FB-PNet: A Semantic Segmentation Model for Automated Plant Leaf and Disease Annotation

Automated Plant Leaf and Disease Annotation

P Dinesh, Ramanathan Lakshmanan*
School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India

***Abstract*—Semantic segmentation is an important operation in computer vision, which is generally plagued by computational resources and the time-consuming process for labor intensive of pixel-wise labeling. As a solution to this issue, the present study introduces a state-of-the-art segmentation system based on the Forward-Backward Propagated Percept Net (FB-PNet) architecture, augmented with Perception Convolution layers designed specifically for this purpose. The suggested method improves segmentation precision and processing the efficiency by capturing fine visual features and reducing some unnecessary data. The performance of the model is tested using key evaluation metrics, including Intersection over Union (IoU), Dice coefficient, Loss, Recall, and Precision. Experimental results indicate that the model works effective in segmenting leaf and disease regions in plant images without requiring full pixel-by-pixel labeling. Data augmentation techniques also greatly improve the capability of the model to handle new situations. A strong partitioning technique of the dataset allows for best performance testing, demonstrating the strength and flexibility of the model with respect to new data in the PlantVillage dataset, even without the employment of annotation masks. The innovation of this research is an efficient and scalable approach to large-scale plant leaf and disease detection, which is able to sustain precision agriculture application cases.**

***Keywords*—Semantic segmentation; forward-backward propagated percept net; intersection over union; data augmentation**

I. INTRODUCTION

The recent advancement in the progress of computer vision and deep learning techniques has transformed the domain of plant phenotyping, which will enhance the capacity for precise and automatic measure of plant characteristics. Accurate leaf annotation should outline and distinguish the various parts of leaves in an image, necessary to examine the health, growth habits, and reaction of plants to environmental factors. While traditional methods can depend on time-consuming or semi-automatic processes that take several days, are labor intensive in bulk, and prone to human mistakes, having more high-resolution plant images to process, as well as advancements in advanced deep learning algorithms, and are motivating researchers to develop a model that can perform this task automatically.

Recent advancements in deep learning architectures and methods have immensely enhanced the process of automatic annotation of leaves. In the existing transformer models, the Pyramid Vision Transformer (PVT) and Swin Transformer have been unparallely effective in handling the dense prediction tasks by refine capturing of multi-scale features as well as long-range dependencies [1] [2]. These models perform very well in handling problems of overlapping leaves and complex leaf edges, which are prevalent in handling plant images. Further, these self-supervised learning techniques such as contrastive learning and data augmentation methods that performs operation such as Mixup, have reduced the reliance on large manually annotated datasets [3] [4]. These advancements bring the possibilities of developing accurate models even when handling limited annotated data into reality, and hence they are highly beneficial in plant phenotyping applications.

TransUNet and TransFuse are two models that combine Transformer-based encoders with CNN-based decoders to get cutting edge results in both medical and plant picture segmentation [5] [6]. Transformers and Transformative of the Convolutional Neural Networks (CNNs) model have made a significant contribution to the improvement of labeling accuracy, since the architectures such as TransUNet and TransFuse and some others combine based on the Transformer and Encoder Decoder Structure-CNN architecture to give better results in the delineation of medical and botanical images [5] [6]. Such designs, in turn, act as a supplement to the global context and comprehension that is obtained using transformers and enhance the power of localized feature extraction using CNNs and are well suited for the annotation of plant leaves. In fact, failures have arisen from the integration of attention mechanisms in certain regions to enhance their performance in complex environments [7]. The novelties of these specialized neural networks, distinct from typical convolutional ones, involve finesse in handling images that possess unstructured or uneven data, thus improving their effectiveness in annotating other classes [4].

In spite of these positive developments, there are many issues that need to be addressed when automating leaf analysis in plant leaves. The variations in shapes, sizes and textures of different leaves, also the presence of leaves above others and obscure backgrounds, make accomplish segmentation hard [8]. Several approaches have been taken to bring an end to these difficulties, such as the use of some data modification methodologies flips, shifts and scaling, and rotation such that the dataset is enhanced in terms of having variety [4]. Moreover, there was a new focus on artificial intelligence (AI) in plants, where additional training data was generated by the generative adversarial network (GAN) methods, and there were reasons to be hopeful that model performance could be better

than with the previously used solutions [9]. These methods are empowering in practical cases in which there is limited annotated data. Moreover, several other studies have also been aimed at making annotated models more efficient and more applicable. For example, the development of SegFormer utilizes the same principles of the Transformers that work best in an encoder but in a relatively simple decoder for reasons of minimizing the computational work while still capturing the necessary segments [10]. In the same way, DenseCL applies contrastive learning to dense prediction tasks, such that the need of self-supervised trainings for segmentation models is no longer required [11]. The application of these methods made it possible to introduce the use of annotation models to actual field practices in agriculture despite the inevitable geographical constraints.

Traditional convolution-based and transformer-based models have issues in separating the leaves especially when they overlap, have innumerable little leaves, and are differential in texturing in consequences. This results in suboptimal annotation due to the large-scale labeled datasets overused by most advanced segmentation models. Additionally, computational efficiency is a big question, the reason being this is that, normally such high-performing models usually demand huge processing power, restricting their use on resource-constrained devices. Moreover, techniques available at a given time may become immediately outdated when applied to different plants under different environmental conditions. This is because the conditions of the environment, lightning, occlusions including background complexity change.

To address these limitations, this research introduces a novel deep learning model called Forward-Backward Percept Net (FB-PNet) for automatic annotation of plant leaf and disease. The model comprises a pre-trained ResNet50-based U-Net architecture to which some convolutional layers are added for more enhanced feature extraction and improved segmentation accuracy. Developed using PyTorch Lightning and Segmentation Models for PyTorch, the proposed architecture was examined with important metrics of performance, including Intersection over Union (IoU), dice coefficient, precision, and recall. During the inference phase, an input image is given to the FB-PNet model. In each forward pass the image is forwarded through all layers to produce feature representation known as Percepts. These Percepts are then selectively filtered with the input image to produce the final rendition of the segmentation. Extensive training and validation confirms that the model is strong and able of generalizing over several plant datasets.

This research automated the plant phenotyping and shows a significant increase in the field of intelligent and effective leaf recognition. In this context, the outcomes of the given study are expected to be useful for such areas as precision farming, automatic plant monitoring, and protection of biodiversity, thus advancing artificial intelligence adoption in the field of plant studies. With the automation of the leaf annotation process, this work aims to increase the productivity level in agriculture while promoting sustainability. The model suggested not only mitigating the shortcomings in conventional methods of annotation, but also utilizes cutting-edge advances in deep learning to attain higher accuracy and working efficiency.

Further this study continues with previous study related to the work in Section II, description of the dataset in Section III, methodology of the proposed work in Section IV, followed by experimental analysis and conclusion in Section V and Section VI respectively.

II. Related Work

The task of automatically annotating properties of plant leaves and disease has gained most significant attention in recent years due to its critical role in plant phenotyping, disease detection, and species identification. The entry of deep learning along with computer vision technologies has revolutionized this area allowing very efficient techniques for leaf, disease segmentation and as well as related processes to be developed and implemented. The paragraph given below discuss about the key advancements and recent research in this domain.

A. Evolution of Deep Learning for Segmentation Models

Conventional methods in annotation of plant leaves depended on image processing techniques such as thresholding, edge detection or region-growing algorithm. These techniques, however, do not work effectively because of the variability of shapes, overlapping structures or complex backgrounds among leaves. The entire advancement starts with fully convolutional networks (FCNs), which allowed end-to-end training for pixel-wise segmentation [12]. They are the predecessors of today's segmentation applications of deep-learning models such as U-Net or DeepLab models. One of them, U-Net, introduced by Ronneberger et al. [13], has come under most use, as its encoder-decoder architecture combined with skip connections makes it a good option for capturing local-global features. The attention U-Net and Residual U-Net are contemporary refinements that are currently enhancing segmentation accuracy through attention mechanisms and residual connections [14].

The model DeepLab is proposed by [15] and followed this atrous convolution and Conditional Random Fields (CRFs) to depict fine details and it is used to facilitate boundary delineation to enhance segmentation. The encoder-decoder part was incorporated in DeepLabv3+ and they achieve state-of-the-art results in plant leaf segmentation. On the other hand, extending Faster R-CNN, Mask-RCNN brought instance segmentation with a branch for pixel-wise mask prediction [16]. The methodology has been frequently employed for plant leaf annotation, especially in conditions of overlapping leaves and dense foliage.

In the last few years, attention mechanisms have been included in segmentation models, which improved performances in noisy or complex scenarios and also focusing on the region of interest. For instance, Partial Convolutions (PConv) have been able to deal with incomplete or damaged data sources, making them very useful for real-life plant leaf annotation tasks [17]. These researches have all revolved around the bulk enhancement and improvement concerning the robustness and efficiency of automatic annotation models. Transformer-based architectures have been pushed forward on

segmentation tasks as they capture global dependencies [18]. In parallel, self-supervised learning models have been reviewed in terms of reducing dependence on widely annotated datasets [19].

B. Advanced Learning Techniques

In [20], the authors propose a method for image segmentation to imply the ability of multi-feature interaction and fusion techniques contained within the cloud framework. This method, through that, added a better segmentation accuracy provided by the interaction of several attributes of an image and their interdependence. Besides, the segmentation is parallelized and optimized using cloud computing resources to push computation efficiency so that the rapid and accurate processing of medical images can occur.

An advanced deep learning network was also developed by [21], aimed at the voxel-level processing and interlayer connections, as well as intra-axis feature extraction. The proposed model, thus, is capable of dynamically learning the 3D spatial properties while complementing fine-edge delineation. Similarly, the work proposed in [22] involves inter-anatomical domain significance, deep reasoning brain tumor segmentation, and implementation based on the Swin-T architecture. This approach primarily consists of a backbone hybrid network (BHN) and a deep micro-texture extraction module (DMTE) for improved segmentation accuracy. Additionally, [23] introduce a CNN-based brain tumor segmentation method that integrates an MIE module to enhance the utilization of multi-modality data.

Zero-label segmentation, as proposed by [24], employs a self-training mechanism in iterative manner, where the model is initially trained on labeled datasets and subsequently generates pseudo-annotation for unannotated data, refining its segmentation accuracy through continuous learning. Meanwhile, Few-shot learning, as discussed by [25], focuses on model development with a minimal training dataset. It aims to segment query images using a limited number of reference samples.

C. Computational Constraints and Challenges in Annotation

Deep learning architectures typically necessitate a substantial set of parameters to attain higher precision, which can lead to prolonged training times and increased computational overhead. To mitigate this issue, researchers have designed various foundational network architectures, known as backbones, which serve as the core framework for different models. In [26], the authors' models incorporate specific backbone networks such as ResNet101, ResNet50 and MobileNetV3. In the domain of segmentation task, MobileNetV3 is widely utilized as a lightweight backbone suitable for embedded and mobile systems, whereas ResNet101 and ResNet50 are preferred in scenarios demanding high accuracy, although they come with greater computational complexity and memory consumption.

Inherent limitations are common among the above mentioned methodologies. The aim to attain absolute accuracy and to build a finer and stronger model relies heavily on a greater number of computational parameters, which in fact greatly adds to the computational load on the system, rendering it resource intensive. The other side of this is that it makes the model computationally expensive, confirmed by memory consumption during execution.

It's the manual laborious and time-consuming label-to-image annotation process which becomes the sought-after annotation for the semantic segmentation process. That is assigning class labels to each pixel of an image and thus requiring a steady control over the way. This should be done because these will require pixel-level accuracy in the mind of the annotators. The interpretation of the semantics of an image may differ from one annotator to the next, leading to inconsistency in annotating it. This subjectivity also causes variations in the dataset and is a challenge in setting a standard ground truth.

Moreover, objects with very complex boundaries or irregular geometric structures, such as trees or animals, require accurate contour delineation, thus adding to the complexity of the whole annotation task. When these complexities are compounded further by the endeavors in handling the large-scale datasets, more computational power and human labor become necessary. Domain experts would have to be recruited to ensure the accuracy of the annotation process. The training and maintenance of consistency of annotations across a large dataset have always been difficult. Human annotators would themselves introduce errors in pixel classification and also inconsistencies in annotations, which would call for intense quality control. The whole manual labeling procedure for semantic segmentation is thus laden with many challenges, such as very high-resolution pixel-wise accuracy, subjective interpretation, tricky object boundaries, scalability of the datasets, and the need for expert annotators and guaranteed quality measures.

Additionally, models such as Percept-CNN (P-CNN) introduce percepts highly activated pixels extracted during forward propagation to focus on salient visual information [27]. While promising, P-CNN struggles with over-segmentation at object borders and requires validation for multi-class tasks. Enhancing the model with multi-scale percept extraction and attention mechanisms could improve its robustness and accuracy.

Various existing segmentation models along with their categorization, advancements and correlations are illustrated in Fig. 1, where the overlapping regions of the models reflect shared characteristics between distinct methodologies. The models highlighted in the figure include traditional segmentation model, deep learning-based models, lightweight model, instance, hybrid, attention-based models as well as self-supervised models. Fig. 1 leads to a reference point for the work as it demonstrates the interrelation of each model and thus provides an insight towards the incorporation of different algorithm in the proposed work to solve the gap of manual annotation in the existing works. This study proposes FB-PNet, a hybrid self-supervised lightweight segmentation model, and demonstrates the efficiency of the proposed model with respect to the segmentation task.

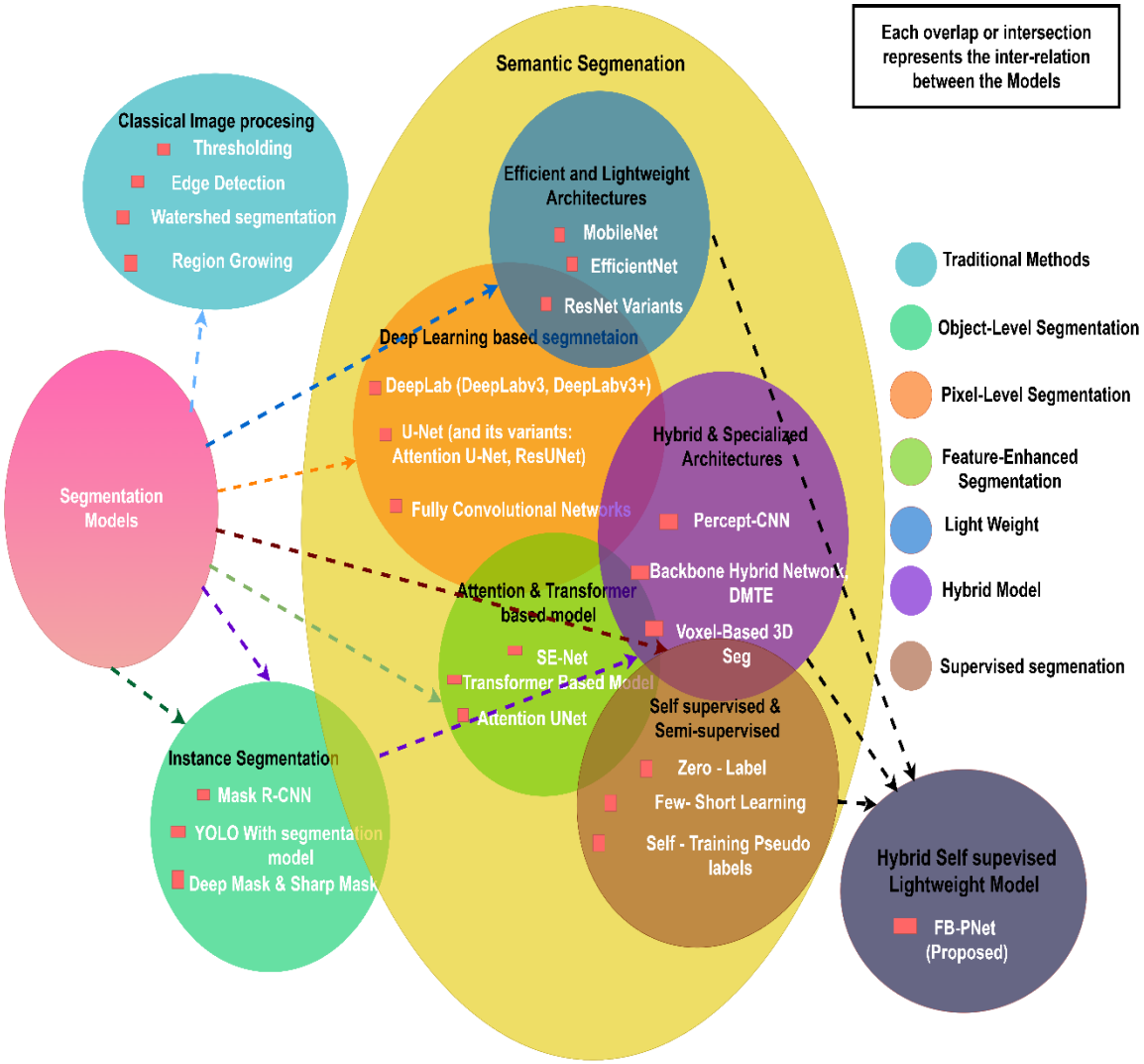

Fig. 1. Evaluation overview of segmentation.

Overall, deep learning-based segmentation techniques have significantly improved plant leaf annotation, challenges related to computational efficiency, annotation labor, and dataset scalability remain critical areas for future research and optimization. These limitations are prevalent across all the discussed methodologies.

III. Dataset Description

The dataset is created from the existing citrus leaf dataset which consists of around 500 images [28]. Further the images of the new dataset is created by adding two distinct mask to support the sematic segmentation process. The proposed segmentation and analysis in the system make leverage of two distinct dataset with images containing leaf mask and disease mask, later the dataset is divided into train, valid, and test splits for the segmentation process. To enhance model generalization and performance, various levels of augmentation were applied, resulting in multiple dataset versions. The dataset used in this work is available in zenodo [29].

Original Datasets of Leaf and Disease: The initial dataset comprises 568 images for both leaf and disease segmentation, respectively [30]. Augmented Datasets of leaf and disease: To increase dataset diversity and to enhance the model's adaptability across diverse conditions, data augmentation were implemented, generating a larger dataset. The augmentation techniques included transformations such as rotation, flipping, scaling, brightness adjustments, and contrast enhancement. This resulted in 1,702 augmented images for both leaf and disease datasets. Fully Augmented Dataset: A final augmentation stage was conducted to further expand the dataset, leading to 3,405 images for both leaf and disease segmentation. These datasets information and splits are mentioned in Table I.

A. Significance of Dataset Augmentation

Initially, the images undergo a preprocessing phase involving augmentation. This technique generates altered versions of the original images by applying various transformations. The progressive augmentation strategy ensures that the model is trained on a diverse set of images, reducing overfitting and enhancing its capability to accurately segment leaves and classify diseases under varied environmental conditions, lighting variations, and occlusions.

By leveraging a structured dataset split, model performance is rigorously evaluated, ensuring robust generalization to unseen data in plant village dataset [31]. Augmentation enhances the model's exposure to diverse training samples, thus enhancing its capability to generalize effectively to unseen or real-world conditions. This dataset framework plays a crucial role in improving segmentation accuracy, thereby contributing to precise automatic plant leaf annotation using deep learning. The sample images on leaf and disease dataset is shown in Fig. 2.

TABLE I. OUTLINES THE VARIOUS DATASET AND ITS SPLIT FOR TRAINING, VALIDATION AND TESTING

Dataset	Total No. of images	Training set	Validation set	Test set
Leaf dataset	568	464	52	52
Disease dataset	568	464	52	52
Augmented Leaf dataset	1702	1393	155	154
Augmented Disease dataset	1702	1393	155	154
Fully Augmented Leaf dataset	3405	2786	310	309
Fully Augmented Disease dataset	3405	2786	310	309

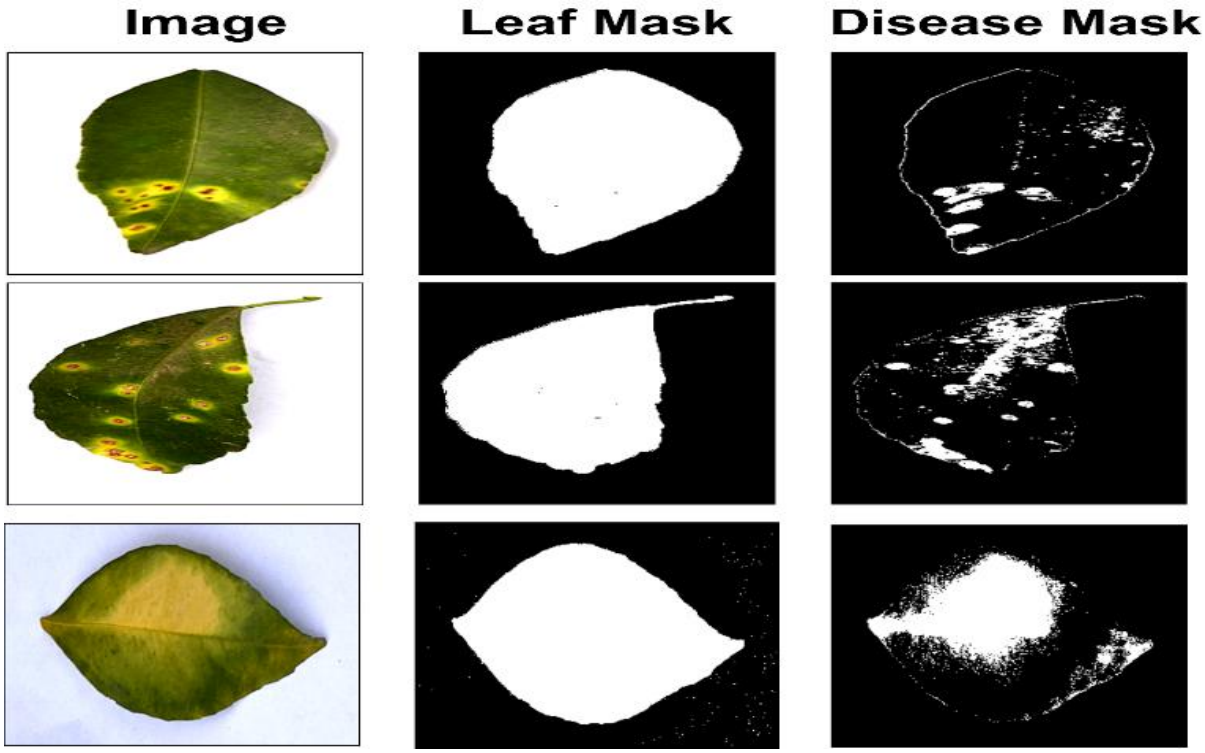

Fig. 2. Data samples with the image in the first column and corresponding annotation mask of leaf and disease in second and third column.

IV. METHODOLOGY

Forward-Backward Propagated Percept U-Net (FB-PNet) is using a Pre-trained U-Net for segmentation. The U-Net encoder extracts multi-scale features of leaf or diseases, while the decoder reconstructs spatially - detailed feature maps for segmentation. The use of SMP U-Net from segmentation models pytorch simplifies development by leveraging a modular, pre-built network. Custom layers perception convolution (PConv) for refinement is used after the U-Net decoder output, and additional custom PConv layers are applied. These layers refine the segmentation predictions, allowing for specific feature transformations beyond the U-Net's capabilities. This model is based on Single-task pipeline and it focuses purely on segmentation with a binary mask output. The forward pass outputs logits for the segmentation task, followed by a loss function BCEWithLogitsLoss for binary segmentation. FB-PNet Model is lighter, leveraging a well-tested U-Net for segmentation, making it easier to train and apply to binary segmentation tasks. It focuses on a streamlined segmentation task with a U-Net backbone, making it simpler but task-specific. The proposed model eliminates the necessity for pixel-wise annotation by leveraging a classification dataset to execute the segmentation task. FB-PNet captures and transfers essential visual characteristics throughout the layers, facilitating their utilization in segmentation processes, where the goal is to get visual characteristics in the input image and it gives binary segmentation output. Fig. 3 describes the complete block diagram of the FB-PNet architecture.

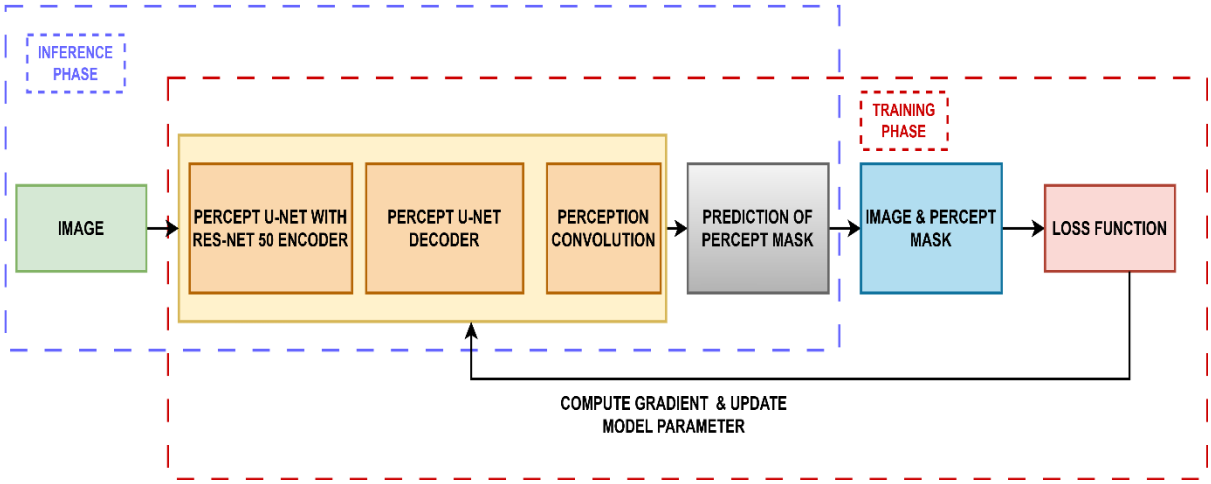

Fig. 3. Block diagram of proposed model.

In the Input Image, where X is an input tensor image of shape [B, C, H, W], where B is batch size, C is the number of channel, H is the height and W is the width. A batch of RGB images with shape (batch_size, 3, height, width).

In the process of the Forward Pass FB-PNet Encoder which is hierarchical, multi-scale Features are extracted at multiple scales using the ResNet50 encoder. Early percept layers focus on low-level features like edges, textures etc. Deeper percept layers capture high-level features that are the object shapes, contextual information. Output from the percepts gives multi-scale feature maps. In these, $F_{enc}^{(i)}$ is the Percept Feature map at stage i of FB-PNet encoder. Then these input X is processed through the encoder, generating a series of Percept feature maps [see Eq. (1)].

$$\{F_{enc}^{(1)},\ F_{enc}^{(2)}, \dots\dots, F_{enc}^{(n)}\} = Encoder\ (X) \quad (1)$$

Models decoder is percept Multi-Scale Feature Combination. The decoder combines these percept multi-scale features to generate segmentation features. Percept Features from deeper layers are upsampled to match the spatial resolution of shallower layers. These upsampled features are concatenated with corresponding encoder features using skip connections. This combination ensures that both low-level percept detailed and high-level semantic information is preserved. In these $F_{dec}^{(i)}$ is the Percept Feature map at stage i of the BPNet decoder. The decoder takes the Percept Feature maps from the encoder and reconstructs the F_{BPNet} feature map [see Eq. (2)].

$$F_{BPNet} = Decoder\ \{F_{dec}^{(1)},\ F_{dec}^{(2)}, \dots\dots, F_{dec}^{(n)}\} \quad (2)$$

Refinement of segmentation features in this Percept Convolution PConv1 and PConv2 further refine the segmentation features. Enhance the representation of fine details of leaf and diseases. Improve the distinction between leaf, diseases and background regions, where W_{LPConv}^{k} is the learnable weights of LPConv and $k \times k$ convolution. σ_l: $Leaky\ relu$ activation function. After obtaining F_{BPNet} from the BPNet, the following convolution operations are applied. In the first 3×3 PConv layer with LeakyRelu activation as in Eq. (3):

$$F_{PConv1} = \sigma_l(W_{PConv1}^{3\times3} * F_{BPNet}) \quad (3)$$

The second PConv layer is another 3×3 convolution followed by LeakyRelu activation which is applied to the output of the first PConv layer [see Eq. (4)].

$$F_{PConv2} = \sigma_l(W_{PConv2}^{3\times3} * F_{PConv1}) \quad (4)$$

Segmentation prediction for leaf or disease is a refined feature in a single-channel segmentation map, the prediction in the model refers to assigning each pixel of the input image a class label or a probability of belonging to a leaf or diseases class. Final segmentation mask is produced by the pred layer using Conv2d. A Conv2d layer reduces the number of channels to 1, corresponding to the binary classification for each pixel in the leaf image. Kernel size of 1x1, which ensures that the spatial dimensions (height, width) remain unchanged. Each pixel has a value between 0 and 1 after applying a sigmoid. Activation: Sigmoid, applied to the output of this layer to produce probabilities for each pixel. A value close to 1 at pixel (i, j) row and column indicate a high probability of the pixel (i, j) row and column belonging to the target leaf or disease class. Values close to 0 indicate a low probability, it belongs to the background. The output shape is (batch_size, 1, height, width).

In the final prediction layer a 1×1 convolution is applied to reduce the number of the channels to 1 for the binary segmentation logits [see Eq. (5)].

$$F_{Pred} = W_{Pred}^{1\times1} * F_{PConv2} \quad (5)$$

Here, the overall FB-PNet models forward pass of combining all steps is summarize [see Eq. (6)].

$$F_{Pred} = W_{Pred}^{1\times1} * \sigma_l\left(W_{PConv2}^{3\times3} * \sigma_l(W_{PConv1}^{3\times3} * F_{BPNet})\right) \quad (6)$$

The prediction Y of the annotation is obtained after applying the sigmoid activation into the final prediction of FB-PNet model. It gets the result of the predicted annotation.

$$Y = \sigma(F_{Pred}) \quad (7)$$

Perception refer to pixels that encapsulate significant visual information extracted after each layer of the network. These pixels correspond to regions with high activation values in the associated feature maps. The output of this operation is

subsequently processed using a sigmoid activation function to enhance the perception of essential visual components. Mathematically, this is represented as Eq. (7). The application of the sigmoid function further emphasizes the most relevant pixels in the input image, which are identified as Perception.

It enables the model to concentrate exclusively on pixels containing critical visual information, as illustrated in Fig. 4. This figure provides a conceptual representation of how an FB-PNet layer processes an image using a specific filter. To illustrate, consider a filter designed to detect the part of leaf. When an image is processed through FB-PNet using this filter, it generates a feature map and a corresponding perception mask. For intuitive understanding, Fig. 4 presents the original image overlaid with the perception mask.

Fig. 5 shows an illustration of perception generation. A leaf image is processed with a filter designed to detect a part of leaf and disease parts. The output includes a feature map and a percept mask, highlighting pixels corresponding to the detected structure.

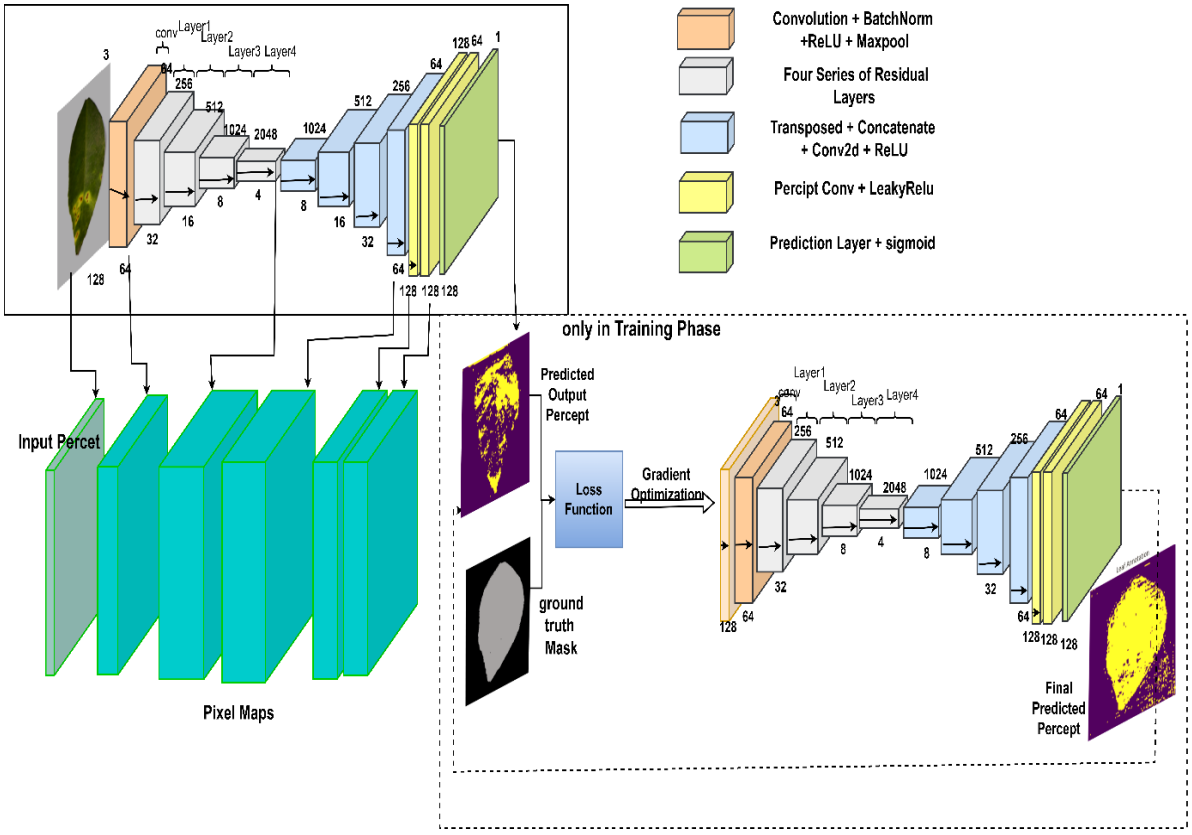

Fig. 4. Forward-backward propagated perception net.

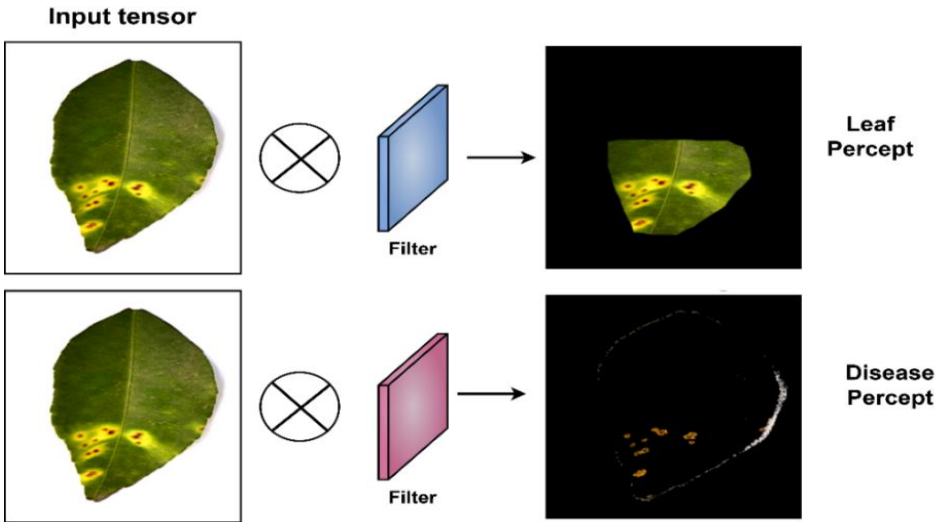

Fig. 5. Perception generation.

Then the output logits F_{Pred} are passed to the binary cross-entropy loss function with the logits, this will help the model to get the details of the models lagging through the loss function. $\mathcal{L}_{BCE}$ represents the binary Cross-Entropy loss, M_i is the ground truth mask of the input leaf or disease and N represents the total count of pixels in the input image tensor [see Eq. (8)].

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[\, M_i \,.\log(Y) + (1 - M_i)\,.\log(1 - Y)] \quad (8)$$

Then the information from the $\mathcal{L}_{BCE}$ loss obtained on the forward pass is getting into backward pass. In backward pass gradients of the loss are computed with respect to model parameters using backpropagation. Here, Adam optimizer updates the model parameter regarding the loss obtained in the previous inference of the model, it is to minimize the loss and increase the predicted annotation mask.

The comprehensive architecture of FB-PNet is illustrated in Fig. 3. The module positioned in the lower right section of the architecture serves as a classification unit, employed solely during the training phase for weight refinement. During the inference stage, the input image undergoes processing through FB-PNet, resulting in the generation of a perceptual mask or output perception, as illustrated in Fig. 4. The final perceptual mask represents the segmentation output.

V. Experimental results and discussion

To evaluate the effective performance of the proposed deep learning model for automatic plant leaf annotation, the FB-PNet model was trained for 100 epochs using the ResNet50 U-Net backbone integrated with Perception Convolution layers (PConv). A standardized training setup was maintained across all datasets to ensure consistency in evaluation for FB-PNet. The parameters IoU, dice coefficient, precision and recall is used for evaluation and comparison, since these metrics are better in determining the quality of segmentation.

This configuration included: all these models training and evaluation process were executed on the GPU available in Google colab. This GPU provides more disk space, which offered enhanced computational resources and sufficient storage capacity for efficient handling of large datasets required on training the model. The model was trained on a single GPU using CUDA version 11.8 to ensure efficient computation. Training was conducted for a maximum of 100 epochs, with early stopping applied if validation performance did not improve for 20 consecutive epochs. To optimize data loading and preprocessing, four worker processes were utilized. A batch size of 16 images was used during training to balance computational efficiency and memory constraints. Initially the learning rate was configured to 0.0001, with a ReduceLROnPlateau scheduler dynamically adjusting the learning rate by a factor of 0.5, if validation loss did not improve for 5 epochs. The Adam optimizer was employed, along with weight decay (0.00001) to enhance generalization. Additionally, gradient clipping with a value of 1.0 was applied to prevent exploding gradients. No warm up strategy was applied, as the model converged effectively with the selected learning rate and optimizer settings.

The trained proposed FB-PNet model is compared in Table II which represents the segmentation performance of different models across various dataset configurations, including original, augmented, and fully augmented leaf and disease datasets. The results demonstrate that our proposed Forward-Backward-Propagated Percept Net consistently outperforms existing architectures, achieving the highest performance across all dataset variations. Notably, our model exhibits a significant improvement in segmentation accuracy on test data, particularly in the fully augmented dataset, where it achieves an IoU of 0.9802 and 0.8680 for leaf and disease datasets, respectively. The superior performance of our approach highlights the effectiveness of incorporating Perception Convolution layers and advanced feature propagation mechanisms. In contrast, traditional models like P-CNN show the lowest performance across all datasets, while other deep learning models such as UNetResNet50 and UNetEfficientNetB0 show moderate improvements but still fall short of FB-PNet's results. This highlights FB-PNet's effectiveness in handling complex segmentation tasks, especially when trained on augmented data. As a result, the model's performance was evaluated using the training and validation datasets, and the findings are discussed.

A. Intersection over Union (IoU)

IoU quantifies the similarity between the predicted and actual masks by measuring their overlap [Eq. (9)].

$$IoU = \frac{|M \cap Y|}{|M \cup Y|} = \frac{TP}{TP+FP+FN} \quad (9)$$

TABLE II. Performance Comparison of Different Segmentation Model Across Various Test Dataset

	Leaf dataset (500+ images)	**Disease dataset (500+ images)**	**Augmented Leaf dataset (1500+ images)**	**Augmented Disease dataset (1500+ images)**	**Fully Augmented Leaf dataset (3000+ images)**	**Fully Augmented Disease dataset (3000+ images)**
P-CNN	0.5119	0.4113	0.4752	0.4839	0.5223	0.5203
UNetResNet34	0.8206	0.5535	0.6874	0.6845	0.9406	0.6535
DeepLabV3PlusResNet50	0.8902	0.6575	0.6307	0.6228	0.9106	0.6307
UNetEfficientNetBo	0.9091	0.6689	0.8959	0.6838	0.9381	0.7268
UNetResNet50	0.9325	0.6816	0.9226	0.7100	0.9506	0.7375
Forward-Backward-Propagated Percept Net **(Ours)**	**0.9391**	**0.7306**	**0.9589**	**0.7174**	**0.9802**	**0.8680**

where,

- TP (True Positives) are correctly predicted leaf or diseased pixels.
- FP (False Positives) are background pixels wrongly predicted as leaf or diseased.
- FN (False Negatives) are leaf or diseased pixels wrongly predicted as background.

B. Dice Coefficient (F1-Score for Segmentation)

The Dice coefficient evaluates the similarity between the predicted and ground truth masks [Eq. (10)].

$$Dice = \frac{2|M \cap Y|}{|M| \cup |Y|} = \frac{2TP}{2TP+FP+FN} \quad (10)$$

Dice is closely related to IoU but gives more weight to correctly predicted pixels.

C. Precision

Precision measures how many of the predicted positive pixels are actually correct [Eq. (11)].

$$Precision = \frac{TP}{TP+FP} \quad (11)$$

D. Recall

Recall measures how many of the actual positive pixels were correctly predicted [Eq. (12)].

$$Recall = \frac{TP}{TP+FN} \quad (12)$$

Table III presents the FB-PNET final training and validation performance metrics at epoch 100, including Loss, IoU, Dice Coefficient, Precision, and Recall.

TABLE III. FB-PNET TRAINING AND VALIDATION METRICS AT EPOCH 100

Metric	Training	Validation
Loss	0.5218	0.5166
IoU	0.8071	0.7010
Dice	0.8928	0.8215
Precision	0.8094	0.7038
Recall	0.9966	0.9938

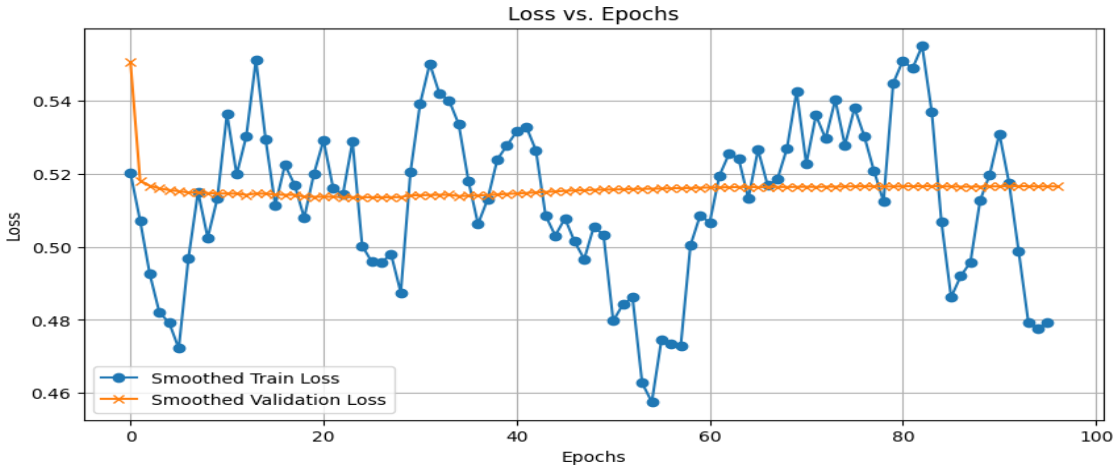

Fig. 6. Training and validation of loss of FB-PNet model for leaf annotation.

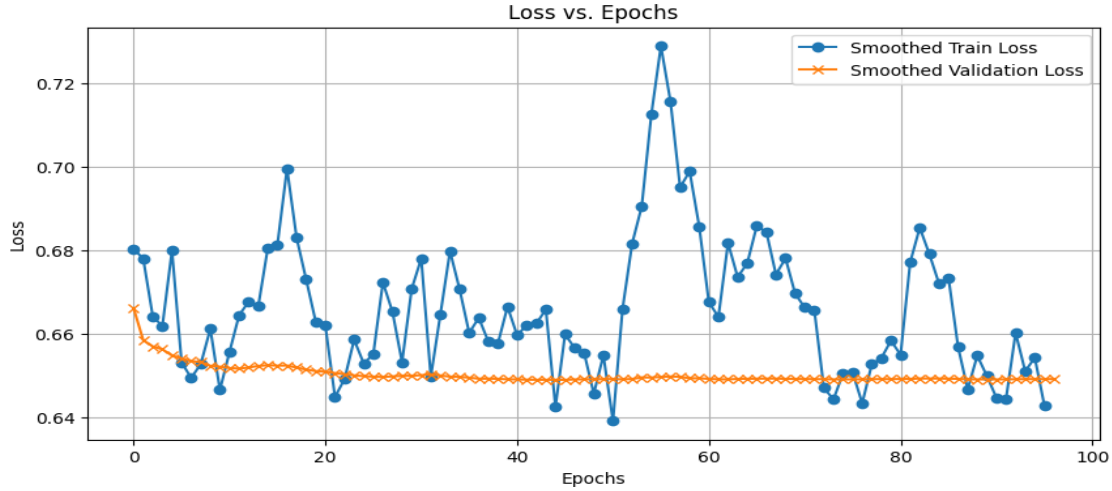

Fig. 7. Training and validation of loss of FB-PNet model for disease annotation.

Fig. 6 and Fig. 7 illustrate the training and validation loss trends for leaf annotation and disease annotation, respectively. The validation loss remains stable, while training loss exhibits fluctuations, indicating the model's learning behavior and adaptation to complex features. Fig. 8 and Fig. 9 show the IoU trends for leaf and disease annotations, where the validation IoU remains relatively stable while training IoU fluctuates, indicating the model's robust generalization.

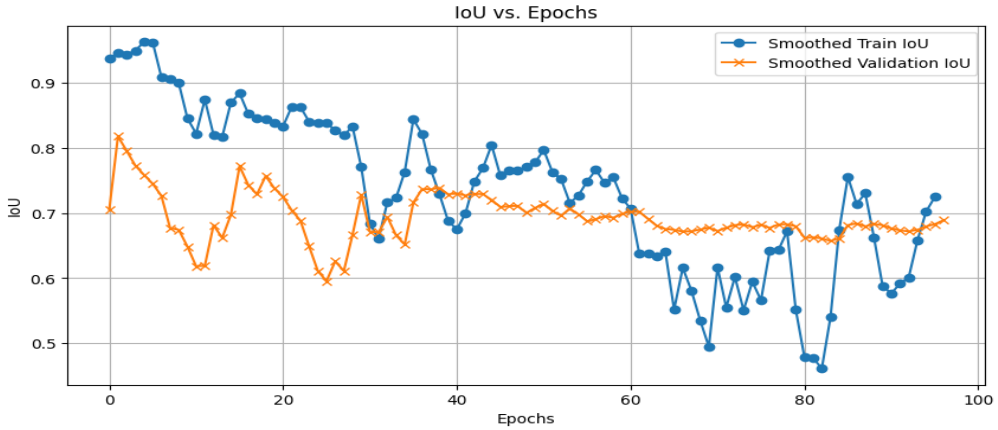

Fig. 8. Training and validation IoU of FB-PNet model for leaf annotation.

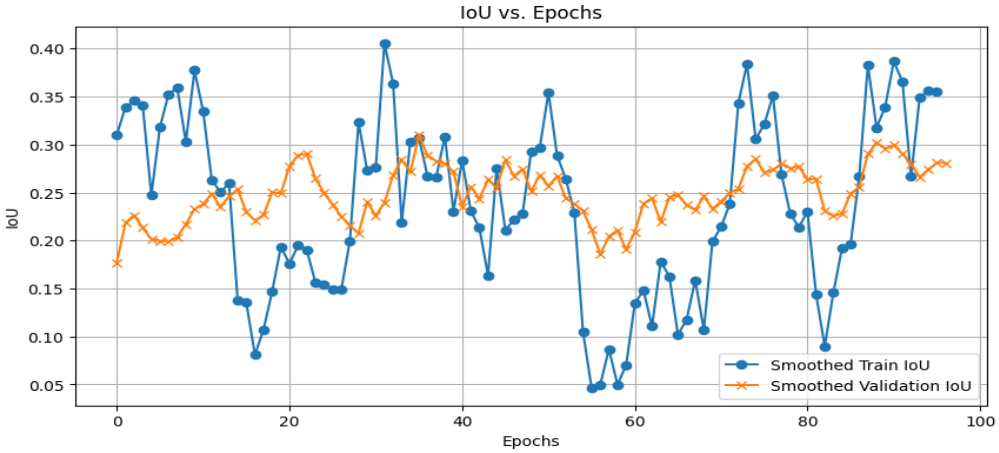

Fig. 9. Training and validation IoU of FB-PNet model for disease annotation.

TABLE IV. COMPARISON WITH RELATED METHODS

Method	Model	Validation IoU Accuracy
IIC [32]	R18+FPN	44.5
PiCIE [33]	R18+FPN	54.2
P-CNN [27]	–	67.2
DINO [34]	ViT-S/8	68.6
FB-PNet For Leaf	FB-PNet (ours)	**70.2**
FB-PNet For Disease	FB-PNet (ours)	**30.5**

Table IV provides a comparative evaluation of different segmentation models in terms of validation IoU accuracy. The proposed Forward-Backward-Propagated Percept Net (FB-PNet) outperforms several benchmark methods, including IIC, PiCIE, and DINO, achieving an IoU of 70.2 for leaf segmentation and 30.5 for disease segmentation. The findings emphasize the effective capability of FB-PNet in accurately capturing intricate plant leaf structures and disease patterns.

Fig. 10 provides a qualitative evaluation of the proposed FB-PNet model. The figure showcases:

- Leaf segmentation results: The predicted segmentation closely aligns with the ground truth, confirming accurate leaf annotation.
- Disease segmentation results: The disease prediction model captures infected regions but exhibits slight over-segmentation in certain areas.

The overall architecture of the Forward-Backward Propagated Percept Net is illustrated in the center of the figure.

The bottom section of the figure displays results from the PlantVillage dataset those are unseen to the model, showcasing the model's ability to accurately annotate both leaves and diseases across various plant species.

From the findings, it is evident that the proposed model is having better efficiency and advantage of avoiding the manual annotation during the segmentation process.

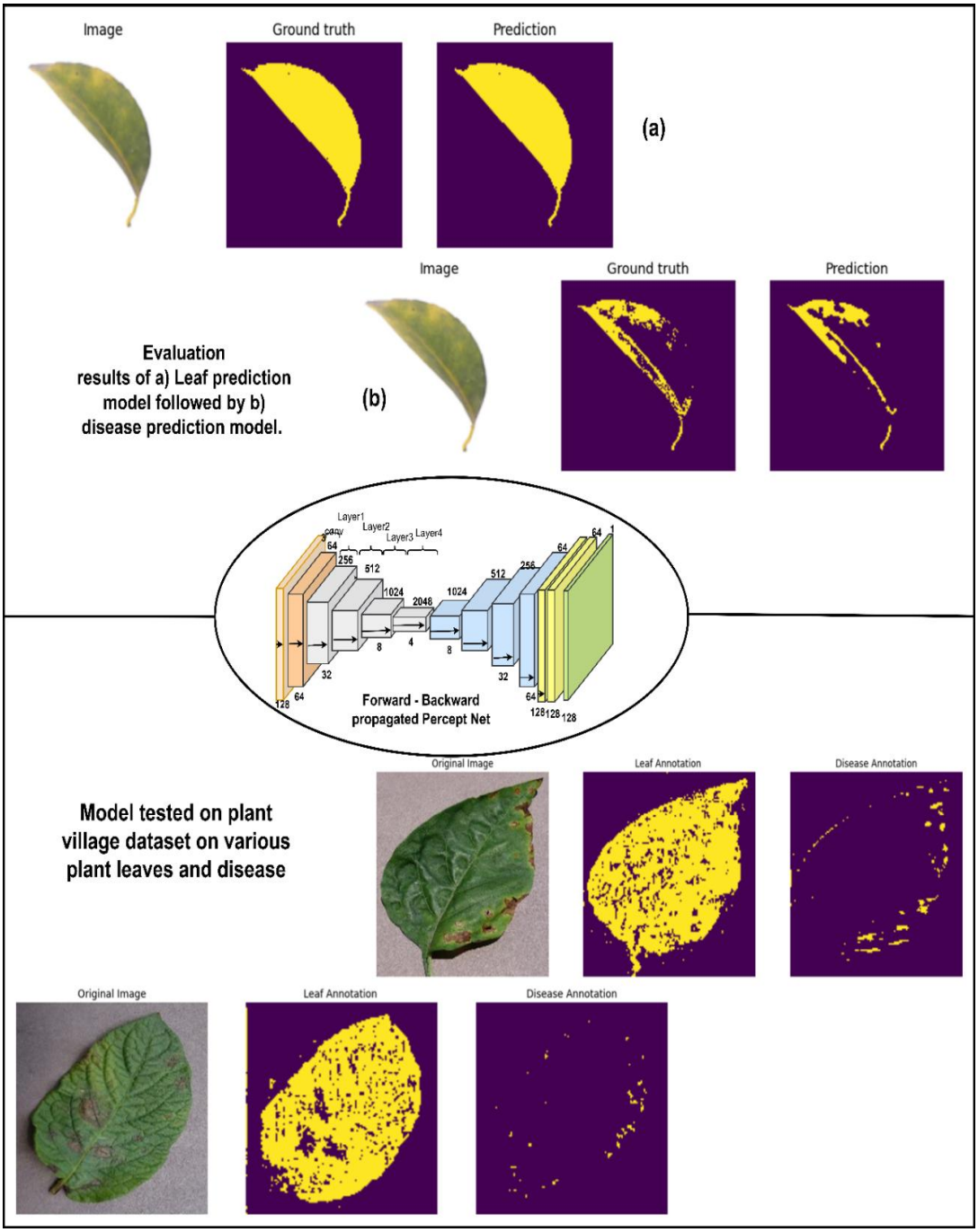

Fig. 10. Test Results of the proposed model with various dataset.

VI. Conclusion

This research describes a new improved deep-learning-based model for plant leaf and disease automatic annotation, FB-PNet. The model improves feature extraction by emphasizing salient visual details by discarding some other computations, thus helpful in enhancing the segmentation accuracy. The model trained using BCEWithLogitsLoss and the optimization was done using the Adam optimizer with ReduceLROnPlateau to stabilize the convergence and improve the generalization. Experimental results also clearly indicate that this approach offers superior performances measured with Intersection over Union (IoU), Dice coefficient, Precision, and Recall. However, the model is capable of over-segmenting in areas close to object borders because some natural parameters impede its effectiveness in accurately segmenting boundaries.

Despite its strengths, the model occasionally struggles with precise boundary segmentation, leading to over-segmentation near object borders. In future studies, we may attempt to combine multi-scale feature extraction and attention mechanisms for eluding trivial features while optimizing some crucial details to further improve the performance of the model. Future attempts would also include extending the framework to

solving multi-label segmentation tasks, improving self-supervised approaches to increase versatility across different datasets. By addressing these issues and fine tuning it, we aim to substantially develop our automatic plant leaf and disease segmentation system for real-world agricultural and biological applications.

REFERENCE

[1] Wang, W., Xie, E., Li, X., Fan, D., Song, K., Liang, D., ... & Shao, L. (2021). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 548-558.

[2] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 9992-10002.

[3] Chen, X., Fan, H., Girshick, R., & He, K. (2020). Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297.

[4] Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2017). Mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412.

[5] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., ... & Zhou, Y. (2021). TransUNet: Transformers make strong encoders for medical image segmentation. Medical Image Analysis, 70, 101996.

[6] Zhang, Y., Liu, H., & Hu, Q. (2021). TransFuse: Fusing transformers and CNNs for medical image segmentation. Medical Image Analysis, 70, 102004.

[7] Guo, M. H., Xu, T. X., Liu, J. J., Liu, Z. N., Jiang, P. T., & Mu, T. J. (2022). Attention mechanisms in computer vision: A survey. Computational Visual Media, 8(3), 331-368.

[8] Pound, M. P., Atkinson, J. A., Wells, D. M., Pridmore, T. P., & French, A. P. (2017). Deep learning for multi-task plant phenotyping. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2055-2063.

[9] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems (NeurIPS), 2672-2680.

[10] Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., ... & Zhang, L. (2021). Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6881-6890.

[11] Wang, X., Zhang, R., Shen, C., Kong, T., & Li, L. (2021). Dense contrastive learning for self-supervised visual pre-training. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3024-3033.

[12] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3431-3440.

[13] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. Medical Image Analysis, 39, 234-241.

[14] Zhang, J., Jiang, Z., Dong, J., Hou, Y., & Liu, B. (2020). Attention gate resU-Net for automatic MRI brain tumor segmentation. IEEE Access, 8, 58533-58545.

[15] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4), 834-848.

[16] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2961-2969.

[17] Liu, G., Reda, F. A., Shih, K. J., Wang, T. C., Tao, A., & Catanzaro, B. (2018). Image inpainting for irregular holes using partial convolutions. Proceedings of the European Conference on Computer Vision (ECCV), 85-100.

[18] Wang, W., Xie, E., Li, X., Fan, D., Song, K., Liang, D., ... & Shao, L. (2021). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 548-558.

[19] Chen, X., Fan, H., Girshick, R., & He, K. (2020). Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297.

[20] He X, Qi G, Zhu Z, Li Y, Cong B, Bai L. 2023. Medical image segmentation method based on multi-feature interaction and fusion over cloud computing. Simul Modell Pract Theory. 126:102769. ISSN 1569-190X. 10.1016/j.simpat.2023.102769.

[21] Zhu Z, Sun M, Qi G, Li Y, Gao X, Liu Y. 2024. Sparse dynamic volume TransUNet with multi-level edge fusion for brain tumor segmentation. Comput Biol Med. 172(108284):ISSN 0010–4825. doi: 10.1016/j.compbiomed.2024.108284.

[22] Xu Y, He X, Xu G, Qi G, Yu K, Yin L, Yang P, Yin Y, Chen H. 2022. A medical image segmentation method based on multi-dimensional statistical features. Front Neurosci. 16. doi: 10.3389/fnins.2022.1009581.

[23] Zhu Z, Wang Z, Qi G, Mazur N, Yang P, Liu Y. 2024. Brain tumor segmentation in MRI with multi-modality spatial information enhancement and boundary shape correction. Pattern Recognit. 153(110553). ISSN 0031–3203. doi: 10.1016/j.patcog.2024.110553.

[24] Pastore G, Cermelli F, Xian Y, Mancini M, Akata Z, Caputo B. 2021. A closer look at self-training for zero-label semantic segmentation. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Nashville, TN, USA.

[25] Kang D, Cho M. 2021. Integrative few-shot learning for classification and segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Nashville, TN, USA.

[26] O. Elharrouss, Y. Akbari, N. Almaadeed, and S. Al-Maadeed, "Backbones-review: Feature extraction networks for deep learning and deep reinforcement learning approaches," 2022.

[27] Hegde, D., & Balaji, G. N. (2024). P-CNN: Percept-CNN for semantic segmentation. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 12(1), 2387458.

[28] H. T. Rauf, B. A. Saleem, M. I. U. Lali, M. A. Khan, M. Sharif, and S. A. C. Bukhari, "A citrus fruits and leaves dataset for detection and classification of citrus diseases through machine learning," *Data Brief*, vol. 26, Oct. 2019, doi: 10.1016/j.dib.2019.104340.SS

[29] D. PONNAMBALAM, "Perception For automatic annotation of plant leaf and disease.". Zenodo, Feb. 19, 2025. doi: 10.5281/zenodo.14898047.

[30] P. Dinesh and R. Lakshmanan, "Deep Learning-Driven Citrus Disease Detection: A Novel Approach with DeepOverlay L-UNet and VGG-RefineNet," International Journal of Advanced Computer Science and Applications, vol. 15, no. 7, pp. 1023–1041, 2024, doi: 10.14569/IJACSA.2024.01507100.

[31] Hughes, D.P. and Salathe (2015) An Open Access Repository of Images on Plant Health to Enable the Development of Mobile Disease Diagnostics.

[32] Ji X, Henriques JF, Vedaldi A. 2019. Invariant information clustering for unsupervised image classification and segmentation. Proceedings of the IEEE/CVF International Conference on Computer Vision; Seoul, Korea (South). p. 9865–9874.

[33] Hyun Cho J, Mall U, Bala K, Hariharan B. 2021. Picie: unsupervised semantic segmentation using invariance and equivariance in clustering. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; Nashville, TN, USA. p. 16794–16804.

[34] Caron M, Touvron H, Misra I, Herve J´E, Mairal J, Bojanowski P, Joulin A. 2021. Emerging properties in self-supervised vision transformers. Proceedings of the IEEE/CVF international conference on computer vision; Virtual Conference. p. 9650–10.