4.

The training paradigm consisted of the Adam optimizer with a learning rate ranging from 1E–4 to 1E–7, batch-size of 16 and 80 epochs. The process was further configured with ReduceLROnPlateau with 5 epochs, a factor of 0.1 and early stopping with a patience of 20 epochs. L2 regularization was also used in the SE module with a factor of 1E–4. Incorporating an ensemble approach, an additional training methodology leveraged progressive refinement of weights, as depicted in Fig. 3.

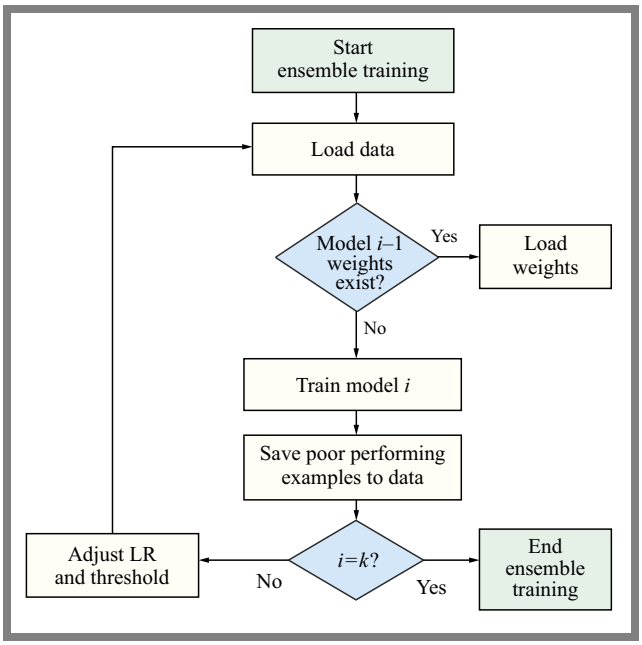

Fig. 3. Flowchart depicting ensemble training with i as the current model index and k representing the total number of models.

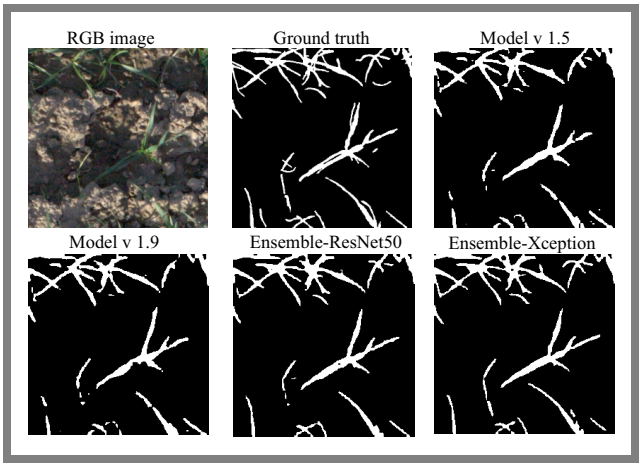

Fig. 4. Evaluation on EWS dataset.

Each subsequent model in the ensemble was initialized with the weights of the previous model. To increase robustness, poor performing examples present in the validation set, with the intersection over union (IoU) value below a certain threshold, were identified and oversampled at the end of each model's training. This intensified the model's subsequent exposure to challenging instances. To expedite model convergence, the learning rate for each new model in the ensemble was decreased (see Eq. (1)), aiding in achieving a dynamic adaptation mechanism. Additionally, the initial IoU threshold was increased, promoting a more deliberate learning process using the following equation:

$$\alpha_i = \frac{\alpha}{2^{(i-1)}}\,, \tag{1}$$

where: α_i – learning rate for model I, α – initial learning rate for the ensemble, and i – index of current model in the ensemble.

Tab. 1. Dataset details.

Dataset	Origin	No. of images	Plants	Imaging equipment	Conditions
EWS [23]	Eschikon, CH	190	Wheat	Canon 5D Mark II	Outdoor
CVPPP [15]	Various	810	Arabidopsis, tobacco	Varied cameras	Indoor
HIL [24]	Aarhus University, DK	144	Plant seedlings	Not specified	Indoor
BMS [19]	Chandigarh, IN	47	Various	Nikon D3300	Outdoor

4. Results

For training and evaluation of the model, three publicly available datasets were used: Eschikon wheat segmentation (EWS) [23], humans in the loop (HIL) plant segmentation [24] and computer vision problems in plant phenotyping (CVPPP) [15]. For a final evaluation with less than-ideal real-world conditions, a custom dataset called botanic mixed set (BMS) [19] was developed. The evaluation performed on the custom dataset aims to test the reliance of the model on the training datasets used, and to observe the model's ability to adapt to changing illumination, angle, and distance to subject within the images.

The datasets detailed in Tab. 1 are annotated with binary masks for soil/background and plant regions. The CVPPP dataset features subsets A1 to A4, offering distinct experimental settings for *Arabidopsis thaliana* and *Nicotiana tabacum* (Tobacco). Additionally, it undergoes a custom subset, combining subsets A1, A2 and A4 (totaling 267 images). The evaluation set for CVPPP consists of 63 images representing all of the subsets.

Data augmentation is a pivotal technique in enhancing the training efficiency of deep neural networks by artificially increasing the size of training set through transformations applied to the existing dataset. In study [25], various data augmentation techniques were tested within the context of CNNs, revealing their substantial impact on model training and evaluation. The chosen augmentation strategy plays a crucial role in improving the model's robustness by exposing it to a diverse range of scenarios. For this paper, data augmentation served the dual purpose of addressing the limited size of sourced datasets and aiding the model in effectively generalizing in response to previously unseen data.

Implemented through the augmentations Python library, the data augmentation process involved applying six transformations to each example in the dataset, resulting in six modified copies alongside the original example. This approach effectively increased the dataset size by a factor of six. The transformations included horizontal and vertical flipping, channel shuffling, random adjustments to brightness, contrast, rotation of the image by 45°, and a random crop. The resultant dataset was resized to 256 by 256 pixels.

It is crucial to note that the input images and masks underwent normalization to the [0, 1] range. Furthermore, the images were standardized using a mean of [0.485, 0.456, 0.406] and a standard deviation of [0.229, 0.224, 0.225] to adhere to ImageNet specifications. This pre-processing ensured compatibility and consistency in model training.

4.1. Metrics

For assessing performance of the model during the training and evaluation phases, intersection over union (IoU) and Dice coefficient were used. IoU, also known as the Jaccard index, is the main performance tool used in this work, and is a popular metric for segmentation tasks. It measures the accuracy of localizing objects by calculating the intersection of the predicted mask and the ground truth mask, dividing that by the union of the two:

$$IoU(A, B) = \frac{A \cap B + x}{A \cup B + x} . \tag{2}$$

Dice coefficient, also known as the Dice similarity index or Dice score, is another metric for image segmentation. It is used to quantify the similarity or overlap between two sets. In the context of image segmentation, it can be used to compare pixel-wise agreement between the predicted mask and its corresponding ground truth:

$$Dice\,(A, B) = \frac{2 \times (A \cap B) + x}{(A \cup B) + x} . \tag{3}$$

Dice loss is the loss function of the proposed model, derived from the Dice coefficient. This loss function encourages the model to produce masks that have a higher area of overlap with ground truth masks:

$$DiceLoss\,(A, B) = 1 - \frac{2 \times (A \cap B) + x}{(A \cup B) + x} . \tag{4}$$

In Eqs. (2) – (4), A is the predicted mask, B stands for the ground truth mask, and x is the smoothing factor.

The focus of the experiments was to enhance the generalization ability of the model across varying conditions. Dice loss, IoU, mean IoU (mIoU) and mean Dice coefficient (mDice) are the evaluation metrics used. It should be noted that the benchmarks presented further are results of training with a ResNet50 and Xception backbone, pretrained with the ImageNet weights.

4.2. Evaluation on EWS Dataset

In the process of hyperparameter and dataset tuning, an array of model versions was developed to obtain the best results. The best validation results achieved, measured based on IoU,

Tab. 2. Performance evaluation on EWS dataset.

Model version	Training	Backbone	IoU	Dice loss	mIoU
1.0	Singular	ResNet50	0.705	0.176	0.578
1.1	Singular	ResNet50	0.741	0.151	0.607
1.2	Singular	ResNet50	0.753	0.143	0.621
1.3	Singular	ResNet50	0.755	0.141	0.628
1.4	Singular	ResNet50	0.763	0.134	0.621
1.5	Singular	ResNet50	0.768	0.131	0.629
1.6	Singular	ResNet50	0.766	0.132	0.625
1.7	Singular	ResNet50	0.762	0.152	0.625
1.7.1	Singular	ResNet50	0.413	0.486	0.643
1.8	Singular	ResNet50	0.758	0.170	0.621
1.9	Singular	ResNet50	0.767	0.143	0.626
-	Ensemble	ResNet50	0.846	0.084	0.828
-	Ensemble	Xception	0.842	0.087	0.826

Tab. 3. Benchmark results and comparison with other papers related to EWS dataset.

Benchmark	IoU
Rico-Fernández *et al.* [26]	0.691
Zenkl *et al.* [23]	0.775
Yu *et al.* [27]	0.666
Sadeghi-Tehran *et al.* [28]	0.638
Proposed model	0.846

and their comparison with other publications, are presented in Tab. 3.

In [26], spatial context is provided in the form of a 5×5 window around individual pixels translated into CIELUV color space and input into a support vector classifier (SVC). Paper [23] used a DeepLabV3+ model with a ResNet50 backbone, feeding extra features as supplementary inputs. SVC was employed in the decision tree with preliminary weather state classification in [27]. The authors of [28] used a random forest classifier with the input having the form of 21 different color features. Note that the methods from [26]–[28] were reverse-engineered and tested by authors of [23] to obtain benchmarks on the EWS dataset. Table 3 presents a comparison of metrics between the aforementioned publications, while Tab. 2 presents a comparison of results between different model variations proposed in this paper.

As one may notice from Tab. 2, there is a significant decrease in the metrics in model v1.7.1. This decrease may be attributed to the number of dilated convolutional layers in the ASPP module. Model v1.7.1 saw an increase to 4 dilated convolution layers, but their respective rates were reduced to 4, 8, 12 and 16, in contrast to the higher dilation rates used in other model variations. This highlights the observation that the choice of dilation rates and the number of dilated convolutions may exert a significant impact on model performance and segmentation quality. The contrast between the Dice loss metric in model v1.5 and v1.9 is visible in that the former was configured with 3 dilated convolutions with rates of 12, 24, 36, and the latter was configured with 5 dilated convolutions with rates of 6, 12, 18, and 24. While IoU between the two remained close (0.768 vs. 0.767), the Dice loss varied (0.131 vs. 0.143). It should be noted that every other adjustable hyperparameter was kept identical for both model variations.

Since ASPP is crucial for capturing spatial information, lower dilation rates may help in paying attention to intricate details in the scene, while higher rates can assist in capturing wider plant variations. Despite the higher loss in model v1.9, it is able to capture the intricate details better, but may require more computation on account of more parallel dilated convolutions.

The ensemble training approach making use of ResNet50 yields superior metrics, as the progressive refinement of weights is coupled with increased exposure to sub-optimal instances. It leads to a final model characterized by superior segmentation quality and heightened resilience to diverse conditions. The Xception-backed ensemble model achieves slightly lower metrics than ResNet50, but was increasingly efficient at capturing the finer-grained details in the scene, paving the way for better segmentation quality.

4.3. Evaluation on HIL Plant Segmentation Dataset

The training process for HIL plant segmentation dataset (HIL-PS) followed the same procedure as EWS. The results of different model variations are presented in Tab. 4. While the model has demonstrated satisfactory performance on other datasets, it is crucial to take note of dataset-specific attributes. The lower performance metrics on the HIL-PS dataset can be attributed to a distinctive characteristic, specifically the relatively smaller size of plant specimens. Despite the model's capability to capture fine-grained details, the diminutive

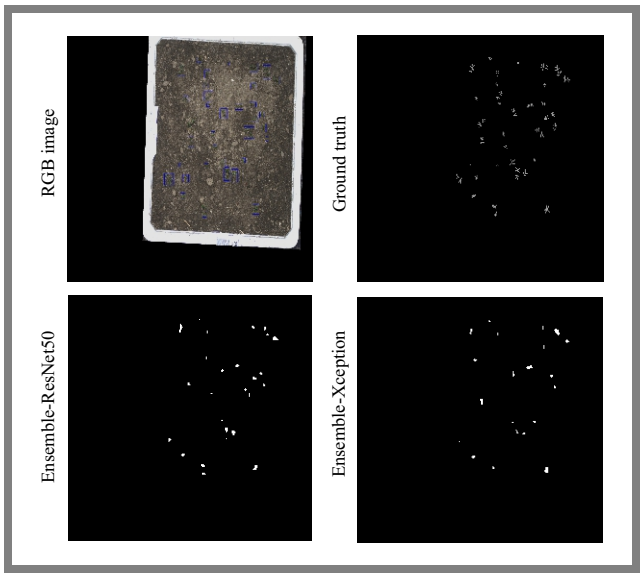

Fig. 5. Evaluation on HIL – diminutive samples.

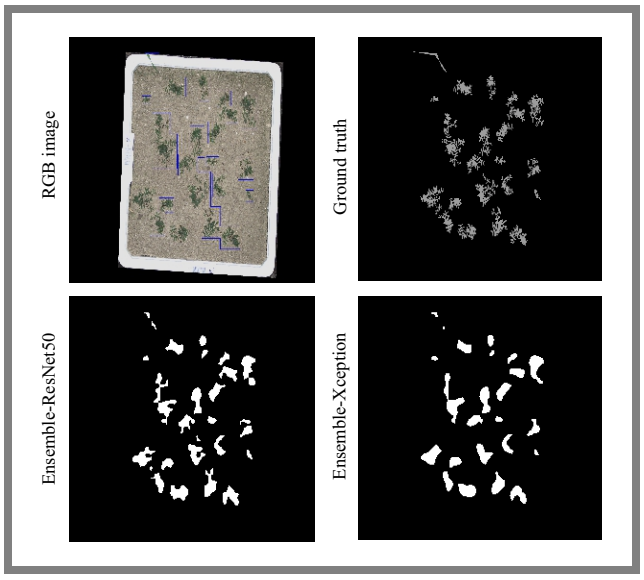

Fig. 6. Evaluation on HIL – optimal size samples.

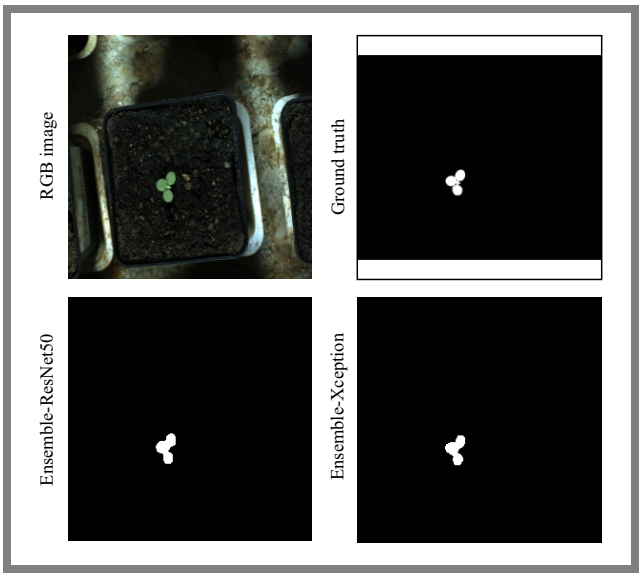

Fig. 7. Evaluation on CVPPP.

Fig. 8. Evaluation on CVPPP dataset.

proportions of the plants result in scenes with limited informative content. The model excels when dealing with scenes abundant with acquirable information, but faces difficulties in preserving segmentation quality for smaller-sized samples. The approach to ensemble training proved to be a pivotal enhancement. It not only led to improvements in performance metrics but also showcased superior segmentation quality for diminutive and optimally sized samples. The segmentation outcomes from the optimal model (ensemble) for this specific dataset are depicted in Figs. 5–6.

4.4. Evaluation on CVPPP Dataset

This subsection presents the results of models trained on different splits (A1, A2, A3, A4, custom split) of the CVPPP dataset. The evaluation split is formulated from a combination of splits known as A1 to A4 (63 images in total) – see Tab. 5. In Tab. 6, the precise dataset splits used for model training determine the observed variance in outcomes. The A1 split was mostly made up of *Arabidopsis thaliana* images, with plant specimens grown in pots, frequently accompanied by a dirt surface/background covered with green colored moss. This unique environment complexity most likely contributed to the poorer IoU and Dice loss metrics. The A3 split, on the other hand, revealed a data restriction with just 27 images of *Nicotiana tabacum* plants. Furthermore, the photos included plants with lower proportions and complex backgrounds, which made proper model training difficult.

Despite these obstacles, the A3-trained model demonstrated exceptional generalization skills on the CVPPP evaluation set, showcasing its adaptability to conditions exceeding the training limits. In contrast, the model trained on the custom split displayed exceptional metrics – as listed in Tab. 6 – for splits A1 to A4. Suitably sized and augmented splits increased its diversity and facilitated effective model generalization. Using an ensemble paradigm for custom split training has led to a small improvement in metrics and better generalization

Tab. 4. Performance evaluation on HIL-PS dataset.

Model version	Training	Backbone	IoU	Dice loss	mIoU
1.0	Singular	ResNet50	0.288	0.567	0.372
1.1	Singular	ResNet50	0.516	0.356	0.420
1.2	Singular	ResNet50	0.550	0.307	0.472
1.3	Singular	ResNet50	0.550	0.306	0.469
1.4	Singular	ResNet50	0.547	0.309	0.458
–	Ensemble	ResNet50	0.665	0.202	0.547
–	Ensemble	Xception	0.646	0.217	0.494

Tab. 5. Performance evaluation on CVPPP dataset.

Dataset	Training	Backbone	IoU	Dice loss	mIoU
A1	Singular	ResNet50	0.454	0.387	0.215
A2	Singular	ResNet50	0.915	0.044	0.635
A3	Singular	ResNet50	0.451	0.362	0.652
A4	Singular	ResNet50	0.921	0.043	0.812
Custom split	Singular	ResNet50	0.957	0.051	0.853
Custom split	Ensemble	ResNet50	0.975	0.013	0.859
Custom split	Ensemble	Xception	0.972	0.015	0.856

Tab. 6. Split-wise evaluation results.

Dataset	Training	Backbone	Evaluation – split	mIoU	mDice
Custom split	Singular	ResNet50	A1	0.930	0.964
			A2	0.816	0.884
			A3	0.880	0.919
			A4	0.915	0.955
Custom split	Ensemble	ResNet50	A1	0.937	0.967
			A2	0.844	0.907
			A3	0.889	0.924
			A4	0.921	0.958
Custom split	Ensemble	Xception	A1	0.922	0.960
			A2	0.787	0.861
			A3	0.898	0.943
			A4	0.911	0.953

across all splits. This shows the impact of dataset variables on model's performance and illustrates the efficiency of strategic augmentation and ensemble techniques in generalization across plant diversity. Figures 7-8 show segmentation results on the optimal model for this dataset.

4.5. Evaluation on BMS Dataset

The best-performing models from singular and ensemble training on EWS, HIL, and CVPPP datasets were further tested on the botanic mixed set (BMS). This was done to assess how well they could adapt to different training sets. Mean intersection over union (mIoU) and mean Dice coefficient (mDice) were the metrics used for this evaluation. The goal was to understand if these models could perform consistently across a diverse range of botanical specimens. The results shed light on the adaptability and reliability of the models in diverse settings. Through the metrics presented in Tab. 7, it is evident the EWS-trained model is the best performer. The model was able to capture a variety of plant structures in the image, such as leaves, long stems, etc. but struggled with capturing diminutive structures and often under- or over-classified the objects. This could be attributed to the nature of the data the model was trained on. The HIL-trained

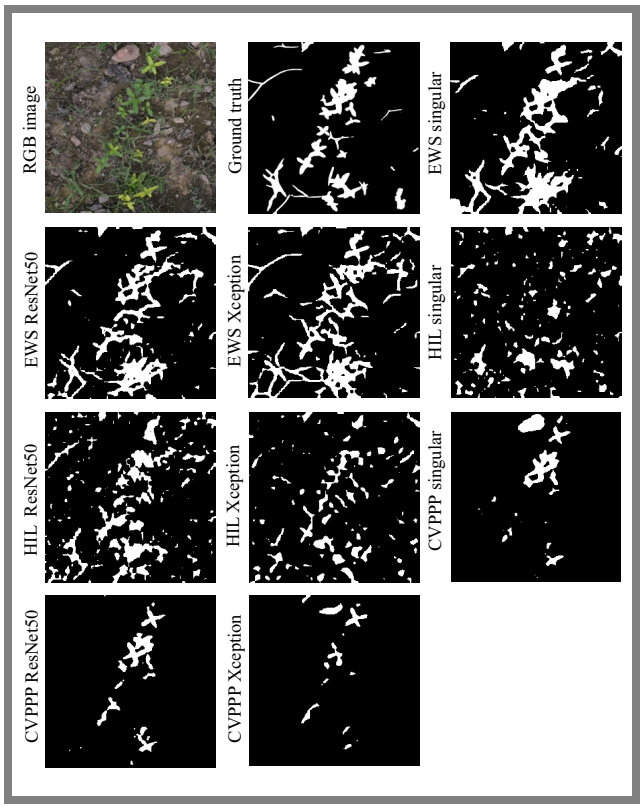

Fig. 9. Evaluation on BMS dataset.

Tab. 7. Performance evaluation on BMS dataset.

Dataset	Training	Backbone	mIoU	mDice
EWS	Singular	ResNet50	0.504	0.642
	Ensemble	ResNet50	0.503	0.643
	Ensemble	Xception	0.508	0.648
HIL	Singular	ResNet50	0.243	0.369
	Ensemble	ResNet50	0.385	0.524
	Ensemble	Xception	0.278	0.423
CVPPP	Singular	ResNet50	0.287	0.394
	Ensemble	ResNet50	0.248	0.332
	Ensemble	Xception	0.184	0.259

model was comparatively better at capturing the diminutive species but struggled with global context, hence could not segment larger plant structures. It was also observed that the model significantly struggled with complex backgrounds in the input images. The CVPPP-trained model was remarkable at capturing leaf structures and finer details but struggled with plants that did not resemble the structure of plants in its training set, i.e. long stems. It was observed that the ensemble models, even when not trained with augmented data, demonstrated superior metrics and segmentation quality in certain scenarios. This underscores their exceptional capacity to generalize across a dataset despite having fewer examples – see Fig. 9.

5. Limitations and Future Work

Deep learning models are capable of fitting large amounts of training data for generalization purposes. Their learning capabilities make them extremely susceptible to overfitting or underfitting in a scenario in which the dataset is smaller in size. The goal is to find a balance between the two and generalize well for a salient task. DeepLabV3+, while being a powerful semantic segmentation model, is relatively large and complex. Training with larger datasets and batch sizes may require significant amounts of computational power. The architecture also requires large datasets for effective training, which can be a bottleneck as datasets with good ecological diversity and quality annotations are difficult to find in the public domain. Adapting the architecture to a particular domain requires significant data and hyper-parameter tuning, both of which are time consuming.

The real-world generalization ability of each specific dataset-trained model was consistently meeting the expected standards. Factors impacting a given scene, such as lightning, background, angle of view, only minimally impacted the model's ability to recognize objects of interest. While each model had the ability to perform segmentation despite scene-related conditions, it is essential to acknowledge that not all models provided decent generalization for real-world scenarios. The training datasets lacked the size and diversity for such a precise task. Addressing such a limitation would require obtaining a dataset with more ecological diversity.

Plant specimens of smaller sizes also posed a limitation for the proposed method. Despite selected approaches to data augmentation and model-tuning, the segmentation of smaller plant structures proved to be challenging. Further research is required to address this constraint and potential directions may include the exploration of scale-aware models. A combined dataset incorporating examples from EWS, HIL, CVPPP and BMS could be leveraged for model training. Based on the results, relevant examples to maintain botanic diversity could be included or excluded from this new dataset to discourage class imbalance.

6. Conclusions

The proposed architecture performed semantic segmentation of plant images by incorporating an extended DeepLabV3+ model with a channel-wise attention mechanism. The work aimed to address generalization- and scene-related variations affecting task of segmenting plant images. The proposed model offers a powerful semantic segmentation solution emphasizing the features and leveraging semantic information with global context.

Several datasets were used to train their own versions of the model, which were further tested on the custom BMS dataset used for evaluation purposes. Adding an attention mechanism offered an increase in the quality of segmentation, when compared to earlier model-variations during dataset

and hyper parameter tuning. The models exhibited robustness to variations in lighting, backgrounds, and subject angles, showcasing their adaptability to real-world applications.

Additionally, even better results could be obtained by the inclusion of ensemble training. Through the application of ensemble approaches, the models showed exceptional resilience to real-world differences in lighting, backdrops, and subject angles, as well as outstanding generalization skills. The segmentation performance was much improved by relying on the ensemble training approaches.

Further research directions will be focused on improvements to the model's architecture and dataset composition. The current efforts are intended to expand on the accomplishments of the past and improve the model's ability to handle various problems that arise in practical implementations.

References

[1] M.A. Castillo-Martinez *et al.*, "Color Index Based Thresholding Method for Background and Foreground Segmentation of Plant Images", *Computers and Electronics in Agriculture*, vol. 178, 2020 (https://doi.org/10.1016/j.compag.2020.105783).

[2] D. Riehle, D. Reiser, and H.W. Griepentrog, "Robust Index-based Semantic Plant/background Segmentation for RGB-images", *Computers and Electronics in Agriculture*, vol. 169, 2020 (https://doi.org/10.1016/j.compag.2019.105201).

[3] D.M. Woebbecke, G.E. Meyer, K.V. Bargen, and D.A. Mortensen, "Color Indices for Weed Identification Under Various Soil, Residue, and Lighting Conditions", *Transactions of the ASAE*, vol. 38, pp. 259–269, 1995 (https://doi.org/10.13031/2013.27838).

[4] E.R. Hunt *et al.*, "Evaluation of Digital Photography from Model Aircraft for Remote Sensing of Crop Biomass and Nitrogen Status", *Precision Agriculture*, vol. 6, pp. 359–378, 2005 (https://doi.org/10.1007/s11119-005-2324-5).

[5] D. Zhang *et al.*, "A Universal Estimation Model of Fractional Vegetation Cover for Different Crops Based on Time Series Digital Photographs", *Computers and Electronics in Agriculture*, vol. 151, pp. 93–103, 2018 (https://doi.org/10.1016/j.compag.2018.05.030).

[6] J. Singh and H. Kaur, "Plant Disease Detection Based on Region-based Segmentation and KNN Classifier", *Proc. of the International Conference on ISMAC in Computational Vision and Bio-Engineering*, pp. 1667–1675, 2018 (https://doi.org/10.1007/978-3-030-00665-5_154).

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 640–651, 2017 (https://doi.org/10.1109/TPAMI.2016.2572683).

[8] L.G. Divyanth, A. Ahmad, and D. Saraswat, "A Two-stage Deep-learning Based Segmentation Model for Crop Disease Quantification Based on Corn Field Imagery", *Smart Agricultural Technology*, vol. 3, 2023 (https://doi.org/10.1016/j.atech.2022.100108).

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation", *Lecture Notes in Computer Science*, vol. 9351, pp. 234–241, 2015 (https://doi.org/10.1007/978-3-319-24574-4_28).

[10] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-decoder Architecture for Image Segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 2481–2495, 2017 (https://doi.org/10.1109/TPAMI.2016.2644615).

[11] L.-C. Chen *et al.*, "Encoder-decoder with Atrous Separable Convolution for Semantic Image Segmentation", *Computer Vision – ECCV*, vol. 11211, pp. 833–851, 2018 (https://doi.org/10.1007/978-3-030-01234-2_49).

[12] D. Ward, P. Moghadam, and N. Hudson, "Deep Leaf Segmentation Using Synthetic Data", *ArXiv*, 2018 (https://doi.org/10.48550/arXiv.1807.10931).

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA, 2016 (https://doi.org/10.1109/CVPR.2016.90).

[14] T.-Y. Lin *et al.*, "Feature Pyramid Networks for Object Detection", *ArXiv*, 2017 (https://doi.org/10.48550/arXiv.1612.03144).

[15] M. Minervini, A. Fischbach, H. Scharr, and S.A. Tsaftaris, "Finely-grained Annotated Datasets for Image-based Plant Phenotyping", *Pattern Recognition Letters*, vol. 81, pp. 80–89, 2016 (https://doi.org/10.1016/j.patrec.2015.10.013).

[16] M. Trivedi and A. Gupta, "Automatic Monitoring of the Growth of Plants Using Deep Learning-based Leaf Segmentation", *International Journal of Applied Science and Engineering*, vol. 18, 2021 (https://doi.org/10.6703/IJASE.202106_18(2).003).

[17] J. Fuentes-Pacheco *et al.*, "Fig Plant Segmentation from Aerial Images Using a Deep Convolutional Encoder-decoder Network", *Remote Sensing*, vol. 11, 2019 (https://doi.org/10.3390/rs11101157).

[18] S. Sharma, K. Verma, and P. Hardaha, "Implementation of Artificial Intelligence in Agriculture", *Journal of Computational and Cognitive Engineering*, vol. 2, pp. 155–162, 2023 (https://doi.org/10.47852/bonviewJCCE2202174).

[19] M.K. Surehli, N. Aggarwal, and G. Joshi, "Botanic Mixed Set", *GitHub*, 2023 (https://github.com/mukund-ks/botanic-mixed-set.git).

[20] J. Hu, L. Shen, G. Sun, "Squeeze-and-Excitation Networks", *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, USA, 2018 (https://doi.org/10.1109/CVPR.2018.00745).

[21] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions", *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1251–1258, 2017 (https://doi.org/10.1109/CVPR.2017.195).

[22] J. Deng *et al.*, "ImageNet: A Large-scale Hierarchical Image Database", *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, USA, 2009 (https://doi.org/10.1109/CVPR.2009.5206848).

[23] R. Zenkl *et al.*, "Outdoor Plant Segmentation with Deep Learning for High-throughput Field Phenotyping on a Diverse Wheat Dataset", *Frontiers in Plant Science*, vol. 12, 2022 (https://doi.org/10.3389/fpls.2021.774068).

[24] Humans in the Loop, "Plant Segmentation Dataset", (https://humansintheloop.org/resources/datasets/plant-segmentation).

[25] E. Pereira, G. Carneiro, and F.R. Cordeiro, "A Study on the Impact of Data Augmentation for Training Convolutional Neural Networks in the Presence of Noisy Labels", *35th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, Natal, Brazil, 2022 (https://doi.org/10.1109/SIBGRAPI55357.2022.9991791).

[26] M. Rico-Fernandez *et al.*, "A Contextualized Approach for Segmentation of Foliage in Different Crop Species", *Computers and Electronics in Agriculture*, vol. 156, pp. 378–386, 2019 (https://doi.org/10.1016/j.compag.2018.11.033).

[27] K. Yu *et al.*, "An Image Analysis Pipeline for Automated Classification of Imaging Light Conditions and for Quantification of Wheat Canopy Cover Time Series in Field Phenotyping", *Plant Methods*, vol. 13, 2017 (https://doi.org/10.1186/s13007-017-0168-4).

[28] E. David *et al.*, "Global Wheat Head Detection (GWHD) Dataset: A Large and Diverse Dataset of High-resolution RGB-labelled Images to Develop and Benchmark Wheat Head Detection Methods", *Plant Phenomics*, 2020 (https://doi.org/10.34133/2020/3521852).

Mukund Kumar Surehli, B.Tech.
School of Computing Science and Engineering
E-mail: mukund.28.k@gmail.com
VIT-Bhopal University, Madhya Pradesh, India
https://vitbhopal.ac.in

Naveen Aggarwal, Ph.D.
University Institute of Engineering and Technology
E-mail: navagg@pu.ac.in
Panjab University, Chandigarh, India
https://puchd.ac.in

Garima Joshi, Ph.D.
University Institute of Engineering and Technology
E-mail: joshigarima5@yahoo.com
Panjab University, Chandigarh, India
https://puchd.ac.in

Harsh Nayyar, Ph.D.
Department of Botany
E-mail: nayarbot@pu.ac.in
Panjab University, Chandigarh, India
https://puchd.ac.in