SegFormer Inspired Multi Head Spectral Attention with Edge Gating light weight model for Leaf Area Segmentation

A. Shamim Banu
National Institute of Technology
S. Deivalakshmi
deiva@nitt.edu

National Institute of Technology

Research Article

Keywords: Leaf segmentation, Precision agriculture, Transformer, Multi head spectral attention, Edge gating, SegFormer, Deep learning

Posted Date: September 5th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7496305/v1

Additional Declarations: No competing interests reported.

SegFormer Inspired Multi Head Spectral Attention with Edge Gating light weight model for Leaf Area Segmentation

A. Shamim Banu, Research Scholar, National Institute of Technology, Tiruchirappalli , Tamilnadu, India
(Lecturer, GPTC, Thuvakudimali, Tiruchirappalli,, Tamilnadu, India)
Email: shamimbanuphd@gmail.com
Dr. S. Deivalakshmi, Associate Professor, National Institute of Technology, Tiruchirappalli,, Tamilnadu, India
Email: deiva@nitt.edu

Abstract

Accurate segmentation of leaf area is a critical task in plant phenotyping and precision agriculture, as it directly impacts yield estimation, disease monitoring, and weed management. Conventional Convolutional Neural Networks (CNNs), such as UNet and its variants, often struggle with capturing long range contextual dependencies and preserving fine structural boundaries, while pure transformer based architectures like the Vision Transformer (ViT) suffer from poor inductive bias and limited data efficiency. To overcome these challenges, we propose a SegFormer inspired model that integrates Edge Gated Multi Head Spectral Attention (EG MHSA) for robust leaf area segmentation. The spectral attention mechanism captures discriminative frequency domain representations across spectral bands, while the edge gating module enhances boundary preservation by adaptively fusing multiscale edge features. Evaluated on the benchmark CWFID dataset, the proposed model achieves superior performance with an F1score of 97.33%, IoU of 95.84%, and the lowest loss of 0.0395, outperforming UNet variants and transformer based baselines. Qualitative analysis further demonstrates its effectiveness in accurately delineating fine leaf boundaries under complex field conditions. The ablation results highlight the complementary contributions of spectral attention and edge gating in boosting segmentation performance. With its lightweight architecture, edge focused refinement, and strong generalization capability, the proposed approach sets a new benchmark for leaf area segmentation and provides a practical, scalable solution for agricultural applications.

Keywords: Leaf segmentation, Precision agriculture, Transformer, Multi head spectral attention, Edge gating, SegFormer, Deep learning.

I. Introduction

Accurate leaf segmentation is fundamental to plant phenotyping, precision agriculture, and automated trait extraction, as it directly contributes to yield estimation, disease monitoring, and resource optimization. Traditional computer vision methods, such as thresholding, edge detection, and region growing were initially used for plant segmentation but often failed under natural field conditions due to variable illumination, complex backgrounds, and overlapping plant structures [1], [2]. The introduction of deep learning marked a major breakthrough, as early dense prediction approaches transitioned from patch based classifiers to end to end pixelwise models. The advent of Fully Convolutional Networks (FCNs) demonstrated, for the first time, the feasibility of training convolutional architectures to generate complete segmentation maps without requiring handcrafted features or separate post processing [3], [4], thereby setting a new standard in semantic segmentation.

Building on the success of FCNs, the UNet architecture emerged as a milestone for biomedical and plant image analysis, owing to its encoder–decoder structure with symmetric skip connections [5]. This design

enabled the simultaneous capture of global contextual information in the downsampling path and precise localization in the upsampling path, making it highly effective for delineating complex leaf boundaries. UNet quickly became the de facto backbone for leaf segmentation tasks, outperforming traditional methods in handling occlusions, scale variations, and noise from natural backgrounds. Its ability to integrate multi scale features while preserving fine structural details made it particularly suitable for agricultural datasets, where accurate segmentation is crucial for distinguishing leaves from stems, soil, and weeds [6], [7]. The original UNet uses plain convolutional layers, which can suffer from vanishing gradients in deeper networks.

To overcome the limitations of the original UNet, numerous variants have been proposed for leaf segmentation and related plant phenotyping tasks. These include ResidualUNet [8], which integrates residual connections to improve gradient flow and training stability; Dense UNet [9], which introduces dense feature reuse to capture richer representations; and Attention UNet [10], which incorporates attention gates to highlight salient regions during decoding. Other notable extensions include R2UNet (Recurrent Residual UNet) [11], which models temporal dependencies within feature maps, and Nested UNet (UNet++) [12], which employs deeply supervised skip pathways for more precise feature fusion. These variants have consistently demonstrated improved performance on leaf segmentation benchmarks, particularly in scenarios involving occluded or overlapping leaves, where conventional UNet struggles to preserve fine boundaries.

A critical enhancement in these UNet variants is the integration of attention mechanisms, which allow the model to focus selectively on informative features [13]. Spatial attention emphasizes important regions in the spatial domain, ensuring that fine leaf edges and contours are preserved during downsampling and upsampling [14]. Channel attention assigns adaptive weights to feature channels, enhancing channels that capture critical textures such as venation or chlorophyll distribution while suppressing irrelevant background responses [15]. Moving beyond RGB inputs, spectral attention leverages correlations across multiple spectral bands (e.g., RGB, NIR, NDVI), thereby enhancing vegetation specific signals and suppressing noise from soil, shadows, or weeds [16]. In addition, wavelet based attention integrates frequency domain analysis, where high frequency sub bands enhance leaf boundaries and fine structures, while low frequency bands retain global shape information [17]. Collectively, these attention augmented UNet variants produce sharper and more continuous segmentation masks, significantly improving the robustness of leaf area segmentation under diverse agricultural conditions.

Accurate segmentation of fine structures such as leaves requires models that can capture both global semantic context and precise boundary details. Transformer based architectures like SegFormer have emerged as powerful backbones because they combine hierarchical feature encoding with lightweight decoders, enabling efficient global–local context modeling [18]. Unlike traditional convolutional models, SegFormer avoids the need for positional encodings and achieves strong multiscale representation learning, making it suitable for complex agricultural images where leaves vary in size, shape, and orientation [19]. Building on this foundation, researchers have explored frequency domain adaptations such as Frequency self attention Network (FsaNet), Frequency Attention Enhanced Network (FAENet), etc., which integrate low frequency self attention to reduce computational overhead while enhancing edge relevant cues [20]. These developments highlight the adaptability of Transformer frameworks to resource constrained and edge sensitive segmentation tasks.

Recent advances in plant phenotyping and agricultural image analysis have increasingly shifted from convolution based backbones to Transformer driven segmentation models such as SegFormer and Vision Transformers (ViTs) [21]. The SegFormer architecture stands out by combining a hierarchical Transformer encoder with a lightweight MLP decoder, allowing it to efficiently capture both global context and fine details without the heavy computational burden of standard ViTs [19]. Several variants have been proposed

for agricultural tasks to address these limitations. For example, Annealing Integrated Sparrow Optimization Algorithm – Salient Segmentation Transformer(AISOA SSformer) adapts SegFormer for rice leaf disease segmentation by introducing sparse linear layers to replace standard Multi Layer Perceptrons (MLPs) and by incorporating attention fusion modules to better integrate multi scale features [22]. Similarly, other SegFormer inspired approaches incorporate spatial attention and edge enhancement modules to tackle fragmented boundaries in thin leaves, producing crisper and more continuous segmentation masks. Denoising Diffusion Probabilistic Model (DDPM) SegFormer, which fuses a denoising diffusion probabilistic model with a vision transformer to overcome information bottlenecks, generating refined multiscale semantic features that boost LULC segmentation accuracy [23].

Parallel to SegFormer, pure Vision Transformer (ViT) approaches have also emerged. Unlike CNNs that rely on local receptive fields, ViTs process entire images as sequences of patches, enabling stronger modeling of long range dependencies, which is critical in capturing inter leaf boundaries and occlusions [24]. Models such as Guided Mask Transformer (GMT) extend this idea by introducing spatial priors and guided positional encodings that emphasize the spatial distribution of leaves, thereby improving the separability of overlapping or clustered foliage [25]. This yields robust instance level leaf segmentation, even in cases of dense overlap. On the other hand, lightweight ViT variants demonstrate that Transformer based methods can achieve state of the art accuracy while remaining computationally efficient, making them suitable for deployment in real world field scenarios where hardware resources are limited [26], [27].

Parallel research has focused on explicitly embedding edge awareness into segmentation networks to address the challenge of fragmented or blurry boundaries. Architectures such as Edge Aware Transformer (EAFormer) employ edge guided cross attention, where boundary maps guide the attention process to sharpen mask predictions [28], while convolutional counterparts like Convolutional Block Attention Module (CBAM) extend this principle through spatial and channel attention modules that emphasize relevant regions and contour information [29]. Such approaches demonstrate that incorporating edge priors into spatial attention mechanisms leads to crisper, more continuous boundaries. In the context of leaf segmentation, these edge aware and SegFormer inspired models provide a promising direction by integrating boundary refinement with powerful global attention, ensuring robust performance under complex field conditions [30].

The main contributions of this study are as follows: (1) A spectral transformer based model was developed for leaf area segmentation, performing FFT along channels per token and splitting frequency components into low, mid, and high bands to capture fine grained spectral features. (2) A multi scale edge guided attention mechanism was proposed, computing Sobel edge tokens at multiple scales and fusing them via learnable weights to enhance the segmentation of thin and intricate leaf boundaries. (3) A SegFormer style transformer encoder architecture was designed, integrating spectral attention with edge guidance in a unified framework, enabling high fidelity leaf area mask prediction with computational efficiency. (4) An end to end evaluation pipeline was established, supporting training with BCE and Dice loss and standard metrics such as IoU and F1, allowing quantitative and qualitative validation of the proposed edge gated spectral attention mechanism.

The remaining sections of this paper are organized as follows: Section II presents the proposed SegFormer inspired multi head spectral attention framework with edge gating for leaf area segmentation, including patch embedding and tokenization, multi scale edge token extraction, edge gated multi head spectral attention with frequency band splitting, and the MLP based residual block with edge aware attention, followed by the lightweight decoder. Section III describes the data set, data augumetation and preprocessing procedures, the loss functions including Dice, binary Cross Entropy, and combined losses, and the implemen-

tation and training details. Section IV presents the results and discussion, including quantitative, qualitative evaluations and ablation studies validating the number of encoders, the number of attention heads, and the effect of edge fusion and edge guidance. Sections V and VI conclude the study with discussion.

II. Proposed SegFormer Inspired Multi Head Spectral Attention with Edge Gating for Leaf Area Segmentation

To address the challenges of accurate and efficient leaf area segmentation, we propose a SegFormer inspired architecture that integrates a novel Multi Head Spectral Attention (MHSA) mechanism with Edge Gating. This work is driven by two primary shortcomings observed in existing approaches: (i) difficulty in capturing long range contextual dependencies, and (ii) poor boundary localization around fine leaf structures.

Our approach draws inspiration from the SegFormer encoder design, which employs lightweight transformers to model global dependencies while maintaining computational efficiency. We extend this paradigm by incorporating a spectral attention module, where features are processed across spectral bands to emphasize informative channels and suppress redundant ones. Multi head attention facilitates the modeling of diverse interactions, thereby enhancing the representation of leaf textures and intensity variations . In parallel, we introduce an Edge Gating mechanism to refine spatial details and enhance boundary preservation. The edge gate extracts edge aware features from the ground truth mask and fuses them adaptively with spectral features. This integration ensures that the model not only attends to global context but also maintains fine grained structural accuracy around leaf boundaries.

As shown in Fig. 1, the proposed architecture consists of a patch embedding module, spectral attention encoder blocks with edge gating, an MLP residual block, and a lightweight decoder for segmentation. This design balances global context modeling with edge aware refinement, making it particularly suitable for plant phenotyping applications where precise leaf area estimation is critical. The architecture is lightweight, modular, and can be trained end to end using standard segmentation loss functions, ensuring both effectiveness and scalability.

A. Patch Embedding and Tokenization

To convert the input image into a sequence of latent feature representations, a patch embedding module is employed. Instead of explicitly dividing the image into non overlapping patches, we use a convolutional layer with a stride of 2 to simultaneously extract local spatial patterns and reduce the resolution. This approach preserves local neighborhood information while projecting the raw image into a higher dimensional feature space.

Given an input image $\mathbf{I}' \in \mathbb{R}^{H\times W\times 3}$, the patch embedding operation is defined in Equation **??**eq:patch$_e$*mbedding*)

$$\mathbf{x_t} = \mathrm{ReLU}\big(\mathrm{Conv}_{3\times 3,\, s=2}(\mathbf{I}')\big), \tag{1}$$

where $\mathrm{Conv}_{3\times 3,\, s=2}$ denotes a 2D convolution with kernel size 3×3, stride $s = 2$, and 64 filters. The resulting feature map $\mathbf{x_t} \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times 64}$ serves as the initial representation for the subsequent stages of spectral attention and the transformer. This operation effectively downsamples the image resolution by a factor of two, while expanding the channel dimension to $C = 64$, thus preparing compact yet discriminative feature tokens for the following processing layers.

B. MultiScale Edge Token Extraction

Edge features are extracted at multiple scales (1, 0.5, and 0.25) to capture fine and coarse boundary details. The original scale preserves fine edges, the half scale provides regional context, and the quarter

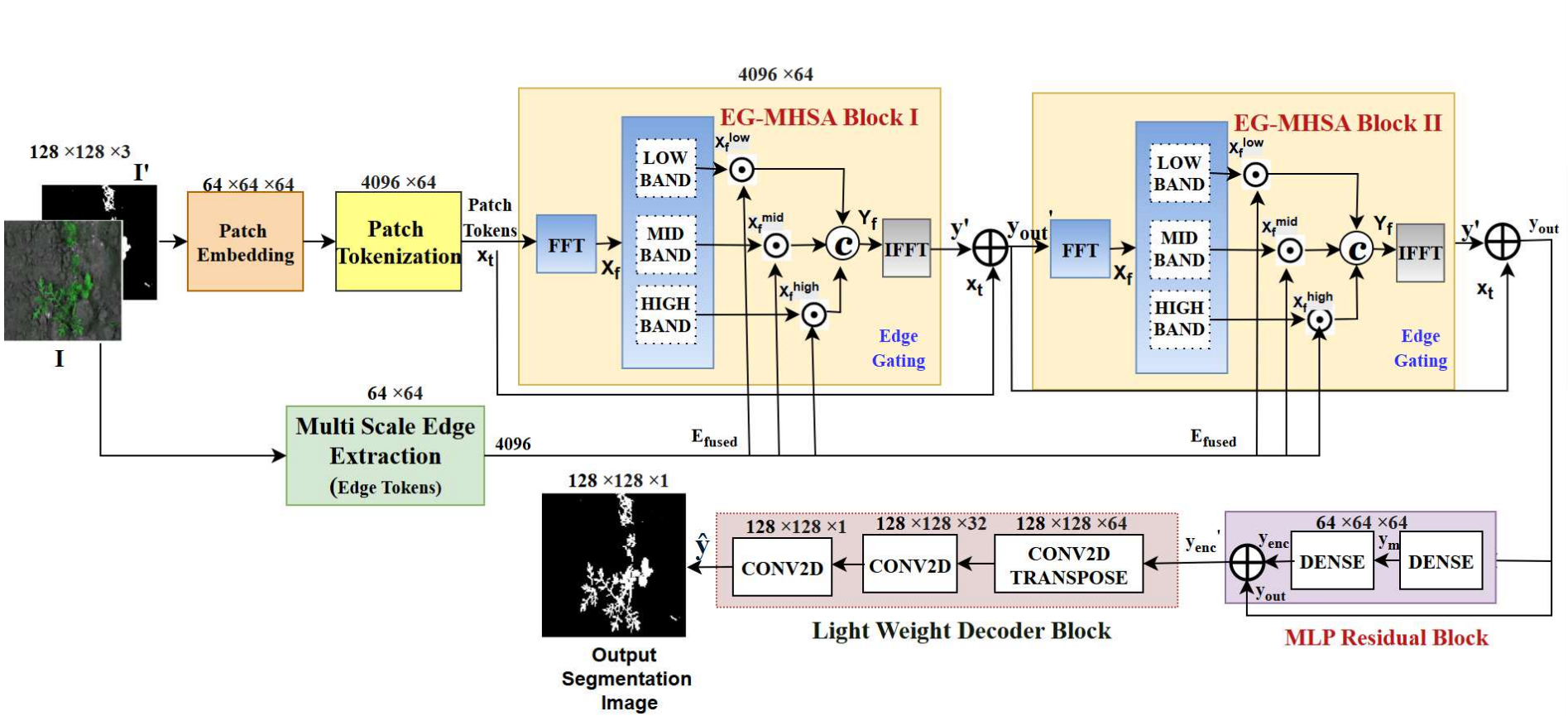


Fig. 1: Architecture of the proposed SegFormer inspired model with Multi Head Spectral Attention (MHSA) and Edge Gating for leaf area segmentation.

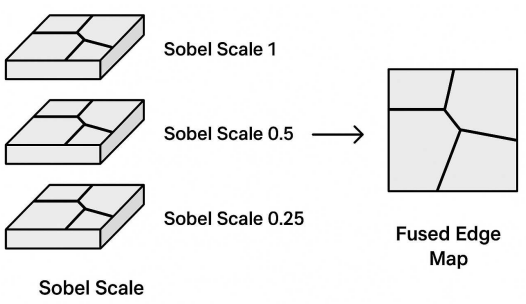


Fig. 2: Multi Scale Edge Detection and Fusion

scale captures global structure. These multiscale edge tokens act as structural priors and supply boundary cues to the network. During spectral multihead attention, they function as gating signals, emphasizing high frequency boundary information while maintaining global consistency. This fine to coarse integration improves segmentation accuracy and preserves vegetation boundary integrity.

1) Single Scale Edge Tokens

Given an input ground truth image $I_g \in \mathbb{R}^{H\times W\times 3}$ is subjected to Sobel filtering to compute the horizontal $\mathrm{G}_x and vertical G_y gradients as shown in Equation$ (2)

$$(G_x, G_y) = \mathrm{Sobel}(I_g), \tag{2}$$

The gradient magnitude is then computed as in Equation (3)

$$M = \sqrt{G_x^2 + G_y^2}, \tag{3}$$

and to account for varying image intensities, the gradient magnitude is normalized by its global maximum represented in Equation (4)

$$M_n = \frac{M}{\max(M) + \epsilon}, \tag{4}$$

where ϵ is a small constant added for numerical stability. The normalized edge map M_n is then flattened into a sequence of real valued tokens suitable for transformer input, as defined in Equation (5)

$$E = \mathrm{Flatten}(M_n) \in \mathbb{R}^{B\times T\times 1}, \quad T = H_t \times W_t, \tag{5}$$

where B denotes the batch size, and H_t and W_t represent the tokenized spatial dimensions.

2) Multi Scale Fusion of Edge Tokens

To capture edges at different levels of granularity, we extract edge tokens at multiple downsampling scales $s \in \{1.0, 0.5, 0.25\}$. For each scale, we compute the edge tokens as described above, resize them to the target resolution, and form a collection of maps $\{E^{(s)}\}$.

A set of learnable weights $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_S]$ governs the relative importance of each scale. To ensure positivity and unit sum, these weights are normalized using a softmax function, as expressed in Equation(6)

$$w_i = \frac{\exp(\alpha_i)}{\sum_{j=1}^{S} \exp(\alpha_j)}, \tag{6}$$

where w_i denotes the normalized contribution of the i th scale. Subsequently, the multi scale edge tokens are fused into a single representation as represented in Fig. 2 through a convex combination, given in Equation (7)

$$E_{\text{fused}} = \sum_{i=1}^{S} w_i \, E^{(s_i)}, \tag{7}$$

where $E^{(s_i)}$ corresponds to the edge tokens extracted at scale s_i. This adaptive fusion mechanism enables the model to emphasize fine grained edges or broader structural boundaries depending on the image content. The fused edge token is then projected as E^b to align with the frequency bands extracted in the Multi Head Spectral Attention block. Thus, the multi scale edge token extractor dynamically emphasizes the edges that are most informative for the segmentation task, providing a robust boundary aware representation.

Algorithm 1: Multi Scale Edge Token Extraction and Fusion

Input: Ground truth image I, predefined scales $S = \{1.0, 0.5, 0.25\}$, learnable weights W

Output: Projected fused edge tokens E^b

Step 1: Single Scale Edge Extraction

Consider I' groundtruth image ;

Apply Sobel filtering to compute horizontal gradients G_x and vertical gradients G_y ;

Compute gradient magnitude: $G = \sqrt{G_x^2 + G_y^2}$;

Normalize G using its global maximum: $\hat{G} = \frac{G}{\max(G)}$;

Flatten $\hat{G}$ into a sequence of tokens E ;

Step 2: Multi Scale Extraction

foreach $s \in S$ **do**

- Downsample I' to scale s: $I_s = \text{Downsample}(I, s)$;
- Extract edge tokens $E^{(s)}$ using Step 1 on I_s ;
- Upsample $E^{(s)}$ back to the original resolution ;

Step 3: Fusion of Edge Tokens

Normalize learnable weights using softmax: $\tilde{W} = \text{Softmax}(W)$;

Fuse tokens: $E_{\text{fused}} = \sum_{s \in S} \tilde{W}_s \cdot E^{(s)}$;

Step 4: Projection for Spectral Attention

Project E_{fused} into spectral attention space: $E^b = \text{Proj}(E_{\text{fused}})$;

return $\mathbf{E}^b$

C. EG MHSA Edge Gated Multi Head Spectral Attention with Frequency Band Splitting and Multi scale edge Fusion

We propose an Edge Gated Multi Head Spectral Attention (EGMHSA) mechanism that includes frequency domain representations and integrates *edge aware gating*. The Spectral MultiHead Attention block enhances spatial spectral representations by learning separate dynamics across low , mid , and high frequency features,strengthening high frequency features with edge aware gating, Maintaining stable optimization through frequency domain normalization. The pipeline consists of four main stages:

1) **Spectral Projection (FFT):** Given an input token sequence $x_t \in \mathbb{R}^{B\times T\times C}$, where B is the batch size, T the number of tokens, and C the channel dimension, the Fourier transform is applied along the channel axis, as shown in Equation(8)

$$X_f = \mathcal{F}(x_t), \quad X_f \in \mathbb{C}^{B\times T\times C}, \tag{8}$$

where $\mathcal{F}$ denotes the Fast Fourier Transform (FFT).

2) **Frequency Band Splitting:** The spectrum is divided into three disjoint frequency bands, formulated in Equations (9)–(11)

$$X_f^{\text{low}} = X_f[:, :, 0 : c_{\text{low}}], \qquad \text{(low-frequency band)} \tag{9}$$

$$X_f^{\text{mid}} = X_f[:, :, c_{\text{low}} : c_{\text{low}} + c_{\text{mid}}], \qquad \text{(mid-frequency band)} \tag{10}$$

$$X_f^{\text{high}} = X_f[:, :, c_{\text{low}} + c_{\text{mid}} : C], \qquad \text{(high-frequency band)} \tag{11}$$

where $c_{\text{low}} + c_{\text{mid}} + c_{\text{high}} = C$.

3) **Band wise Attention with Edge Gating:** For each frequency band $X_f^b \in \mathbb{C}^{B\times T\times c_b}$, the real and imaginary parts are explicitly separated and concatenated, as shown in Equation(12)

$$Z^b = \text{Concat}\big(\Re(X_f^b), \Im(X_f^b)\big), \quad Z^b \in \mathbb{R}^{B\times T\times 2c_b}. \tag{12}$$

This representation is projected, processed with Multi Head Attention (MHA), and projected back and give in Equation(13)

$$E^b = W_b^{\text{out}} \cdot \text{MHA}\Big(W_b^{\text{in}} \cdot \text{Concat}\big(\Re(X_f^b), \Im(X_f^b)\big)\Big), \tag{13}$$

where $W_b^{\text{in}}, W_b^{\text{out}}$ are learnable projection matrices.

Edge awareness is incorporated using an edge embedding $E^b \in \mathbb{R}^{B\times T\times c_b}$ with a gating mechanism in Equation(14)

$$\hat{E}^b = E^b \odot \Big(1 + \sigma(\gamma_b) \cdot \sigma(W_b^e E^b)\Big), \tag{14}$$

where $\odot$ denotes element wise multiplication, W_b^e is the edge projection, γ_b is a learnable scalar, and σ is the sigmoid activation. Finally, the gated output is mapped back to a complex valued form as per the Equation(15)

$$Y_f^b = \Re(\hat{E}^b) + i \cdot \Im(\hat{E}^b). \tag{15}$$

4) **Recombination and Inverse FFT:** The band wise outputs are concatenated and projected back to the frequency domain as per the Equation (16),(17),(18)

$$Y_f = \text{Concat}\big(Y_f^{\text{low}}, Y_f^{\text{mid}}, Y_f^{\text{high}}\big). \tag{16}$$

Applying the inverse Fourier transform recovers the spatial representation:

$$y = \mathcal{F}^{-1}(Y_f), \quad y \in \mathbb{R}^{B\times T\times C}. \tag{17}$$

A Layer Normalization (LN) is then applied to stabilize training

$$y' = \text{LN}(y). \tag{18}$$

To preserve the original spatial context from the patch embeddings, the normalized output y' is fused

with the input token sequence x_t via a residual skip connection as per Equation(19)

$$y_{\text{out}} = y' + x_t. \tag{19}$$

This residual connection ensures that spectral attention strengthens the feature representation while preserving the essential spatial information embedded in the patch tokens, thereby improving the robustness of downstream segmentation. To further clarify the working mechanism, Fig. 3 presents a flowchart of the proposed Edge Gated Multi Head Spectral Attention (EG MHSA) module. The diagram illustrates the sequential process of spectral projection, frequency band decomposition, band wise attention with edge gating, and recombination, followed by residual fusion with the original tokens to yield enhanced representations.

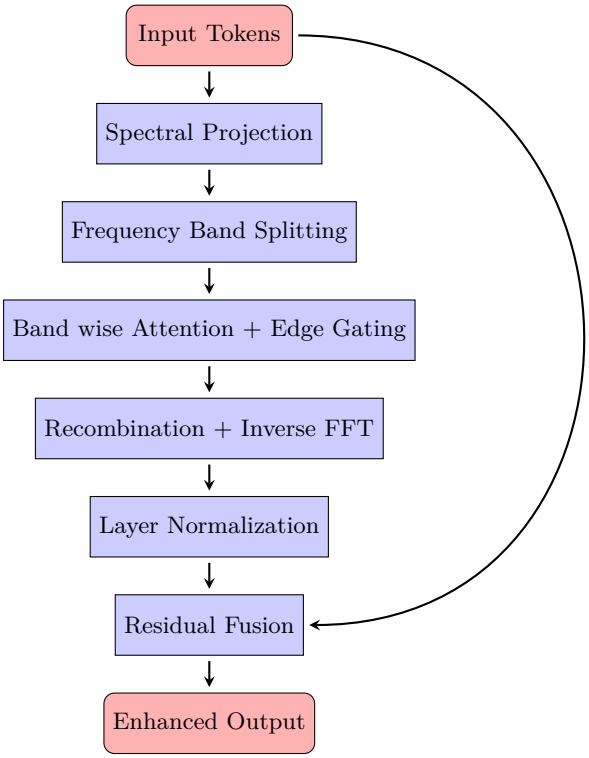


Fig. 3: Flowchart of Edge Gated Multi-Head Spectral Attention (EGMHSA).

D. MLP Block and Transformer Encoder with Edge Aware Attention

The Multi Layer Perceptron (MLP) block acts as a feed forward module following the spectral attention and edge gating mechanism. Given an input sequence y_{out}, the MLP block transforms the features as depicted in Equations (20),(21)

$$y_m = \text{Dropout}\big(\text{GeLU}(y_{\text{out}} W_1 + b_1)\big), \tag{20}$$

$$y_{\text{enc}} = \text{Dropout}(y_m W_2 + b_2), \tag{21}$$

where $W_1 \in \mathbb{R}^{C \times d_h}$ and $W_2 \in \mathbb{R}^{d_h \times C}$ are learnable weight matrices, b_1, b_2 are bias terms, and d_h represents the hidden layer dimension. The non linearity is introduced using the Gaussian Error Linear Unit

(GeLU).

The Transformer encoder integrates this MLP block with the Multi Head Spectral Attention mechanism output y_{out}. The encoder layer output is then expressed as Equation(22)

$$y'_{\text{enc}} = y_{\text{out}} + y_{\text{enc}}. \tag{22}$$

where residual connections ensure stable gradient propagation and better feature preservation. This design effectively couples global spectral dependencies with localized edge aware features while maintaining multi scale robustness.

E. Lightweight Decoder

To reconstruct the segmentation mask from the encoded feature representation, we employ a lightweight decoder that focuses on efficient upsampling and refinement. The decoder first applies a transposed convolution to upsample the spatial resolution, followed by a standard convolution to refine local structures. Finally, a 1×1 convolution with sigmoid activation generates the predicted segmentation mask. This design ensures minimal computational overhead while preserving segmentation accuracy. The final prediction is expressed as per the Equation (23)

$$\hat{y} = \sigma\big(\text{Conv}_{1\times1}\big(\text{ReLU}(\text{Conv}_{3\times3}(\text{ReLU}(\text{Conv}^{T}_{3\times3,\, s=2}(\mathbf{y'}_{\text{enc}}))))\big)\big), \tag{23}$$

where $\text{Conv}^{T}_{3\times3,\, s=2}$ denotes a transposed convolution with kernel size 3×3 and stride 2, $\text{Conv}_{3\times3}$ is a standard convolution with 32 filters, $\text{Conv}_{1\times1}$ projects to the number of classes, and $\sigma(\cdot)$ is the sigmoid activation. Thus, $\hat{y} \in \mathbb{R}^{H\times W\times C}$ represents the predicted segmentation mask.

III. Implementation

A. Dataset Description, Data Augumentation and Preprocessing

The proposed methodology was evaluated on benchmark dataset designed to capture diverse and challenging field conditions the **CWFID dataset** [31] . The CWFID dataset comprises 60 high resolution field images with detailed pixel level annotations, collected using the Bonirob agricultural robot on an organic carrot farm. These images show carrot plants at the early stage of true leaf growth, where segmentation is particularly challenging due to dense vegetation, overlapping leaves, and complex double compound structures.

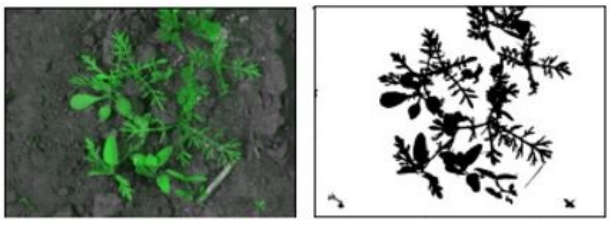


Fig. 4: Sample Images of RGB image and Mask from CWFID dataset

For evaluation, both data sets were divided into 80% for training and 20% for testing. Representative samples along with their corresponding ground truth masks are presented in Fig 4(a)–(d). To facilitate uniform training, all images and masks were resized to 128×128 pixels as per Equation(24)and normalized as represented in Equation (25)

To improve generalization and prevent overfitting, data augmentation techniques were applied during training. These included random horizontal and vertical flips, rotations within a fixed range, and random scaling to simulate variability in plant orientation and size. Additionally, brightness and contrast adjustments were performed to account for differences in illumination conditions across the dataset. Such augmentations ensured that the model was exposed to diverse variations in appearance, thereby enhancing its robustness and segmentation performance.

$$I' = \text{Resize}(I, 128 \times 128) \tag{24}$$

where X denotes the original image and I' represents the resized version. The pixel values were normalized to the range $[0, 1]$ as

$$I_{norm}(x, y) = \frac{I'(x, y)}{255} \tag{25}$$

B. Loss Functions and Evaluation Metrics

To train and evaluate the segmentation network, we employ a hybrid loss function that combines Binary Cross Entropy (BCE) with Dice loss, and assess the performance using standard evaluation metrics such as Intersection over Union (IoU) and F1score.

1) Dice Loss

The Dice loss is derived from the Dice Similarity Coefficient (DSC), which measures the overlap between the predicted mask $\hat{Y}$ and the ground truth mask Y. It is defined as per the Equation(26)

$$\text{DSC}(Y, \hat{Y}) = \frac{2|Y \cap \hat{Y}|}{|Y| + |\hat{Y}|} = \frac{2\sum_{i=1}^{N} y_i \hat{y}_i}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{y}_i}, \tag{26}$$

where $y_i \in \{0, 1\}$ denotes the ground truth label, $\hat{y}_i \in [0, 1]$ denotes pixel i the predicted segmentation mask and N the total number of pixels. The Dice loss is then formulated as per the Equation (27)

$$\mathcal{L}_{\text{Dice}} = 1 - \text{DSC}(Y, \hat{Y}), \tag{27}$$

2) Binary Cross Entropy Loss

The BCE loss evaluates the pixel wise binary segmentation performance between ground truth and predicted mask, The formula is given by the Equation(28)

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right]. \tag{28}$$

3) Combined Loss

The overall training objective integrates both losses as given in Equation (29) to exploit their complementary benefits

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{Dice}}. \tag{29}$$

This formulation ensures that the model not only minimizes pixel level loss but also maximizes region wise overlap between predictions and ground truth, leading to robust segmentation performance.

C. Implementation and Training Details

The proposed model was implemented in `TensorFlow 2.x` and trained using Google Colab with an NVIDIA Tesla T4 GPU (16 GB RAM). The input RGB images and their corresponding ground truth masks were resized to 128×128 pixels.

TABLE I: Training configuration for the proposed model.

Hardware	Input size	Batch size	Optimizer	Learning rate	Loss function	Metrics	Epochs	Heads	MLP Dim	Activation
Google Colab (Tesla T4 GPU, 16 GB RAM)	$128 \times 128 \times 3$	4	Adam	1×10^{-4}	BCE + Dice loss	IoU, F1score	100	2	128	ReLU

Preprocessing steps included normalization of pixel intensities to the range $[0, 1]$ and data augmentation (random flips and rotations) to improve model generalization. The model was compiled with the Adam optimizer, using a learning rate of 1×10^{-4}. A hybrid *binary cross entropy + Dice loss* function was employed to balance pixel level accuracy and overlap based similarity. The evaluation metrics used during training were the Intersection over Union (IoU) and F1score. We trained the model for 100 epochs with a batch size of 4. An overview of the training configuration is given in Table I.

IV. Results and Discussion

A. Quantitative Evaluation

To assess the performance of the proposed model, several evaluation metrics were computed, including Intersection over Union (IoU), F1score, and Combined loss.

1) Comparison of proposed model with existing methods

Table II presents a comprehensive comparison of state-of-the-art segmentation models, including conventional UNet variants, transfer learning-based architectures, and transformer-based networks, evaluated on standard metrics such as F1-score, Intersection over Union (IoU), and loss.

TABLE II: Comparison of Segmentation Models on Evaluation Metrics

S. No	Method	F1-score (%)	IoU (%)	Loss
1	UNet++ [12]	77.56	63.37	0.2274
2	Inception UNet [32]	62.31	45.26	0.6211
3	Dense UNet [9]	91.49	84.34	0.1446
4	VGG16 UNet [33]	90.70	82.99	0.0954
5	VGG19 UNet [34]	94.19	85.16	0.0634
6	ResNet UNet [8]	60.72	75.50	0.1730
7	SDUNet [35]	83.97	72.39	0.1650
8	INCSAUNet [36]	94.34	89.30	0.1204
9	AKDUNet [37]	96.75	93.75	0.0811
10	AWUNet [38]	95.55	91.49	0.0789
11	SegNeXt [39]	92.00	83.97	0.1310
12	SegFormer [19]	96.98	93.21	0.0489
13	Vision Transformer (ViT) [24]	60.76	24.10	0.8462
14	**Proposed EG-MHSA Model**	**97.33**	**95.84**	**0.0395**

Among the evaluated models, UNet++ achieves a moderate performance with an F1score of 77.56% and IoU of 63.37%. Its performance is limited by the relatively shallow feature extraction layers and nested skip connections, which, while improving gradient flow, are insufficient to capture complex leaf textures and overlapping structures. Inception UNet performs poorly F1score 62.31%, IoU 45.26% due to its reliance on parallel convolutions of different kernel sizes that fail to aggregate enough high-level contextual information for precise segmentation, particularly for thin or occluded leaves. Dense UNet and VGG19 UNet achieve significantly higher F1scores 91.49% and 94.19% and IoU 84.34% and 85.16%, as Dense UNet benefits from dense connectivity, enhancing feature reuse and gradient flow, while VGG19 UNet leverages a deep pre-trained backbone providing rich hierarchical feature representations. VGG16 UNet also performs well F1score 90.70%, IoU 82.99% due to transfer learning. ResNet UNet shows lower F1score (0.72% despite a higher IoU 75.50% as residual blocks capture dominant features at the cost of fine local details, leading to fragmented segmentation. Attention-based models, including SDUNet, INCSAUNet, AKDUNet, and AWUNet, demonstrate improved segmentation, with AKDUNet achieving F1score 96.75% and IoU 93.75%, highlighting the effectiveness of combining dense connections with channel-spatial attention. Transformer-based SegFormer attains F1score 96.98% and IoU 93.21%, effectively capturing global context while preserving local details. In contrast, Vision Transformer (ViT) underperforms F1score 60.76%, IoU 24.10%, illustrating the limitations of pure transformers without convolutional inductive biases for dense prediction tasks. The proposed EG-MHSA SegFormer model surpasses all other architectures with F1score 97.33%, IoU 95.84%, and the lowest loss 0.0395, due to the integration of enhanced edge-guided multi-head self-attention, which preserves fine boundary details and captures long-range dependencies, enabling precise segmentation of complex leaf structures while minimizing false positives. Overall, models with attention mechanisms and deep pre-trained backbones perform better, while models lacking local feature inductive bias underperform, confirming the effectiveness of the proposed EG-MHSA approach in achieving superior segmentation performance.

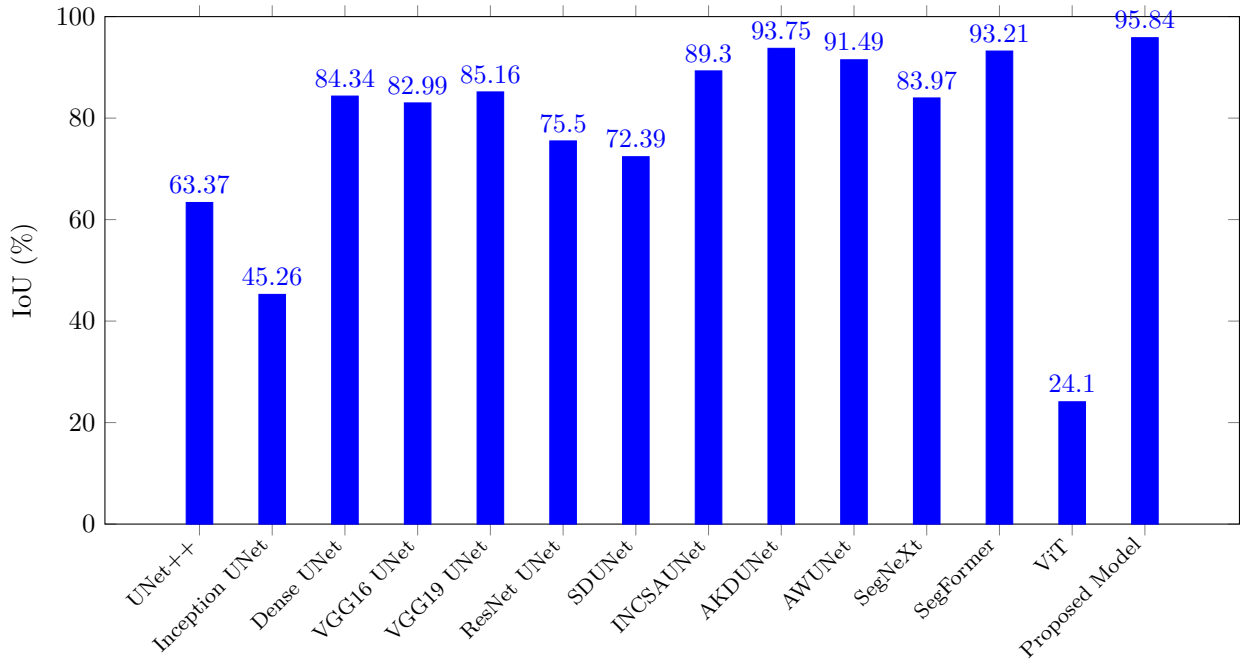


Fig. 5: Comparison of IoU (%) for different segmentation models.

TABLE III: Comparison of Segmentation Models and Their Parameters (in Millions)

S.No	Model	Parameters (M)
	Transfer UNet	
1	Inception UNet	0.177
2	Dense UNet	4.085
3	VGG19 UNet	16.24
4	ResNet UNet	24.29
5	VGG16 UNet	25.85
	UNet Variants	
6	UNet++	0.149
7	INSCA UNet	0.177
8	SDUNet	0.498
9	AWUNet	2.354
	Transformer-based Nets	
10	SegFormer (base)	0.157
11	SegNext	3.021
12	ViT	6.318
13	**Proposed EG-MHSA Model**	**0.158**

2) Comparison of Segmentation Models Parameters

Table III presents a comparison of various semantic segmentation models and their respective parameter counts (in millions). The models are categorized into three groups: Transfer UNet architectures, UNet variants and Transformer-based networks. Transfer UNet models, which incorporate pre-trained backbones such as VGG16, VGG19, ResNet, DenseNet, and Inception, exhibit a wide range of parameters from 0.177M (Inception U-Net) to 25.85M (VGG16 UNet), reflecting the complexity introduced by deep feature extractors. In contrast, UNet variants including UNet++, INSCA UNet, SDUNet, AWUNet, and the proposed EG-MHSA model maintain relatively lower parameter counts, ranging from 0.149M to 2.354M, while introducing architectural innovations such as nested skip connections, spatial attention, and multi-head attention for enhanced segmentation performance. Transformer-based networks, including SegFormer, SegNext, ViT, and the proposed model demonstrate parameter counts between 0.124M and 6.318M, indicating that lightweight transformer designs can achieve efficiency comparable to optimized U-Net variants. Overall, the table highlights a clear trade-off between model complexity and parameter efficiency. Models with fewer parameters, such as INSCA UNet and proposed model, are expected to offer faster training and inference times with reduced memory requirements, making them suitable for deployment on resource-constrained platforms, while larger models like VGG16 UNet and ViT provide higher representational capacity at the cost of increased computational load.

B. Qualitative Evaluation

Fig. 6 presents a qualitative comparison between the ground truth annotations and the predicted segmentation masks from the proposed model. The results demonstrate that the model effectively captures leaf boundaries and fine details, reducing over segmentation and under segmentation errors.

The qualitative analysis illustrated provides a visual comparison of leaf segmentation results obtained from UNet variants, transformer based models, and the proposed EG MHSA SegFormer. The baseline SDUNet and INSCA UNet exhibit noisy predictions with incomplete segmentation, particularly failing to capture fine

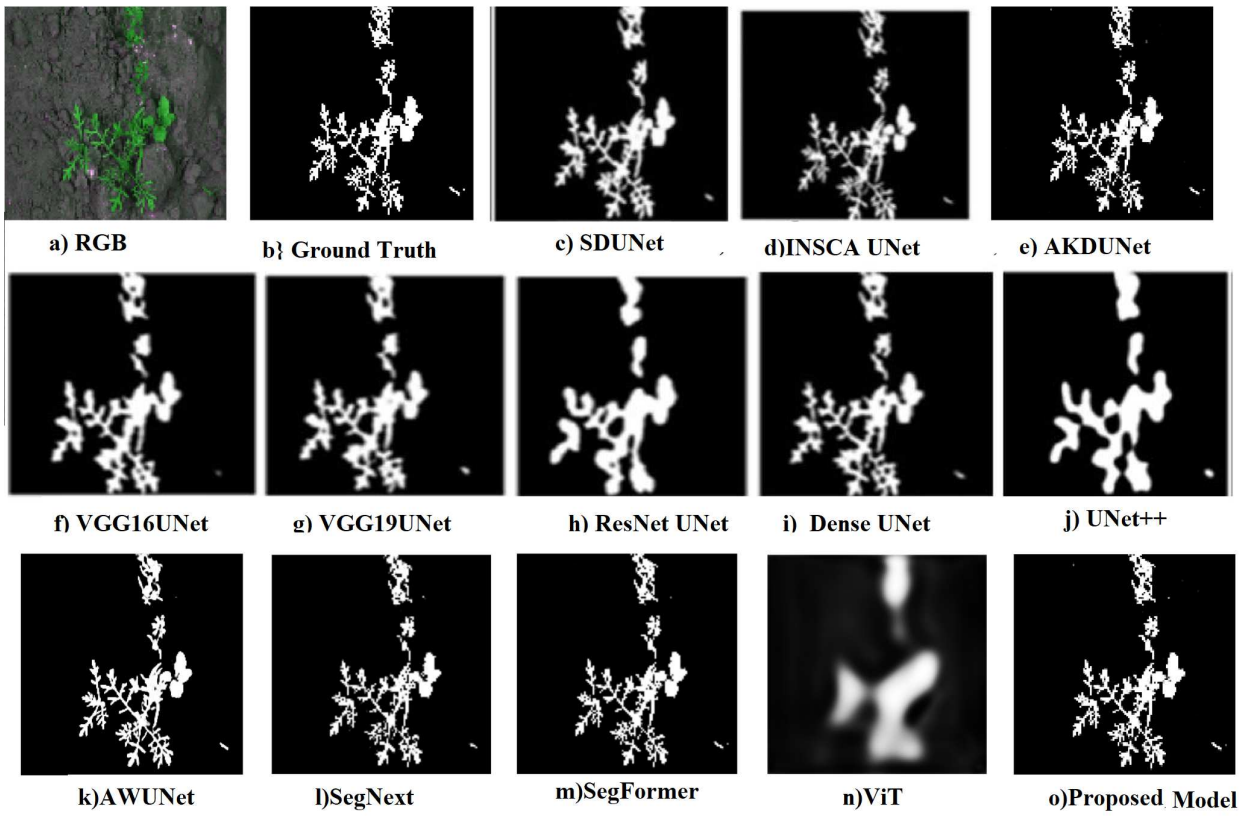


Fig. 6: Comparison of leaf segmentation performance between UNet variants, Transfer UNets, Transformer based models, and the proposed method

leaf structures and boundary details. Enhanced UNet variants such as AKDUNet and AWUNet improve segmentation quality by refining edge details and better capturing leaf regions, although some inaccuracies remain in densely overlapping vegetation. Traditional UNet variants such as VGG16 UNet, VGG19 UNet, and ResNet UNet capture the overall structure of the crop–weed regions, but often struggle with preserving fine details, resulting in smoother outputs with blurred or incomplete boundaries. Dense UNet provides improved connectivity and better structural retention but introduces slight noise, while UNet++ produces oversimplified masks that fail to delineate thin leaf structures.SegNext achieves intermediate performance, producing relatively clean segmentation masks but still lacking the precision required for accurate boundary delineation. Transformer based models, particularly SegFormer, show remarkable improvements, generating sharper and more accurate outputs that closely match the ground truth, thus demonstrating the strength of hybrid transformer convolutional architectures in learning both global context and local details. However, the Vision Transformer (ViT) performs poorly, yielding fragmented and inconsistent predictions, which underscores its limitations in agricultural segmentation tasks without inductive biases. Most notably, the proposed EG MHSA SegFormer produces the most reliable results, accurately capturing leaf boundaries with minimal false positives and delivering masks that closely resemble the ground truth annotations. Overall, the visual comparisons confirm the robustness of the proposed model and establish its superiority over both UNet variants and other transformer based approaches for precision leaf segmentation.

C. Ablation Study

To systematically evaluate the contribution of different components in the proposed architecture, we conducted a series of ablation studies. Each experiment isolates a specific design choice, such as the number of encoders, the number of attention heads, and the incorporation of edge related modules, while keeping other parameters fixed. This analysis provides insights into the trade offs between accuracy, model complexity, and computational efficiency, and highlights the role of spectral spatial feature extraction and edge aware refinement in achieving robust segmentation performance.

1) Validating Number of Encoders

The first ablation investigates the effect of varying the number of encoders in the proposed architecture. As shown in Table IV, using two encoders achieves the best balance between accuracy and computational efficiency, yielding the highest IoU of 95.84% and the lowest loss of 0.0369. A single encoder reduces the parameter count and training time but compromises accuracy. Increasing the number of encoders beyond two leads to a sharp rise in the number of parameters and training time (477–637 ms per step) without any gain in performance. In fact, with three or four encoders, IoU drops to 94.17% and 94.22%, respectively, while the loss increases.

TABLE IV: Effect of varying the number of encoders.

No of Encoders	**1**	**2**	**3**	**4**
IoU (%)	95.38	95.84	94.17	94.22
Loss	0.0499	0.0369	0.0675	0.0494
Parameters ($\times 10^3$)	108	158	208	259
Training Time (ms/step)	188	319	477	637

2) Validating Number of Heads

The second study evaluates the influence of varying the number of attention heads. As shown in Table V, a configuration with two heads achieves the best performance with an IoU of 95.84% and a minimal loss of 0.0369. Increasing the number of heads beyond two results in a drop in IoU, for example 94.04% with four heads, and a rise in loss ranging between 0.0503 and 0.0509.

TABLE V: Effect of varying the number of heads.

Heads	**2**	**4**	**8**	**16**
IoU(%)	95.84	94.04	94.64	94.22
Loss	0.0369	0.0503	0.0501	0.0509

3) With Edge Fusion vs Without Fusion

Table VI compares the performance of the model with and without the edge fusion module. Incorporating edge fusion significantly improves segmentation performance, yielding an IoU of 95.84 and a loss of 0.0369, compared to 93.23 IoU and 0.0529 loss without fusion.

TABLE VI: Comparison of edge fusion with and without fusion.

Setting	**IoU**(%)	**Loss**
With Edge Fusion	95.84	0.0369
Without Edge Fusion	93.23	0.0529

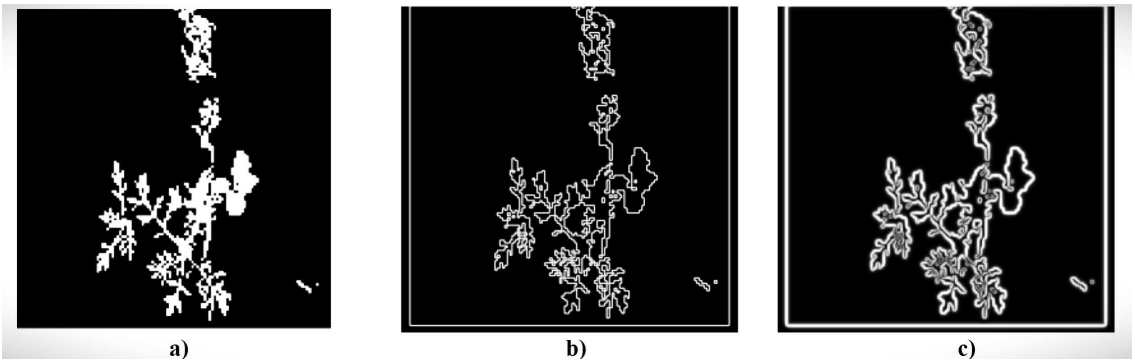


Fig. 7: Comparison of edge extraction methods: (a) Original binary mask, (b) conventional Sobel operator, and (c) proposed multiscale Sobel fusion.

4) With Edge vs Without Edge

In this experiment, the contribution of explicit edge information is examined. As shown in Table VII, including edge information achieves superior results with an IoU of 95.84 % and a loss of 0.0369, whereas the absence of edge information lowers IoU to 94.30% and increases the loss to 0.0507.

To validate the effectiveness of multiscale fusion in edge detection, we compared the conventional Sobel operator with the proposed multiscale Sobel method. Fig 7 shows the qualitative results. The conventional Sobel operator (middle panel) detects only fine boundary details but often produces discontinuous and weak edges, especially in thin leaf regions. In contrast, the multiscale Sobel method (right panel) integrates edge responses across multiple scales, resulting in more continuous, well defined, and structurally consistent boundaries. Compared to the original binary mask (left panel), the multiscale approach better preserves the fine venation and overall plant structure, thereby providing a richer edge representation that can be exploited in subsequent segmentation and refinement stages.

TABLE VII: Performance with and without edge information.

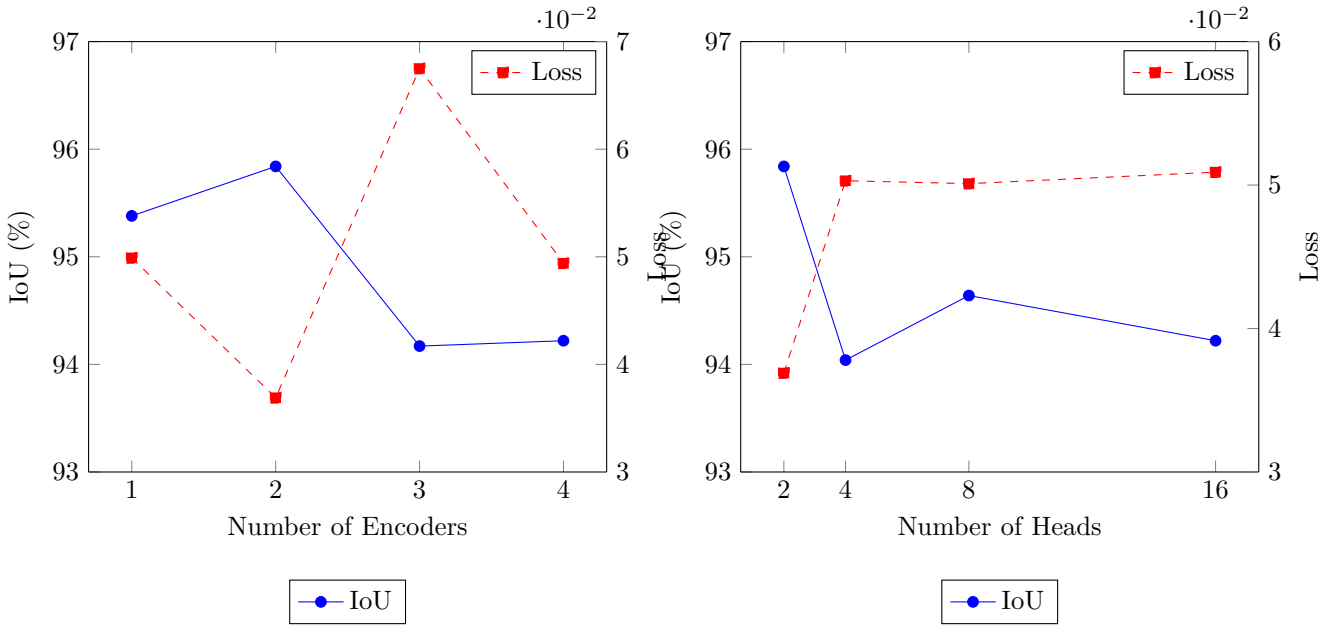
Setting	**IoU**(%)	**Loss**
With Edge	95.84	0.0369
Without Edge	94.30	0.0507

The best performing configuration across all ablations consistently aligns with the dual encoder, two head, edge aware model with fusion, which achieves the highest IoU of 95.84% and the lowest loss of 0.0369. This configuration balances efficiency and accuracy, demonstrating the synergistic effect of spectral spatial feature learning and edge aware refinement.

V. Discussion

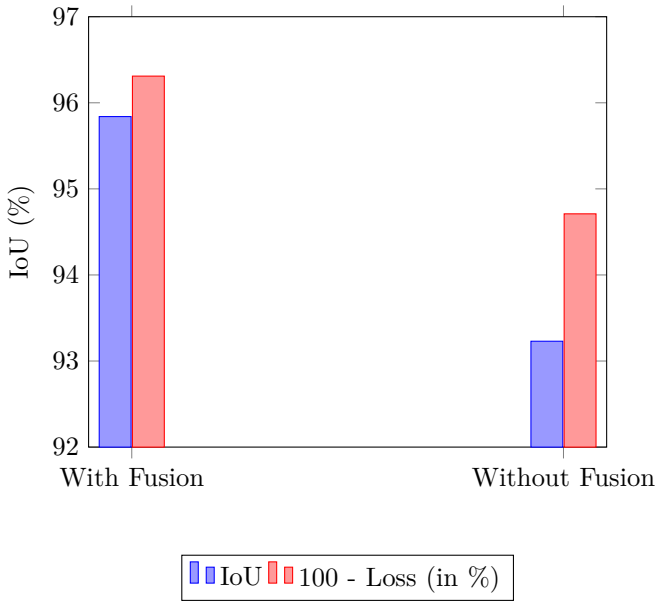
The ablation study results Fig. 8 highlight the contributions of various architectural components to leaf segmentation performance. Increasing the number of encoders from one to two improved the IoU from 95.38% to 95.84%, indicating that a dual encoder design effectively captures richer multiscale features. However, adding more than two encoders slightly degraded performance, suggesting that excessive depth may introduce redundancy or overfitting.

The number of attention heads also influenced segmentation accuracy. Two heads achieved the highest IoU, while increasing beyond this point slightly reduced performance. This suggests that, for our dataset, a modest number of attention heads is sufficient to model inter-leaf dependencies without unnecessarily increasing computational complexity. Furthermore, experiments on feature fusion and edge awareness demonstrate

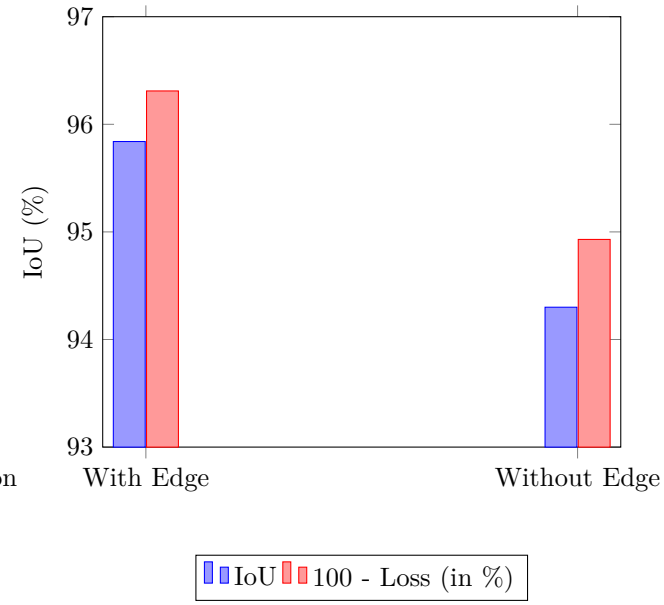


(a) Encoders vs IoU/Loss

(b) Heads vs IoU/Loss



(c) With vs Without Fusion



(d) With vs Without Edge

Fig. 8: Ablation study results: comparison across encoders, heads, fusion, and edge integration.

the importance of explicitly integrating multiscale information and boundary priors. Multi-scale feature fusion increased IoU by over 2%, while incorporating edge information further improved boundary delineation. These results are consistent with prior work showing that edge-guided attention mechanisms produce crisper and more continuous masks, particularly in scenarios with overlapping leaves or challenging illumination.

In addition to the ablation results, comparisons with state-of-the-art segmentation models provide further insights. UNet variants (e.g., VGG16-UNet, Dense-UNet, and UNet++) capture the overall leaf structure but often produce smooth outputs with fragmented or blurred edges. Advanced UNet extensions

such as SDUNet and AWUNet improve fine-grained segmentation but still show limitations in accurately delineating thin boundaries. Transformer-based models such as SegFormer and SegNext demonstrate stronger performance in preserving thin leaflets and venation patterns due to their ability to capture long-range dependencies; however, they may generate over-smoothed masks in dense foliage. In contrast, our proposed architecture produces sharper and more structurally consistent outputs, effectively preserving both global context and intricate boundary details.

Overall, the combination of dual encoders, optimal attention heads, feature fusion, and edge-guided mechanisms provides a balanced architecture that effectively captures both global context and fine leaf boundaries. This design is robust under complex agricultural conditions, where structural complexity, secondary leaflets, and intricate venation patterns pose significant challenges for accurate segmentation. By outperforming state-of-the-art baselines in both qualitative outputs and quantitative metrics, the proposed model demonstrates strong potential for precision agriculture applications such as early weed detection, crop monitoring, and yield estimation.

VI. Conclusion

Accurate leaf segmentation is critical for plant phenotyping, precision agriculture, and automated trait extraction. This study systematically evaluated the impact of encoder depth, attention heads, feature fusion, and edge awareness on segmentation performance. Key findings include Dual encoders provide optimal feature representation without overfitting. A modest number of attention heads is sufficient to capture inter leaf dependencies efficiently. Feature fusion and edge integration significantly enhance boundary precision and mask continuity. Furthermore, Transformer inspired modules, when combined with attention and edge mechanisms, achieve robust segmentation performance under challenging field conditions.

These insights provide guidelines for designing lightweight, efficient, and accurate leaf segmentation networks suitable for real world agricultural applications. Future work may explore the integration of multi-spectral inputs, dynamic attention allocation, and graph based refinement to further improve segmentation accuracy in heterogeneous crop environments.

References

[1] A. Upadhyay, N. S. Chandel, K. P. Singh, *et al.*, "Deep learning and computer vision in plant disease detection: a comprehensive review of techniques, models, and trends in precision agriculture," *Artificial Intelligence Review*, vol. 58, p. 92, 2025.

[2] S. Zhang and C. Zhang, "Modified u-net for plant diseased leaf image segmentation," *Computers and Electronics in Agriculture*, vol. 204, p. 107511, 2023.

[3] J. Tagnamas, H. Ramadan, A. Yahyaouy, and H. Tairi, "Sca-inceptionunext: A lightweight spatial-channel-attention-based network for efficient medical image segmentation," *Knowledge-Based Systems*, vol. 311, p. 113166, 2025.

[4] L. Qian, H. Huang, X. Xia, *et al.*, "Automatic segmentation method using fcn with multi-scale dilated convolution for medical ultrasound image," *Visual Computing*, vol. 39, pp. 5953–5969, 2023.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, 2015.

[6] R. Gu and L. Liu, "An agricultural leaf disease segmentation model applying multi-scale coordinate attention mechanism," *Applied Soft Computing*, vol. 172, p. 112904, 2025.

[7] G. Singh, A. A. Al-Huqail, A. Almogren, *et al.*, "Enhanced leaf disease segmentation using u-net architecture for precision agriculture: A deep learning approach," *Food Science & Nutrition*, vol. 13, no. 7, p. e70594, 2025.

[8] M. Heinrich, M. Stille, and T. Buzug, "Residual u-net convolutional neural network architecture for low-dose ct denoising," *Current Directions in Biomedical Engineering*, vol. 4, pp. 297–300, 2018.

[9] Y. Zhou, H. Chang, X. Lu, and Y. Lu, "Denseunet: Improved image classification method using standard convolution and dense transposed convolution," *Knowledge-Based Systems*, vol. 254, p. 109658, 2022.

[10] O. Oktay, J. Schlemper, L. Le Folgoc, *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[11] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation," *arXiv preprint arXiv:1802.06955*, 2018.

[12] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," *arXiv preprint arXiv:1807.10165*, 2018.

[13] B. Mustapha, Y. Zhou, B. Nawel, S. Chunyan, and X. Zhitao, "Optimized attention u-net for enhanced lung and area of infection segmentation in chest x-rays and ct scans," *Journal of Radiation Research and Applied Sciences*, vol. 18, no. 3, p. 101650, 2025.

[14] X. Yang, H. Li, W. Zhu, and Y. Zuo, "Rshrnet: Improved hrnet-based semantic segmentation for uav rice seedling images in mechanical transplanting quality assessment," *Computers and Electronics in Agriculture*, vol. 234, p. 110273, 2025.

[15] W. Tong, W. Chen, W. Han, X. Li, and L. Wang, "Channel-attention-based densenet network for remote sensing image scene classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 4129–4140, 2020.

[16] H. Zheng, X. Zhou, J. He, *et al.*, "Early season detection of rice plants using rgb, nir–g–b and multispectral images from uav," *Computers and Electronics in Agriculture*, vol. 169, p. 105223, 2020.

[17] Z. Tao, T. Wei, and J. Li, "Wavelet multi-level attention capsule network for texture classification," *IEEE Signal Processing Letters*, vol. 28, pp. 1–1, 2021.

[18] S. Zhao, X. Wu, K. Tian, *et al.*, "A transformer-based hierarchical hybrid encoder network for semantic segmentation," *Neural Processing Letters*, vol. 57, p. 66, 2025.

[19] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *arXiv preprint arXiv:2105.15203*, 2021.

[20] J. Zhong, T. Zeng, Z. Xu, *et al.*, "A frequency attention-enhanced network for semantic segmentation of high-resolution remote sensing images," *Remote Sensing*, vol. 17, p. 402, 2025.

[21] Z. Tang, J. Sun, Y. Tian, J. Xu, *et al.*, "Cvrp: A rice image dataset with high-quality annotations for image segmentation and plant phenomics research," *Plant Phenomics*, vol. 7, no. 1, p. 100025, 2025.

[22] W. Dai, W. Zhu, G. Zhou, G. Liu, *et al.*, "Aisoa-ssformer: An effective image segmentation method for rice leaf disease based on the transformer architecture," *Plant Phenomics*, vol. 6, p. 0218, 2024.

[23] J. Fan, Z. Shi, Z. Ren, Y. Zhou, and M. Ji, "Ddpm-segformer: Highly refined feature land use and land cover segmentation with a fused denoising diffusion probabilistic model and transformer," *International Journal of Applied Earth Observation and Geoinformation*, vol. 133, p. 104093, 2024.

[24] O. Elharrouss, Y. Himeur, Y. Mahmood, *et al.*, "Vits as backbones: Leveraging vision transformers for feature extraction," *Information Fusion*, vol. 118, p. 102951, 2025.

[25] F. Chen, S. A. Tsaftaris, and M. V. Giuffrida, "Gmt: Guided mask transformer for leaf instance segmentation," *arXiv preprint arXiv:2406.17109*, 2025.

[26] R. H. Chowdhury and S. Ahmed, "Mangoleafvit: Leveraging lightweight vision transformer with runtime augmentation for efficient mango leaf disease classification," in *27th International Conference on Computer and Information Technology (ICCIT)*, pp. 699–704, 2024.

[27] S. Hemalatha and J. J. B. Jayachandran, "A multitask learning-based vision transformer for plant disease localization and classification," *Int J Comput Intell Syst*, vol. 17, p. 188, 2024.

[28] H. Yu, T. Fu, B. Li, and X. Xue, "Eaformer: Scene text segmentation with edge-aware transformers," in *Computer Vision – ECCV 2024. Lecture Notes in Computer Science*, vol. 15083, Springer, Cham, 2025.

[29] X. Wang, W. Lu, H. Liu, W. Zhang, and Q. Li, "Daft-net: Dual attention and fast tongue contour extraction using enhanced u-net architecture," *Entropy (Basel)*, vol. 26, no. 6, p. 482, 2024.

[30] B. Yan, C. Wang, and X. Hao, "Research on the performance of the segformer model with fusion of edge feature extraction for metal corrosion detection," *Scientific Reports*, vol. 15, p. 8134, 2025.

[31] S. Haug and J. Ostermann, "A crop/weed field image dataset for the evaluation of computer vision based precision agriculture tasks," in *Computer Vision - ECCV 2014 Workshops*, pp. 105–116, 2015.

[32] A. Sharma and P. K. Mishra, "Dri-unet: Dense residual-inception unet for nuclei identification in microscopy cell images," *Neural Computing and Applications*, vol. 35, pp. 19187–19220, 2023.

[33] Y. Miao, R. Wang, Z. Jing, *et al.*, "Ct image segmentation of foxtail millet seeds based on semantic segmentation model vgg16-unet," *Plant Methods*, vol. 20, p. 169, 2024.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition." arXiv preprint, 2014. http://arxiv.org/abs/1409.1556.

[35] M. Yang, Y. Yuan, and G. Liu, "Sdunet: Road extraction via spatial enhanced and densely connected unet," *Pattern Recognition*, vol. 126, p. 108549, 2022.

[36] I. Delibasoglu, "Incsa-unet: Spatial attention inception unet for aerial images segmentation," *Computing and Informatics*, vol. 40, pp. 1244–1262, 2022.

[37] A. S. Banu and S. Deivalakshmi, "Enhancing leaf area segmentation by using attention gates and knowledge distillation in unet architecture," *JTIT*, vol. 101, pp. 51–62, Aug 2025.
[38] A. S. Banu and S. Deivalakshmi, "Awunet: Leaf area segmentation based on attention gate and wavelet pooling mechanism," *SIViP*, vol. 17, pp. 1915–1924, 2023.
[39] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu, "Segnext: Rethinking convolutional attention design for semantic segmentation," *arXiv preprint arXiv:2209.08575*, 2022.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Contributions

All authors carried out the literature review. Ms. A. Shamim Banu conducted the experiments. Dr. S. Deivalakshmi reviewed the manuscript and approved the final version of the manuscript after peer review. All the authors approved submission of the manuscript to the journal titled *Signal, Image and Video Processing*.

Ethics declarations

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethical approval

Not applicable.

Data availability

The datasets used in the study are publicly available in the repository: https://github.com/cwfid/dataset