Attention-guided Multi-Scale Deep Learning Approach for Tree Health Detection Using Very High-Resolution Aerial Imagery

Khatereh Meshkini[1], Mirela Beloiu[1], Zhongyu Xia[1], Verena C. Griess[1]

[1] Department of Environmental Systems Science, Institute of Terrestrial Ecosystems, ETH Zurich, 8092 Zurich, Switzerland
khatereh.meshkini@usys.ethz.ch, mirela.beloiu@usys.ethz.ch, zhongyu.xia@usys.ethz.ch, verena.griess@usys.ethz.ch

Keywords: Tree Health Monitoring, Deep Learning, Forest Management, Attention Mechanism, Remote Sensing

Abstract

Monitoring tree health is essential for detecting early signs of stress, defoliation, and potential mortality, supporting effective forest management, ecosystem conservation, and early warning systems. Advances in deep learning have enabled automated analysis of trees in remote sensing imagery through object detection methods that leverage both spectral and spatial information. However, assessing tree defoliation remains challenging, as subtle differences between defoliation levels make accurate classification difficult. To address this, we propose the hybrid ResNet-Swin Transformer, an object detection architecture built on a Faster R-CNN framework, incorporating a fused ResNet and Swin Transformer backbone with attention-based feature fusion. This design captures rich, multiscale representations by combining convolutional and transformer-based features and progressively refines them through channel-wise attention blocks for robust detection and classification. The architecture was evaluated on a very high-resolution aerial dataset from Switzerland, partially annotated with five classes: Conifer (healthy), Conifer (defoliated), Broadleaf (healthy), Broadleaf (defoliated) and Dead. Comparative experiments with state-of-the-art object detection and classification methods demonstrate that the proposed approach achieves higher accuracy and robustness, highlighting its potential for precise and reliable automated tree health monitoring.

1. Introduction

Monitoring forest health is increasingly important as climate-driven disturbances, such as droughts, pests, and wildfires, put individual trees and entire ecosystems at risk (Gazol et al., 2025). Although large-scale forest inventories such as the International Cooperative Program on Assessment and Monitoring of Air Pollution Effects on Forests (ICP Forests) provide valuable information, they are often too sparse and infrequent to capture early signs of tree stress or localized mortality events (Michel et al., 2020). Remote sensing offers a practical solution that allows for an automated assessment of forest vitality in large areas (Xulu et al., 2018). Recent advances in Earth observation have expanded the range of satellite sensors available for forest monitoring. Multispectral, hyperspectral, thermal, radar, and LiDAR observations now enable improved detection of forest disturbances, vegetation traits, and ecosystem dynamics across spatial scales (Holzwarth et al., 2020). These satellite sensors can detect deforestation, forest degradation, insect outbreaks, and fire events. Various studies have demonstrated the usage of medium- and coarse-resolution imagery, such as Landsat, Sentinel-2, and MODIS, to track changes in forest cover and disturbance regimes over time (Gao et al., 2019; Li et al., 2024; Molnár and Király, 2024; Meshkini and Bovolo, 2025). However, while these datasets provide broad-scale information, detecting subtle or early-stage tree-level defoliation requires much higher resolution images.

Very high-resolution aerial (cm scale, e.g. 10 cm) and satellite data ($\leq$ 10 m), including Unmanned Aerial Vehicles (UAV) imagery and WorldView or Pleiades satellites, have been used to study individual tree crowns, canopy segmentation, and fine-scale forest structure (Waser et al., 2014; Akbari and Kalbi, 2017). A study investigates the use of bitemporal WorldView-2 and WorldView-3 satellite images to map the dominant urban tree species in Beijing, China (Li et al., 2015). The authors apply object-based Support Vector Machine (SVM) and Random Forest (RF) classifiers under different temporal and spectral configurations to evaluate their potential for accurate identification of urban tree species. In another work, a Local Maximum (RLM) algorithm was developed to automatically detect individual tree crowns using high-resolution satellite imagery, replacing moving-window methods with directional transect searches and adaptive windowing (Xu et al., 2021). The research showed the potential of very high spatial resolution data for a detailed analysis of forest structure. Although these approaches effectively analyze individual trees even in dense and overlapping canopies, they mainly rely on classical image analysis techniques, such as object-based image analysis or vegetation indices, without leveraging the representational power of deep learning.

The ability of deep learning to automatically extract hierarchical and abstract features has made it widely used in various fields such as medical imaging, agriculture, and urban analysis, where detailed spatial understanding is required (Elizar et al., 2022). Similarly, in the forest, these capabilities allow deep learning models to perform a fine-grained analysis of the tree and forest structure (GRASS Development Team, 2017). Deep learning has the power to learn complex spatial and spectral patterns directly from high-resolution images, effectively capturing the rich spatial arrangements and spectral characteristics of vegetation. It can learn different variations in texture, color, and structural patterns that differentiate tree species, canopy densities, and general forests (Zhong et al., 2024). A lightweight convolutional neural network (CNN) applied to high-resolution UAV RGB images to accurately classify tree species under varying illumination and phenological conditions (Egli and Höpke, 2020). The method takes advantage of spatial and structural information of high-resolution images to achieve strong classification performance, even with a minimal ground sampling density. In Yue et al. (2019), a deep learning framework was developed that integrates

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-1-2026-303-2026

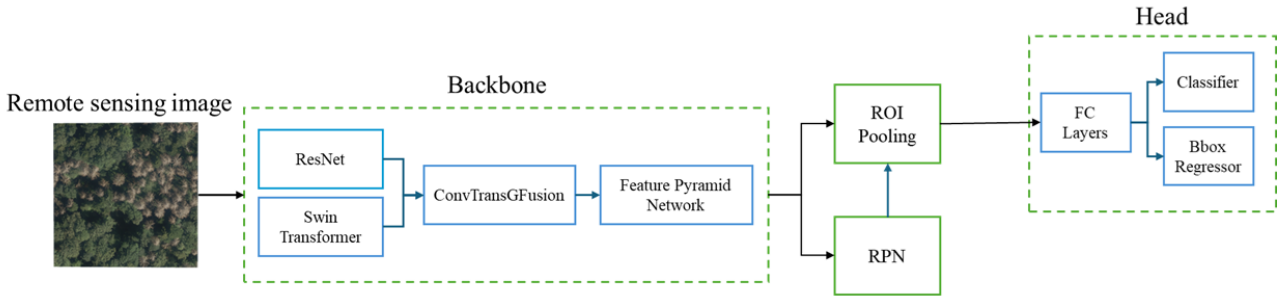

Figure 1. General block scheme of the proposed approach.

adaptive Tree-CNN blocks and multiscale feature fusion. The framework has been shown to improve pixel-level classification and better differentiate easily confused classes in subdecimeter aerial imagery. Region-based Convolutional Neural Networks (R-CNNs) that have been widely used in several studies (Xie et al., 2021; Hmidani and Alaoui, 2022) are a class of deep learning models designed for object detection. These deep learning models first propose candidate regions (region proposals) in an image, and then classify each region into object categories while refining the bounding-box coordinates. R-CNN and its variants have been applied to tree and forest analysis for tasks such as individual tree detection, crown delineation, and species classification, showing the ability to handle complex and overlapping canopies in high-resolution imagery (Yue et al., 2019). In this context, a mask version of R-CNN combined with UAV imagery has been effectively applied for instance segmentation at the level of olive tree crown detection. It provided an accurate estimate of the biovolume of individual trees in different spectral bands and vegetation indices (Safonova et al., 2021). The deep learning methods used for tree-level object detection have a strong potential to analyze tree health and defoliation conditions, an application that remains relatively less explored in the current literature.

As forests face increasingly frequent and severe disturbances, there is a growing need to use very high-resolution images to automatically extract subtle patterns of stress, canopy defoliation, and mortality at the individual tree level. Deep learning approaches can significantly improve the timeliness, accuracy, and scalability of forest health assessments, complementing traditional inventory methods and medium-resolution satellite monitoring. Recent research has demonstrated the potential of deep learning, particularly CNNs, to localize tree stress and defoliation patterns from aerial or UAV imagery (Anwander et al., 2024). Ecke et al. (2024) presents a standardized pipeline that integrates ground-based and UAV data with deep learning to classify tree species and crown conditions. The pipeline demonstrates the potential of deep learning-based monitoring approaches to streamline large-scale forest assessments. Beloiu et al. (2024) developed an automated approach to monitor tree defoliation and mortality in Germany and Switzerland using deep learning applied to freely available aerial imagery. Their CNN-based model, trained on a large dataset of annotated tree crowns, demonstrated that RGB-based networks can effectively detect and classify tree health conditions. They provided a scalable and accurate alternative to traditional field surveys for the identification of tree health. However, both works focus on presegmented individual trees, which may limit their performance when applied to full, unsegmented forest images. Object detection frameworks such as YOLO (Zhang et al., 2024) and Faster R-CNN (Hou et al., 2023) can also be used to detect individual trees and estimate the severity of defoliation, as they are able to extract more representative features. However, current backbones still need to be adapted specifically for the challenges of tree defoliation detection, especially under conditions of imbalanced or partially labeled data.

To effectively use the strengths of the Faster R-CNN object detection for tree defoliation scenarios, we proposed our deep learning architecture, the hybrid ResNet-Swin Transformer, which aims to improve tree detection and defoliation classification by combining the complementary strengths of CNNs and transformers within an object detection framework. The idea is to understand how the network structure for feature extraction can lead to improvements in object detection accuracy for tree health analysis, even though the dataset itself is not perfect. By combining a ResNet backbone with a Swin Transformer, the hybrid ResNet-Swin Transformer captures both local spatial patterns and long-range contextual relationships in high-resolution aerial imagery. ResNet excels at extracting fine-grained, low- and midlevel features such as leaf texture, crown edges, and small structural differences, while Swin Transformer provides robust multiscale representations and global dependencies that help distinguish subtle variations between healthy and defoliated trees. In addition, channel attention blocks are integrated to selectively emphasize the most informative feature channels, suppress irrelevant background signals, and improve the networks ability to focus on defoliation signals within each tree crown. This combination allows the hybrid ResNet-Swin Transformer to more accurately detect individual trees even in dense canopies and classify them into detailed health categories, including defoliated classes, outperforming traditional CNN-only or transformer-only detection pipelines. Using convolutional and transformer features with attention-guided refinement, the model achieves more robust discriminative representations that are critical for subtle tree health assessment. We also evaluated the model using both RGB and RGB–NIR images (including the NIR infrared band) to determine the contribution of the infrared channel to tree detection performance. Furthermore, the performance of our hybrid ResNet-Swin Transformer was compared with other object detection approaches, including YOLO and a Faster R-CNN variant with an EfficientNet backbone, which is widely used in the research community.

2. Proposed Approach

The proposed architecture builds upon the standard Faster R-CNN pipeline that, based on recent studies, remains a practical choice for accurate and robust detection across diverse tasks (Edozie et al., 2025). Faster R-CNN which operates in sequential stages for object detection, efficiently combines both region proposal and classification results for object localization. The

backbone network extracts hierarchical features from the input aerial imagery using a ResNet, generating spatially rich feature maps. These features are then processed by the Region Proposal Network (RPN), which identifies potential object regions through anchor boxes of multiple scales and aspect ratios. To enhance the discriminative power of the backbone, the proposed model replaces the standard ResNet backbone with a hybrid ResNet–Swin Transformer, as illustrated in Figure 1. In this design, multi-level convolutional features from ResNet and hierarchical self-attention features from Swin Transformer are projected into a shared embedding space and fused using learnable weighting parameters. A channel attention module further refines the fused representations by emphasizing informative feature channels and suppressing less relevant responses, ensuring that defoliation signs and fine-scale canopy textures are effectively captured. The resulting feature maps are then fed into the Feature Pyramid Network (FPN) and RPN stages for a robust detection and classification of individual tree crowns across diverse spatial scales.

2.1 Faster R-CNN

Faster R-CNN is a state-of-the-art deep learning framework designed for real-time object detection, capable of identifying both the location and category of objects within an image. The architecture operates through a sequence of interconnected stages (Girshick, 2015). First, the backbone network—typically a ResNet—extracts hierarchical visual features from the input image, forming a feature map. As shown in Figure 2,the backbone follows a FPN structure with five convolutional stages (Conv1–Conv5) forming the bottom-up pathway that extracts increasingly abstract features (C2–C5). Each feature map passes through a 1×1 convolution for dimensional alignment and participates in a top-down fusion process, where higher-level maps are successively upsampled and merged with lower-level ones through lateral connections. The fused outputs are refined by 3×3 convolutions to produce multiscale feature maps P2-P5, while P6 is generated by downsampling P5. This hierarchical design enables the network to capture both fine-grained and high-level information. Next, the RPN scans this feature map to generate candidate object regions using anchor boxes of varying scales and aspect ratios. Each proposed region is then processed through ROI Pooling (or ROI Align) to produce fixed-size feature representations, which are passed to the detection head. Finally, the head network classifies each region and refines its bounding box coordinates via fully connected layers responsible for object classification and bounding box regression.

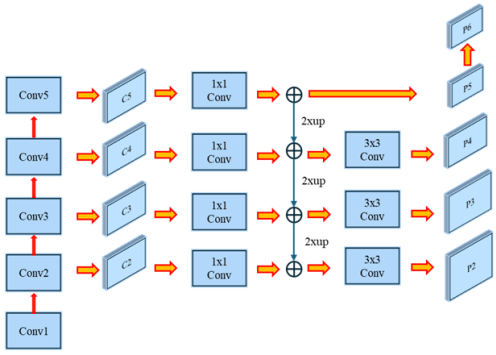

Figure 2. Faster R-CNN Backbone Architecture.

2.2 Swin transformer

The Swin Transformer is a hierarchical vision transformer designed for efficient image representation (Han et al., 2020). Unlike traditional transformers that compute attention globally, it divides the image into local windows and applies self-attention within each window, significantly reducing the computational cost (Liu et al., 2022). To enable information exchange between windows, the model shifts the position of the windows in alternating layers, allowing features of neighboring regions to interact without the need for heavy computation (see Figure 3).

Within each Swin Transformer block, self-attention based on the window and shifted window is combined with normalization, linear transformations, activation, and dropout to refine the features. Between blocks, neighboring patches are merged to form higher-level representations with larger receptive fields, capturing both fine-grained details and broader contextual information. This architecture is particularly effective for dense prediction tasks such as object detection and segmentation.

2.3 Hybrid ResNet–Swin Transformer

To enhance the feature representation capability of Faster R-CNN detector, we propose a hybrid backbone that fuses a Swin Transformer with a ResNet-50 network (see Figure 1). The Swin Transformer introduces hierarchical self-attention on image patches to capture long-range contextual dependencies. Specifically, given an input feature map $X \in \mathbb{R}^{C\times H\times W}$, window-based multihead self-attention (W-MSA) computes:

$$\text{Attention}(Q,K,V) = \text{Softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \tag{1}$$

where $Q, K, V \in \mathbb{R}^{N\times d_k}$ are the query, key and value matrices derived from X, and d_k is the dimension of each attention head. The shifted window mechanism (SW-MSA) further allows cross-window interactions without incurring the quadratic complexity of full global attention. This property is particularly beneficial for high-resolution forest imagery where objects appear at multiple scales.

The ResNet branch complements the Swin Transformer by providing strong local convolutional feature extraction, which is effective for detecting small or texture-rich tree structures. To combine these two modalities, we introduce a ConvTransG fusion module, which aligns the spatial dimensions of the Swin and ResNet feature maps, projects them into a shared embedding space, and applies channel-wise attention to emphasize informative features.

$$F_{\text{fused}} = \text{BN}\Big(\text{CA}\big(\alpha F_{\text{swin}} + \beta F_{\text{res}}\big)\Big) \tag{2}$$

where α and β are learnable scaling factors, F_{swin} and F_{res} are projected feature maps, CA denotes channel attention computed by global average pooling followed by bottleneck MLP and sigmoid activation, and BN represents batch normalization. Finally, the fused multiscale features are processed through an FPN to generate a set of enriched, multiresolution feature maps for the Faster R-CNN detector. This hierarchical fusion leverages both the long-range contextual modeling of the Swin Transformer and the local feature extraction of ResNet, improving detection performance on small and rare tree classes in high-resolution imagery.

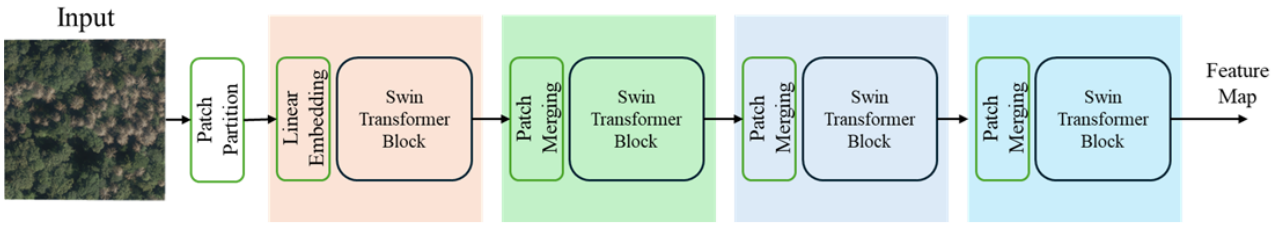

Figure 3. Swin Transformer Architecture.

3. Materials

3.1 Study Area

The study area covers forest regions across Switzerland, encompassing a broad ecological and altitudinal gradient from mixed deciduous forests in the colline and montane zones to conifer-dominated stands in the subalpine belt. The dominant tree species include Norway spruce *(Picea abies),* European beech *(Fagus sylvatica), and* silver fir *(Abies alba),* with smaller proportions of *Quercus* spp., *Acer* spp., *Larix decidua,* and *Pinus sylvestris*. Forests occupy roughly one third of the national territory. The Swiss climate is temperate with warm summers and no distinct dry season. In 2018 and again in 2022, Switzerland experienced severe summer droughts that caused widespread canopy stress and elevated tree mortality, particularly at low- and mid-elevation sites. These events led to extensive decline in Norway spruce and increased crown defoliation in European beech, underscoring the growing vulnerability of Swiss forests to recurring drought stress.

3.2 Dataset

Two complementary datasets were used in this study: (a) aerial orthophotos from the Swiss national mapping program and (b) ground-based forest condition observations from national monitoring networks. Together, they provide a spatially explicit reference for modeling tree defoliation and mortality at the crown level.

Aerial imagery was obtained from the Swiss Federal Office of Topography (Swisstopo) using the Leica ADS line sensor, which captures four spectral bands (red, green, blue, and near-infrared). The spatial resolution ranges from 10 cm in lowland and valley areas to 25 cm in mountainous terrain. All selected images were acquired under leaf-on conditions (May–September, 2017–2023) to ensure optimal crown visibility, and to exclude natural autumn senescence, which occurs in autumn.

Field reference data originated from the Swiss Long-term Forest Ecosystem Research (LWF) the Swiss program part of ICP Forests networks. Each tree was visually assessed for defoliation in 5% increments following standardized ICP Forests protocols (Michel et al., 2019), with accompanying information on species, diameter at breast height (DBH), social status, and geographic coordinates. Tree crowns were spatially linked to field observations through GPS-based crown centroids and buffers proportional to crown radius; when required, crowns were manually delineated in the aerial imagery to ensure alignment.

Most tree crowns were manually delineated in the aerial imagery based on GPS-referenced tree locations from national forest inventories to ensure precise spatial correspondence between field data and image annotations. A buffer-based method was subsequently applied for a subset of trees with well-defined crown geometries, mainly conifers, to streamline the delineation process. Defoliation levels were classified into three categories: healthy (0–25% defoliation), defoliated (30–95%), and dead (99–100%). After data harmonization and quality control, the Swiss dataset comprised approximately 29,065 annotated tree crowns. Table 1 shows the number of annotated trees per defoliation classes and Figure 4 presents the distribution of annotations in Switzerland.

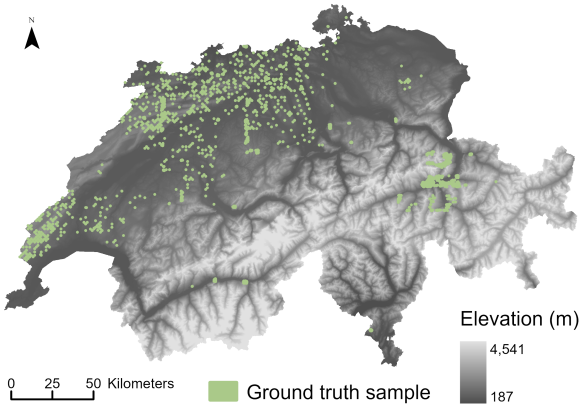

Figure 4. The distribution of tree crown annotations in Switzerland.

Defoliation class	Healthy 0–25%	Defoliated 30–95%	Dead 99–100%	Total
Conifer	8 178	903		9 081
Broadleaf	4 983	744		5 727
Dead tree			15 501	15 501
Total	**13 161**	**1 647**	**15 501**	**29 065**

Table 1. Number of annotated trees per defoliation class.

4. Experimental Results

4.1 Model Training

Image chips of 320 × 320 pixels were generated using ArcGIS Pro, converted from 16-bit to 8-bit RGB, and preprocessed with a 0.5 m masking buffer and border cleaning to avoid edge artifacts. The dataset was divided into 70 % training, 20 % validation, and 10 % testing subsets using spatial block cross-validation (102.4 m grid) to ensure independence between neighboring crowns and prevent overestimation of model performance. In this study, multiple backbone architectures were integrated into the Faster R-CNN detection framework to evaluate their effectiveness in tree species and defoliation classification. Specifically, the standard ResNet backbone was first employed as the baseline model.

For comparative evaluation, EfficientNet (Tan and Le, 2019) and Swin Transformer backbones were subsequently tested under the same detection framework. Additionally, YOLOv11, one of the most frequently used methods for object detection, was included using standard YOLO parameters to provide a benchmark for comparison with our results (Jocher and Qiu, 2024). Finally, the proposed hybrid ResNet-Swin model was trained and evaluated in both RGB and RGBI to demonstrate its advantages in multiscale feature representation and detection accuracy. The detailed training configuration and experimental setup are summarized in Table 2.

Parameter	Value
Optimizer	AdamW
Epochs	100
Initial LR	1e-4
Weight Decay	1e-4
LR Scheduler	ReduceLROnPlateau (Al-Kababji et al., 2022)
Patience	5
LR Factor	0.5
Batch Size	8
Hardware	2× RTX 2080 Ti (11GB)
CUDA	12.8

Table 2. Training configuration

Since the defoliation class contained a substantially smaller number of samples compared to other classes, weighted sampling was applied during training. This technique ensures that underrepresented classes contribute proportionally to the loss computation. Moreover, it prevents the model from developing bias toward majority classes and improving generalization across all categories.

4.2 Evaluation Metrics

To assess and compare the performance of the models, metrics such as precision, recall, F1 score, and mean average precision (mAP) were used. True positive (TP), true negative (TN), false positive (FP) and false negative (FN) values were calculated for each image to calculate precision and recall. These calculations were subsequently used to derive the F1-score as the harmonic mean of precision and recall. The average precision (AP) and mean average precision (mAP), which served as the principal metrics for evaluating detection performance across all classes.In our evaluation, we report mAP@50, which considers a detection correct if IoU is greater than 0.5. These metrics enabled a comprehensive and quantitative analysis of the detection capacity of the models.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

$$IoU = \frac{\text{Area of Overlap}}{\text{Area of Union}} \tag{6}$$

$$AP = \int_0^1 p(r)\, dr \tag{7}$$

$$mAP = \frac{1}{C} \sum_{i=1}^{C} AP_i \tag{8}$$

In Equation (6), the IoU quantifies the overlap ratio between the predicted and ground-truth bounding boxes, and it is used as a threshold criterion to define true positives. In Equation (7), $p(r)$ represents the maximum precision at a given recall level, while C in Equation (8) denotes the total number of object classes.

4.3 Quantitative analysis

Table 3 provides a detailed overview of how different backbone architectures perform on tree species and defoliation classification. As illustrated in the table, the models were evaluated after training on different backbones, YOLO, and the proposed approach, across multiple classes, using metrics including Precision, Recall, F1-Score, and mAP@50, along with model complexity indicators such as the number of parameters and FLOPs. It is important to note that the number of samples per class is limited, especially for defoliated trees, which introduces inherent variability in the evaluation metrics. Additionally, the dataset is only partially labeled, which means that not all trees in the images have annotations. Therefore, these results should be interpreted as indicative of the models capacity to improve feature representation rather than as definitive performance benchmarks.

All metrics were computed after excluding predictions that did not overlap with any ground truth (GT) bounding box. This adjustment was necessary because the dataset is only partially annotated, meaning that some objects in the images lack corresponding GT labels. Including predictions for such unlabeled objects would artificially increase the number of false positives, thus lowering the precision and mAP measured and leading to an inaccurate evaluation of the true performance of the model. The remaining bounding boxes were further processed using Non-Maximum Suppression (NMS) (Hosang et al., 2017). It is a standard post-processing step in object detection that removes redundant bounding boxes corresponding to the same object. It works by retaining only the box with the highest confidence score among a set of highly overlapping detections and suppressing the rest.

From the table, it is evident that the hybrid Resnet-Swin (RGB) backbone delivers consistently strong performance across multiple classes. For Conifer (Healthy), Broadleaf (Defoliated), and Dead trees, it achieves the highest F1 score of 0.6010, 0.6320 and 0.7930, respectively. It also demonstrates competitive F1 score performance for Conifer (Defoliated) with an F1 score of 0.5050 and Broadleaf (Healthy) with 0.5420, highlighting its ability to handle the detection of both defoliated and healthy trees. In terms of mAP_{50}, the hybrid Resnet-Swin (RGB) outperforms other backbones for Broadleaf (Defoliated) and Dead trees. However, Yolo achives higher mAP_{50} for healthy trees, indicating that its fully convolutional architecture preserves the leaf-level texture information of healthy canopies. The average result for all classes indicates that the hybrid architecture

Backbone / Class	Precision	Recall	F1	mAP_{50}	Parameters	FLOPs
YOLOv11						
Conifer (Healthy)	**0.74**	0.51	**0.60**	**0.63**	47.50 M	105.00 GMac
Conifer (Defoliated)	0.31	0.27	0.29	0.26		
Broadleaf (Healthy)	**0.75**	0.41	0.52	**0.60**		
Broadleaf (Defoliated)	0.31	0.44	0.36	0.29		
Dead	0.72	**0.86**	0.78	**0.86**		
All	0.49	0.46	0.47	0.45		
EfficientNet-B2						
Conifer (Healthy)	0.60	0.55	0.57	0.43	9.10 M	32.30 GMac
Conifer (Defoliated)	0.85	0.33	0.48	0.33		
Broadleaf (Healthy)	0.57	0.51	0.54	0.38		
Broadleaf (Defoliated)	0.57	0.40	0.47	0.33		
Dead	0.72	0.78	0.75	0.71		
All	0.66	0.51	0.56	0.44		
ResNet-50						
Conifer (Healthy)	0.70	0.52	**0.60**	0.41	41.10 M	134.30 GMac
Conifer (Defoliated)	0.85	**0.42**	**0.57**	**0.41**		
Broadleaf (Healthy)	0.56	0.52	0.54	0.37		
Broadleaf (Defoliated)	0.72	0.51	0.59	0.45		
Dead	**0.78**	0.78	0.78	0.73		
All	0.72	0.55	**0.62**	0.47		
Swin Transformer						
Conifer (Healthy)	0.66	0.42	0.51	0.34	44.77 M	33.93 GMac
Conifer (Defoliated)	**1.00**	0.27	0.43	0.27		
Broadleaf (Healthy)	0.53	**0.56**	0.54	0.44		
Broadleaf (Defoliated)	0.62	0.52	0.56	0.42		
Dead	0.74	0.76	0.75	0.70		
All	0.71	0.51	0.56	0.43		
Hybrid Resnet-Swin (RGB)						
Conifer (Healthy)	0.61	**0.59**	**0.60**	0.48	56.47 M	82.73 GMac
Conifer (Defoliated)	0.83	0.36	0.51	0.35		
Broadleaf (Healthy)	0.58	0.51	0.54	0.38		
Broadleaf (Defoliated)	0.81	**0.52**	**0.63**	**0.47**		
Dead	**0.78**	0.80	**0.79**	0.74		
All	0.72	**0.56**	**0.62**	**0.48**		
Hybrid Resnet-Swin (RGBI)						
Conifer (Healthy)	0.66	0.39	0.49	0.31	56.95 M	83.10 GMac
Conifer (Defoliated)	0.75	0.09	0.16	0.08		
Broadleaf (Healthy)	0.63	0.36	0.46	0.28		
Broadleaf (Defoliated)	**0.92**	0.35	0.51	0.34		
Dead	0.77	0.75	0.76	0.69		
All	**0.74**	0.39	0.48	0.34		

Table 3. Per-class detection metrics for different backbones on tree species and defoliation classification. Bold values indicate the best method for each class and metric.

effectively uses local features from ResNet, global contextual information from the Swin Transformer, as well as multi-scale structural patterns to accurately detect both healthy and defoliated trees from different species.

EfficientNet-B2, despite having significantly fewer parameters (9.1 M) and FLOPs (32.3 GMac), achieves competitive F1 scores for Dead trees (0.7462) and Conifer (Defoliated) (0.4783), demonstrating that lightweight models can still extract meaningful features for these classes. Swin Transformer alone also performs well across several classes, showing self-attention’s advantage in capturing intricate canopy patterns. In addition, the hybrid ResNet-Swin RGBI model does not consistently improve performance. Its F1 score for Dead trees is 0.7576, which is slightly below the RGB-only variant, where the F1 score for Dead trees is 0.7930. Across all classes, the RGB-only model consistently achieves higher F1 scores than the RGBI variant, indicating that the inclusion of NIR provides limited additional discriminative power for crown-level defoliation detection in this dataset. Overall, these results emphasize that hybrid backbone architectures can incorporate complementary feature representations to improve detection, particularly in scenarios with limited labeled data and class imbalance.

4.4 Qualitative analysis

As part of our qualitative analysis, we selected the best-performing models—our proposed hybrid ResNet–Swin architecture, the Swin Transformer, ResNet, and EfficientNet—to visualize their performance on the test set, as illustrated in

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-1-2026-303-2026 | © Author(s) 2026. CC BY 4.0 License.

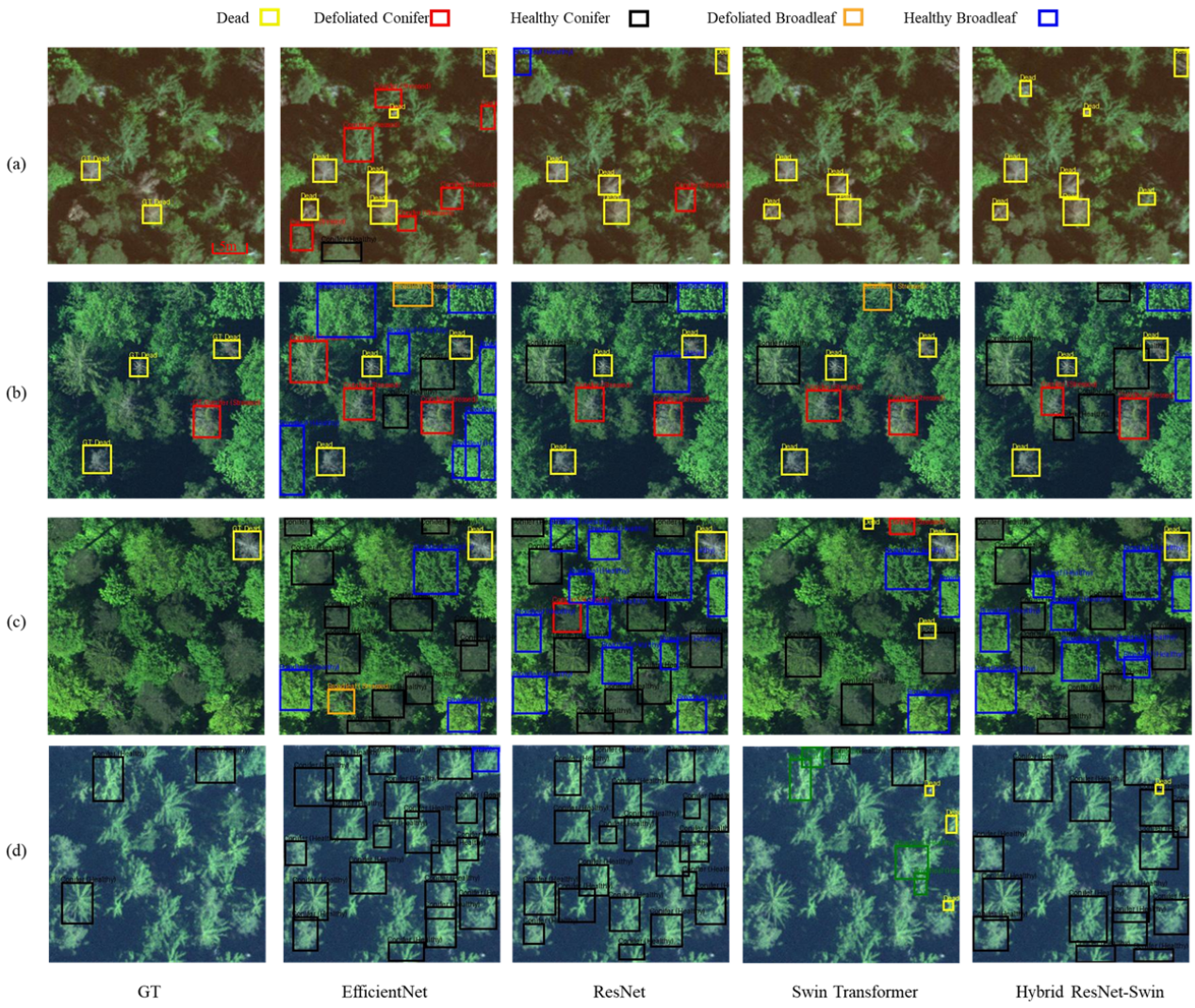

Figure 5. Sample prediction images for the different backbone models, focusing primarily on defoliated and dead tree areas to highlight the models effectiveness in detecting defoliated trees. Images (a) and (d) are from 2020, while (b) and (c) are from 2021.

Fig. 5. In the qualitative analysis, all predictions were retained for visualization of sample test images. The results show how effectively the model detects not only the GT-labeled trees but also additional bounding boxes corresponding to trees from various classes that were not annotated in the dataset.

Taking into account Fig. 5, in sample (a), the Hybrid ResNet–Swin model detects dead trees more precisely, whereas the Swin Transformer tends to identify a larger number of dead trees. However, based on photointerpretation, the detections from the Hybrid ResNet–Swin appear more accurate, as the dead trees are consistently identified across all methods. In sample (b), the Hybrid ResNet–Swin model demonstrates more accurate detection of defoliated trees, consistently identifying the instances that are also detected by other models. In contrast, EfficientNet and the Swin Transformer tend to over-detect defoliated trees, with several predictions appearing inconsistent with the ground truth, suggesting lower reliability. In sample (c), while the ResNet backbone performs well in detecting almost all tree crowns in the scene, the Hybrid ResNet–Swin model similarly captures nearly all crowns, indicating that it inherits the strong local feature extraction capability of ResNet. In sample (d), which predominantly contains healthy trees according to the ground truth, the proposed Hybrid ResNet–Swin model correctly classifies the majority of trees as healthy, while also identifying a small dead tree present in the scene. This further demonstrates its ability to maintain high precision in both dominant and minority classes.

5. Conclusions

In this study, we introduced a hybrid deep learning architecture that integrates ResNet and Swin Transformer backbones within a Faster R-CNN framework. The proposed approach was evaluated using both RGB and RGBI very high resolution aerial imagery, together with more than 29,000 manually annotated tree crowns from Swiss forests. Across a series of comparative experiments against the YOLO, EfficientNet, ResNet, and Swin Transformer models, our hybrid model achieved competitive and consistently strong performance across evaluation metrics. The results confirm the advantage of combining convolutional and transformer-based feature extraction for tree-level health assessment. Despite the inherent imperfections of the ground-truth data, such as class imbalance and partial labeling, our proposed model reached an overall F1-score of 0.61 and 0.79 for the Dead class, indicating strong robustness and generalization. By enabling accurate, large-scale, and automated detection of individual tree defoliation and mortality from very-high-resolution

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-1-2026-303-2026 | © Author(s) 2026. CC BY 4.0 License.

imagery, this approach provides a scalable tool for operational forest monitoring. This capability represents an important step toward integrating fine-grained crown-level health condition detection for forest management and ecological research, particularly under increasing climate change and disturbance regimes.

References

Akbari, H., Kalbi, S., 2017. Determining Pleiades satellite data capability for tree diversity modeling. *Iforest - Biogeosciences and Forestry*, 10, 348-352. 10.3832/ifor1884-009.

Al-Kababji, A., Bensaali, F., Dakua, S., 2022. Scheduling Techniques for Liver Segmentation: ReduceLROnPlateau Vs OneCycleLR. 204-212. 10.1007/978-3-031-08277-1$_1$7.

Anwander, J., Brandmeier, M., Paczkowski, S., Neubert, T., Paczkowska, M., 2024. Evaluating Different Deep Learning Approaches for Tree Health Classification Using High-Resolution Multispectral UAV Data in the Black Forest, Harz Region, and Göttinger Forest. *Remote Sensing*, 16(3). https://www.mdpi.com/2072-4292/16/3/561. 10.3390/rs16030561.

Beloiu, M. S., Berger, E., Ecke, S., Xia, Z., Reichmuth, C., Gessler, A., Klemmt, H.-J., Glatthorn, J., Stillhard, J., Griess, V., Waser, L. T., 2024. Tree Crown Mortality and Defoliation Assessment in Temperate Forests Using Aerial Imagery and Deep Learning. *SSRN Electronic Journal*. https://ssrn.com/abstract=5360617. 10.2139/ssrn.5360617.

Ecke, S., Stehr, F., Frey, J., Tiede, D., Dempewolf, J., Klemmt, H.-J., Endres, E., Seifert, T., 2024. Towards operational UAV-based forest health monitoring: Species identification and crown condition assessment by means of deep learning. *Computers and Electronics in Agriculture*, 219, 108785. https://doi.org/10.1016/j.compag.2024.108785.

Edozie, E., Shuaibu, A. N., John, U. K., Sadiq, B. O., 2025. Comprehensive review of recent developments in visual object detection based on deep learning. *Artificial Intelligence Review*, 58(9), 277. 10.1007/s10462-025-11284-w.

Egli, S., Höpke, M., 2020. CNN-Based Tree Species Classification Using High Resolution RGB Image Data from Automated UAV Observations. *Remote Sensing*, 12(23). 10.3390/rs12233892.

Elizar, E., Zulkifley, M. A., Muharar, R., Zaman, M. H. M., Mustaza, S. M., 2022. A Review on Multiscale-Deep-Learning Applications. *Sensors (Basel, Switzerland)*, 22. 10.3390/s22197384.

Gao, Y., Quevedo, A., Szantoi, Z., Skutsch, M., 2019. Monitoring forest disturbance using time-series MODIS NDVI in Michoacán, Mexico. *Geocarto International*, 36, 1768 - 1784. 10.1080/10106049.2019.1661032.

Gazol, A., Pizarro, M., Hammond, W. M., Allen, C. D., Camarero, J. J., 2025. Droughts preceding tree mortality events have increased in duration and intensity, especially in dry biomes. *Nature Communications*, 16(1), 5779. https://doi.org/10.1038/s41467-025-60856-5.

Girshick, R., 2015. Fast R-CNN. 2015 IEEE International Conference on Computer Vision (ICCV). *Santiago*, 7–13.

GRASS Development Team, 2017. Geographic Resources Analysis Support System (GRASS) Software. Open Source Geospatial Foundation. grass.osgeo.org (20 September 2017).

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y., Tao, D., 2020. A Survey on Vision Transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP, 1-1. 10.1109/tpami.2022.3152247.

Hmidani, O., Alaoui, E. M. I., 2022. A comprehensive survey of the R-CNN family for object detection. *2022 5th International Conference on Advanced Communication Technologies and Networking (CommNet)*, 1-6. 10.1109/commnet56067.2022.9993862.

Holzwarth, S., Thonfeld, F., Abdullahi, S., Asam, S., Canova, E. D. P., Gessner, U., Huth, J., Kraus, T., Leutner, B. F., Kuenzer, C., 2020. Earth Observation Based Monitoring of Forests in Germany: A Review. *Remote. Sens.*, 12, 3570. 10.3390/rs12213570.

Hosang, J., Benenson, R., Schiele, B., 2017. Learning Non-maximum Suppression. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6469-6477. 10.1109/cvpr.2017.685.

Hou, J., Yang, C., He, Y.-D., Hou, B., 2023. Detecting diseases in apple tree leaves using FPN–ISResNet–Faster RCNN. *European Journal of Remote Sensing*, 56. 10.1080/22797254.2023.2186955.

Jocher, G., Qiu, J., 2024. Ultralytics yolo11.

Li, D., Ke, Y., Gong, H., Li, X., 2015. Object-Based Urban Tree Species Classification Using Bi-Temporal WorldView-2 and WorldView-3 Images. *Remote. Sens.*, 7, 16917-16937. 10.3390/rs71215861.

Li, Z., Ota, T., Mizoue, N., 2024. Monitoring tropical forest change using tree canopy cover time series obtained from Sentinel-1 and Sentinel-2 data. *International Journal of Digital Earth*, 17. 10.1080/17538947.2024.2312222.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2022. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9992-10002. 10.1109/iccv48922.2021.00986.

Meshkini, K., Bovolo, F., 2025. Transformer-based spatio-temporal change detection network using satellite image time series: A case study of forest disturbance in trentino, italy following the vaia storm. *IGARSS 2025 - 2025 IEEE International Geoscience and Remote Sensing Symposium*, 6985–6988. 10.1109/IGARSS55030.2025.11314053.

Michel, A., Prescher, A.-K., Schwärzel, K., 2019. Forest condition in europe: 2019 technical report of icp forests. Technical report, ICP Forests.

Michel, A., Prescher, A.-K., Schwärzel, K., 2020. Forest condition in europe: 2019 technical report of icp forests. report under the unece convention on long-range transboundary air pollution (air convention).

Molnár, T., Király, G., 2024. Forest Disturbance Monitoring Using Cloud-Based Sentinel-2 Satellite Imagery and Machine Learning. *Journal of Imaging*, 10. 10.3390/jimaging10010014.

Safonova, A., Guirado, E., Maglinets, Y., Alcaraz-Segura, D., Tabik, S., 2021. Olive Tree Biovolume from UAV Multi-Resolution Image Segmentation with Mask R-CNN. *Sensors*, 21(5). 10.3390/s21051617.

Tan, M., Le, Q., 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. K. Chaudhuri, R. Salakhutdinov (eds), *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 97, PMLR, 6105–6114.

Waser, L., Küchler, M., Jütte, K., Stampfer, T., 2014. Evaluating the Potential of WorldView-2 Data to Classify Tree Species and Different Levels of Ash Mortality. *Remote. Sens.*, 6, 4515-4545. 10.3390/rs6054515.

Xie, X., Cheng, G., Wang, J., Yao, X., Han, J., 2021. Oriented R-CNN for Object Detection. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 3500-3509. 10.1109/iccv48922.2021.00350.

Xu, X., jing Zhou, Z., Tang, Y., Qu, Y., 2021. Individual tree crown detection from high spatial resolution imagery using a revised local maximum filtering. *Remote Sensing of Environment*. 10.1016/j.rse.2021.112397.

Xulu, S., Gebreslasie, M., Peerbhay, K., 2018. Remote sensing of forest health and vitality: a South African perspective. *Southern Forests: a Journal of Forest Science*, 81, 102 - 91. 10.2989/20702620.2018.1512787.

Yue, K., Yang, L., Li, R., Hu, W., Zhang, F., Li, W., 2019. TreeUNet: Adaptive Tree convolutional neural networks for subdecimeter aerial image segmentation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 156, 1-13. https://doi.org/10.1016/j.isprsjprs.2019.07.007.

Zhang, Z., Han, C., Wang, X., Li, H., Li, J., Zeng, J., Sun, S., Wu, W., 2024. Large field-of-view pine wilt disease tree detection based on improved YOLO v4 model with UAV images. *Frontiers in Plant Science*, 15. 10.3389/fpls.2024.1381367.

Zhong, L., Dai, Z., Fang, P., Cao, Y., Wang, L., 2024. A Review: Tree Species Classification Based on Remote Sensing Data and Classic Deep Learning-Based Methods. *Forests*. 10.3390/f15050852.