Integration of U-Net and FastSAM for Accurate Leaf Image Segmentation in Complex Backgrounds

Parichat Sermwuthisarn
Department of Electrical Engineering, Faculty of Engineering and Industrial Technology, Silpakorn University, Thailand
sermwuthisarn_p@su.ac.th

Sopon Phumeechanya
Department of Electrical Engineering, Faculty of Engineering and Industrial Technology, Silpakorn University, Thailand
phumeechanya_s@su.ac.th (corresponding author)

Received: 1 September 2025 | Revised: 6 October 2025 | Accepted: 22 October 2025

 DOI: https://doi.org/10.48084/etasr.14464

ABSTRACT

Leaf segmentation plays a crucial role in plant phenotyping and precision agriculture, enabling the monitoring of growth, disease detection, and informed crop management. However, accurate segmentation in natural environments is challenging due to complex backgrounds, overlapping structures, irregular boundaries, and varying illumination. This paper proposes a hybrid six-stage framework that integrates U-Net with the Fast Segment Anything Model (FastSAM) to achieve accurate and efficient leaf segmentation. The pipeline consists of initial U-Net segmentation, largest component filtering, contour extraction with convex hull transformation, bounding box derivation via distance transform, promptable refinement with FastSAM, and final contour selection. The experiments conducted used 633 images from the Pl@ntLeaves database: 333 images for model development with a train/validation split of 266/67 (20% validation), and a held-out test set of 300 images. On the 300-image test set, the proposed framework achieved superior results (Precision = 0.966, Recall = 0.945, Intersection over Union (IoU) = 0.917, Dice = 0.953, HD95 = 27.859), outperforming DeepLabV3 and CLIPSeg. These findings confirm that combining U-Net's fine-grained feature extraction with FastSAM's efficient prompt-based refinement provides a robust and scalable solution for plant phenotyping and precision agriculture, particularly by enhancing boundary accuracy in complex natural scenes.

Keywords-leaf segmentation; U-Net; FastSAM; precision agriculture; plant phenotyping

I. INTRODUCTION

Leaf segmentation is a fundamental task in plant phenotyping and precision agriculture, as it enables the extraction of morphological traits, non-destructive monitoring of growth, and early detection of diseases [1]. Accurate segmentation facilitates automated classification, quantitative trait analysis, and crop management, thereby contributing to sustainable agriculture and food security [2, 3]. However, segmentation in natural environments is highly challenging due to cluttered and heterogeneous backgrounds, overlapping structures, irregular boundaries, and variations in illumination [4]. These challenges reduce the robustness of conventional and learning-based approaches, highlighting the need for more adaptive and scalable solutions.

Early research relied on traditional image processing methods, such as thresholding, clustering, and active contour models. For instance, the Chan–Vese model combined with Sobel operators was used for overlapping leaves [5], while contour extraction techniques enabled the separation of single and occluded leaves [6]. Factorization-based active contour methods were proposed for cotton leaves [7], and robust shape descriptors were introduced to support plant species identification [8]. Although these approaches achieved reasonable accuracy in controlled conditions, their performance degraded under real-world backgrounds [9].

With the advent of deep learning, Convolutional Neural Networks (CNNs) significantly advanced segmentation. Fully Convolutional Networks (FCN) [10] and U-Net [11] pioneered end-to-end pixel-level prediction, inspiring numerous agricultural applications. Variants have been applied to persimmon leaf diseases [12], cucumber leaves with bilayer networks [13], and pear leaves with Mask R-CNN [14]. Improved U-shaped networks addressed under-segmentation [15], while semantic segmentation pipelines with ResNet

backbones [16], LinkNet architectures [17], and FB-PNet [18] improved efficiency and adaptability. Pyramid CNNs further targeted dense foliage [19], and deep learning pipelines extended to classification tasks under complex backgrounds [20]. Despite these advances, CNN-based methods remain constrained by data dependency and computational cost.

To overcome these limitations, hybrid and refinement strategies were developed. Local refinement mechanisms enhanced segmentation in cluttered environments [21], and hybrid automatic methods combined contour and marker-based strategies for overlapping leaves [22]. Mobile-oriented systems for Tuta absoluta damage segmentation highlighted field applicability [23]. Unsupervised methods, such as maximum mutual information with Tsallis entropy [24] and G-mutual information–based fusion [25], improved robustness when annotations were scarce. Detection-driven frameworks, such as YOLOv5/6 [4] and YOLOv10 [26], further demonstrated potential for disease detection and real-time segmentation.

Foundation models and transformer-based approaches have reshaped segmentation research. DeepLab with atrous convolution [27] improved multi-scale context capture, while promptable and zero-shot models, including Segment Anything (SAM) [28], CLIPSeg [29], and FastSAM [30], enabled flexible segmentation with text, image, or box prompts. These approaches reduce reliance on annotated datasets and offer scalability, though they remain limited in fine boundary precision and occlusion-heavy scenarios [1, 2]. Despite these significant advances across traditional, deep learning, and hybrid paradigms, common bottlenecks remain unresolved. Severe occlusion among overlapping leaves, heterogeneous background clutter, and illumination variation continue to degrade performance across methods. These persistent challenges emphasize the necessity for a more robust and adaptive framework. To illustrate these difficulties, Figure 1 presents representative samples from the Pl@ntLeaves database [31, 32], highlighting cluttered backgrounds, overlapping leaves, and non-uniform illumination.

Motivated by these limitations, the present study proposes a novel six-stage hybrid framework that integrates U-Net for initial localization with FastSAM for prompt-driven refinement. The contributions of the current study are:

- Unified pipeline: Integration of U-Net and FastSAM into a single framework, combining fine-grained feature extraction with prompt-based refinement.
- Geometry-aware prompting: Use of convex hull transformation and distance transform to derive bounding box prompts, thereby stabilizing segmentation and enhancing boundary accuracy.
- Progressive refinement: Stepwise enhancement from coarse localization to precise contour delineation, reducing error propagation and improving robustness in complex backgrounds.

Collectively, these contributions establish a robust and scalable solution for plant phenotyping and precision agriculture, with potential applications in crop monitoring, disease detection, and automated agricultural management.

Fig. 1. Sample images in natural scenes with complex background from the Pl@ntLeaves database [31, 32].

II. PROPOSED FRAMEWORK

The proposed framework comprises six sequential stages, as shown in Figure 2. (1) First, leaf regions were localized using a U-Net model trained to produce a binary segmentation mask of leaf versus background. (2) To focus analysis/Analysis focused on a single salient specimen, with only the largest connected component in the mask retained, suppressing small spurious regions. (3) The contour of this component was then extracted, and its convex hull was computed to regularize boundary irregularities and capture the leaf's global outline. (4) From this outline, an inscribed rectangular box within the contour was derived to serve as a box prompt for the subsequent stage. (5) Promptable segmentation was performed with FastSAM, guided by this box to refine the mask of the dominant leaf. (6) Finally, the largest contour from the FastSAM output was selected as the definitive leaf segmentation produced by the pipeline.

A. Step 1: Initial Segmentation with U-Net

U-Net is a CNN architecture originally proposed for biomedical image segmentation [11]. Its design follows a symmetric encoder–decoder structure, where the encoder progressively downsamples the input image to extract hierarchical features, while the decoder upsamples these features to recover spatial resolution. A key innovation of U-Net is the introduction of skip connections, which directly transfer feature maps from the encoder to the corresponding decoder layers. This mechanism preserves fine-grained spatial information that might otherwise be lost during downsampling, enabling the network to generate precise pixel-level predictions even from relatively small training datasets. Due to its efficiency and accuracy, U-Net and its variants have been widely adopted in agricultural applications, particularly for tasks requiring accurate object delineation in complex natural environments. Building on these strengths, this study employed a U-Net model as the foundation for the initial segmentation stage of the proposed framework.

To generate the initial segmentation masks, a U-Net model was trained using 333 paired images from the leaf database [31, 32], each consisting of an original input sample and its corresponding ground-truth mask. All images were resized to

256 × 256 pixels prior to training to ensure uniform input dimensions. Of these 333 samples, 266 were used for training and 67 for validation, yielding a reproducible twenty-percent validation split determined by a fixed random seed.

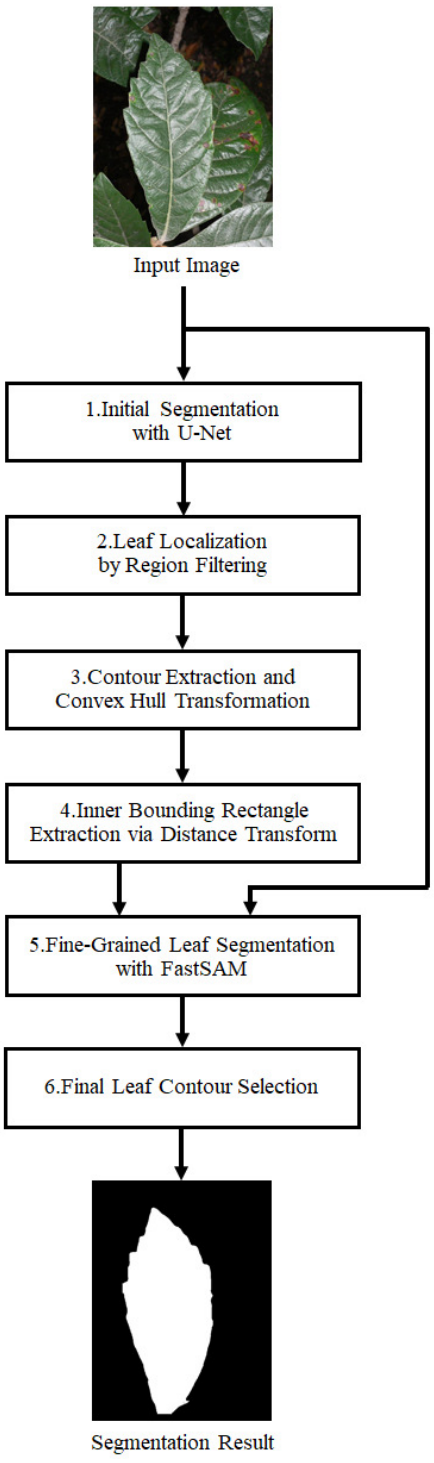

Fig. 2. Proposed framework.

The network employed ResNet-34 as its backbone encoder, providing a balanced depth–efficiency trade-off that enables effective feature extraction from relatively small datasets while limiting overfitting. Leveraging ImageNet-pretrained weights further accelerated convergence and improved generalization to natural leaf images.

For optimization, training was conducted using the Binary Cross-Entropy with Logits Loss (BCEWithLogitsLoss) function, which is widely employed in binary segmentation tasks because it directly compares predicted probabilities against ground-truth pixel labels. By incorporating the sigmoid activation internally. This loss function ensures numerical stability and effectively handles class imbalance between foreground (leaf) and background pixels. Additionally, the Adam optimizer was adopted due to its adaptive learning rate strategy, which combines the advantages of momentum and RMSProp. This allows Adam to maintain stable and efficient convergence, even in scenarios with noisy gradients or heterogeneous data, making it particularly suitable for training deep networks, such as the U-Net, in complex agricultural imaging contexts.

The trained model produced binary segmentation outputs, where leaf regions were represented by white pixels and background areas by black pixels. This stage provided a robust foundation for the subsequent refinement and precise localization of the target leaf.

B. Step 2: Leaf Localization by Region Filtering

The segmentation masks generated by the U-Net frequently contained multiple candidate regions due to the presence of overlapping structures, complex backgrounds, or noise. To isolate the primary structure of interest, only the largest connected region within the binary mask was retained, based on the assumption that the dominant leaf occupies the greatest area in the image. This filtering procedure, referred to as leaf localization, eliminated minor or irrelevant components and ensured that subsequent processing steps focused exclusively on the most salient leaf structure.

C. Step 3 Contour Extraction and Convex Hull Transformation

After obtaining the approximate location of the leaf, the contour of the identified region was extracted and subsequently transformed into its convex hull. This transformation eliminates local irregularities and concavities along the contour while preserving the overall geometry of the leaf. The convex hull can be mathematically defined as the smallest convex polygon enclosing all points of a given set.

Formally, let $S = \{p_1, p_2, \ldots, p_n\} \subset R^2$ represent the set of contour points. The convex hull of S, denoted as $\mathrm{Conv}(S)$, is given by:

$$\mathrm{Conv(S)} = \{\textstyle\sum_{i=1}^{n} \lambda_i p_i \,|\lambda_i \geq 0, \sum_{i=1}^{n} \lambda_i = 1\} \qquad (1)$$

where λ_i are non-negative weights representing a convex combination of the points. This formulation ensures that every point in $\mathrm{Conv}(S)$ lies within the convex boundary formed by the original contour points. By applying the convex hull, the resulting boundary becomes a robust and smoothed

representation of the leaf region, serving as a reliable basis for subsequent bounding-box extraction in Step 4.

The use of the convex hull is particularly important in this task because natural leaf contours often exhibit irregularities, noise, or concavities due to overlapping foliage, lighting variation, or complex backgrounds. These local distortions can lead to unstable or fragmented bounding regions if processed directly. The convex hull mitigates this issue by eliminating unnecessary concavities while still preserving the global shape of the leaf, thereby ensuring geometric consistency. As a result, the subsequent bounding-box extraction becomes more reliable, facilitating accurate prompt generation for FastSAM and ultimately improving the precision of the final segmentation.

D. Step 4: Inner Bounding Rectangle Extraction via Distance Transform

After obtaining the convex hull of the leaf contour, an inner bounding rectangle was extracted to serve as a reliable box prompt for the subsequent segmentation stage. This was achieved by applying a distance transform to the binary mask derived from the convex hull.

The distance transform assigns to each foreground pixel a value equal to the Euclidean distance to the nearest background pixel. Formally, for a binary mask $B(x, y)$, where foreground pixels are defined as $B(x, y) = 1$ and background pixels as $B(x, y) = 0$, the distance transform $D(x, y)$ at a pixel (x, y) is given by:

$$D(x, y) = \min_{(u,v)\in\Omega_0} \sqrt{(x-u)^2 + (y-v)^2} \qquad (2)$$

where Ω_0 is the set of all background pixels (u, v) such that $B(x, y) = 0$. This transformation produces a map in which pixels at the center of the leaf region have higher values, while those near the boundary have lower values. By identifying the location of the maximum distance value, the most deeply embedded pixel within the leaf region was determined. Around this point, a rectangular window was iteratively expanded until its boundary reached the convex hull, thereby forming the largest inner rectangle entirely contained within the leaf region. This inner bounding rectangle provides a robust and geometry-aware box prompt, ensuring that the subsequent segmentation step (using FastSAM) focuses on the correct leaf region while avoiding background interference.

E. Step 5: Fine-Grained Leaf Segmentation with FastSAM

The FastSAM [30] is a lightweight, CNN-based alternative to the Transformer-driven SAM [28]. Although SAM introduced promptable segmentation with strong generalization ability, its Transformer backbone requires intensive computation, limiting practical deployment in resource-constrained environments such as agriculture. FastSAM overcomes this challenge by reformulating the task as instance segmentation, delivering up to 50 times faster inference while maintaining competitive accuracy. Built upon YOLOv8-seg with an instance segmentation branch adapted from YOLACT [33], FastSAM operates through two stages: all-instance segmentation, where category-agnostic masks for all objects are generated, and prompt-guided selection, where point, bounding box, or text prompts (via CLIP embeddings) are applied to refine and isolate the target object. This architecture ensures seamless integration with bounding box prompts from earlier stages of the proposed framework, enhancing both efficiency and robustness. With these advantages, FastSAM serves as an effective refinement module that balances accuracy, scalability, and real-world applicability in plant phenotyping and precision agriculture, particularly in scenarios requiring rapid and accurate segmentation of individual leaves.

Leveraging these strengths, FastSAM was integrated into Step 5 of the proposed framework. With the inner bounding rectangle obtained in Step 4, the segmentation process was refined using FastSAM, which is a promptable segmentation framework that can incorporate external guidance, such as bounding boxes, points, or text prompts, to localize objects more effectively. In this study, the bounding rectangle derived from the convex hull served as the box prompt, ensuring that the segmentation was constrained to the localized leaf region.

Formally, given an input image I and a bounding box prompt R, the FastSAM predicts a segmentation mask $\widehat{M}$ as:

$$\widehat{M} = f(I, R) \qquad (3)$$

where f represents the FastSAM model. By utilizing the bounding box as a prior, FastSAM effectively separated the target leaf from surrounding elements, such as other leaves, stems, or soil. Compared to unconstrained segmentation, this box-guided strategy improved both accuracy and robustness, yielding fine-grained masks that preserved the natural boundaries of the leaf while reducing noise. This step provided a precise segmentation of the leaf, enabling reliable downstream phenotyping and analysis.

F. Step 6: Final Leaf Contour Selection

In the final stage, the segmentation results produced by FastSAM were consolidated to obtain the definitive leaf mask. Among the candidate regions detected, the largest contour was selected as the final representation of the dominant leaf, following the assumption that the target specimen occupies the largest visible area within the image. This straightforward selection criterion effectively excluded minor or fragmented regions that may arise from background structures or noise. The resulting binary mask provided a reliable and accurate delineation of the leaf, preserving its overall geometry and serving as the final output of the proposed framework.

III. EXPERIMENTS AND RESULTS

A. Dataset

The dataset used in this study was derived from the Pl@ntLeaves database [31, 32], which contains plant leaf images captured in natural environments. All available images from this database, each paired with its corresponding ground-truth mask, were employed in the experiments conducted. The dataset was randomly divided into the training, validation, and testing subsets using a fixed seed to ensure reproducibility and include a wide variety of leaf shapes and background conditions in each subset.

From the total of 633 images and their corresponding ground-truth masks, 333 paired samples were randomly selected for Step 1: Initial Segmentation with U-Net, of which 266 were used for training and 67 were reserved for validation. Each pair consisted of an original RGB image and its manually annotated ground-truth mask, which precisely delineated the leaf regions from the background. To ensure consistency and computational efficiency, all images were resized to 256 × 256 pixels prior to model training.

The remaining 300 images and their corresponding ground-truth masks were reserved as a test set for independent evaluation. These test samples, which were not included in the training or validation process, were used exclusively to assess segmentation performance. This separation between training and testing ensured a fair and unbiased evaluation of the proposed framework, allowing the reported metrics to accurately reflect the model's ability to generalize to unseen data.

The dataset is particularly challenging because the leaf images were captured against complex natural backgrounds. These include other overlapping leaves, branches, stems, fruits, soil, and mulch, as well as variations in lighting and shadows that obscure clear boundaries. Such complexity makes segmentation difficult for conventional methods and provides a rigorous benchmark for evaluating the robustness of the proposed framework. Figure 3 displays representative examples of original images and their corresponding ground-truth masks.

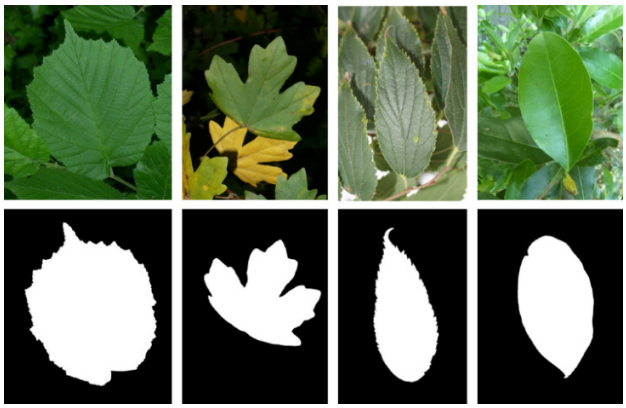

Fig. 3. Examples of leaf images used for testing and their corresponding ground truth masks from the Pl@ntLeaves database [31, 32].

B. U-Net Training

A U-Net with a ResNet-34 encoder, pre-trained on ImageNet, was trained to produce a single-channel segmentation mask. The dataset comprised 333 pairs of RGB images and binary masks. All images were resized to 256 × 256 pixels and normalized to ImageNet statistics, while the masks were resized with nearest-neighbor interpolation and binarized at a threshold of 0.5. The data were randomly split into 266 training samples and 67 validation samples, corresponding to an eighty–twenty split.

The network was optimized with the Adam optimizer. The learning rate was 1×10^{-4}, and the batch size was 8. Training ran for 25 epochs under the BCEWithLogitsLoss objective. During inference, a sigmoid was applied to the logits and a threshold of 0.5 was used to obtain binary masks. The performance was monitored after each epoch, and the checkpoint that achieved the lowest validation loss was retained.

Figure 4 shows the training and validation losses across the 25 epochs. Both curves fall steeply during the first few epochs and then decline more gradually. From around epoch 10 and onward, the validation curve remains slightly above the training curve, with a small and stable generalization gap. Minor ripples in the middle of training reflect variation across mini-batches rather than true degradation. By the final epoch, the training loss was about 0.04, and the validation loss was about 0.08, which indicates good convergence without overfitting.

This U-Net model serves as Step 1 of the proposed six-stage framework, namely the initial segmentation stage, and its masks provide the starting point for the downstream refinement steps.

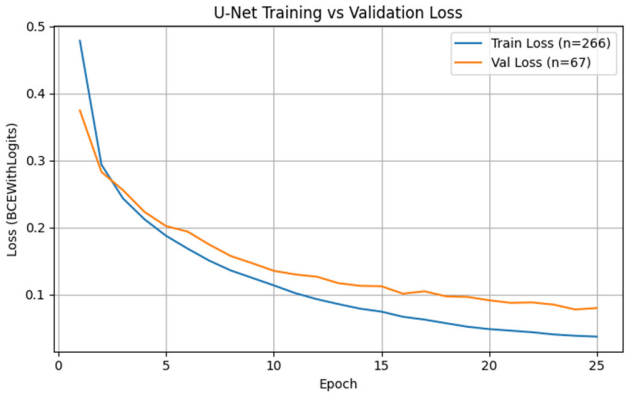

Fig. 4. Training and validation loss of the U-Net model.

C. Step-by-Step Results of the Proposed Framework

To demonstrate the progressive refinement achieved by the proposed six-stage framework, Figure 5 illustrates the intermediate outputs of a representative test image at each step. In the first row, the original input RGB image is shown as the starting point for the segmentation process. The second row presents the initial segmentation generated by the U-Net model (Step 1). While the leaf regions are successfully distinguished from the background, multiple components and noise are sometimes present due to overlapping leaves and complex surroundings.

The third row depicts Steps 2 and 3. In Step 2, region filtering is applied to retain only the largest connected component, thereby isolating the dominant leaf and discarding small or irrelevant regions. Step 3 then extracts the contour of the isolated leaf and applies a convex hull transformation to regularize the shape and eliminate local irregularities along the boundary. The fourth row corresponds to Step 4, where the distance transform is applied to the convex hull mask. From

this, the largest inner bounding rectangle is derived to serve as a reliable box prompt for the subsequent stage.

The fifth row demonstrates Step 5, where the FastSAM performs prompt-guided segmentation using the bounding box. This produces refined leaf masks that capture accurate boundaries while suppressing interference from the background. Finally, the sixth row presents Step 6, where the largest contour is selected from the FastSAM output to generate the definitive binary mask of the leaf. This ensures a clean and geometrically consistent segmentation result. The framework incrementally improves the segmentation quality step by step, leading to accurate isolation of the dominant leaf from complex natural backgrounds.

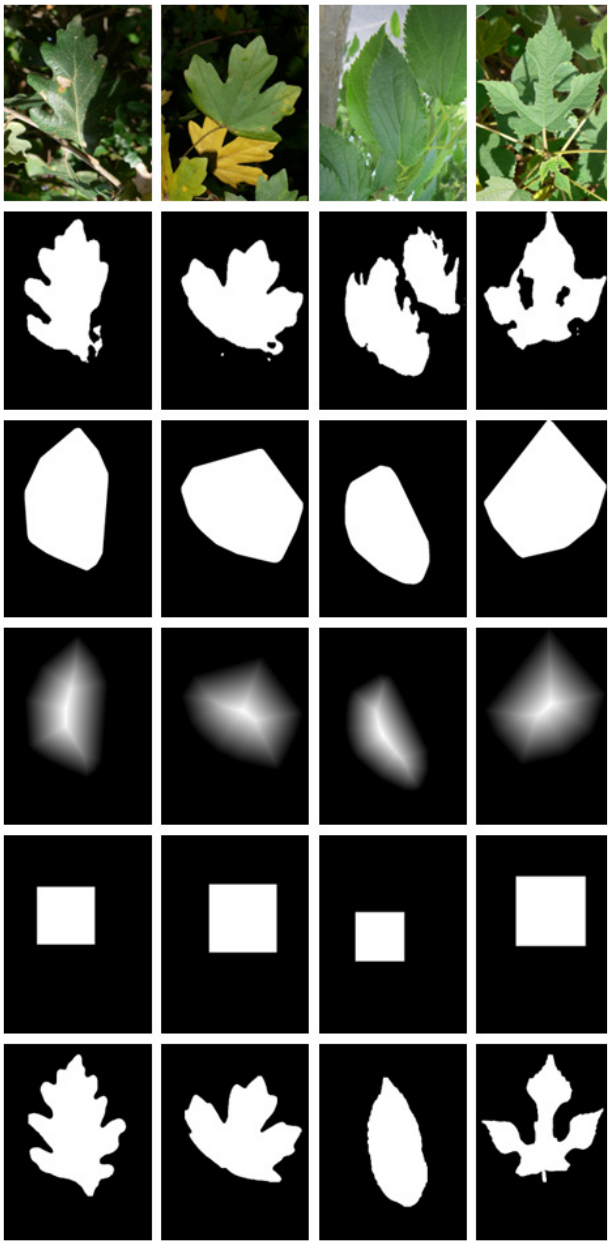

Fig. 5. Step-by-step results of the proposed leaf segmentation framework.

D. Evaluation Metrics

To evaluate the performance of the proposed segmentation framework, four standard quantitative metrics were employed: Precision, Recall, IoU, and the Dice Coefficient. These measures capture complementary aspects of segmentation quality and together provide a comprehensive assessment. In addition to these pixel-level metrics, boundary accuracy was further assessed using the 95th Percentile Hausdorff Distance (HD95) [34], offering a more robust evaluation of contour alignment. In (4)-(7), True Positives (TP), represent the leaf pixels correctly identified by the model, False Positives (FP) refer to background pixels misclassified as leaf pixels, and False Negatives (FN) denote actual leaf pixels that were missed in the prediction.

Precision measures the fraction of the predicted leaf pixels that are truly correct. A higher precision value indicates fewer FP, meaning that the model rarely misclassifies background regions as leaves:

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

Recall quantifies the proportion of actual leaf pixels that were successfully detected. A higher recall means fewer FN, indicating that most of the true leaf regions were captured:

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

IoU evaluates the overlap between the predicted segmentation and the ground truth mask, normalized by their union. It provides an interpretable measure of spatial consistency and is widely used as a benchmark in segmentation research:

$$IoU = \frac{TP}{TP+FP+FN} \quad (6)$$

The Dice coefficient, equivalent to the F1-score in binary segmentation, computes the harmonic mean of precision and recall. It provides a balanced measure that is especially useful when dealing with class imbalance or small object regions:

$$Dice = \frac{2TP}{2TP+FP+FN} \quad (7)$$

In addition to the conventional pixel-level metrics (Precision, Recall, IoU, and Dice), the HD95 was further employed [34] to assess boundary accuracy, which is particularly important in leaf segmentation tasks. The conventional Hausdorff Distance (HD) measures the maximum deviation between the predicted contour and the ground truth contour; however, it is highly sensitive to outliers, as even a single misclassified boundary pixel can disproportionately inflate the metric. To mitigate this issue, HD95 replaces the maximum operator with the 95th percentile, capturing the typical worst-case error while discarding extreme deviations. This provides a more robust and stable estimate of boundary alignment, which is especially relevant for natural leaf images that often contain irregular shapes, occlusion, and noisy background structures.

Formally, given the predicted contour A and the ground truth contour B, HD_{95} is defined as:

$$HD_{95}(A,B) = max\begin{pmatrix} P_{95}\left\{\underset{b\in B}{min}\|a-b\| \,\middle|\, a\in A\right\}, \\ P_{95}\left\{\underset{a\in A}{min}\|b-a\| \,\middle|\, a\in B\right\} \end{pmatrix} \quad (8)$$

where $\|a-b\|$ denotes the Euclidean distance between boundary points, and $P_{95}\{.\}$ is the 95th percentile operator. The metric yields values in the range $(0,\infty)$, with lower values indicating closer alignment of predicted and ground truth boundaries. To ensure consistency with the proposed framework, in cases where multiple contours appeared in either the prediction or the ground truth mask, only the largest connected component was retained for HD95 computation, reflecting the evaluation of the dominant leaf.

Precision ensures that predictions are not contaminated with background noise, while Recall guarantees that most of the leaf region is successfully detected. IoU provides a direct measure of the overall spatial overlap between the predicted mask and the ground truth, while the Dice Coefficient balances the trade-off between Precision and Recall by computing their harmonic mean. All four metrics take values within the range of (0, 1), where values closer to 1 indicate superior segmentation performance. In addition to these pixel-level measures, boundary accuracy was further assessed using the HD95, which captures the typical worst-case boundary error while mitigating the influence of outliers. Lower HD95 values indicate that the predicted contours more closely align with the ground truth, providing a robust assessment of geometric fidelity. By employing this combination of metrics, the evaluation captures both pixel-level accuracy and boundary-level precision, which are crucial for reliable segmentation in complex natural scenes.

E. Baseline Comparison

To validate the effectiveness of the proposed framework, its performance was compared against two representative and widely recognized segmentation models: DeepLabV3 [27] and CLIPSeg [29]. These baselines were chosen because they reflect two distinct paradigms in modern segmentation research: convolution-based multi-scale feature extraction and prompt-based segmentation leveraging vision-language models.

DeepLabV3 is a CNN-based segmentation model that employs Atrous (dilated) convolution to enlarge the receptive field without reducing spatial resolution. Its core component, the Atrous Spatial Pyramid Pooling (ASPP) module, applies multiple parallel atrous convolutions with different dilation rates to effectively capture multi-scale contextual information. DeepLabV3 has achieved strong performance on benchmark datasets such as PASCAL VOC and Cityscapes. In this study, for a fair comparison, DeepLabV3-ResNet50 was implemented and trained using PyTorch. The pretrained model was adopted, and the final classifier layer was modified from a 21-class output to a single-channel binary mask (foreground: leaf/background). Training and evaluation were performed on the same dataset used in Step 1 (U-Net). The training phase employed the Adam optimizer with a learning rate of 1×10^{-3}, batch size = 2, and 10 epochs. The BCEWithLogitsLoss function was used as the objective, and all Batch Normalization layers were fixed in evaluation mode during optimization to stabilize training. The model was trained end-to-end using mixed-precision computation on the GPU. This configuration ensured that DeepLabV3 was trained and tested under the same experimental settings as the proposed framework, thereby allowing a fair and reproducible comparison of segmentation performance.

CLIPSeg extends the multimodal CLIP model by incorporating a lightweight transformer-based decoder for dense pixel-wise prediction, enabling segmentation guided by either text or image prompts. This architecture leverages the pretrained vision–language alignment capability of CLIP, originally trained on hundreds of millions of image–text pairs collected from the internet, to interpret semantic cues directly from natural language without additional fine-tuning. The CLIPSeg decoder was trained in [29] on the extended PhraseCut+ dataset, which includes over 340,000 phrase–image–mask pairs, augmented with visual prompts, negative samples, and text–image interpolation to enhance zero-shot and one-shot generalization. This study did not retrain or modify the model; instead, it utilized the pretrained CLIPSeg weights released by the authors to fully exploit its zero-shot segmentation capability. During inference, the model was applied to current test images using the text prompt "a salient leaf", allowing it to automatically focus on the dominant leaf region in each image. This configuration demonstrates how CLIPSeg performs segmentation through prompt-based reasoning rather than supervised retraining. CLIPSeg was selected as a baseline because it represents the prompt-driven segmentation paradigm, providing a meaningful comparison to both convolutional approaches and the proposed box-prompt-guided framework.

F. Results

The results, as shown in Table I, demonstrate that the proposed framework consistently outperformed all baseline models across all evaluation metrics. In terms of Precision, the proposed method achieved 0.966, which is higher than U-Net (0.899), DeepLabV3 (0.907), and CLIPSeg (0.832), indicating fewer FP and more accurate isolation of the leaf regions. Regarding Recall, CLIPSeg (0.942) showed relatively strong sensitivity in detecting leaf pixels, but the proposed framework still achieved the highest value (0.945), ensuring both sensitivity and completeness compared with the other models.

TABLE I. PERFORMANCE EVALUATION OF THE PROPOSED FRAMEWORK, U-NET ONLY (AT STEP 1), DEEPLABV3, AND CLIPSEG IN TERMS OF PRECISION, RECALL, IoU, DICE, AND HD95

Method	Precision	Recall	IoU	Dice	HD95
DeeplabV3 [27]	0.907	0.887	0.808	0.887	58.686
CLIPSeg [29]	0.832	0.942	0.791	0.871	74.617
U-Net only (at Step 1)	0.899	0.923	0.885	0.936	33.615
Proposed	0.966	0.945	0.917	0.953	27.859

For IoU and Dice, which are holistic measures of the segmentation quality, the proposed framework achieved 0.917 (IoU) and 0.953 (Dice), outperforming U-Net only (0.885, 0.936), DeepLabV3 (0.808, 0.887), and CLIPSeg (0.791, 0.871). These improvements highlight the effectiveness of combining U-Net–based initial segmentation with convex hull refinement and FastSAM-guided box prompting.

In addition, boundary accuracy evaluated by HD95 further confirmed the superiority of the proposed framework. The proposed method achieved the lowest HD95 value (27.859), indicating more precise boundary alignment compared with U-Net only (33.615), DeepLabV3 (58.686), and CLIPSeg (74.617), all of which exhibited larger deviations along leaf edges. This substantial reduction in boundary error underscores the robustness of the proposed multi-stage refinement strategy, particularly under challenging conditions involving irregular leaf shapes, occlusion, and complex background clutter.

Overall, the proposed framework demonstrated superior segmentation accuracy, achieving high pixel-level correctness (Precision/Recall), robust geometric consistency (IoU/Dice), and reliable boundary preservation (HD95). The integration of the convex hull transformation further enhanced geometric stability by regularizing irregular contours before prompt-based refinement. These findings confirm that combining conventional modeling, convex hull–based shape regularization, and prompt-guided refinement provides significant advantages over relying solely on CNN-based or prompt-only segmentation approaches.

Figure 6 illustrates a qualitative comparison of the leaf segmentation results obtained from two representative baseline methods, DeepLabV3 and CLIPSeg, and the proposed multi-stage framework evaluated at Step 1, Step 3, and Step 6 (final output), with ground truth masks as reference. Consistent with the quantitative metrics reported earlier, the superiority of the proposed approach is evident across all test cases. DeepLabV3 tends to under-segment the leaf area, leading to incomplete masks and background leakage, whereas CLIPSeg often produces fragmented results with spurious regions, particularly under cluttered scenes or uneven illumination. In contrast, the proposed framework shows progressive refinement: Step 1 (U-Net) achieves high recall but with rough or eroded boundaries; Step 3 (convex hull and contour consolidation) suppresses noise and stabilizes the outline, albeit sometimes over-smoothing highly non-convex structures; and Step 6 (FastSAM refinement) restores fine geometry and produces masks that closely approximate the ground truth in both completeness and boundary fidelity.

A row-by-row examination highlights these improvements. In the first row, showing a simple oval leaf, DeepLabV3 captures only the left half of the leaf, while CLIPSeg covers most of the structure but misses a small portion near the lower edge. In contrast, the proposed framework refines step by step from an initial coarse mask to a clean and nearly perfect segmentation at Step 6, closely matching the ground truth. In the second row, which presents a partially occluded leaf against a textured background, both baselines misclassify surrounding elements, but the proposed framework incrementally corrects these errors until Step 6 provides a precise outline without leakage. The third row, with a multicolored leaf in a heterogeneous scene, exposes baseline weaknesses. DeepLabV3 misses interior regions, and CLIPSeg produces fragmented masks, whereas the proposed framework gradually eliminates noise, achieving a smooth and accurate final mask. In the fourth row, featuring an elongated leaf with highlights and shadows, baseline methods yield jagged or spurious results, while the proposed framework successfully delivers a slender and geometrically faithful segmentation at Step 6. In the fifth row, depicting a circular leaf under uneven illumination, DeepLabV3 under-segments the structure, while CLIPSeg produces a result closer to the ground truth, performing comparably well to the proposed framework. The sixth row, containing a deeply lobed leaf, is among the most challenging: baselines generate incomplete or broken lobes, and even Step 3 temporarily oversimplifies the contour; nevertheless, Step 6 successfully recovers the lobed morphology with high fidelity, underscoring the advantage of the refinement stage. Finally, in the seventh row, with a serrated leaf under shadow, both baselines suffer from background leakage and loss of serration details, whereas the proposed method gradually improves until Step 6 achieves the best trade-off, preserving serration patterns while maintaining a clean silhouette.

Taken together, these qualitative results complement the quantitative findings by confirming that the proposed multi-stage framework consistently outperforms established baselines. Its design, starting from coarse but reliable localization, followed by progressive background suppression and prompt-guided refinement, enables robust segmentation across a variety of challenging scenarios, including cluttered environments, occlusion, shadow variation, and diverse leaf morphologies. This validates the effectiveness and generalizability of the approach in practical agricultural settings.

Nevertheless, the framework still inherits an important limitation. Its performance strongly depends on the quality of the initial segmentation produced by the U-Net. If the U-Net fails to generate a sufficiently accurate mask, such as when the dominant leaf is heavily occluded, blurred, or poorly distinguished from the background, the subsequent region filtering may be unable to correctly localize the salient leaf. In such cases, the bounding box derived in later stages may not represent the true leaf position, which in turn constrains FastSAM to an inaccurate prompt and degrades the quality of the refined segmentation. This dependency highlights the critical role of reliable initial localization and suggests that enhancing the robustness of the first stage is significant for future improvement.

In addition, the reliance on a single dataset for training may restrict the generalization capability of the framework, particularly when applied to diverse plant species or field conditions not represented in the training images. Furthermore, this study relied solely on the Pl@ntLeaves database, which, while challenging, provides limited coverage of natural variability across crops, seasons, and imaging modalities. To address these constraints, future work should incorporate larger and more diverse datasets, ideally spanning multiple species and environmental conditions, to improve robustness and ensure broader applicability. Evaluating the framework across multiple independent databases would also strengthen its generalizability and provide deeper insights into its performance under heterogeneous agricultural scenarios.

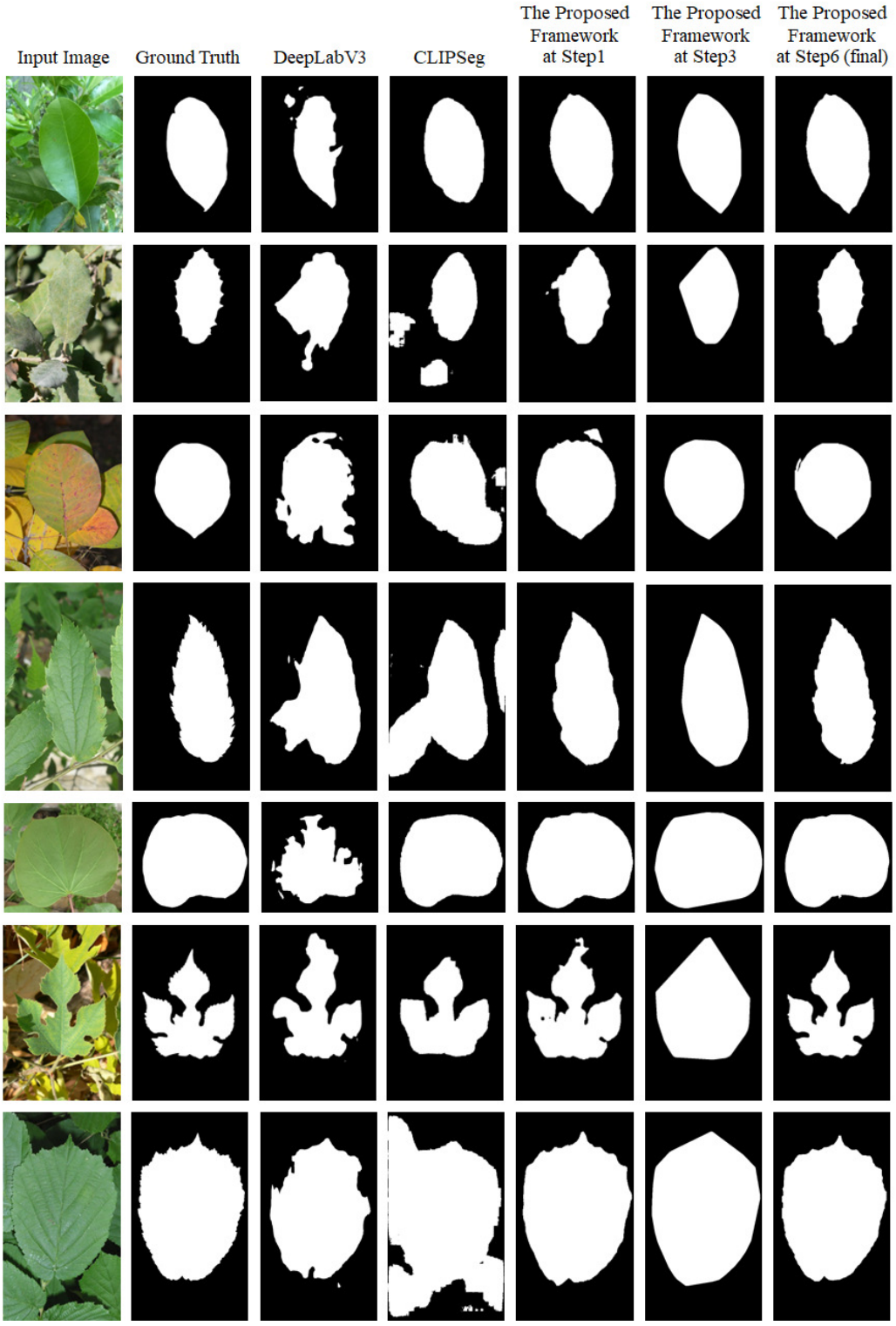

Fig. 6. Performance comparison of the proposed framework against DeepLabV3 and CLIPSeg.

IV. CONCLUSION

This study introduced a hybrid six-stage framework that integrates U-Net and the Fast Segment Anything Model (FastSAM) to address the challenges of leaf segmentation in complex natural environments. The framework combines U-Net's fine-grained feature extraction with FastSAM's prompt-driven refinement, progressively enhancing segmentation from

initial localization to precise boundary delineation. Convex hull and distance transform–based bounding box generation further stabilized the process, improving robustness against clutter, occlusion, and irregular leaf geometries.

Experiments on the Pl@ntLeaves dataset showed that the proposed method consistently outperformed DeepLabV3 and CLIPSeg across all metrics, achieving higher Precision, Recall, Intersection over Union (IoU), and Dice, with lower 95th Percentile Hausdorff Distance (HD95). The qualitative analysis confirmed that stepwise refinement enables reliable segmentation under challenging conditions, such as uneven lighting, overlapping leaves, and heterogeneous scenes.

Overall, the framework offers a robust, accurate, and scalable solution for plant phenotyping and precision agriculture, with applications in crop monitoring, disease detection, and automated management. Future work may extend to multi-leaf segmentation, real-time field deployment, and integration with broader phenotyping pipelines.

ACKNOWLEDGMENT

This research was financially supported by the Silpakorn University Research, Innovation, and Creative Fund.

DATA AVAILABILITY STATEMENT

The dataset used in this study, the Pl@ntLeaves database, is publicly available at:

https://liris.univ-lyon2.fr/reves/content/en/databases.php.

REFERENCES

[1] J. W. Abe, J. Ilao, and G. Foliente, "Promptable Leaf Segmentation in Plant Phenotyping: Research Perspectives and Challenges," in *2024 30th International Conference on Mechatronics and Machine Vision in Practice*, Leeds, UK, Oct. 2024, pp. 1–6, https://doi.org/10.1109/M2VIP62491.2024.10745998.

[2] A. Lyasmine, F. Idir, and B. Samia, "Plant Leaf Image Segmentation in Natural Scenes: A Multi-layer Graph Queries Propagation Approach," *Pattern Analysis and Applications*, vol. 28, no. 1, Mar. 2025, Art. no. 1, https://doi.org/10.1007/s10044-024-01380-y.

[3] L. Gao and X. Lin, "Fully Automatic Segmentation Method for Medicinal Plant Leaf Images in Complex Background," *Computers and Electronics in Agriculture*, vol. 164, Sep. 2019, Art. no. 104924, https://doi.org/10.1016/j.compag.2019.104924.

[4] E. Iren, "Comparison of YOLOv5 and YOLOv6 Models for Plant Leaf Disease Detection," *Engineering, Technology & Applied Science Research*, vol. 14, no. 2, pp. 13714–13719, Apr. 2024, https://doi.org/10.48084/etasr.7033.

[5] Z. Wang, K. Wang, F. Yang, S. Pan, and Y. Han, "Image Segmentation of Overlapping Leaves Based on Chan–Vese Model and Sobel Operator," *Information Processing in Agriculture*, vol. 5, no. 1, pp. 1–10, Mar. 2018, https://doi.org/10.1016/j.inpa.2017.09.005.

[6] R. Khan and R. Debnath, "Segmentation of Single and Overlapping Leaves by Extracting Appropriate Contours," in *6th International Conference on Computer Science, Engineering and Information Technology*, Chennai, India, Nov. 2019, pp. 287–300, https://doi.org/10.5121/csit.2019.91323.

[7] B. M. Patil and B. Amarapur, "Cotton Leaf Image Segmentation using Modified Factorization-Based Active Contour Method," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 9, pp. 516–521, 2020, https://doi.org/10.14569/IJACSA.2020.0110962.

[8] C. Yang, L. Fang, Q. Yu, and H. Wei, "A Learning Robust and Discriminative Shape Descriptor for Plant Species Identification," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 20, no. 1, pp. 39–51, Jan. 2023, https://doi.org/10.1109/TCBB.2022.3148463.

[9] L. Gao and X. Lin, "A Method for Accurately Segmenting Images of Medicinal Plant Leaves With Complex Backgrounds," *Computers and Electronics in Agriculture*, vol. 155, pp. 426–445, Dec. 2018, https://doi.org/10.1016/j.compag.2018.10.020.

[10] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, Jun. 2015, pp. 3431–3440, https://doi.org/10.1109/CVPR.2015.7298965.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, vol. 9351, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham, Switzekland: Springer International Publishing, 2015, pp. 234–241, https://doi.org/10.1007/978-3-319-24574-4_28.

[12] Z. Jia, A. Shi, G. Xie, and S. Mu, "Image segmentation of persimmon leaf diseases based on UNet," in *2022 7th International Conference on Intelligent Computing and Signal Processing*, Xi'an, China, Apr. 2022, pp. 2036–2039, https://doi.org/10.1109/ICSP54964.2022.9778390.

[13] T. Qian *et al.*, "Cucumber Leaf Segmentation Based on Bilayer Convolutional Network," *Agronomy*, vol. 14, no. 11, Nov. 2024, Art. no. 2664, https://doi.org/10.3390/agronomy14112664.

[14] W. Mu, Z. Jia, Y. Liu, W. Xu, and Y. Liu, "Image Segmentation Model of Pear Leaf Diseases Based on Mask R-CNN," in *2022 International Conference on Image Processing and Media Computing*, Xi'an, China, May 2022, pp. 41–45, https://doi.org/10.1109/ICIPMC55686.2022.00016.

[15] J. Kan, Z. Gu, C. Ma, and Q. Wang, "Leaf Segmentation Algorithm Based on Improved U-shaped Network under Complex Background," in *2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference*, Chongqing, China, Jun. 2021, pp. 87–92, https://doi.org/10.1109/IMCEC51613.2021.9482382.

[16] Q. Wang, W. Du, C. Ma, and Z. Gu, "Leaf Image Semantic Segmentation Based on Deep Learning," in *2021 IEEE 2nd International Conference on Information Technology, Big Data and Artificial Intelligence*, Chongqing, China, Dec. 2021, pp. 1147–1151, https://doi.org/10.1109/ICIBA52610.2021.9688307.

[17] L. Zhang and X. Liang, "Image Segmentation of Plant Leaves in Natural Environments Based on LinkNet," *Journal of Computing and Electronic Information Management*, vol. 11, no. 3, pp. 67–72, Dec. 2023, https://doi.org/10.54097/jceim.v11i3.15.

[18] P. Dinesh and R. Lakshmanan, "FB-PNet: A Semantic Segmentation Model for Automated Plant Leaf and Disease Annotation," *International Journal of Advanced Computer Science and Applications*, vol. 16, no. 5, 2025, https://doi.org/10.14569/IJACSA.2025.0160549.

[19] D. Morris, "A Pyramid CNN for Dense-Leaves Segmentation," in *2018 15th Conference on Computer and Robot Vision*, Toronto, ON, Canada, May 2018, pp. 238–245, https://doi.org/10.1109/CRV.2018.00041.

[20] K. Yang, W. Zhong, and F. Li, "Leaf Segmentation and Classification with a Complicated Background Using Deep Learning," *Agronomy*, vol. 10, no. 11, Nov. 2020, Art. no. 1721, https://doi.org/10.3390/agronomy10111721.

[21] R. Ma *et al.*, "Local Refinement Mechanism for Improved Plant Leaf Segmentation in Cluttered Backgrounds," *Frontiers in Plant Science*, vol. 14, Aug. 2023, Art. no. 1211075, https://doi.org/10.3389/fpls.2023.1211075.

[22] J. Bala, H. B. Salau, I. J. Umoh, A. J. Onumanyi, S. A. Tijani, and B. Yahaya, "Development of Hybrid Automatic Segmentation Technique of a Single Leaf from Overlapping Leaves Image," *Journal of ICT Research and Applications*, vol. 14, no. 3, Mar. 2021, https://doi.org/10.5614/itbj.ict.res.appl.2021.14.3.4.

[23] L. Loyani and D. Machuve, "A Deep Learning-based Mobile Application for Segmenting Tuta Absoluta's Damage on Tomato Plants," *Engineering, Technology & Applied Science Research*, vol. 11, no. 5, pp. 7730–7737, Oct. 2021, https://doi.org/10.48084/etasr.4355.

[24] N. Nikbakhsh, Y. Baleghi, and H. Agahi, "Maximum Mutual Information and Tsallis Entropy for Unsupervised Segmentation of Tree

Leaves in Natural Scenes," *Computers and Electronics in Agriculture*, vol. 162, pp. 440–449, Jul. 2019, https://doi.org/10.1016/j.compag.2019.04.038.

[25] N. Nikbakhsh, Y. Baleghi, and H. Agahi, "A Novel Approach for Unsupervised Image Segmentation Fusion of Plant Leaves Based on G-mutual Information," *Machine Vision and Applications*, vol. 32, no. 1, Jan. 2021, Art. no. 5, https://doi.org/10.1007/s00138-020-01130-0.

[26] R. Alanazi, "A YOLOv10-based Approach for Banana Leaf Disease Detection," *Engineering, Technology & Applied Science Research*, vol. 15, no. 3, pp. 23522–23526, Jun. 2025, https://doi.org/10.48084/etasr.11138.

[27] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation." arXiv, 2017, https://doi.org/10.48550/ARXIV.1706.05587.

[28] A. Kirillov *et al.*, "Segment Anything," in *2023 IEEE/CVF International Conference on Computer Vision*, Paris, France, Oct. 2023, pp. 3992–4003, https://doi.org/10.1109/ICCV51070.2023.00371.

[29] T. Luddecke and A. Ecker, "Image Segmentation Using Text and Image Prompts," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, Jun. 2022, pp. 7076–7086, https://doi.org/10.1109/CVPR52688.2022.00695.

[30] X. Zhao *et al.*, "Fast Segment Anything." arXiv, 2023, https://doi.org/10.48550/ARXIV.2306.12156.

[31] H. Goeau *et al.*, "The ImageCLEF 2011 plant images classification task." 2012, [Online]. Available: https://liris.univ-lyon2.fr/reves/content/en/databases.php.

[32] H. Goeau *et al.*, "The ImageCLEF 2012 Plant Identication Task." 2012, [Online]. Available: https://liris.univ-lyon2.fr/reves/content/en/databases.php.

[33] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT: Real-Time Instance Segmentation," in *2019 IEEE/CVF International Conference on Computer Vision*, Seoul, South Korea, Oct. 2019, pp. 9156–9165, https://doi.org/10.1109/ICCV.2019.00925.

[34] A. A. Taha and A. Hanbury, "Metrics for Evaluating 3D Medical Image Segmentation: Analysis, Selection, and Tool," *BMC Medical Imaging*, vol. 15, no. 1, Dec. 2015, Art. no. 29, https://doi.org/10.1186/s12880-015-0068-x.