

Diffusion–CLIP-Guided Dual-Branch Framework for Infected Region Segmentation in Fruit and Leaf Images

Poornima Basatti Hanuma Gowda,
Basavanna Mahadevappa
0009-0007-0569-2527,
0000-0001-9404-4183
Department of Studies in Computer Science,
Davangere University, Davanagere, India
Email: {poornima.bh, basavanna.m}
@davanagereuniversity.ac.in

Shivakumara Palaiahnakote,
Muhammad Hammad Saleem, Ali Alameer,
Mohamad Saraee
0000-0001-9026-4613, 0000-0002-3625-3021,
0000-0002-7969-3609, 0000-0002-3283-1912
School of Science, Engineering, and Environment, University
of Salford, Salford M5 4WT, United Kingdom.
Data Science and Artificial Intelligence Hub, University of
Salford, Salford M5 4WT, United Kingdom. Email:
{S.Palaiahnakote, M.H.Saleem, A.Alameer1, M.Saraee}
@salford.ac.uk

Niranjan Mallappa Hanumanthu
0000-0002-8440-9270,
Department of Studies in Biotechnology,
Davangere University, Davanagere, India
Email: niranjanmh.bt@davanagereuniversity.ac.in

Abstract—Accurate segmentation of plant diseases caused by viruses, bacteria, and fungi is a critical step for automated crop monitoring and precision agriculture. However, variations in disease appearance, illumination, and background clutter make pixel-level delineation highly challenging. In this paper, we propose a novel diffusion–CLIP-guided dual-branch segmentation framework that explicitly models the relationship between healthy and infected tissue regions. The proposed architecture integrates a diffusion-based normality deviation module to highlight abnormal regions, a CLIP-based semantic interpretation module to provide high-level contextual guidance, and a dual-branch encoder that separately learns representations for healthy and infected regions in the input image. A cross-region reasoning mechanism further enhances relational understanding between normal and infected regions, while multi-scale feature fusion ensures robust segmentation across varying disease scales. Experimental evaluations on fruit and leaf disease datasets demonstrate that the proposed method consistently outperformed existing approaches and achieved Dice scores of 93.10% on fruit images and 90.08% on leaf images.

Index Terms—Plant disease segmentation, diffusion models, CLIP, dual-branch networks, cross-region reasoning, precision agriculture.

I. INTRODUCTION

PLANT diseases caused by viruses, fungi, and bacteria pose a persistent threat to agricultural productivity, crop quality, and global food security [1, 2]. As demand for sustainable and precision agriculture grows, automated image-based disease analysis has attracted significant attention as a non-invasive and scalable solution. Among computer vision tasks, accurate segmentation of diseased regions in fruit and

leaf images is a critical prerequisite for downstream applications, including disease severity estimation, yield loss assessment, treatment planning, and post-harvest quality inspection.

Early research in plant disease analysis primarily focused on image-level classification, where deep convolutional neural networks were employed to identify disease categories from entire images. While these approaches achieved promising classification accuracy, the models struggled to achieve stable and reliable performance. This is due to the presence of an uninfected complex background, limiting their applicability in real-world agricultural decision-making [3–6].

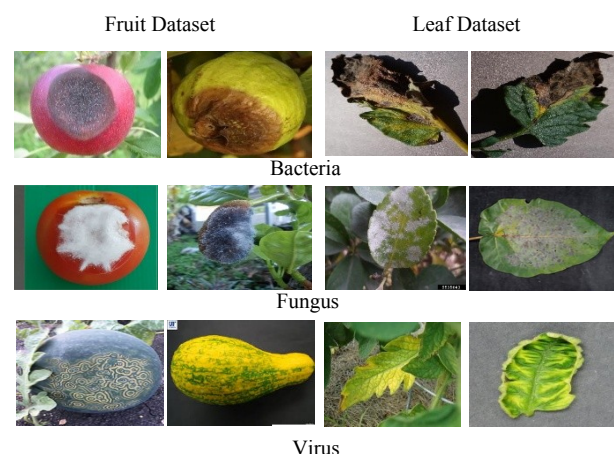


Fig 1. Representative samples from the fruit and leaf disease dataset.

Despite notable advances in deep learning–based segmentation, several challenges remain unresolved in practical agricultural scenarios. Fruit and leaf images often exhibit complex backgrounds, varying illumination conditions, occlusions, and large intra-class variability. It is evident from the samples shown in Fig. 1 that one can see fruit and leaf with a complex background, and similarity between the infected regions of different categories. These challenges become significantly more severe in the case of infections, where viral, fungal, and bacterial diseases cause extensive necrosis, tissue deformation, discoloration, and overlapping lesions.

Moreover, symptoms across different pathogen types frequently share similar visual characteristics at advanced stages, making it difficult for conventional segmentation models to distinguish healthy and diseased regions reliably. There are segmentation methods, including encoder–decoder architectures and multi-scale convolutional models, that often rely heavily on local texture and color cues [7–10]. As a result, their performance degrades when disease boundaries are ambiguous or when infected regions exhibit irregular morphology and diffuse edges. In addition, most existing methods focus on particular datasets and diseases but not on fruit and leaves of different domains.

In this work, we focus on the problem of the infected region in fruit and leaves affected by viruses, bacteria and fungi. Motivated by the excellent performance of diffusion and language models for classification of objects, we propose a diffusion–CLIP-guided dual-branch framework designed to jointly model healthy tissue distribution and disease-specific semantic characteristics [11–14]. By leveraging diffusion-based normality modelling and vision–language semantic alignment, the proposed approach aims to enhance boundary delineation and improve robustness against severe symptom overlap and visual ambiguity.

The main contributions of this work are summarized as follows.

- We introduce a diffusion–CLIP-guided dual-branch segmentation framework for virus, bacteria, and fungus-infected fruit and leaf images.
- Vision–language semantic guidance is integrated to enhance discrimination among visually similar viral, fungal, and bacterial infection patterns.
- A dual-branch architecture with cross-region reasoning explicitly contrasts healthy and infected features to improve boundary delineation and generalization.

The remainder of this paper is organized as follows. Section II reviews related work on plant disease segmentation, diffusion-based modeling, and vision–language integration. Section III presents the proposed methodology in detail, including the overall pipeline, diffusion-based normality modeling, semantic guidance, dual-branch feature learning, and uncertainty-guided fusion. Section IV describes the experimental setup, datasets, evaluation metrics, and implementation details. Section V reports and discusses the quantitative

and qualitative segmentation results, including comparative and ablation studies.

II. BACKGROUND

With the rapid advancement of deep learning, segmentation techniques have evolved from traditional hand-crafted feature–based approaches to end-to-end learning frameworks. Fully Convolutional Networks (FCNs) introduced dense pixel-wise prediction by replacing fully connected layers with convolutional layers, forming the foundation of modern semantic segmentation models [15]. Encoder–decoder architectures, particularly U-Net, further enhanced segmentation accuracy by integrating skip connections that fuse low-level spatial features with high-level semantic representations [16].

To address scale variation and complex background interference, multi-scale feature learning strategies were proposed. Models such as SegNet, PSPNet, and the DeepLab family employed pyramid pooling and atrous (dilated) convolutions to capture contextual information at multiple receptive fields, significantly improving segmentation robustness [17–20]. These architectures have become standard benchmarks across various application domains. Several comprehensive surveys have emphasized that effective multi-scale context aggregation is essential for accurate semantic segmentation [7, 8].

More recently, attention mechanisms and transformer-based architectures have gained prominence in segmentation research. By modeling long-range dependencies and global contextual relationships, transformer-driven models such as SETR and SegFormer overcome the locality constraints of convolutional neural networks, achieving strong performance on complex segmentation tasks [21–23]. These approaches demonstrate effectiveness in handling irregular structures and visually ambiguous regions.

In the agricultural domain, semantic segmentation has been widely adopted for plant disease analysis, enabling precise delineation of infected regions in leaf and fruit images. Deep learning–based segmentation models facilitate quantitative disease severity estimation and support precision agriculture applications. U-Net and DeepLab variants have been extensively applied to segment plant diseases across diverse crops, including tomato, apple, grape, and citrus [24–28]. Prior studies and reviews indicate that segmentation-based approaches provide more informative disease characterization than image-level classification methods alone [16].

Despite these advances, accurately segmenting final-stage plant diseases caused by viruses, fungi, and bacteria remains challenging. Advanced-stage infections are characterized by extensive necrosis, lesion overlaps, and reduced visual contrast between healthy and diseased tissues. Moreover, symptoms across different pathogen types often exhibit high visual similarity, further complicating segmentation. Existing models frequently fail to generalize under these severe conditions, highlighting the need for segmentation frameworks

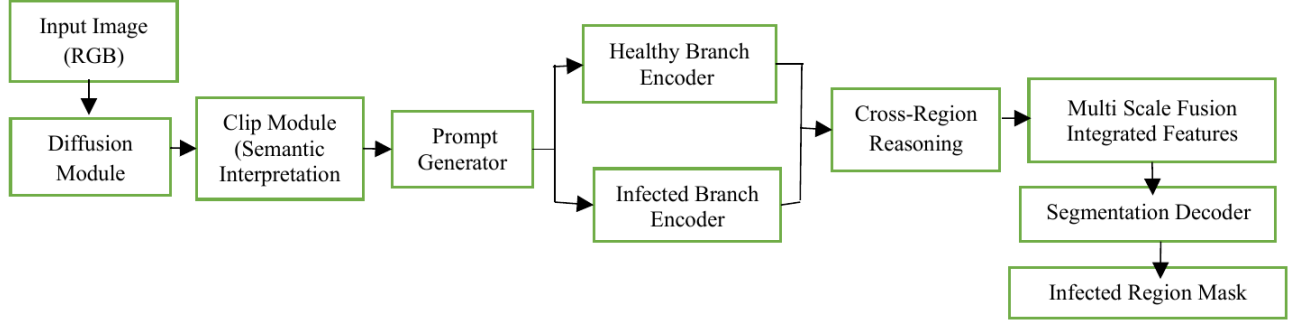


Fig 2. Overall block diagram of the proposed diffusion-CLIP-guided dual-branch segmentation pipeline. The framework takes an input RGB disease image and processes it through diffusion-based normality modeling, CLIP-based semantic interpretation, and a dual-branch encoder consisting of healthy and infected feature learning paths. The extracted features are refined through cross-region reasoning and multi-scale fusion modules, followed by a segmentation decoder to generate the final pixel-level infected region mask.

that incorporate richer contextual modeling and semantic guidance [29–33].

Recent segmentation methods have explored advanced feature learning strategies for accurate disease region localization. TSNUNet introduced a nested U-Net architecture for improved feature representation [34], while Bhatti et al. proposed a plant disease segmentation framework combining dimensionality reduction, fuzzy C-means clustering, and deep learning [35]. These methods demonstrate the effectiveness of enhanced segmentation approaches and are considered for comparison in this study. In addition, most existing models aim to identify diseases or classify the images without segmenting the infected regions in the images. To the best of our knowledge, the proposed segmentation is the first of its kind.

Therefore, this study aims to develop a Diffusion-CLIP-Guided Dual-Branch Segmentation Framework for images infected by viruses, bacteria, and fungi.

III. RESEARCH METHODOLOGY

As mentioned in the previous section, there are models for disease identification and classification. However, these models do not focus on segmenting the region of interest for classification. The models use global features for classification. The global features are indeed insensitive to local regions, and therefore, the performance of the methods may not be stable and reliable. Thus, this work aims to segment the infected regions accurately into fruit and leaf images affected by viruses, bacteria, and fungi. Unlike conventional encoder-decoder architectures that treat all image regions uniformly, the proposed approach decomposes the segmentation process into normality learning, semantic conditioning, and relational feature contrast, enabling robust boundary delineation under severe necrosis, lesion overlap, and visual ambiguity. An overview of the complete pipeline is shown in Fig. 2.

A. Overall Pipeline

Given an RGB input image $I \in \mathbb{R}^{H \times W \times 3}$, the framework processes the image through three sequential phases:

- **Normality and Semantic Analysis** Healthy tissue distribution is learned via a diffusion-based module, while global disease semantics are extracted using a CLIP-based vision-language encoder.
- **Dual-Branch Feature Learning with Cross-Region-Reasoning**: Feature representations are explicitly separated into healthy and infected branches, followed by cross-attention-based relational modeling.
- **Uncertainty-Guided Multi-Scale Fusion and Decoding**. Deviation-aware features are adaptively fused and decoded to generate the final infected region mask.

Each stage is designed to progressively refine feature representations, transitioning from normality estimation to semantic discrimination and finally to precise spatial segmentation. The detailed architecture of the proposed method is presented in Fig. 3.

B. Diffusion-Based Normality Modeling

The infected region segmentation is challenging due to extensive necrosis and blurred boundaries. To address this, we introduce a diffusion-based normality modeling module that learns the distribution of healthy tissue appearance. The diffusion module follows an encoder-decoder structure trained exclusively on healthy samples. Given an input image I , the encoder maps it to a latent representation as defined in Equation (1).

$$z \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times 32} \quad (1)$$

The decoder reconstructs the expected healthy appearance I_{recon} . The normality deviation map is computed as defined in Equation (2).

$$\delta = \|I - I_{\text{recon}}\|_2 \quad (2)$$

which highlights regions that deviate from healthy patterns and serves as a spatial prior for disease localization. The effect of diffusion-based deviation modeling is visualized in Fig. 4 and Fig. 5 (Step 2), where infected regions are selec-

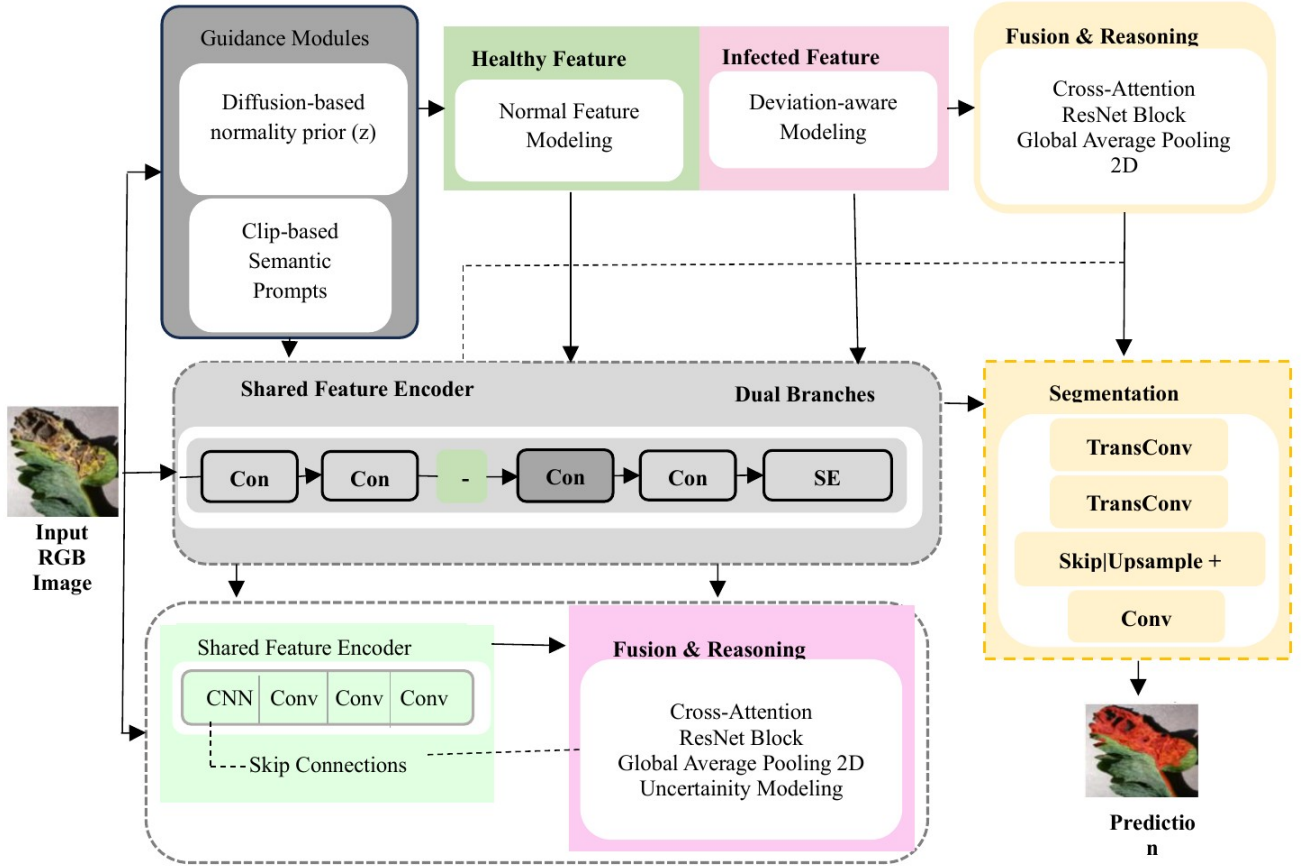


Fig 3. Detailed architecture of the proposed diffusion-CLIP-guided dual-branch segmentation network. The architecture consists of diffusion and CLIP guidance modules for extracting complementary normality and semantic information, a shared feature encoder with skip connections for hierarchical representation learning, separate healthy and infected feature branches for disease-aware feature modeling, a fusion and reasoning module for cross-attention-based feature interaction, and a decoder for producing the final segmentation output

tively emphasized while healthy structures are reconstructed for both fruits and leaves, respectively.

C. CLIP-Based Semantic Interpretation

While diffusion modeling provides spatial deviation cues, it lacks semantic discrimination between disease types. To incorporate high-level contextual understanding, we employ a CLIP-based vision-language semantic interpretation module. The CLIP image encoder maps the input image into a shared embedding space, where it is aligned with text prompts representing healthy and infected tissue as defined in Equation (3).

$$T_{\text{healthy}}, T_{\text{infected}} \quad (3)$$

Cosine similarity between image and text embedding produces semantic confidence scores as defined in Equation (4).

$$P = \text{softmax}(\text{sim}(E(I), T_{\text{healthy}}), \text{sim}(E(I), T_{\text{infected}})) \quad (4)$$

These scores are transformed via lightweight MLPs into channel-wise modulation vectors, which guide downstream feature learning without enforcing explicit spatial constraints.

D. Dual-Branch Feature Encoding

To explicitly model the contrast between healthy and diseased tissue, we introduce a dual-branch encoder architecture. First, shared low-level features are extracted by concatenating the RGB input with diffusion latent features as defined in Equation (5).

$$F_{\text{shared}} = \text{Conv}(\text{Concat}[I, z]) \quad (5)$$

Two parallel branches then process these shared features:

Step 1: Healthy Branch: The healthy branch learns structural patterns of normal tissue as defined in Equation (6).

$$F_h = \text{Encoder}_h(F_{\text{shared}}) \odot \text{Adapt}(P_{\text{healthy}}) \quad (6)$$

Step 2: Infected Branch: The infected branch focuses on abnormal regions by integrating semantic guidance and deviation modulation as defined in Equation (7).

$$F_i = \text{Encoder}_i(F_{\text{shared}}) \odot \text{Adapt}(P_{\text{infected}}) \odot \delta \quad (7)$$

This separation allows the network to learn complementary representations, rather than entangling healthy and infected features within a single latent space. Intermediate branch outputs are visualized in Fig. 4 and Fig. 5 (Steps 6–7), showing distinct specialization of healthy and infected representations.

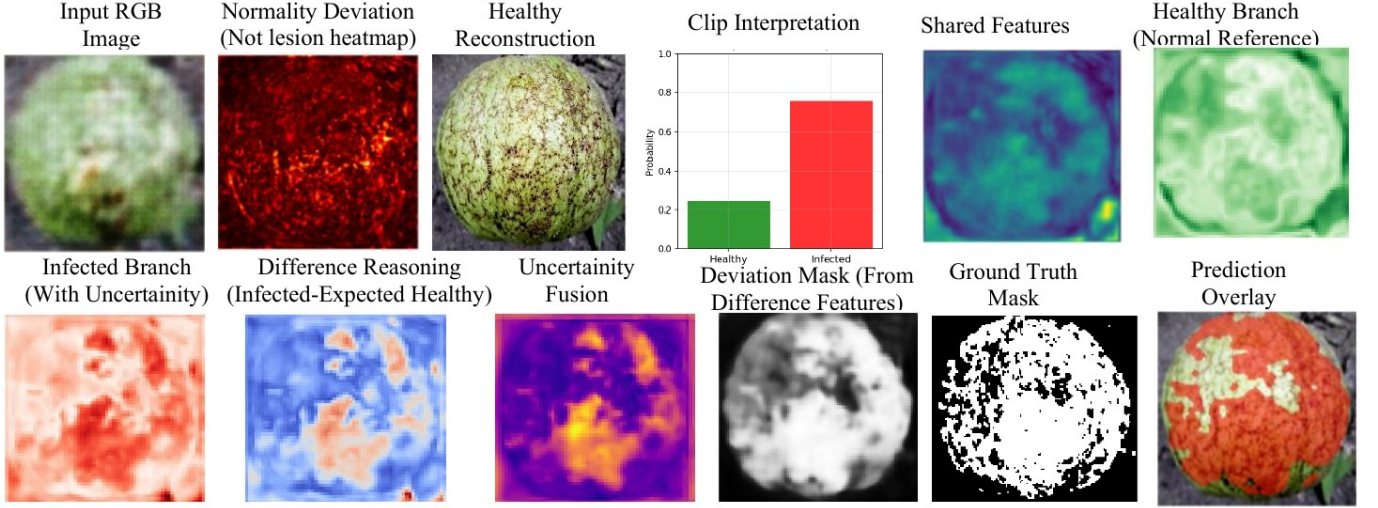


Fig 4. Intermediate feature visualization of the proposed diffusion-CLIP-guided dual-branch framework on a fruit disease sample. The visualization shows the progressive feature refinement process starting from the original fruit input image, followed by diffusion-based healthy appearance reconstruction, deviation map generation for identifying infected regions, CLIP-based semantic interpretation, healthy and infected branch feature representations, and the final predicted segmentation mask highlighting the infected area.

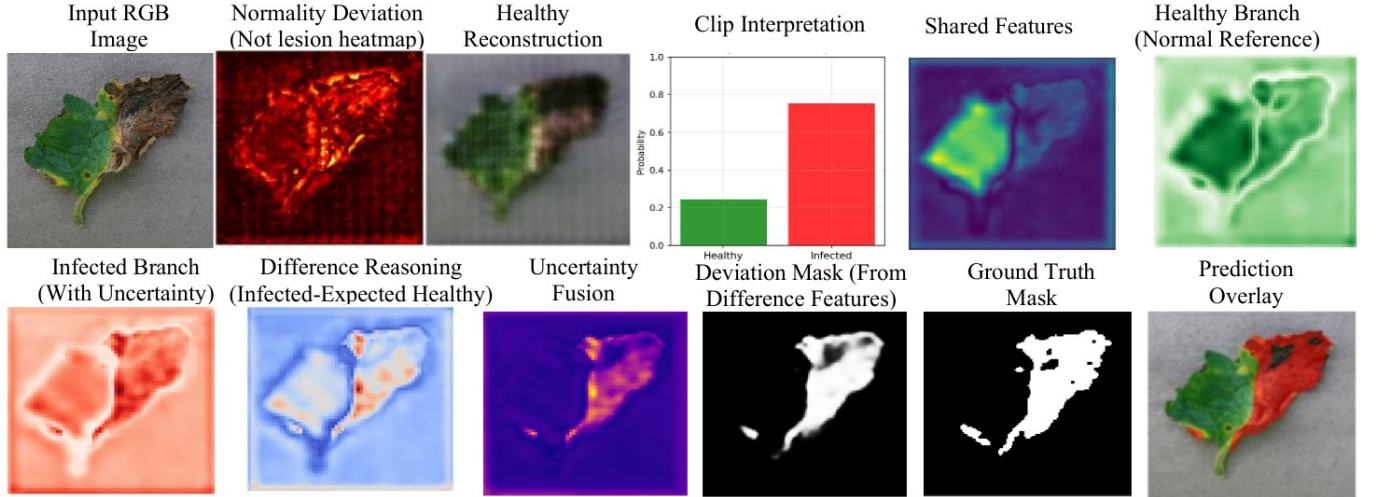


Fig 5. Intermediate feature visualization of the proposed diffusion-CLIP guided dual-branch framework on a leaf disease sample. The visualization demonstrates the feature transformation process from the original leaf image through diffusion-based reconstruction, abnormality deviation estimation, CLIP-based semantic feature extraction, healthy and infected branch representations, and the final segmentation output showing the detected infected region.

E. Cross-Region Reasoning

To explicitly model *what makes a region abnormal*, we introduce a cross-region reasoning module that contrasts infected features against healthy references. Cross-attention is applied using infected features as queries and healthy features as keys and values, as defined in Equation (8).

$$F_{\text{diff}} = F_i - \gamma \cdot \text{Attention}(Q = F_i, K = F_h, V = F_h) \quad (8)$$

This formulation encourages the network to focus on deviations from expected healthy appearance, sharpening boundaries and suppressing false positives.

F. Uncertainty-Guided Multi-Scale Fusion and Decoding

Disease appearance varies significantly in size and morphology. To address this, we employ uncertainty-guided multi-scale fusion. An uncertainty map is estimated from the diffusion deviation, and fused features are computed as defined in Equation (9).

$$F_{\text{fused}} = u \cdot \text{DilatedConv}(F_{\text{diff}}) + (1 - u) \cdot \text{Conv}(F_{\text{diff}}) \quad (9)$$

This adaptive fusion allows large receptive fields in uncertain regions and fine-grained processing in confident areas. The segmentation decoder upsamples fused features using transposed convolutions and skip connections from both

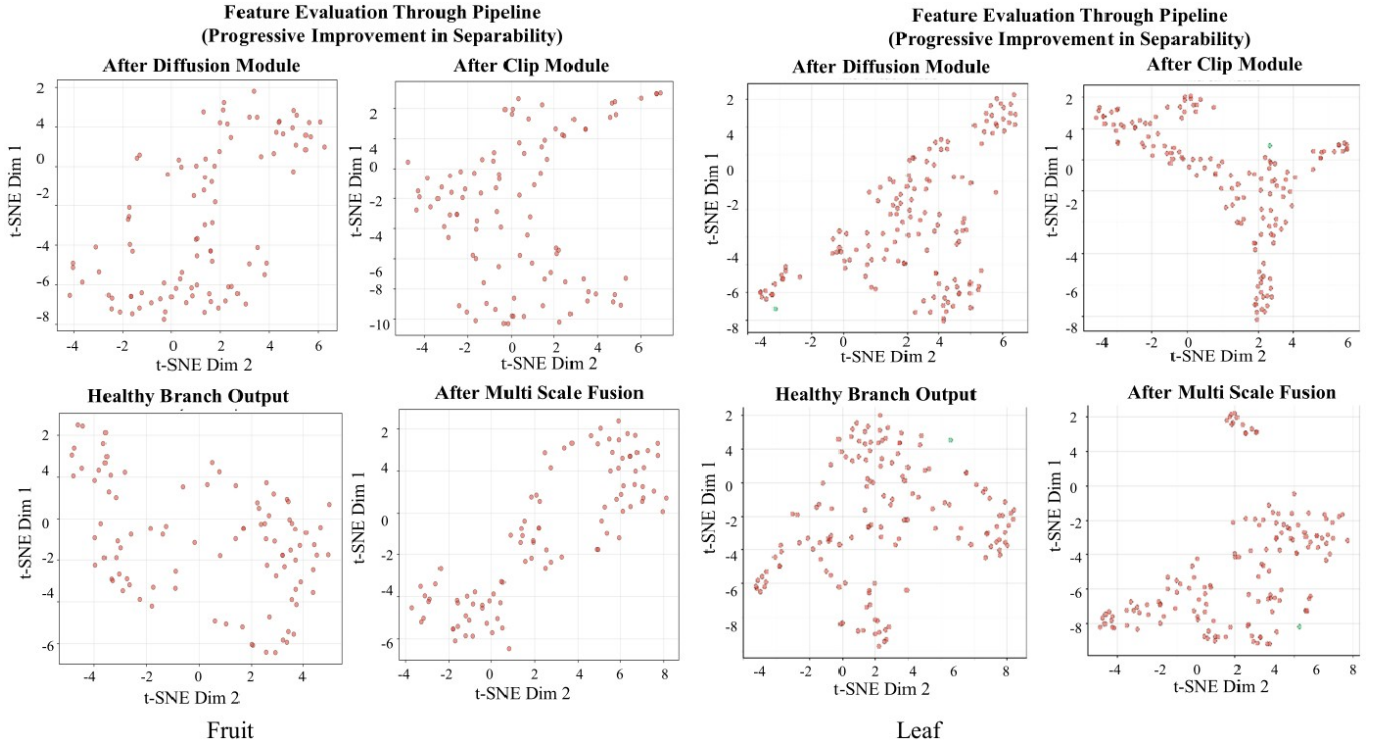


Fig. 6. t-SNE visualization of feature distributions at different stages of the proposed diffusion-CLIP guided dual-branch framework for fruit and leaf disease datasets. The plots illustrate the progressive feature refinement process through diffusion-based normality modeling features, representing the initial feature distribution after healthy appearance reconstruction, CLIP-based semantic feature representations obtained using visual-semantic guidance, dual-branch encoder features after separating healthy and infected representations, and final multi-scale fusion features used for segmentation prediction. The left and right groups correspond to fruit and leaf datasets, respectively, showing the improvement in feature separability across the proposed pipeline.

branches, producing the final prediction as defined in Equation (10).

$$M = \sigma(\text{Decoder}(F_{\text{fused}})) \quad (10)$$

G. Progressive Feature Evolution Analysis

To analyze the contribution of each module, t-SNE visualizations are presented in Fig. 6. Features extracted after the diffusion module exhibit diffuse distributions, reflecting non-discriminative normality modeling. CLIP-guided features show improved global structure but remain intermixed. After dual-branch encoding and cross-region fusion, feature distributions become increasingly compact and structured, confirming progressive semantic separation. These visualizations demonstrate that discriminative segmentation emerges only through the joint integration of diffusion deviation, semantic prompts, and relational reasoning, validating the design of the proposed framework.

Training Strategy: Training is performed in two phases:

Diffusion Pre-training on healthy images to learn normality priors. End-to-End Optimization of the full network using a composite loss as defined in Equation (11).

$$L = L_{\text{seg}} + \lambda_{\text{cons}} L_{\text{cons}} \quad (11)$$

where segmentation loss combines BCE and Dice loss, and consistency loss enforces accurate reconstruction of healthy regions.

IV. EXPERIMENTAL RESULTS

A. Datasets

The proposed method is evaluated on fruit and leaf disease image datasets containing annotated ground-truth masks, focusing specifically on final-stage virus-, fungus-, and bacteria-infected samples, where lesion spread, necrosis, and discoloration are prominent. Sample images with disease-affected regions (bacteria, Fungus and Virus) can be seen in Fig. 7. The dataset was collected from the Biotechnology Department, Davangere University, Karnataka, India. The images were verified by domain experts and manually annotated to generate pixel-level infected region masks. The dataset contains bacterial, fungal, and viral disease samples from both fruit and leaf categories. Only final-stage infection images were selected because they provide visible lesion patterns suitable for accurate segmentation.

Pixel-level annotations were prepared by manually identifying infected regions corresponding to bacterial, fungal, and viral symptoms. The infected areas were marked as foreground pixels, while healthy regions were considered background pixels. The generated masks preserve lesion boundaries, irregular infection patterns, and affected tissue regions, and were visually verified before being used for model training and evaluation.

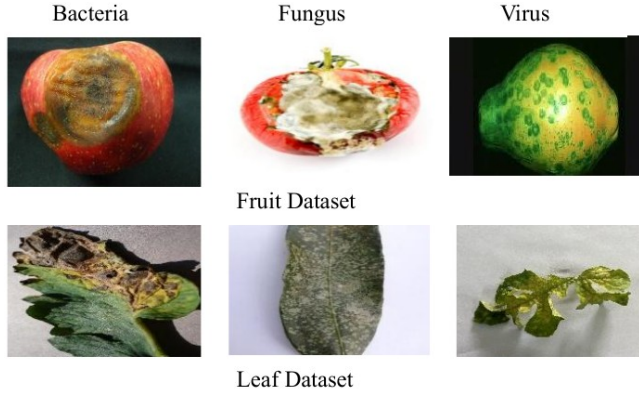


Fig 7. Sample images from the fruit and leaf disease datasets used in this study.

TABLE I. DATASET DESCRIPTION AND PIXEL-LEVEL ANNOTATION DETAILS FOR FRUIT AND LEAF DISEASE SEGMENTATION

Dataset	Disease Categories	Number of Images	Total	Annotation
Fruit	Bacteria	250	600	Pixel-level infected region masks
	Fungus	200		
	Virus	150		
Leaf	Bacteria	300	700	Pixel-level infected region masks
	Fungus	200		
	Virus	200		

The sample images exhibit significant variability in terms of illumination conditions, background complexity, and infection severity, making the segmentation task challenging. The datasets include diverse disease types, lighting conditions, and background variations. A few samples of the dataset are presented in Fig. 7. More details of the dataset are listed in Table I.

The dataset was randomly divided into training and validation subsets using an 80:20 split ratio. The same split configuration was maintained for all experiments to ensure a fair comparison between the proposed framework and existing methods. The training subset was used for optimization, while the validation subset was used to monitor model performance during training.

B. Evaluation Metrics

To quantitatively evaluate segmentation performance, five widely used metrics are employed: Dice coefficient, Intersection over Union (IoU), Precision, Recall, and F1-score. These metrics are computed using the number of true positives (TP), false positives (FP), and false negatives (FN) obtained by comparing the predicted segmentation mask with the ground-truth annotation.

The **Dice coefficient** measures the similarity between the predicted infected region and the ground-truth mask and is defined as defined in Equation (12).

$$\text{Dice} = 2TP / (2TP + FP + FN) \quad (12)$$

A higher Dice score indicates greater overlap between predicted and true infected regions. This metric is particu-

larly useful for medical and agricultural image segmentation tasks where accurate boundary overlap is important.

The **Intersection over Union (IoU)**, also known as the Jaccard Index, evaluates the ratio between the intersection and union of predicted and ground-truth regions as defined in Equation (13).

$$\text{IoU} = TP / (TP + FP + FN) \quad (13)$$

IoU measures the accuracy of region localization and provides a stricter evaluation compared to Dice since it penalizes false predictions more strongly.

Precision quantifies the proportion of correctly predicted infected pixels among all pixels predicted as infected, as defined in Equation (14).

$$\text{Precision} = TP / (TP + FP) \quad (14)$$

High precision indicates that the model produces fewer false-positive detections, meaning healthy regions are rarely misclassified as diseased.

Recall (also known as sensitivity) measures the proportion of actual infected pixels that are correctly detected, as defined in Equation (15).

$$\text{Recall} = TP / (TP + FN) \quad (15)$$

High recall indicates that most infected regions are successfully identified without being missed.

The **F1-score** represents the harmonic mean of Precision and Recall as defined in Equation (16).

$$F1 = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (16)$$

The F1-score provides a balanced evaluation when both false positives and false negatives must be considered simultaneously. Together, these metrics provide a comprehensive evaluation of segmentation performance by assessing region overlaps, localization accuracy, and the balance between false detections and missed infected regions.

Implementation Details: All experiments were implemented using Python and PyTorch. The input images were resized to 256×256 pixels before training. The network was optimized using the Adam optimizer with a learning rate of 0.0001 for 60 epochs with a batch size of 4. Binary Cross-Entropy and Dice-based optimization were used for segmentation learning. GPU acceleration was used during training and evaluation. All compared methods were trained under the same experimental settings.

The training and validation loss curves for fruit and leaf datasets are shown in Fig. 8. The decreasing trend of both training and validation losses indicates stable convergence of the proposed framework during optimization.

C. Ablation Study

To evaluate the contribution of individual architectural components, extensive ablation experiments were conducted, as summarized in Table II. Removing key modules such as the diffusion normality deviation, CLIP-based semantic interpretation, dual-branch encoding, or cross-region reasoning consistently leads to noticeable performance degradation. For instance, eliminating both diffusion and

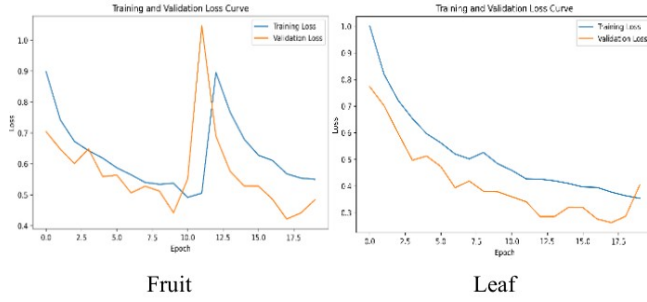


Fig 8. Training and Loss curves for the fruit and leaf datasets.

CLIP guidance results in a Dice score drop of more than **3.5%** on fruit images compared to the full model.

The results demonstrate that diffusion-based normality modeling plays a critical role in highlighting abnormal regions by learning healthy tissue distributions. Similarly, the absence of CLIP interpretation or semantic prompt generation reduces the model's ability to disambiguate visually similar healthy and diseased regions, particularly in advanced infection stages. Single-branch variants and models without cross-region reasoning further confirm that explicitly modeling interactions between healthy and infected features significantly improves boundary localization and segmentation consistency. Overall, the ablation analysis validates that the proposed framework benefits from the complementary integration of diffusion modeling, vision-language semantic guidance, and dual-branch relational learning.

To provide a comprehensive comparison, conventional encoder-decoder and transformer-based segmentation archi-

tectures were implemented using the same dataset split and experimental configuration. The baseline models include U-Net, DeepLabV3+, and SegFormer, representing convolutional, multi-scale, and transformer-based segmentation approaches, respectively.

The overall segmentation performance of the proposed method is summarized using the Dice coefficient, Intersection over Union (IoU), Precision, Recall, and F1-score. As reported in Table I, the proposed framework achieves a Dice score of 93.10% on fruit images and 90.08% on leaf images, indicating strong pixel-level agreement with ground-truth annotations. High IoU values of 87.29% for fruits and 82.34% for leaves further demonstrate accurate region delineation, even under severe infection conditions. The balanced Precision and Recall values suggest that the model effectively minimizes both false positives and false negatives. This is particularly important for final-stage disease segmentation, where excessive necrosis and overlapping symptoms often cause conventional models to over-segment infected regions or miss subtle boundaries. The consistent F1-scores across both datasets confirm the robustness of the proposed framework across different plant organs and disease appearances.

D. Comparative Study with Existing Methods

Table III presents a comparative analysis between the proposed approach and recent state-of-the-art plant disease segmentation methods. The comparison methods were selected based on their relevance to plant disease segmentation and their ability to model spatial and semantic information. These methods represent recent CNN-based, transformer-based, and feature-enhanced segmentation approaches. The

TABLE II. DICE, IOU, PRECISION, RECALL, AND F1-SCORE OF THE KEY STEPS OF THE PROPOSED METHOD FOR FRUIT AND LEAF DISEASE SEGMENTATION

#	Key Steps	Fruit					Leaf				
		Dice	IoU	Precision	Recall	F1	Dice	IoU	Precision	Recall	F1
1	No Diffusion + No CLIP	89.58	81.73	87.49	91.88	89.58	85.56	80.55	87.25	83.14	84.56
2	RGB + CLIP Only	90.21	84.60	91.30	89.22	90.21	84.43	79.33	83.20	82.22	84.43
3	No Diffusion Module	91.52	85.45	90.42	89.77	91.52	87.00	82.42	88.52	87.03	87.00
4	Only RGB + Decoder	88.26	81.98	89.92	88.68	88.26	83.82	72.78	85.46	82.76	83.82
5	Dual Branch w/o Cross Reasoning	89.58	81.63	89.34	89.95	89.58	83.76	72.66	81.12	86.79	83.76
6	No CLIP Interpretation	89.55	81.51	88.97	90.23	89.55	83.78	72.82	85.19	82.84	83.78
7	Single Branch w/ All Features	89.37	81.25	89.28	89.55	89.37	83.85	72.86	84.32	83.71	83.85
8	No Prompt Generator	89.34	81.29	89.06	89.95	89.34	83.65	72.55	80.09	87.86	83.65
9	No Deviation Modulation	89.27	81.22	88.87	89.75	89.27	83.45	72.27	83.79	83.61	83.45
10	No CLIP + No Prompts	89.05	80.82	88.65	89.79	89.05	83.55	72.55	88.35	79.90	83.55
11	No Cross-Reasoning	88.86	80.46	89.87	88.13	88.86	84.31	73.47	84.97	83.98	84.31
12	RGB + Diffusion Only	88.73	80.25	89.75	88.00	88.73	83.30	72.04	83.06	84.01	83.30
13	No Dual Branch	88.70	80.24	90.17	87.49	88.70	85.29	74.94	86.34	84.64	85.29
14	U-Net	90.59	83.04	93.99	87.59	90.59	84.76	74.49	92.39	79.20	84.76
15	DeepLabV3+	82.88	71.86	87.21	80.53	82.88	79.93	67.67	87.63	74.76	79.93
16	SegFormer	80.61	68.55	79.81	82.66	80.61	71.83	57.27	61.74	88.52	71.83
17	Proposed Method	93.10	87.29	94.29	91.85	93.10	90.08	82.34	90.21	90.18	90.08

Table III. DICE, IOU, PRECISION, RECALL, AND F1-SCORE OF THE PROPOSED AND EXISTING METHODS ON THE FRUIT AND LEAF DATASET.

Method	Fruit					Leaf				
	Dice	IoU	Precision	Recall	F1	Dice	IoU	Precision	Recall	F1
Proposed Method	93.10	87.29	94.29	91.85	93.10	90.08	82.34	90.21	90.18	90.08
Sun et al. (2026) [34]	86.53	76.26	78.79	95.96	86.53	82.50	70.22	84.17	80.90	82.50
Bhatti et al. (2024) [35]	81.62	81.11	87.47	91.77	89.57	75.14	65.72	77.37	81.44	79.31
Zhang et al. (2024) [25]	90.21	82.35	93.10	87.75	90.21	88.47	79.68	90.85	86.20	88.47

proposed framework consistently outperforms existing methods across all evaluation metrics for both fruit and leaf datasets. Compared to Sun et al. [34], the proposed method improves Dice scores by over 6.5% on fruit images and 7.5% on leaf images. Similar improvements are observed when compared to Bhatti et al. [35], particularly in terms of IoU and F1-score. These performance gains highlight the limitations of conventional multi-scale convolutional models when applied to final-stage infections. In contrast, the proposed diffusion-CLIP-guided architecture effectively captures global semantic context and structural deviations, enabling more reliable segmentation under complex visual conditions.

E. Qualitative Results for Visual Analysis

To visually demonstrate the effectiveness of the proposed method, Fig. 9 and Fig. 10 present representative segmentation results for the samples shown in Section 4.1. The figure compares the input images, ground-truth masks, predicted masks, and overlay visualizations of the infected regions.

Fig. 11 presents representative failure cases of the proposed segmentation framework. As observed, the model struggles to accurately delineate infected regions in scenarios characterized by severe necrosis, irregular lesion boundaries, and low contrast between healthy and diseased tissues. In certain cases, the infected regions exhibit diffuse and ambiguous edges, leading to under-segmentation or over-segmentation. Additionally, visual similarities between advanced-stage infections and surrounding healthy tissue result in misclassification. Complex backgrounds, illumination variations, and overlapping symptoms further contribute to performance degradation. These challenges highlight the inherent difficulty of segmenting final-stage plant diseases and indicate potential directions for improvement, such as enhanced boundary refinement and better modeling of subtle texture variations.

V. CONCLUSION AND FUTURE WORK

This paper proposed a diffusion-CLIP-guided dual-branch framework for accurate segmentation of infected regions in fruits and leaves affected by viruses, bacteria, and fungi. By jointly modeling healthy tissue appearance through diffusion-based normality learning and incorporating high-level semantic guidance via a vision-language model, the proposed approach effectively addresses challenges posed by severe necrosis, overlapping lesions, and visual ambiguity at advanced disease stages. The dual-branch design with cross-region reasoning enables comple-

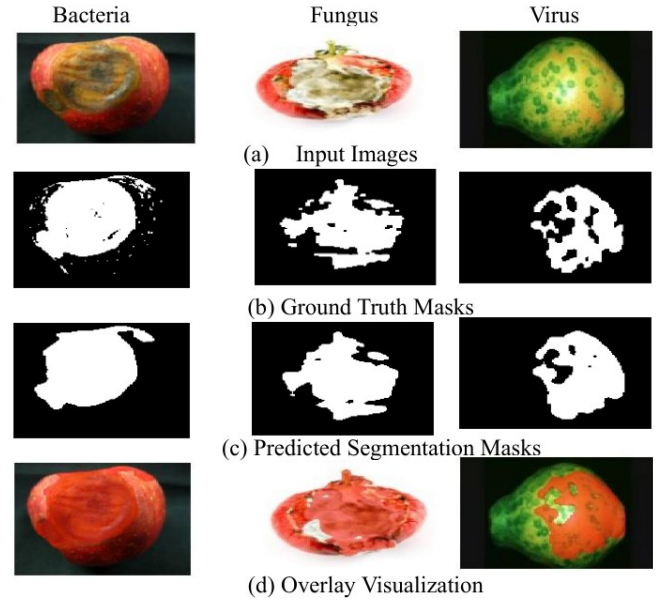


Fig 9. Sample segmentation results for the fruit images using the proposed model.

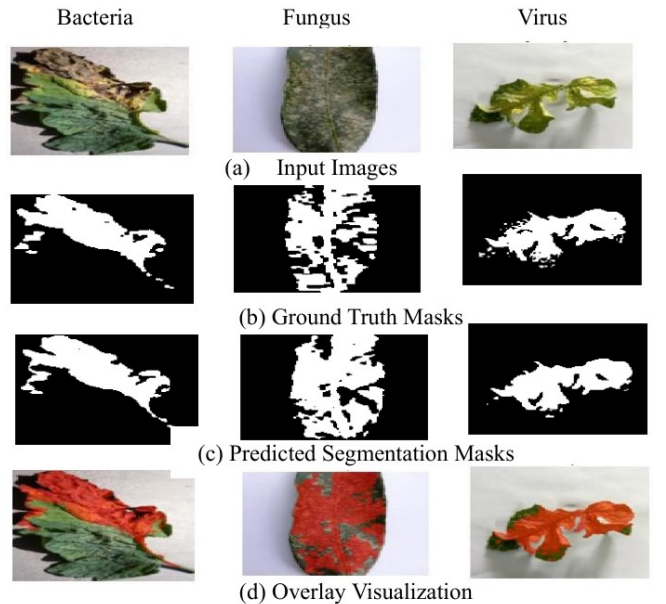


Fig 10. Sample segmentation results for Leaf Dataset images using our model

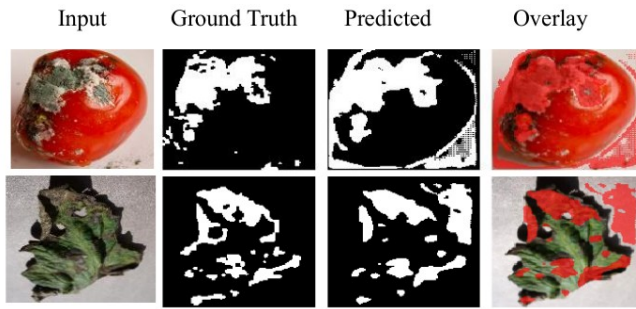


Fig 11. Failure case analysis of the proposed framework. The samples include challenging input images with complex disease patterns, corresponding ground-truth masks, predicted segmentation results, and error visualization highlighting missed or incorrectly segmented infected regions.

mentary feature learning, resulting in improved boundary delineation and robust segmentation performance compared to existing methods.

Although the proposed framework introduces additional computation due to diffusion-based refinement, CLIP-based semantic guidance, and dual-branch feature learning, these components improve infected region localization and feature discrimination. The current implementation is suitable for workstation-based agricultural analysis. Future work will focus on lightweight optimization and efficient deployment on edge devices.

This work considers the disease images at the final stage for segmentation. Therefore, the proposed method may not be effective for images at early and intermediate stages, where the infected regions may not provide visible cues for segmentation. Thus, there is scope for improving the proposed method to address the challenges of early and intermediate stage images.

The proposed segmentation framework can support practical agricultural applications by providing visual localization of infected regions, which can assist farmers, agricultural researchers, and plant disease experts in disease assessment. The segmentation outputs can be integrated into smart farming systems or image-based disease monitoring platforms to enable automated analysis. Future improvements will focus on developing lightweight versions and user-friendly interfaces for practical field deployment.

Further, this work does not include classification results to show the advantages of the segmentation, as it is beyond the scope of this work. For the next work, the results will include classification experiments with and without segmentation. Since the dataset was collected from a specific agricultural environment, variations related to crop types, imaging conditions, and geographical regions may influence generalization. Future studies will include larger multi-location datasets to improve robustness.

GENERATIVE AI DISCLOSURE STATEMENT

The authors confirm that Generative AI was not used in this research and in the preparation of this manuscript.

REFERENCES

- [1] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Advances in Neural Information Processing Systems*, vol. 34, pp. 12077–12090, 2021, doi: 10.48550/arXiv.2105.15203.
- [2] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 1290–1299, doi: 10.1109/CVPR52688.2022.00135.
- [3] F. Li, H. Zhang, H. Xu, S. Liu, L. Zhang, L. M. Ni, and H. Y. Shum, "Mask DINO: Towards a unified transformer-based framework for object detection and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 3041–3051, doi: 10.1109/CVPR52729.2023.00297.
- [4] Y. Wu, Z. Xi, F. Liu, W. Hu, H. Feng, and Q. Zhang, "A deep semantic network-based image segmentation of soybean rust pathogens," *Front. Plant Sci.*, vol. 15, Art. no. 1340584, 2024, doi: 10.3389/fpls.2024.1340584.
- [5] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, 2017, doi: 10.1109/TPAMI.2016.2644615.
- [6] G. Singh, A. A. Al-Huqail, A. Almgren, S. Kaur, K. Joshi, and S. Bharany, "Enhanced leaf disease segmentation using U-Net architecture for precision agriculture: A deep learning approach," *Food Sci. Nutr.*, vol. 13, no. 7, Art. no. e70594, 2025, doi: 10.1002/fsn3.70594.
- [7] G. Zheng, Z. Jiang, X. Zhang, and D. Jiang, "Multi-scale feature fusion-based semantic segmentation network for agricultural remote sensing images," *Chem. Biol. Technol. Agric.*, vol. 12, 2025, doi: 10.1186/s40538-025-00833-8.
- [8] K. Sahu, S. Tiwari, M. K. Singh, J. S. P., H. K. Shakya, and G. Kumar, "Plant disease detection using a hybrid dilated CNN with attention mechanisms and optimized Mask R-CNN segmentation," *Sci. Rep.*, vol. 15, Art. no. 42008, 2025, doi: 10.1038/s41598-025-26192.
- [9] M. K. Surehli, N. Aggarwal, G. Joshi, and H. Nayyar, "Semantic segmentation of plant structures with deep learning and channel-wise attention mechanism," *J. Telecommun. Inf. Technol.*, vol. 99, no. 1, pp. 56–66, 2025, doi: 10.26636/jtit.2025.1.1853.
- [10] H. Xu, J. Song, and Y. Zhu, "Evaluation and comparison of semantic segmentation networks for rice identification based on Sentinel-2 imagery," *Remote Sens.*, vol. 15, no. 6, Art. no. 1499, 2023, doi: 10.3390/rs15061499.
- [11] D. Sotoude, M. Hoseinkhani, and A. Amiri Tehranizadeh, "Context-aware fusion of transformers and CNNs for medical image segmentation," *Informat. Med. Unlocked*, vol. 43, Art. no. 101396, 2023, doi: 10.1016/j.imu.2023.101396.
- [12] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3523–3542, 2022, doi: 10.1109/TPAMI.2021.3059968.
- [13] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818, doi: 10.1007/978-3-030-01234-2_49.
- [14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, doi: 10.48550/arXiv.1409.1556.
- [15] L. Lei, Q. Yang, L. Yang, T. Shen, R. Wang, and C. Fu, "Deep learning implementation of image segmentation in agricultural applications: A comprehensive review," *Artif. Intell. Rev.*, vol. 57, Art. no. 149, 2024, doi: 10.1007/s10462-024-10775-6.
- [16] T. Zhou and W. Wang, "Prototype-based semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 10, pp. 6858–6872, Oct. 2024, doi: 10.1109/TPAMI.2024.3387116.
- [17] S. Zhang, Y. Zhao, and E. Y. Lam, "Semi-supervised semantic segmentation for light field images using disparity information," *IEEE Trans. Image Process.*, 2024, doi: 10.1109/TIP.2024.3441930.
- [18] Z. Zheng, G. Huang, X. Yuan, C.-M. Pun, H. Liu, and W.-K. Ling, "Quaternion-valued correlation learning for few-shot semantic segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 5, pp. 2102–2115, May 2023, doi: 10.1109/TCSVT.2022.3223150.

- [19] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, et al., "Segment Anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Paris, France, 2023, pp. 3992–4003, doi: 10.48550/arXiv.2304.02643.
- [20] T. Yang, Y. Wang, and J. Lian, "Plant diseased lesion image segmentation and recognition based on improved multi-scale attention network," *Appl. Sci.*, vol. 14, no. 5, Art. no. 1716, 2024, doi: 10.3390/app14051716.
- [21] C. Zhao, C. Li, X. Wang, X. Wu, Y. Du, H. Chai, T. Cai, H. Xiang, and Y. Jiao, "Plant disease segmentation networks for fast automatic severity estimation under natural field scenarios," *Agriculture*, vol. 15, no. 6, Art. no. 583, 2025, doi: 10.3390/agriculture15060583.
- [22] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 7262–7272, doi: 10.48550/arXiv.2105.05633.
- [23] Y. Rao, W. Zhao, G. Chen, Y. Tang, Z. Zhu, G. Huang, J. Zhou, and J. Lu, "DenseCLIP: Language-guided dense prediction with context-aware prompting," *arXiv preprint arXiv:2112.01518*, 2022, doi: 10.48550/arXiv.2112.01518.
- [24] A. Das, F. Pathan, J. R. Jim, M. M. Kabir, and M. F. Mridha, "Deep learning-based classification, detection, and segmentation of tomato leaf diseases: A state-of-the-art review," *Artif. Intell. Agric.*, vol. 15, pp. 192–220, 2025, doi: 10.1016/j.iaia.2025.02.006.
- [25] X. Zhang, F. Li, H. Zheng, and W. Mu, "UPFormer: U-shaped perception lightweight transformer for segmentation of field grape leaf diseases," *Expert Syst. Appl.*, vol. 249, Art. no. 123546, 2024, doi: 10.1016/j.eswa.2024.123546.
- [26] H. Mao and J. Lazaro, "Remote sensing image segmentation methods based on deep learning models," *J. Comput. Signal Syst. Res.*, vol. 2, no. 2, pp. 124–133, 2025, doi: 10.71222/15ra5x93.
- [27] J. Cheng, H. Li, D. Li, S. Hua, and V. S. Sheng, "A survey on image semantic segmentation using deep learning techniques," *Comput. Mater. Continua*, vol. 74, no. 1, pp. 1941–1957, 2023, doi: 10.32604/cmc.2023.032757.
- [28] A. K. Singh, A. Rao, P. Chattopadhyay, R. Maurya, and L. Singh, "Effective plant disease diagnosis using vision transformer trained with leafy-generative adversarial network-generated images," *Expert Syst. Appl.*, vol. 254, Art. no. 124387, 2024, doi: 10.1016/j.eswa.2024.124387.
- [29] Z. Xu, D. Wu, C. Yu, X. Chu, N. Sang, and C. Gao, "SCTNet: Single-branch CNN with transformer semantic information for real-time segmentation," *IEEE Access*, vol. 12, pp. 98765–98784, 2024, doi: 10.1609/aaai.v38i6.28457.
- [30] Y. Zhang, Z. Qiu, T. Yao, D. Liu, and T. Mei, "Fully convolutional adaptation networks for semantic segmentation," *arXiv preprint arXiv:1804.08286*, 2018, doi: 10.48550/arXiv.1804.08286.
- [31] J. Wu, R. Fu, H. Fang, Y. Zhang, Y. Yang, H. Xiong, H. Liu, and Y. Xu, "MedSegDiff: Medical image segmentation with diffusion probabilistic model," *arXiv preprint arXiv:2211.00611*, 2023, doi: 10.48550/arXiv.2211.00611.
- [32] H. Yu, H. Li, W. Yang, L. Yu, and G.-S. Xia, "Detecting line segments in motion-blurred images with events," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 5, pp. 2866–2881, May 2024, doi: 10.1109/TPAMI.2023.3334877.
- [33] Z. Huang, X. Wang, L. Huang, S. Huang, Y. Wei, and W. Liu, "CC-Net: Criss-cross attention for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5489–5503, 2022, doi: 10.1109/TPAMI.2020.3007032.
- [34] L. Sun, Z. Chen, X. Zhu, Y. Bi, C. Liu, and Q. M. J. Wu, "TSNUNet: Two-stage nested U-network for salient object detection," *J. Vis. Commun. Image Represent.*, vol. 116, Art. no. 104737, 2026, doi: 10.1016/j.jvcir.2026.104737.
- [35] M. A. Bhatti, Z. Zeeshan, S. M. Syam, U. Bhatti, A. Khan, Y. Ghadi, S. Alsenan, Y. Li, M. Asif, and T. Afzal, "Advanced plant disease segmentation in precision agriculture using optimal dimensionality reduction with fuzzy C-means clustering and deep learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, 2024, doi: 10.1109/JS-TARS.2024.3437469.