IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 3 June 2026, accepted 3 July 2026, date of publication 10 July 2026, date of current version 16 July 2026.

*Digital Object Identifier 10.1109/ACCESS.2026.3712179*

RESEARCH ARTICLE

# SilvaFormer: Transformer-Based Segmentation for Forest Health Monitoring in UAV Imagery

**SHAFIULLAH SOOMRO[1,2], RYTIS MASKELIŪNAS[3], ROBERTAS DAMAŠEVIČIUS[4], AND ARIANIT KURTI[1], (Member, IEEE)**

[1]Department of Computer Science and Media Technology, Linnaeus University, 351 95 Växjö, Sweden
[2]Faculty of Computer Science and Informatics, Berlin School of Business and Innovation, 12043 Berlin, Germany
[3]Center of Real Time Computer Systems, Kaunas University of Technology, 44249 Kaunas, Lithuania
[4]CoE Forest 4.0, Vytautas Magnus University, 44248 Akademija, Lithuania

Corresponding author: Shafiullah Soomro (shafiullah.soomro@lnu.se)

This work was supported by the Horizon Europe Framework Programme (HORIZON), Call Teaming for Excellence (HORIZON-WIDERA-2022-ACCESS-01-two-stage)—Creation of the Center of Excellence in Smart Forestry "Forest 4.0" under Grant 101059985.

**ABSTRACT** Forest health monitoring from unmanned aerial vehicle (UAV) imagery is essential for the timely detection of pine wilt disease (PWD), yet accurate pixel-wise delineation remains challenging because diseased crown regions are often small, fragmented, and visually confounded by surrounding canopy texture, shadows, and background clutter. This paper presents **SilvaFormer**, a transformer-based semantic segmentation framework for diseased-crown mapping in high-resolution RGB aerial imagery. The proposed model combines a hierarchical MiT/SegFormer-style encoder with a lightweight multi-scale fusion decoder, enabling the network to capture long-range contextual dependencies while preserving fine spatial detail. The final formulation uses a segmentation head for dense PWD prediction and an auxiliary boundary head during training to improve crown-edge delineation, together with differential fine-tuning of the pretrained encoder and task-specific decoding layers. We evaluate SilvaFormer on the public PWD-master UAV benchmark using a fixed train/validation/test split and compare it with representative segmentation baselines, including U-Net, PSPNet, DeepLabV3+, ADE-Net, and two recent transformer-based baselines, SegFormer and UNetFormer. Against the classical CNN and attention-enhanced baselines, SilvaFormer achieves the strongest result, with an IoU of **0.7701**, Dice/F1 of **0.8474**, precision of **0.8449**, and recall of **0.8730**. In the additional transformer-focused comparison, SegFormer and UNetFormer achieve higher IoU values of 0.7842 and 0.8053, respectively, while SilvaFormer remains substantially more compact in parameter count (6.11M versus 28.15M and 30.21M). These results show that SilvaFormer provides a competitive and compact transformer-based solution for UAV-based forest health monitoring, while also clarifying its trade-off relative to larger recent transformer segmentation models.

**INDEX TERMS** Transformer, CNN, forest health, UAV dataset, image segmentation.

## I. INTRODUCTION

Forest ecosystems provide critical ecological and socio-economic services, including biodiversity conservation, carbon sequestration, timber production, and landscape resilience. These services are increasingly threatened by forest pests and diseases, which can spread rapidly and cause severe stand-level damage if not detected and managed in time. Among such threats, *pine wilt disease* (PWD) is one of the most destructive, with substantial ecological and economic consequences in affected pine forests.

PWD is associated with the pinewood nematode *Bursaphelenchus xylophilus* and has caused extensive pine mortality in several regions, creating major challenges for forest management and biosecurity [1], [2]. Beyond direct timber losses, large outbreaks can alter stand structure, ecosystem functioning, and long-term forest composition, while climate variability may further modify the potential distribution and

The associate editor coordinating the review of this manuscript and approving it for publication was Ayman El-Baz.

outbreak risk of the disease [3]. These characteristics make timely, scalable, and reliable forest-health monitoring an important component of sustainable forestry.

Traditional field-based forest health surveys remain valuable, but they are labor-intensive, time-consuming, and difficult to scale across large and heterogeneous landscapes. Remote sensing has therefore become a key tool for forest condition assessment, enabling repeatable observations over much larger areas than can be covered by ground surveys alone [4]. Among remote sensing platforms, unmanned aerial vehicles (UAVs) are especially attractive for forest health monitoring because they offer very high spatial resolution, flexible acquisition schedules, and comparatively low deployment cost [5], [6]. Their utility for forest-health applications has been emphasized in recent reviews, which note both the growing maturity of UAV-based monitoring systems and the continuing need for robust, task-specific AI models for operational deployment [7], [8], [9].

More broadly, forest health monitoring has increasingly relied on UAV and remote-sensing technologies for detecting crown decline, pest outbreaks, defoliation, and disease symptoms across different forest ecosystems. Recent studies have shown that high-resolution aerial imagery combined with machine learning and deep learning can support early stress detection, crown-condition assessment, and stand-level health mapping. However, robust pixel-wise delineation of disease-affected canopy regions remains challenging, particularly when symptoms are sparse, fragmented, and visually similar to surrounding background structures.

The current literature on UAV-based PWD monitoring spans several methodological levels. Earlier studies focused mainly on tree-level detection or image-level classification, where candidate crowns or patches are labeled as healthy or diseased [10], [11]. These approaches are useful for rapid screening but do not provide the detailed spatial delineation needed for crown-level damage assessment or downstream intervention planning. More recent studies have begun to explore deep learning more systematically for PWD-related image analysis, including object detection, segmentation, and attention-enhanced disease recognition [12], [13], [14]. At the same time, broader remote sensing segmentation research has rapidly moved from CNN-only designs toward vision transformers and hierarchical transformer architectures, including ViT, SETR, Swin Transformer, SegFormer, and UNetFormer [15], [16], [17], [18], [19]. More recently, state-space and Mamba-style models such as RTMamba and Samba have also been introduced for remote-sensing semantic segmentation to improve long-range context modeling and computational efficiency [20], [21].

Despite these advances, two gaps remain relevant to this study. First, relatively few works address *pixel-wise* PWD segmentation from UAV RGB imagery under a unified and reproducible evaluation protocol. Second, although recent transformer and Mamba-style models have demonstrated strong results in generic remote sensing segmentation, their design principles have not yet been sufficiently contextualized for the specific challenges of UAV-based forest disease mapping, where infected crowns are sparse, fragmented, and easily confused with heterogeneous canopy texture and background clutter.

Motivated by these observations, this paper presents **SilvaFormer**, a transformer-based semantic segmentation model for delineating PWD-affected canopy regions in UAV RGB imagery. SilvaFormer combines a hierarchical MiT/SegFormer-style encoder with a lightweight multi-scale fusion decoder, enabling the model to exploit both local visual detail and broader spatial context while retaining a compact prediction pipeline. We evaluate SilvaFormer on the public PWD UAV benchmark and compare it with representative CNN-based baselines, including U-Net, PSPNet, DeepLabV3+, and ADE-Net, using a fixed train/validation/test split and a standardized evaluation protocol. To address recent developments in segmentation architectures, we further include a focused comparison with SegFormer and UNetFormer under the same test split.

The main contributions of this study are fourfold. First, we introduce **SilvaFormer**, a transformer-based segmentation model for UAV forest-disease mapping that combines a hierarchical MiT/SegFormer-style encoder with a lightweight multi-scale fusion decoder. Second, we formulate the final method around a compact dense-prediction pathway with auxiliary boundary supervision during training, emphasizing stable fine-tuning and compact parameterization. Third, we provide a reproducible evaluation setup on a public UAV PWD benchmark, including consistent data splitting and standardized reporting of segmentation metrics. Fourth, we expand the experimental analysis with modern transformer-based baselines, component-level interpretation, and computational-complexity reporting, including parameters, FLOPs, inference time, and FPS.

## II. MATERIALS AND DATA

### A. STUDY AREA AND UAV IMAGE ACQUISITION

This study uses a publicly available UAV-based RGB dataset for *pine wilt disease* (PWD) crown segmentation introduced by Xia et al. [12]. The imagery was collected in the Laoshan District, Qingdao (Shandong Province, China), an artificial coniferous forest landscape dominated by black pine and *Pinus densiflora*. Xia et al. report four acquisition subareas (A1–A4) within Laoshan used for data collection and annotation.

UAV imaging was conducted from 6–14 October 2018 using a fixed-wing UAV (DB-2, Dabai Technology Co. Ltd., China) equipped with a Sony Alpha 7R II RGB camera (maximum native resolution 7952 × 4472 pixels). Flights were carried out below 700 m altitude with at least 75% forward overlap and 50% side overlap, yielding an approximate ground sampling distance (GSD) of 8 cm per pixel [12]. These acquisition settings were selected to maximize visibility of late-stage PWD symptoms in autumn when infected crowns appear clearly discolored compared to surrounding healthy canopy.

**TABLE 1. Dataset composition and fixed split used in all experiments.**

| Subset | #Tiles | Tile size | Channels | Mask |
|---|---|---|---|---|
| Train | 140 | 256×256 | RGB (3) | binary |
| Validation | 30 | 256×256 | RGB (3) | binary |
| Test | 30 | 256×256 | RGB (3) | binary |
| Total | 200 | 256×256 | RGB (3) | binary |

### B. FIELD SURVEY AND ANNOTATION PROTOCOL

To support accurate interpretation of PWD symptoms in aerial images, Xia et al. conducted a field campaign (27–31 October 2018) and used it to validate manual interpretation of infected trees in UAV imagery (reported interpretation accuracy ≈97%) [12]. For pixel-wise labeling, 45 representative high-resolution UAV images were selected and annotated using the ENVI region-of-interest tool; in total, 2352 infected pine crowns were identified over diverse backgrounds (e.g., water, rocks, farmland, buildings, and mixed vegetation) [12].

### C. DATASET CONSTRUCTION

From the annotated high-resolution UAV images, the dataset is constructed by clipping fixed-size tiles and producing paired binary masks. Following the dataset protocol in [12], **200 RGB tiles** of size 256 × 256 pixels were clipped from the 7952 × 4472 orthophotos such that each tile contains pixels of infected crowns. Each tile is paired with a **binary mask** where label 1 indicates PWD-affected crown pixels and 0 denotes background/healthy vegetation.

Given the reported GSD of ∼8 cm, each 256 × 256 tile corresponds to an area of roughly 20.5 m × 20.5 m on the ground. The dataset is strongly imbalanced because infected crowns occupy a small fraction of pixels compared to background canopy; therefore, we report overlap-based and class-imbalance-aware metrics (e.g., IoU, Dice/F1, precision, recall) in all experiments.

### D. TRAIN/VALIDATION/TEST SPLIT

To ensure a fair and repeatable comparison across all architectures, we use a **fixed split** stored as index files. This prevents accidental re-splitting and guarantees that every model is trained and evaluated on the same samples.

We adopt a **70%/15%/15%** split over the 200 tiles: **140** training tiles, **30** validation tiles, and **30** test tiles. Note that we used a source-image-level split to avoid spatial leakage: all tiles extracted from a given orthophoto are assigned to only one subset (train, validation, or test). Table 1 summarizes the dataset composition used in our experiments.

### E. DATA VISUALIZATION AND SUMMARY STATISTICS

For transparency, we provide qualitative and quantitative summaries of the dataset: (i) representative RGB tiles and their corresponding ground-truth masks, and (ii) per-tile foreground (infected-pixel) ratio distribution to illustrate class imbalance. Figure 1 shows example image–mask pairs. Figure 2 reports dataset-level statistics computed over the 200 tiles and the fixed split counts.

**FIGURE 1. Randomly sampled RGB tiles from the PWD test set with ground-truth crown masks overlaid (red).**

### F. TASK FORMULATION AND SCOPE

Although the available annotations originate from infected tree crowns, the released benchmark and training protocol are organized as a binary mask segmentation task on clipped UAV image tiles rather than as a fully instance-labeled crown dataset with instance identifiers and an associated evaluation protocol. For this reason, the present study formulates the problem as semantic segmentation. This choice is also aligned with the practical objective of estimating the spatial extent of diseased canopy regions, which is directly useful for visual damage assessment and subsequent forest-health mapping. We agree that instance segmentation is an important and promising direction for future work, particularly for tree-level inventory and decision support. However, constructing a reliable instance-segmentation benchmark requires consistent crown-instance annotations and evaluation procedures beyond the scope of the currently released dataset.

## III. METHODOLOGY

### A. OVERVIEW

SilvaFormer is a transformer-based semantic segmentation framework designed to delineate pine wilt disease (PWD)

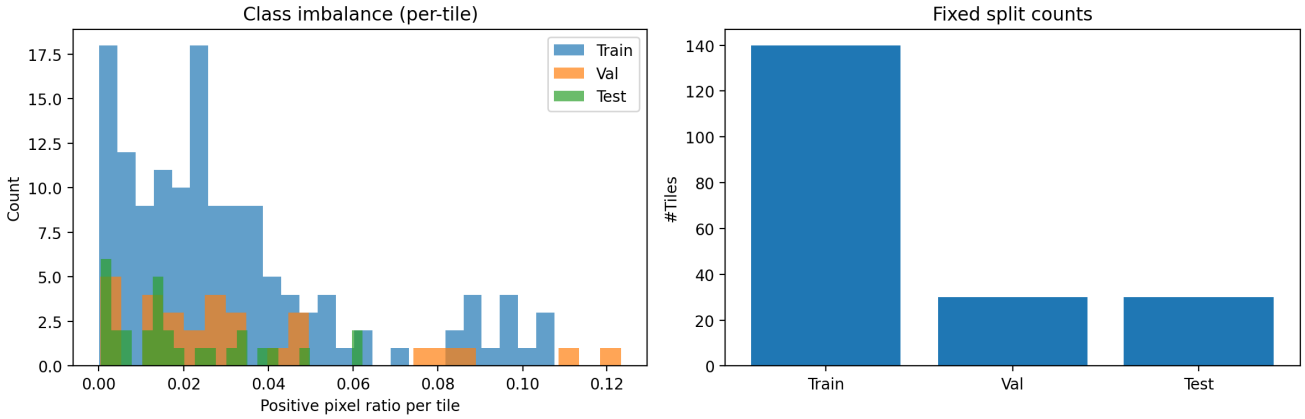



**FIGURE 2.** Dataset statistics. Left: distribution of infected-pixel ratios per tile, highlighting strong class imbalance. Right: fixed split counts used for all model comparisons.

symptoms from high-resolution UAV RGB tiles. The key challenge in this task is that infected crown regions usually occupy only a small fraction of the scene and are embedded in visually heterogeneous backgrounds. To address this, SilvaFormer adopts a compact encoder–decoder design in which a hierarchical transformer encoder provides rich multi-scale contextual features and a lightweight decoder aggregates them for dense prediction.

The final formulation of SilvaFormer consists of four principal components: (i) a hierarchical MiT/SegFormer-style encoder, (ii) a lightweight multi-scale fusion decoder, (iii) a binary segmentation head, and (iv) an auxiliary boundary head used during training. The segmentation head provides the final semantic prediction, whereas the boundary head supplies an additional supervision signal for crown-edge delineation. This design keeps the inference output aligned with the semantic segmentation task while still allowing boundary-aware optimization during training. The overall architecture is illustrated in Fig. 3.

### B. RELATION TO SEGFORMER AND ARCHITECTURAL NOVELTY

SilvaFormer inherits the hierarchical transformer-encoder principle from SegFormer-style architectures rather than proposing a new transformer block. The novelty of the present model lies in its task-specific adaptation for UAV-based PWD mapping: (i) a compact multi-scale fusion decoder tailored to sparse diseased-crown masks, (ii) an auxiliary boundary-supervision branch used to sharpen crown-edge delineation during training, (iii) differential fine-tuning between pretrained encoder and task-specific layers for stable optimization on a small UAV forest dataset, and (iv) an application-specific evaluation protocol for PWD-affected crown segmentation. Thus, SilvaFormer should be understood as a specialized and compact SegFormer-style segmentation framework for forest-health monitoring, not as a replacement for the general SegFormer architecture.

### C. TRANSFORMER ENCODER: SEGFORMER BACKBONE

Let the input UAV tile be denoted by $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, where in our experiments $H = W = 256$. SilvaFormer employs a hierarchical Mix Transformer (MiT) encoder following the SegFormer design principle [18]. Unlike purely convolutional backbones, the transformer encoder can model longer-range spatial relationships more directly, which is particularly beneficial for UAV forest imagery where diseased canopy regions can be fragmented, spatially sparse, and visually similar to shadows or background vegetation.

The encoder maps the input image to a pyramid of multi-scale feature tensors:

$$\{\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3, \mathbf{F}_4\} = \mathrm{Enc}(\mathbf{I}), \tag{1}$$

where $\mathbf{F}_i \in \mathbb{R}^{h_i\times w_i\times C_i}$ denotes the feature representation at stage $i$. The successive stages progressively reduce spatial resolution while increasing semantic abstraction, thereby supplying both local texture information and broader contextual cues that are valuable for dense disease delineation.

### D. LIGHTWEIGHT MULTI-SCALE DECODER (FPN-STYLE FUSION)

A central design objective of SilvaFormer is to keep the decoder lightweight. Each encoder feature map is first projected to a common embedding dimension $D$ using a $1 \times 1$ convolution:

$$\hat{\mathbf{F}}_i = \phi_i(\mathbf{F}_i), \qquad i \in \{1, 2, 3, 4\}, \tag{2}$$

where $\phi_i(\cdot)$ denotes the stage-specific projection and $D = 256$ in our implementation.

The projected feature maps are then spatially aligned by bilinear upsampling to the resolution of the highest-resolution stage:

$$\tilde{\mathbf{F}}_i = \mathrm{Up}(\hat{\mathbf{F}}_i, \mathrm{size}(\hat{\mathbf{F}}_1)). \tag{3}$$

The aligned features are concatenated and fused as

$$\mathbf{Z} = \psi\left([\tilde{\mathbf{F}}_1 \parallel \tilde{\mathbf{F}}_2 \parallel \tilde{\mathbf{F}}_3 \parallel \tilde{\mathbf{F}}_4]\right), \tag{4}$$

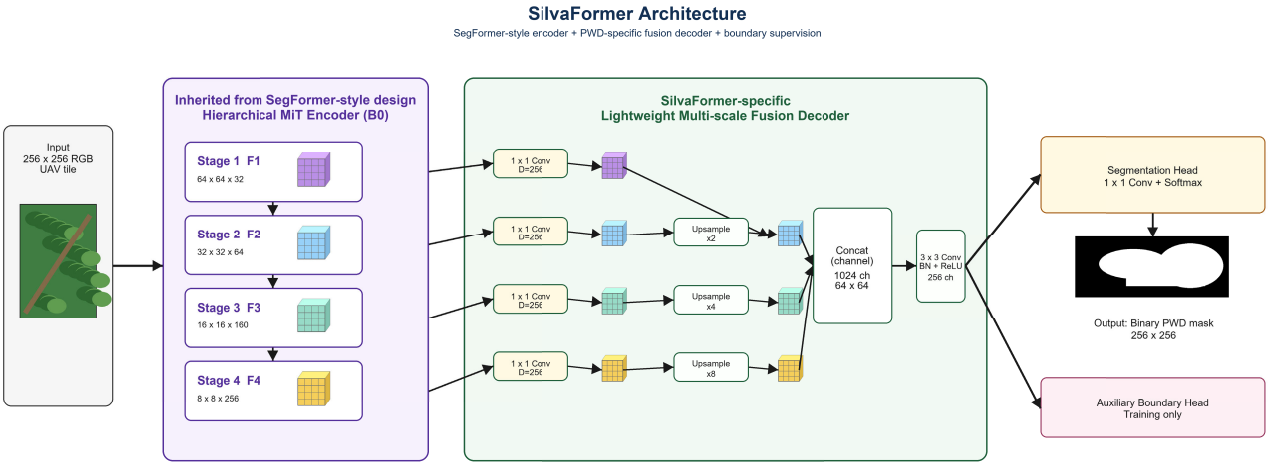



**FIGURE 3. Architecture of SilvaFormer. A hierarchical MiT/SegFormer-style encoder extracts four feature stages, which are projected to a shared embedding dimension and fused by a lightweight multi-scale decoder. The segmentation head predicts the binary PWD mask, while an auxiliary boundary head is used during training to encourage sharper crown-edge delineation. Differential learning rates are used for stable fine-tuning of pretrained encoder and task-specific layers.**

where $[\cdot \parallel \cdot]$ denotes channel-wise concatenation and $\psi(\cdot)$ is a lightweight fusion block composed of a convolution, normalization, and nonlinearity. This simple decoder preserves multi-scale information while maintaining computational efficiency.

### E. PREDICTION AND AUXILIARY BOUNDARY HEADS

The fused representation $\mathbf{Z}$ is passed through a segmentation head to produce class logits:

$$\mathbf{S} = \text{Head}_{seg}(\mathbf{Z}). \tag{5}$$

For binary segmentation, the head outputs logits corresponding to background and diseased canopy. The logits are then upsampled to the input resolution,

$$\hat{\mathbf{Y}} = \text{Up}(\mathbf{S}, (H, W)), \tag{6}$$

from which the final PWD mask is obtained by pixel-wise classification. The semantic segmentation logits therefore remain the primary prediction used for all quantitative evaluation.

During training, SilvaFormer also uses an auxiliary boundary head to predict a binary crown-boundary map from the same fused representation:

$$\mathbf{B} = \text{Head}_{bnd}(\mathbf{Z}). \tag{7}$$

This auxiliary output is used only as an additional training signal to encourage sharper disease-region boundaries. The primary output used for semantic segmentation evaluation remains the class-logit prediction $\hat{\mathbf{Y}}$.

### F. TRAINING OBJECTIVE

Let $y_u \in \{0, 1\}$ denote the ground-truth label of pixel $u$, and let $p_{u,c}$ denote the predicted softmax probability for class $c \in \{0, 1\}$. SilvaFormer is optimized using a compound segmentation objective that combines focal loss, Dice loss, and an auxiliary boundary loss:

$$\mathcal{L} = \lambda_f \mathcal{L}_{focal} + \lambda_d \mathcal{L}_{dice} + \lambda_b \mathcal{L}_{bnd}, \tag{8}$$

where $(\lambda_d, \lambda_f, \lambda_b) = (0.5, 0.3, 0.2)$ in our experiments, matching the configuration in Table 2. The focal term reduces the dominance of the background class, the Dice term directly optimizes foreground overlap, and the boundary term encourages sharper delineation of infected crown edges. This formulation is used consistently in the final training configuration; the ablation study in Section V-E evaluates the effect of removing the Dice or boundary terms.

### G. OPTIMIZATION AND IMPLEMENTATION DETAILS

A practical feature of SilvaFormer is its differential fine-tuning strategy. Because the encoder is initialized from a pretrained backbone while the decoder and segmentation head are task-specific, they are optimized with different learning rates:

$$\eta_{\text{enc}} = \lambda \eta, \qquad \eta_{\text{head}} = \eta, \tag{9}$$

where $\eta$ denotes the base learning rate and $0 < \lambda < 1$ scales the encoder update magnitude. This helps stabilize optimization on the relatively small PWD benchmark while allowing the task-specific layers to adapt more rapidly.

To ensure fair comparison, all models are trained using identical fixed train/validation/test splits (Section II) and evaluated with the same metrics (IoU, Dice/F1, Precision, Recall, Accuracy). Data augmentation is applied only during training, while validation and test images are evaluated without augmentation.

### H. INFERENCE

At inference time, an input UAV RGB tile is forwarded through the encoder, decoder, and segmentation head

to obtain the predicted binary mask. The boundary head is not used for the final semantic mask, although it is retained in the trained network as an auxiliary branch. The reported segmentation metrics are computed from the segmentation logits only.

## IV. EXPERIMENTAL SETUP

### A. PROTOCOL AND REPRODUCIBILITY

All models are trained and evaluated on an identical fixed split of the PWD UAV RGB tiles. The split is generated deterministically with seed 42, using proportions of 70%/15%/15% for train/validation/test. Data augmentation is enabled only for training and is disabled for validation and test evaluation.

### B. MODELS COMPARED

We compare SilvaFormer against representative semantic segmentation baselines commonly used in dense prediction and remote-sensing studies:

- **U-Net** [22]
- **DeepLabV3+** [23]
- **PSPNet** [24]
- **ADE-Net** [25]
- **SilvaFormer (ours)**

These baselines span classical encoder–decoder designs, pyramid-context aggregation, and attention-enhanced CNN segmentation, providing a meaningful reference set for assessing the proposed transformer-based approach. In response to recent advances in transformer-based remote sensing segmentation, we also conduct a focused additional comparison with **SegFormer** [18] and **UNetFormer** [19]. This additional experiment is reported separately because its purpose is to position SilvaFormer against modern transformer-style segmentation architectures rather than to replace the original classical baseline comparison.

### C. TRAINING SETUP

All compared models are trained within a unified experimental framework to ensure methodological consistency. The dataset split is fixed across all models, and the same preprocessing and augmentation pipeline is used during training. SilvaFormer follows the final formulation described in Section III: the model is optimized using the compound segmentation and boundary-aware objective in Table 2, while the pretrained encoder and the task-specific decoder/head layers are fine-tuned with differential learning rates.

**TABLE 2. Training configuration used across all models.**

| Setting | Value |
|---|---|
| Split (train/val/test) | 0.70 / 0.15 / 0.15, seed 42 |
| Native tile size | 256×256 pixels |
| Augmentation | train: on, val/test: off |
| Epochs / batch size | 100 / 8 |
| Optimizer | AdamW, weight decay $10^{-4}$ |
| Base learning rate | $6 \times 10^{-3}$ |
| Scheduler | CosineAnnealingLR, $T_{\max} = 100$ |
| Gradient clipping | $\|\nabla\|_2 \leq 1.0$ |
| Loss (U-Net/DeepLab/PSP) | Focal: $\alpha = 0.25, \gamma = 2.0$ |
| Loss (ADE-Net/SilvaFormer) | Dice+Focal+Boundary: $(0.5, 0.3, 0.2)$ |
| SilvaFormer LR groups | $\eta_{\text{enc}} = 0.1\eta,\ \eta_{\text{head}} = \eta$ |

### D. EVALUATION METRICS

All evaluation metrics are computed at pixel level for the disease class. Let $TP$, $FP$, $FN$, and $TN$ denote the numbers of true positives, false positives, false negatives, and true negatives, respectively, and let $\epsilon$ be a small constant.

#### 1) INTERSECTION-OVER-UNION (IOU)

$$\text{IoU} = \frac{TP}{TP + FP + FN + \epsilon}. \tag{10}$$

#### 2) DICE COEFFICIENT

$$\text{Dice} = \frac{2TP}{2TP + FP + FN + \epsilon}. \tag{11}$$

#### 3) PRECISION, RECALL, AND F1

$$\text{Precision} = \frac{TP}{TP + FP + \epsilon}, \qquad \text{Recall} = \frac{TP}{TP + FN + \epsilon}, \tag{12}$$

$$\text{F1} = \frac{2\,\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall} + \epsilon}. \tag{13}$$

#### 4) ACCURACY

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN + \epsilon}. \tag{14}$$

#### 5) FOREGROUND RATIO DIAGNOSTICS

To monitor behavior under severe class imbalance, we further report the predicted and ground-truth positive ratios:

$$r_{pred} = \frac{1}{HW} \sum_u \mathbb{1}[\hat{y}_u = 1], \qquad r_{gt} = \frac{1}{HW} \sum_u \mathbb{1}[y_u = 1]. \tag{15}$$

### E. INFERENCE SPEED

We report inference throughput on the test set. Inference time is measured with GPU synchronization when available; the average milliseconds per image and FPS are computed from the recorded forward-pass timings:

$$\text{ms/img} = 1000 \cdot \frac{\bar{t}}{B}, \qquad \text{FPS} = \left(\frac{\bar{t}}{B}\right)^{-1}, \tag{16}$$

where $\bar{t}$ is the mean per-batch forward time and $B$ is the batch size used during evaluation.

#### 1) FLOATING-POINT OPERATIONS.

To complement runtime measurements, we also report floating-point operations (FLOPs). For the main comparison, FLOPs are estimated for a single forward pass at the native tile size $1 \times 3 \times 256 \times 256$. For the focused transformer comparison, FLOPs are reported at $1 \times 3 \times 512 \times 512$, matching the input resolution used in that additional experiment. Since FLOP estimates depend on operator definitions and implementation details, they are

interpreted together with parameter counts and measured inference time rather than as an isolated efficiency indicator.

## V. RESULTS

### A. QUANTITATIVE COMPARISON ON THE PWD TEST SET

Table 3 summarizes the test-set performance of the classical and attention-enhanced baselines under the fixed split protocol. SilvaFormer achieves the best segmentation performance in this comparison, reaching an IoU of **0.7701** and a Dice/F1 score of **0.8474**. The closest competing baseline is DeepLabV3+, which attains an IoU of 0.7494 and a Dice score of 0.8338, while ADE-Net, U-Net, and PSPNet remain below these values. This result indicates that SilvaFormer improves the PWD segmentation accuracy relative to the classical CNN-based and attention-enhanced baselines.

SilvaFormer also attains the highest recall (0.8730) and accuracy (0.9968) among the models in Table 3, demonstrating a strong ability to recover sparse diseased pixels while preserving background regions. In addition, its predicted positive ratio remains close to the ground-truth foreground ratio ($r_{pred} = 0.0193$ versus $r_{gt} = 0.0192$), indicating stable behavior despite the severe class imbalance of the dataset.

From an efficiency perspective, SilvaFormer is slower than the CNN-based baselines in Table 3, requiring 164.86 ms per image and achieving 6.07 FPS. This indicates that the improved disease delineation is obtained at the cost of increased inference time, which is expected for a transformer-based encoder. The result is therefore best interpreted as an accuracy-oriented improvement over the classical baselines rather than as a throughput-oriented design.

These findings support the central claim of this work: the combination of hierarchical transformer features and a lightweight multi-scale decoder is effective for UAV-based forest-disease segmentation. The gains over strong CNN baselines, particularly DeepLabV3+, suggest that richer contextual modeling is beneficial for recovering sparse, irregular, and visually ambiguous diseased crown regions.

### B. EFFICIENCY ANALYSIS

Table 3 also reports inference speed. DeepLabV3+ is the fastest strong baseline at 36.42 ms/image (27.46 FPS), whereas SilvaFormer requires 164.86 ms/image (6.07 FPS). This reduction in throughput is consistent with the higher representational cost of transformer-based feature extraction. Nevertheless, SilvaFormer provides the best segmentation accuracy, making it attractive for offline or near-offline forest-health assessment workflows where delineation quality is prioritized.

Table 4 reports parameter counts and FLOPs for the main comparison models. SilvaFormer has the smallest parameter count (6.11M), substantially below U-Net (31.04M), DeepLabV3+ (40.35M), ADE-Net (41.34M), and PSPNet (70.18M). Its FLOPs at the native tile size are comparable to DeepLabV3+ and ADE-Net and much lower than U-Net. These results indicate that the model is parameter-efficient, even though its measured runtime is slower in the current implementation.

### C. COMPARISON WITH RECENT TRANSFORMER-BASED BASELINES

To further address recent progress in segmentation architectures, we conducted an additional focused comparison against SegFormer [18] and UNetFormer [19]. Both models were trained and evaluated using the same PWD test split as SilvaFormer. Table 5 summarizes the results, including parameter count, FLOPs, inference time, and FPS.

The transformer-focused comparison provides a more nuanced view of the proposed method. SegFormer and UNetFormer achieve IoU values of 0.7842 and 0.8053, respectively, which are higher than SilvaFormer's 0.7701 in this experiment. UNetFormer obtains the strongest overlap-based scores, with an IoU of 0.8053 and Dice/F1 of 0.8730. However, SilvaFormer uses substantially fewer parameters: 6.11M compared with 28.15M for SegFormer and 30.21M for UNetFormer. Thus, while the larger transformer baselines achieve higher accuracy and faster measured inference in the current implementation, SilvaFormer remains a compact transformer-based alternative with competitive accuracy and a much smaller parameter footprint.

We did not include a full Mamba-based experimental benchmark because stable and directly comparable Mamba segmentation implementations for this binary UAV PWD setup are less standardized than mature transformer baselines. Nevertheless, recent Mamba-style remote sensing segmentation models such as RTMamba and Samba are discussed in the related work and future-work context [20], [21]. The present comparison therefore focuses on reproducible transformer-based baselines while acknowledging Mamba-based forest-health segmentation as an important direction for subsequent work.

### D. CONFUSION MATRIX

Pixel-level confusion matrices further clarify the behavior of the compared methods. SilvaFormer achieves the most balanced trade-off between disease sensitivity and background preservation. Specifically, it correctly identifies 93.42% of disease pixels while misclassifying only 6.58% of them as background, and at the same time preserves background regions with 99.73% accuracy, with only 0.27% false positive leakage into the disease class. These results indicate that SilvaFormer is able to recover diseased areas reliably without noticeably over-segmenting surrounding healthy canopy.

In comparison, U-Net attains a slightly higher disease recall (95.99%) but does so at the expense of a much larger false positive rate, incorrectly labeling 21.71% of background pixels as diseased. DeepLabV3+ is considerably more conservative, preserving background well (99.60%) but recovering only 71.02% of disease pixels. ADENet provides a relatively strong compromise, with 90.71% disease recovery and 3.07% background false positives, but still

**TABLE 3.** Test-set performance on the UAV PWD segmentation benchmark using the fixed split. Best values are shown in bold.

| Model | IoU | Dice | Precision | Recall | Accuracy | $r_{pred}$ | $r_{gt}$ | ms/img | FPS |
|---|---|---|---|---|---|---|---|---|---|
| U-Net | 0.6726 | 0.7552 | 0.8224 | 0.7484 | 0.9956 | 0.0188 | 0.0192 | 90.36 | 11.07 |
| DeepLabV3+ | 0.7494 | 0.8338 | 0.8424 | 0.8481 | 0.9966 | 0.0195 | 0.0192 | 36.42 | 27.46 |
| PSPNet | 0.5642 | 0.6829 | 0.7628 | 0.6598 | 0.9930 | 0.0180 | 0.0192 | 32.40 | 30.87 |
| ADE-Net | 0.7156 | 0.8020 | 0.7816 | 0.8387 | 0.9959 | 0.0199 | 0.0192 | 34.30 | 29.15 |
| **SilvaFormer (ours)** | **0.7701** | **0.8474** | **0.8449** | **0.8730** | **0.9968** | 0.0193 | 0.0192 | 164.86 | 6.07 |

**TABLE 4.** Computational complexity of the main comparison models. FLOPs are estimated for one forward pass with input size 1 × 3 × 256 × 256.

| Model | Params (M) | FLOPs (G) |
|---|---|---|
| U-Net | 31.04 | 92.05 |
| DeepLabV3+ | 40.35 | 23.27 |
| PSPNet | 70.18 | **16.22** |
| ADE-Net | 41.34 | 25.70 |
| SilvaFormer | **6.11** | 23.09 |

remains inferior to SilvaFormer in disease sensitivity. PSPNet performs worst in this analysis, identifying only 1.43% of disease pixels and therefore missing the vast majority of diseased regions.

The qualitative examples in Fig. 5 are consistent with these observations. SilvaFormer more accurately delineates the spatial extent of diseased crowns, especially when symptoms are fragmented or visually subtle. Compared with U-Net, it markedly suppresses over-segmentation in cluttered backgrounds; compared with DeepLabV3+, it produces more complete disease masks; and compared with ADENet, it captures a larger proportion of diseased pixels while maintaining very low background confusion. Overall, the confusion matrices and qualitative results together confirm that SilvaFormer offers the most favorable balance between sensitivity to diseased regions and robustness against false alarms.

### E. COMPONENT ABLATION OF SILVAFORMER

To better understand the contribution of the principal design components, we conducted a component-level ablation study under the same fixed split. Starting from the full SilvaFormer training configuration in this controlled ablation protocol, we evaluated three variants: removing boundary supervision, removing the Dice term, and replacing the transformer encoder with a CNN backbone. Because the ablation runs were used to assess relative component effects, Table 6 reports performance changes relative to the full ablation configuration rather than repeating a second absolute SilvaFormer benchmark score.

The results show that removing boundary supervision causes a modest reduction in IoU and Dice/F1, suggesting that boundary-aware supervision contributes to delineation quality but is not the sole driver of performance. Removing the Dice term leads to a larger drop in IoU and Boundary F1, indicating that overlap-aware optimization is helpful under severe foreground imbalance. The largest degradation occurs when the transformer encoder is replaced with a CNN backbone: IoU decreases by 0.0329 and HD95 increases by 75.17, confirming that the hierarchical transformer representation is central to robust PWD delineation.

### F. STABILITY ACROSS RANDOM SEEDS

Given the relatively small dataset size, we also examined the stability of SilvaFormer across three random seeds while keeping the dataset split fixed. In this controlled stability check, the IoU was $0.6691 \pm 0.0068$, Dice/F1 was $0.8017 \pm 0.0049$, precision was $0.9159 \pm 0.0082$, recall was $0.7130 \pm 0.0122$, Boundary F1 was $0.4163 \pm 0.0055$, and HD95 was $106.77 \pm 3.50$. These results indicate relatively low run-to-run variability for the proposed model under the fixed split. However, because this stability check was conducted for SilvaFormer rather than for every baseline model, we do not use it to rank models; instead, it is reported to clarify the robustness of the proposed training procedure on the small PWD benchmark.

## VI. DISCUSSION

### A. WHY DOES SILVAFORMER OUTPERFORM CNN BASELINES?

The experimental results show that SilvaFormer consistently surpasses the compared CNN-based baselines on the PWD benchmark. Relative to DeepLabV3+, which is the strongest competing model in the main benchmark, SilvaFormer provides higher IoU, Dice, precision, recall, and accuracy. This suggests that the proposed architecture is better aligned with the visual characteristics of UAV forest imagery, where disease symptoms occupy only a small proportion of the scene and are embedded within visually heterogeneous canopy patterns.

A primary explanation is the hierarchical transformer encoder. By capturing broader spatial context more effectively than standard convolutional backbones, it helps the model recognize diseased regions that may be fragmented or only weakly contrasted against the background. The lightweight multi-scale decoder complements this property by aggregating features from multiple encoder stages without introducing an unnecessarily heavy reconstruction stage.

### B. BEHAVIOR UNDER SEVERE CLASS IMBALANCE

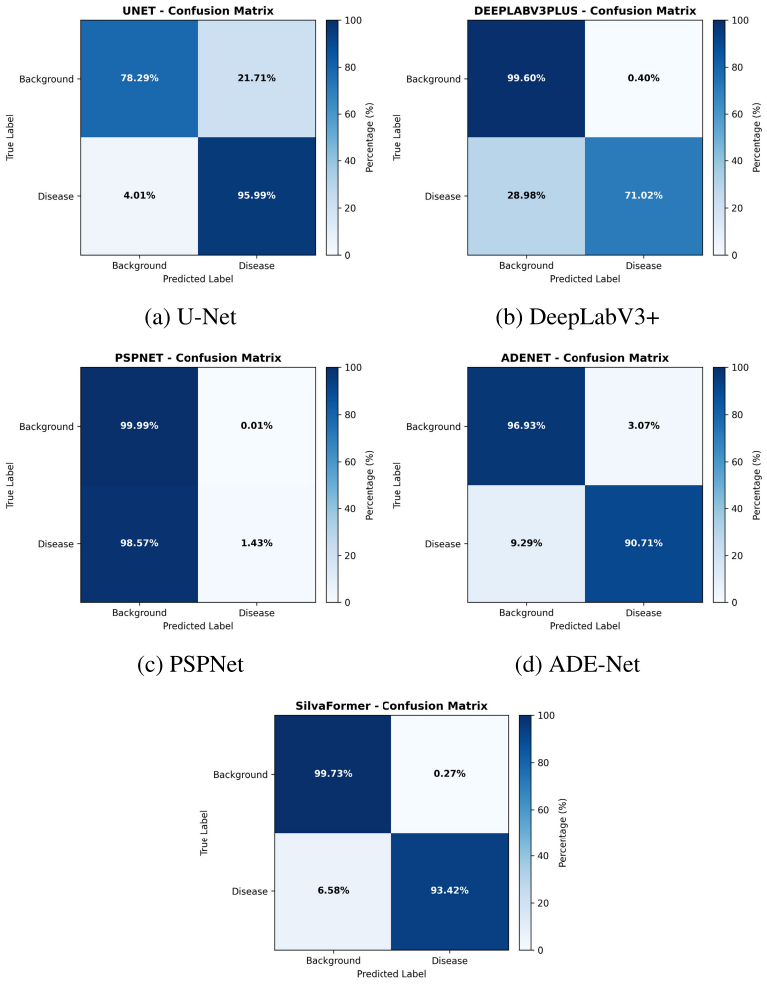
The PWD benchmark is highly imbalanced, with the diseased class occupying only about 1.92% of image pixels.

**TABLE 5.** Focused comparison with recent transformer-based segmentation baselines on the fixed PWD test split. FLOPs are estimated for a single forward pass with input size 1 × 3 × 512 × 512. Best accuracy values are shown in bold; lower values are better for Params, FLOPs, and ms/img.

| Model | Params (M) | FLOPs (G) | IoU | Dice/F1 | Precision | Recall | Accuracy | ms/img | FPS |
|---|---|---|---|---|---|---|---|---|---|
| SilvaFormer | **6.11** | 93.94 | 0.7701 | 0.8474 | 0.8449 | 0.8730 | 0.9968 | 119.44 | 8.37 |
| SegFormer | 28.15 | **61.17** | 0.7842 | 0.8577 | **0.8629** | 0.8606 | 0.9972 | 46.44 | 21.53 |
| UNetFormer | 30.21 | 62.76 | **0.8053** | **0.8730** | 0.8582 | **0.9034** | **0.9975** | **32.28** | **30.98** |

**TABLE 6.** Component ablation of SilvaFormer under the fixed split. Values are changes relative to the full controlled ablation configuration; negative ΔIoU, ΔDice/F1, and ΔBoundary F1 indicate performance degradation, while positive ΔHD95 indicates worse boundary-distance error.

| Variant | ΔIoU | ΔDice/F1 | ΔBoundary F1 | ΔHD95 | Primary effect tested |
|---|---|---|---|---|---|
| Without boundary supervision | -0.0112 | -0.0080 | -0.0039 | -14.06 | Boundary cue contribution |
| Without Dice term | -0.0185 | -0.0133 | -0.0482 | -10.86 | Overlap-aware loss contribution |
| CNN-backbone variant | -0.0329 | -0.0239 | -0.0654 | +75.17 | Transformer encoder contribution |




(a) U-Net (b) DeepLabV3+

(c) PSPNet (d) ADE-Net

(e) SilvaFormer (ours)

**FIGURE 4.** Pixel-level confusion matrices on the UAV PWD test set. Rows correspond to ground-truth classes and columns to predicted classes.

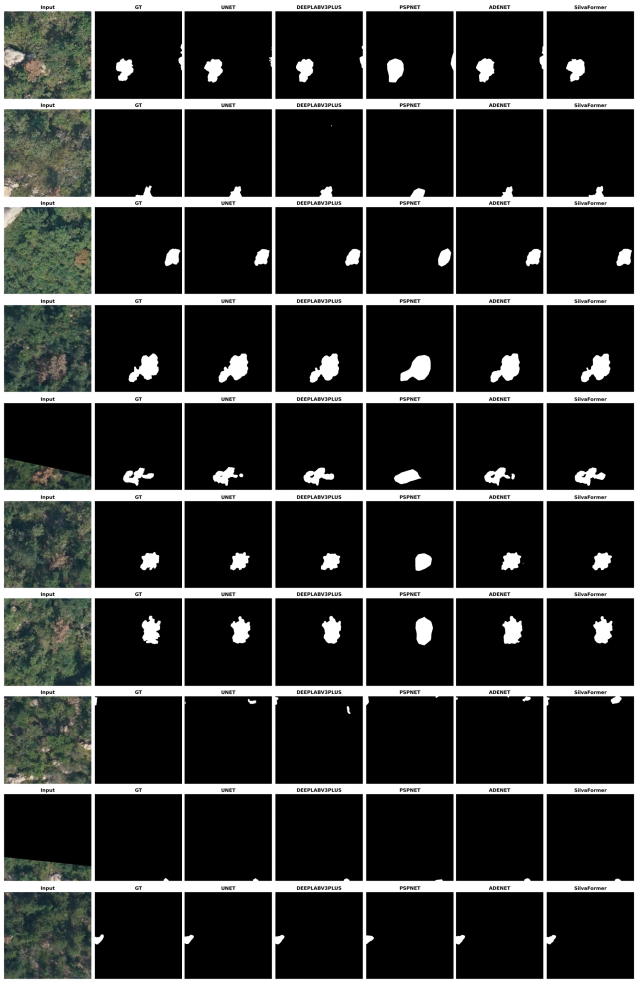



**FIGURE 5.** Qualitative comparison on representative test tiles. In each panel, one test sample is visualized, including the RGB input, the ground-truth mask, and the predictions from all compared methods.

Under such conditions, segmentation models may either collapse toward background-dominated solutions or overcompensate by producing excessive false positives. In the main baseline comparison, SilvaFormer remains well calibrated, with a predicted positive ratio closely matching the ground-truth foreground ratio, while also achieving the highest recall among the CNN-based and attention-enhanced baselines. This indicates strong sensitivity to sparse diseased canopy regions.

### C. INSIGHTS FROM THE COMPONENT ABLATION

The component ablation clarifies how the proposed design behaves when key elements are removed. Removing boundary supervision produces a modest degradation, whereas removing the Dice term reduces both overlap-based accuracy and boundary quality. This confirms that loss design matters in the presence of severe class imbalance. The most informative variant is the CNN-backbone replacement: when the transformer encoder is removed, IoU, Dice/F1, and Boundary F1 all decrease, while HD95 increases substantially. This supports the interpretation that the hierarchical transformer representation is the most important architectural contributor, while boundary and overlap-aware losses provide complementary refinement.

### D. ACCURACY–EFFICIENCY TRADE-OFF

SilvaFormer is not the fastest model in the benchmark, and its inference cost is higher than that of the CNN baselines. However, this additional cost corresponds to a measurable gain in segmentation quality over the classical comparison set. The focused transformer comparison further shows that larger recent transformer baselines can outperform SilvaFormer in IoU and throughput in the current implementation. At the same time, SilvaFormer has a considerably smaller parameter count, which may be advantageous when model storage, transfer, or fine-tuning stability are important. Therefore, the revised interpretation is not that SilvaFormer dominates all modern transformer baselines, but that it offers a compact and competitive transformer-based alternative for UAV PWD segmentation.

### E. POSITION RELATIVE TO TRANSFORMER AND MAMBA-BASED SEGMENTATION

The additional SegFormer and UNetFormer experiments strengthen the positioning of this work relative to recent segmentation architectures. UNetFormer achieves the highest IoU in the focused comparison, suggesting that larger transformer encoder–decoder designs remain highly competitive for UAV forest-disease mapping. SilvaFormer nevertheless remains relevant because it is parameter-efficient and specifically formulated around sparse diseased-crown delineation. Mamba-based models, including recent remote-sensing architectures such as RTMamba and Samba, are promising because state-space modeling may provide long-range context with favorable scaling properties. However, directly comparable and stable implementations for this specific binary PWD benchmark are not yet as standardized as SegFormer- or UNetFormer-style transformer baselines. We therefore treat Mamba-based forest-health segmentation as an important future benchmarking direction rather than as a fully resolved comparison in the present study.

### F. LIMITATIONS AND FUTURE WORK

This study is limited by the relatively small size of the public PWD benchmark and by its focus on binary semantic segmentation from RGB UAV imagery. Although the source annotations are based on infected tree crowns, the released benchmark does not provide a standardized instance-level protocol with instance identifiers and instance-segmentation metrics. Future work should therefore investigate instance segmentation for single-tree disease attribution, larger and more diverse forest-health datasets, multi-modal inputs, domain adaptation across acquisition conditions, Mamba-style segmentation backbones, and lighter transformer variants for computationally constrained deployment.

### G. PRACTICAL CONSEQUENCES FOR FOREST HEALTH MONITORING

The results suggest that SilvaFormer can serve as an effective segmentation model for operational UAV-based forest-health analysis. Accurate delineation of diseased crown regions can support stand-level mapping, damage quantification, and targeted intervention planning. More broadly, this study demonstrates that transformer-based dense prediction can provide tangible benefits in forestry applications where subtle canopy symptoms must be localized precisely.

## VII. CONCLUSION

This paper presented SilvaFormer, a transformer-based semantic segmentation framework for UAV-based pine wilt disease monitoring. The method combines a hierarchical MiT/SegFormer-style encoder with lightweight multi-scale feature fusion to delineate sparse diseased-crown regions in high-resolution RGB imagery. On the fixed PWD benchmark split, SilvaFormer improves over representative CNN-based and attention-enhanced baselines, achieving an IoU of 0.7701 and Dice/F1 of 0.8474 while maintaining a predicted foreground ratio close to the ground truth. A component-level ablation further indicates that the transformer encoder is the most influential architectural component, while Dice and boundary-related terms provide additional refinement under severe class imbalance.

To address recent developments in semantic segmentation, the revised study also included SegFormer and UNetFormer baselines and reported parameter counts, FLOPs, inference time, and FPS. The additional comparison shows that the larger transformer baselines, particularly UNetFormer, can achieve higher IoU on this dataset; however, SilvaFormer remains substantially more compact in parameter count and competitive in segmentation quality. These findings position SilvaFormer as a compact transformer-based alternative for forest-health segmentation while highlighting the need for continued benchmarking against emerging transformer and Mamba-based architectures. Future work will extend the study toward instance-level infected-tree detection, larger multi-site datasets, and deployment-oriented model compression.

## DECLARATION OF COMPETING INTEREST

The authors declare no competing interests.

## DATA AVAILABILITY

The data used in this study are publicly available from the open pine wilt disease dataset and related project repository released by Xia et al. The implementation resources and trained models are available at the project GitHub repository. In this work, we used the publicly available subset containing 200 labeled RGB images derived from that source dataset.

## REFERENCES

[1] Y. Mamiya, "Pathology of the pine wilt disease caused by *Bursaphelenchus xylophilus*," *Annu. Rev. Phytopathol.*, vol. 21, no. 1, pp. 201–220, Sep. 1983.

[2] D. N. Proença, G. Grass, and P. V. Morais, "Understanding pine wilt disease: A comprehensive overview," *MicrobiologyOpen*, vol. 6, no. 1, 2017, Art. no. e00415.

[3] A. Hirata, K. Nakamura, K. Nakao, Y. Kominami, N. Tanaka, H. Ohashi, K. T. Takano, W. Takeuchi, and T. Matsui, "Potential distribution of pine wilt disease under future climate change scenarios," *PLoS ONE*, vol. 12, no. 8, Aug. 2017, Art. no. e0182837.

[4] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[5] I. Colomina and P. Molina, "Unmanned aerial systems for photogrammetry and remote sensing: A review," *ISPRS J. Photogramm. Remote Sens.*, vol. 92, pp. 79–97, Jun. 2014.

[6] M. D. Iordache, V. M. Mantas, E. Baltazar, K. Pauly, and N. Lewyckyj, "A machine learning approach to detecting insect damaged trees from UAV imagery," *Remote Sens.*, vol. 12, no. 14, p. 2280, 2020.

[7] S. Ecke, J. Dempewolf, J. Frey, A. Schwaller, E. Endres, H.-J. Klemmt, D. Tiede, and T. Seifert, "UAV-based forest health monitoring: A systematic review," *Remote Sens.*, vol. 14, no. 13, p. 3205, Jul. 2022.

[8] A. Manase, A. Manyevere, M. A. M. A. Elbasit, and C. Mashamaite, "The use of UAV-based systems in monitoring forest health: Potentials and challenges," *Sci. Afr.*, vol. 28, Jun. 2025, Art. no. e02724.

[9] J. Amoah-Nuamah, B. Child, E. Y. Okyere, O. Adams, and J. A. Danquah, "Applications of artificial intelligence in forest health surveillance and management," *Discover Forests*, vol. 1, no. 1, p. 61, Dec. 2025.

[10] M. Syifa, S.-J. Park, and C.-W. Lee, "Detection of the pine wilt disease tree candidates for drone remote sensing using artificial intelligence techniques," *Engineering*, vol. 6, no. 8, pp. 919–926, Aug. 2020.

[11] Z. Sun, Y. Wang, L. Pan, Y. Xie, B. Zhang, R. Liang, and Y. Sun, "Pine wilt disease detection in high-resolution UAV images using object-oriented classification," *J. Forestry Res.*, vol. 33, no. 4, pp. 1377–1389, Aug. 2022.

[12] L. Xia, R. Zhang, L. Chen, L. Li, T. Yi, Y. Wen, C. Ding, and C. Xie, "Evaluation of deep learning segmentation models for detection of pine wilt disease in unmanned aerial vehicle images," *Remote Sens.*, vol. 13, no. 18, p. 3594, Sep. 2021.

[13] Z. Wang, X. Su, X. Li, M. Cai, and X. Liao, "Mapping large-scale pine wilt disease trees with a first open-sourced PWD segmentation dataset based on UAV images," *Int. J. Remote Sens.*, vol. 45, no. 8, pp. 2786–2808, 2024.

[14] X. Yuan, G. Zhou, Y. Yan, and X. Yan, "Detection of pine wilt disease in UAV remote sensing images based on SLMW-net," *Plants*, vol. 14, no. 16, p. 2490, Aug. 2025.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. 9th Int. Conf. Learn. Represent. (ICLR)*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[16] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. T. Xiang, P. H. S. Torr, and L. Zhang, "Rethinking semantic segmen from a sequence-to-sequence perspective with transformers," 2021, *arXiv:2012.15840*.

[17] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002, doi: 10.1109/ICCV48922.2021.00986.

[18] 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 *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2021.

[19] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, and P. M. Atkinson, "UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 190, pp. 196–214, Aug. 2022.

[20] H. Ding, B. Xia, W. Liu, Z. Zhang, J. Zhang, X. Wang, and S. Xu, "A novel mamba architecture with a semantic transformer for efficient real-time remote sensing semantic segmentation," *Remote Sens.*, vol. 16, no. 14, p. 2620, Jul. 2024.

[21] Q. Zhu, Y. Cai, Y. Fang, Y. Yang, C. Chen, L. Fan, and A. Nguyen, "Samba: Semantic segmentation of remotely sensed images with state space model," *Heliyon*, vol. 10, no. 19, Oct. 2024, Art. no. e38495.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2015, pp. 234–241.

[23] 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. 833–851.
[24] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.
[25] S. Kong, J. Deng, L. Yang, and Y. Liu, "An attention-based dual-encoding network for fire flame detection using optical remote sensing," *Eng. Appl. Artif. Intell.*, vol. 127, Jan. 2024, Art. no. 107238.

**SHAFIULLAH SOOMRO** received the Ph.D. degree. He is currently a Postdoctoral Fellow with Linnaeus University, Växjö, Sweden. He has over a decade of experience in teaching and research in artificial intelligence, machine learning, computer vision, and software engineering. His research interests include deep learning for remote sensing and forestry applications, medical image analysis, semantic segmentation, and explainable AI. He has authored numerous peer-reviewed publications and serves as a reviewer for leading journals and conferences in AI and image processing.

**RYTIS MASKELIŪNAS** is currently a Professor with the Faculty of Informatics, Kaunas University of Technology. He does research in next generation collaborative intelligence and multimodal human–machine signal processing. He is the co-author of more than 200 research papers on the topic and has participated in more than 20 research themes and projects related. He has supervised six Ph.D. students working in this topic.

**ROBERTAS DAMAŠEVIČIUS** received the Ph.D. degree in informatics engineering from Kaunas University of Technology, Lithuania, in 2005. He is currently a Professor with the Department of Applied Informatics, Vytautas Magnus University, and a Key Researcher at the CoE Forest 4.0. His research interests include deep learning, human–computer interface, and the Internet of Things (IoT). He has authored over 500 scientific papers. He serves as the Editor-in-Chief of the journal *Information Technology and Control*.

**ARIANIT KURTI** (Member, IEEE) is currently a Professor in informatics at Linnaeus University. He has over 20 years of international academic experience, he has worked at universities and research institutions in Sweden, Kosovo, and North Macedonia. In addition to his academic roles, he has four years of research leadership experience as the Studio Director and a Senior Researcher at the RISE Research Institutes of Sweden, Norrköping.