Sensors and AI
https://www.sciltp.com/journals/sai

Article

A Dual-Channel Pine Wilt Disease Recognition Method with Discrete Wavelet Transform

Zimo Zhou and Simon X. Yang *

College of Engineering, University of Guelph, 50 Stone Rd E, Guelph, ON N1G 2W1, Canada
* Correspondence: syang@uoguelph.ca

How To Cite: Zhou, Z.; Yang, S.X. A Dual-Channel Pine Wilt Disease Recognition Method with Discrete Wavelet Transform. *Sensors and AI* **2025**, *1*(1), 30–44.

Received: 29 May 2025
Revised: 20 June 2025
Accepted: 15 July 2025
Published: 6 August 2025

Abstract: Pine wilt disease is a significant global plant epidemic and a management priority for numerous countries worldwide. Pine wood nematodes can parasitize a wide range of pine species, making early detection of infected trees essential for preventing further spread of the disease. Recent advances in deep learning and remote sensing technologies have enabled efficient automated detection of diseased trees. Most existing methods rely on convolutional neural network layers for feature extraction and spatial dimension reduction, which may cause the loss of fine-grained texture details and lead to misdetection of background elements and visually similar objects. To enhance diseased tree recognition accuracy, this paper proposes an object detection model using images captured by unmanned aerial vehicles. The proposed method incorporates discrete wavelet transform (DWT) to reduce spatial resolution while preserving critical information for further analysis, and integrates a cross-modal channel enhancement module within a two-stream feature extraction network. Furthermore, the method incorporates a RoI-based similarity constraint that applies cosine similarity loss and classification supervision to ensure coherent feature representations between processing branches. This approach achieves 89.2% accuracy on the pine wilt disease dataset and outperforms advanced methods on the VisDrone dataset. Several object detection models are compared based on the mean average precision (mAP) metric. Results demonstrate that the DWT-based detection algorithm achieves superior performance in detecting individual small targets and clustered infected pine trees.

Keywords: pine wilt disease; unmanned aerial vehicle; object detection; deep learning; image sensors; discrete wavelet transform

1. Introduction

Accurate detection and recognition of plant disease occurrences is essential for forest ecosystem management, as diseases pose significant threats to both ecological health and economic stability. Invasive pathogens cause particularly severe damage due to the lack of natural biological controls. Pine wilt disease (PWD) is one of the most destructive plant diseases and has spread rapidly across much of southern China since it was first detected in Nanjing in 1982 [1]. It not only causes massive economic losses amounting to billions of Chinese yuan [2], but also disrupts local pine forest ecosystems and reduces carbon sequestration capacity [3].

The pine wilt nematode (PWN), a plant parasitic nematode belonging to *Bursaphelenchus* genus, was first identified as a causal agent of pine forest deterioration in 1971, leading to widespread forest decline throughout southwestern Japan [4]. The following year, the sawyer beetle species *Monochamus alternatus* was recognized by Mamiya and Enda as the primary vector of PWN [5]. Multiple control strategies have been developed to manage PWN, each with distinct advantages and limitations. Chemical control, while effective, requires careful pesticide selection and adherence to application principles to minimize risks such as the unintended death of non-target insects, contamination of water resources, and ecosystem disruption [6]. Biological control utilizes natural enemies and microbial agents as environmentally friendly alternatives [7–9], although it requires long-term commitment to

develop suitable control agents. Physical and mechanical control methods, including tree removal, are simple to operate and cost-effective. Currently, using Unmanned Aerial Vehicles (UAVs) to capture remote sensing images and employing methods based on artificial intelligence to detect target trees represents an effective and important approach for early detection. Consequently, integrated pest management approaches that combine diverse control techniques with advanced detection technologies are increasingly favored due to their complementary advantages.

Machine learning and deep neural networks (DNNs) have demonstrated remarkable capabilities in computer vision tasks, particularly object detection and image classification. Unlike manually designed filters constrained by human knowledge, modern neural networks can automatically learn hierarchical feature representations [10], which has greatly promoted the development of other industries and provided possibilities for combining neural networks with various tasks such as robotic path planning, manufacturing defect detection, and forestry or agricultural applications [11–14]. The development of advanced DNN architectures [15–18] has significantly improved the processing of high-dimensional visual data, making them well suited for remote sensing applications.

Remote sensing techniques significantly enhance target detection and localization by utilizing high-precision sensors and automated global positioning system (GPS) navigation systems. Researchers have been using this powerful technology with post-processing procedures to address challenging recognition problems [19–21]. Conventional deep learning methods often struggle with small target detection in high-altitude images, where infected trees appear as small objects against complex forest backgrounds. Feature extraction from low-resolution imagery frequently results in the loss of critical texture information, leading to misclassifications between healthy trees and diseased trees. Additionally, existing object detection frameworks lack effective multi-scale feature fusion mechanisms to handle the varying sizes of individual and clustered infected trees. Background noise and visual similarity between diseased trees and other forest elements further complicate accurate detection. To address these issues, we proposed an object detection model with extra feature processing path leveraging wavelet transform.

The main contributions of this paper with the wavelet-based detection model can be summarized as follows: (1) We propose a novel dual-channel feature extraction architecture that integrates a conventional neural network pathway with a discrete wavelet transform channel, creating a complementary feature fusion mechanism that leverages both spatial and frequency-domain representations. (2) A Haar wavelet-based multilevel feature fusion module that systematically enhances cross-scale feature integration through hierarchical frequency-domain decomposition, enabling superior feature refinement across multiple resolution levels. (3) An object-level similarity loss mechanism is designed that combines cosine similarity constraints with classification supervision, ensuring that both feature extraction channels learn coherent and discriminative feature representations while improving model robustness, and reducing false detections in complex backgrounds. (4) A robust framework for the detection of pine wilt disease is constructed by integrating the proposed feature extraction and refinement modules, demonstrating superior detection precision relative to existing baseline models.

The structure of this paper is organized as follows: Section 2 reviews related works on traditional PWD detection, deep learning object detection models, and intelligent PWD recognition models. Section 3 presents the proposed pine wilt tree detection models, along with its submodules, in detail. In Section 4, results are presented comparing the proposed model and the state-of-the-art object detection models on the pine trees dataset and VisDrone dataset [22]. Section 5 provides detection examples and discusses the findings, while Section 6 concludes the paper.

2. Related Works

Significant efforts by researchers have facilitated the transition from conventional detection methods to DNNs-based detection approaches leveraging remote sensing techniques. This section surveys relevant work, including studies on PWD detection, object detection models, and their applications.

2.1. Conventional PWD Detection

Traditional PWD detection methods include resin secretion assessment, morphological analysis under microscopy, and molecular techniques such as PCR-based genetic identification [23–26]. Although highly accurate, these laboratory-based methods are time-consuming and impractical for large-scale forest monitoring.

2.2. Deep Learning Based Recognition

Since the introduction of LeNet [27] and AlexNet [28], convolutional neural networks (CNNs) have revolutionized computer vision tasks including object detection, multimodal image-text understanding, intelligent robotics, image segmentation, and pose estimation [29–32], significantly reducing detection time and enhancing recognition accuracy. CNNs leverage hierarchical feature extraction through convolutional layers and employ specialized heads

for various recognition tasks. Several key innovations have substantially enhanced neural network capabilities. Through gradient-based optimization, back-propagation [33] enables networks to learn features and update weights automatically. Batch normalization [34] accelerates training convergence and improves model robustness by addressing internal covariate shift. Perhaps most significantly, residual connections introduced by He et al. [35] revolutionized network architecture by creating shortcut identity mapping paths that mitigate gradient vanishing problems, thereby enabling the development of powerful architectures for complex visual recognition tasks.

Object detection has been a popular computer vision technique in recent years, involving the recognition and localization of objects in images or videos. It transcends basic image classification by providing not only the categories of objects but also their precise locations, represented as bounding boxes. Region-based Convolutional Neural Networks (RCNN) [36] and You Only Look Once (YOLO) [37] have been pivotal in the evolution of object detection algorithms, driving research toward greater accuracy, speed, and efficiency in target detection. In PWD detection, we employ standard metrics derived from the confusion matrix shown in Figure 1. The confusion matrix categorizes detection results into four types: True Positives (TP) represent correctly detected diseased trees, False Positives (FP) indicate healthy trees incorrectly classified as diseased, False Negatives (FN) denote diseased trees that were missed by the model, and True Negatives (TN) correspond to healthy trees correctly identified as non-diseased. The equations for Precision and Recall are described as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{2}$$

Precision and recall are two key metrics used to evaluate classification performance. Recall, also known as sensitivity, measures the model's ability to identify all relevant instances. In disease detection applications, high recall is particularly critical as it ensures that infected trees are not missed, preventing potential disease spread.

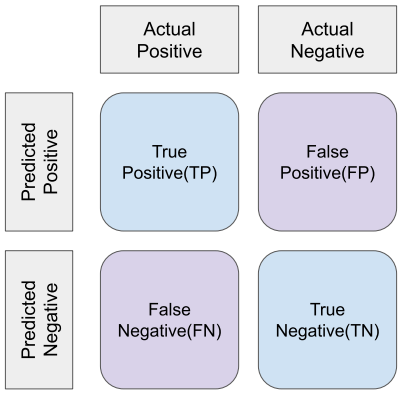


Figure 1. Confusion matrix for object detection evaluation.

Deep learning-based object detection approaches can be classified into anchor-based or anchor-free algorithms, depending on whether predefined anchor boxes are used to align bounding boxes with the extracted feature maps. Alternatively, the algorithm can be categorized based on whether the model includes a distinct stage for generating proposal regions, known as one-stage or two-stage methods. Examples of one-stage methods include the YOLO series [38–40] and Single Shot Multibox Detector (SSD), while two-stage methods include Fast R-CNN [41] and Faster R-CNN [42]. In contrast to these anchor-based approaches, anchor-free models directly predict boundaries without predefined anchors [43,44].

Beyond CNN-based architectures, researchers have explored integrating methodologies from Natural Language Processing (NLP). The self-attention mechanism and Transformer architecture [17,18] have driven evolution from task-specific language models to large language models. These attention-based approaches have begun to rival CNN models in computer vision applications. The Vision Transformer (ViT) is a pioneering model that applies the principles of the Transformer architecture to computer vision tasks, with a greater global context understanding and a larger receptive field [45]. An end-to-end object detection model with Transformer (DETR) was presented by Carion et al. [46], which combines a CNN backbone network and Transformer detectors. By treating objects as a fixed set of learned queries and using Hungarian matching to align predictions with ground-truth boxes, it eliminates the need for non-maximum suppression. Recently, Zhao et al. [47] proposed a real-time Transformer detector achieving higher inference speed than YOLO models on Tesla T4 GPU, incorporating intra-scale feature

interaction and hybrid encoder for reduced latency.

Despite delivering state-of-the-art performance on object detection benchmarks, Transformer-based approaches require substantial GPU memory and lengthy training schedules, which limits their applicability to practical downstream tasks and employment to mobile detection devices.

2.3. Deep Learning Based PWD Detection

Recent advances in remote sensing and deep learning have enabled automated PWD detection from UAV imagery, offering scalable solutions for large-scale forest monitoring. Current approaches primarily utilize two types of data modalities: hyperspectral and RGB imagery. Hyperspectral images provide rich spectral information across multiple wavelengths, enabling detailed analysis of chemical properties for disease classification [48–51]. However, hyperspectral data acquisition is constrained by high costs, technical complexity, and the requirement for lower flight altitudes, resulting in reduced coverage efficiency. In contrast, RGB imagery offers a practical alternative that can be efficiently captured at higher altitudes while maintaining sufficient information for disease detection.

Several deep learning approaches have been developed for PWD detection using RGB imagery. Deng et al. [52] used a Faster-RCNN model with ResNet [35] and predefined some improved anchor boxes on an augmented dataset containing 1700 pine wilt images. A study was conducted by Oide et al. [53] to investigate the distribution of the three channels in RGB images of infected trees and applied artificial neural network and support vector machine on their dataset. Another research [54] built a deep learning model with attention-based fusion of the frequency domain features and refined RGB imge features, which outperformed the advanced YOLO v5 [55] algorithm on their pine trees data from Liaoning Province, China. Other researchers also evaluated the performance of commonly used object detection models on recognizing UAV-captured images [56–59]. These studies have demonstrated significant improvements in implementing pine disease detection systems, but several limitations remain. First, some approaches rely on relatively small datasets, which may lead to overfitting and limit generalizability to diverse forest environments. Additionally, two-stage models such as Faster R-CNN, despite their high accuracy, exhibit slow inference speeds that hinder their deployment in real-time operational forest monitoring systems.

Detecting small objects in UAV imagery poses significant challenges, especially in forest scenarios where infected trees manifest as small-scale targets within cluttered backgrounds. Current approaches have explored various strategies to address scale variation and feature preservation challenges. Feature Pyramid Networks (FPN) [60] provide multi-scale feature representation, with subsequent works enhancing fusion mechanisms through adaptive weighting schemes [61]. Ye et al. [62] also enhanced their model by designing a fusion network that integrates global and local features, with the latter extracted through a self-attention-based backbone network. Recent developments have incorporated additional prediction heads for small objects [63] and attention mechanisms to focus on relevant regions [64], with other architectural improvements further demonstrating advances in small object detection [65–67].

However, challenges remain in applying object detection to forest disease detection. Traditional CNN-based feature extraction methods often lose critical high-frequency texture information essential for distinguishing subtle disease symptoms. Hierarchical processing architectures further exacerbate this issue, as fine-grained spatial details are progressively lost through pooling operations, while high-level semantic information becomes diluted during multi-scale feature integration. To address these feature preservation challenges, we deploy the discrete wavelet transform, which enables simultaneous capture of both spatial and frequency domain features while maintaining invertibility. Our approach reconstructs multi-scale features by selectively preserving specific frequency components through inverse wavelet transforms, thereby retaining critical texture information essential for disease classification.

3. Materials and Methods

To improve the performance of the proposed object detection model, we integrate multiple specialized modules into the backbone and optimize the neck architecture. Each module is discussed in detail below.

3.1. The Dual-Channel Feature Extraction Network

In most object detection methods, CNNs take the pivotal role of extracting features from the initial images. The semantic information becomes richer while the model goes deeper, meanwhile, the representation of texture details starts to diminish. In image processing, objects and details usually exhibit different properties at different scales. Intuitively, edge information is more prominent in high-frequency details, whereas shape contours or large structures are more visible in low frequencies. Therefore, we add a branch to the backbone network and use wavelet transform to extract the information of high- and low-frequency information in the image. Research has demonstrated that wavelet transforms are powerful tools for extracting useful features in the frequency domain [68,69]. The Haar

transform is a compactly supported, dyadic, and orthonormal wavelet transform, which makes it particularly efficient and suitable for real-time or embedded applications in image processing [70]. Due to its sensitivity to edges and abrupt changes, the Haar wavelet provides more distinctive representations when applied to tree-like structures in images.

We incorporate a Haar wavelet transform down-sampling module into one branch of the CNN as an auxiliary structure to construct feature maps. The architecture of the dual-channel network is shown in Figure 2. The convolutional pathway uses the CNN module from FasterNet [71] which employs a 2×2 convolutional kernel with a stride of 2, to perform convolutional operations, reduce the spatial dimensions, and provide the deeper features for subsequent layers. The basic block consists of a convolutional layer (Conv), followed by batch normalization (BN) and Gaussian Error Linear Unit (GELU) activation. To preserve the distinctiveness between the two backbone streams, the first Conv layer employs two separate sets of parameters. Following that, a partial convolutional layer (P_Conv) [71] is employed for efficient feature extraction. Additionally, the Patch Embedding and Patch Merging modules are implemented as (Conv_Down) modules with zero padding and strides of 4 and 2, respectively. In the other branch, the Wavelet stream, the Haar discrete wavelet transform is applied as down-sampling operation and for capturing features in frequency domain using a high frequency filter and a low frequency filter. The operation of two-dimensional DWT of function $f(x, y)$ with size $M \times N$ can be calculated as

$$\begin{aligned} W_\Phi(j, m, n) &= \frac{1}{\sqrt{MN}} \sum_x \sum_y f(x, y) \Phi_{j,m,n}(x, y), \\ W_\Psi^i(j, m, n) &= \frac{1}{\sqrt{MN}} \sum_x \sum_y f(x, y) \Psi_{j,m,n}^i(x, y), \end{aligned} \tag{3}$$

where $W_\Phi(j, m, n)$ calculates an approximation of input; $W_\Psi^i(j, m, n)$ represents the wavelet coefficients in specific orientation i, with $i \in \{H, V, D\}$ corresponding to horizontal, vertical, and diagonal respectively; x and y are defined as $0 \leq x, y < M, N$; m and n are the translation factors; and j indicates the level of wavelet decomposition. In Equation (3), the scaling function and the other three wavelet function can be given as

$$\begin{aligned} \Phi(x, y) &= \varphi(x)\varphi(y), \\ \Psi^H(x, y) &= \psi(x)\varphi(y), \\ \Psi^V(x, y) &= \varphi(x)\psi(y), \\ \Psi^D(x, y) &= \psi(x)\psi(y), \end{aligned} \tag{4}$$

where the $\varphi(x)$ stands for mother scaling function, and is utilized to create a series of approximation at low frequency. The function $\psi(x)$ is the mother wavelet function, which is responsible for analyzing and encoding the differences between adjacent approximations, with a focus on high-frequency variations. To perform 2D Haar transform, these two filter will perform the convolution operation along horizontal and vertical dimension respectively, producing four different feature maps. The DWT module improves feature representation of the model and enhances robustness to noise without introducing any new learnable parameters. The Cross-Modal Channel Enhancement (CMCE) fusion module is applied after the basic block at stage S_2, S_3, and S_4, which will be sent to the refinement module.The details of CMCE fusion model are illustrated in Section 3.2.

3.2. Cross-Modal Channel Enhancement Module

The CMCE module serves as a feature enhancement mechanism, which is designed to improve the representation of multi-scale features in deep networks. It enhances the discriminative power of feature maps by leveraging channel-wise interactions and mutual information sharing. The CMCE module can be defined as follows.

$$f_a, f_b = \text{Split}(x), \quad x \in \mathbb{R}^{B \times 2C \times H \times W}, \tag{5}$$

where Split() denotes that the input tensor x is equally split along the channel dimension into two feature maps, f_a and f_b, each with shape $\mathbb{R}^{B \times C \times H \times W}$. Here, B, C, H, and W represent the batch size, number of channels, height, and width, respectively. To obtain global contextual information, average pooling is applied to both branches

$$f_a^{\text{pool}} = \frac{1}{H \cdot W} \sum_{i=1}^{H} \sum_{j=1}^{W} f_a(:, :, i, j), \quad f_b^{\text{pool}} = \frac{1}{H \cdot W} \sum_{i=1}^{H} \sum_{j=1}^{W} f_b(:, :, i, j). \tag{6}$$

The pooled features are then passed through a shared attention subnetwork composed of two convolutional layers and nonlinearities

$$\mathcal{A}(a) = \sigma\left(\text{Conv}\left(\text{ReLU}\left(\text{Conv}\left(a\right)\right)\right)\right), \tag{7}$$

where Conv denotes a convolutional neural network layer, and σ is the sigmoid function that maps the output to the range between 0 and 1. The cross-modal enhancement is performed by

$$f_a^{\text{enhanced}} = f_a + f_b \cdot \mathcal{A}\left(f_a^{\text{pool}}\right), \quad f_b^{\text{enhanced}} = f_b + f_a \cdot \mathcal{A}\left(f_b^{\text{pool}}\right), \tag{8}$$

in which $\cdot$ denotes element-wise multiplication with broadcasting. Finally, the enhanced features and the fused representation are combined with a residual structure to form the module output

$$f_{\text{fused}} = \text{ReLU}\left(\text{BN}\left(\text{Conv}\left(x\right)\right)\right), \tag{9}$$

$$\text{Out} = \text{Concat}\left(f_a^{\text{enhanced}} + f_{\text{fused}}, f_b^{\text{enhanced}} + f_{\text{fused}}\right), \tag{10}$$

where $\text{Concat}(f_a, f_b)$ operation concatenates two elements as one output.

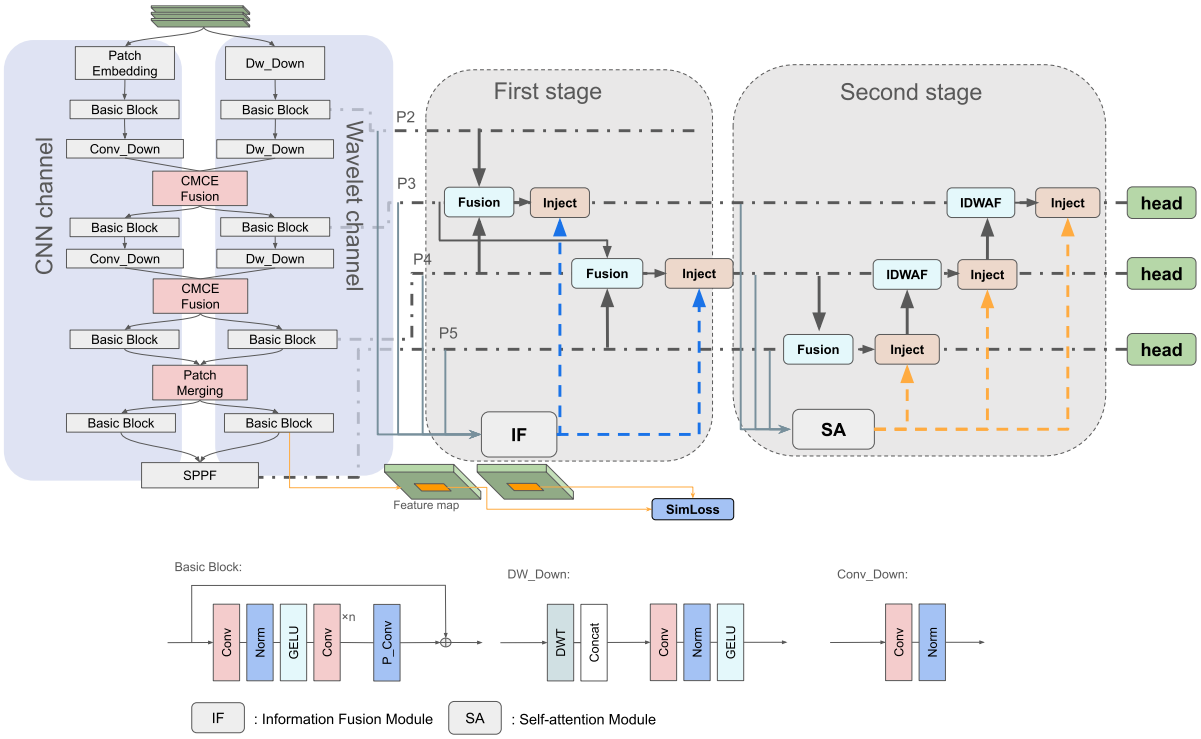
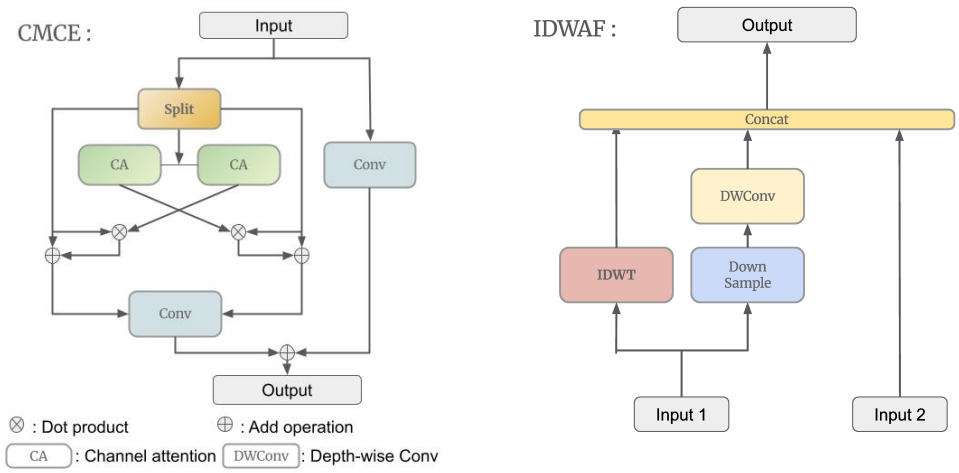


Figure 2. The proposed dual-channel feature extraction model containing a convolutional flow and a wavelet transform flow. The CMCE and IDWAF modules are detailed in Figure 3.

The structure of the CMCE module is depicted in the left part of Figure 3. By employing a feature enhancement mechanism, the module enables the model to maintain consistency between the two channels while extracting deep features, particularly when detecting the same object or processing the same region. As described in Equation (5), features from a lower-level extraction layer are divided into CNN part and wavelet part along the channel dimension. To enhance both streams, the module generates attention weights based on the counterpart branch, enabling cross-modal fusion along the channel axis. This process, detailed in Equation (7), utilizes a 'squeeze' operation inspired by the Convolutional Block Attention Module (CBAM) [72]. While our approach shares the element-wise multiplication principle with attention mechanisms like CBAM, our CMCE module differs in that it specifically targets cross-modal feature enhancement rather than single-channel attention. This operation implements selective information gating that enhances relevant features while suppressing irrelevant information. A bottleneck-like convolutional layer is employed as a projection shortcut to compute the fused features from the original input. Subsequently, an additional layer is used to align the fused features with the enhanced representations, enabling effective integration of both streams.



(**a**) CMCE module architecture (**b**) IDWAF module architecture

Figure 3. The architecture of proposed modules: (**a**) CMCE module and (**b**) IDWAF module.

3.3. Multi-Scale Wavelet Alignment Fusion

In the most CNN-based object detection model, an FPN structure and Path Aggregation Network (PAN) [73] module are used for feature refinement with top-down and bottom-up procedures, where features from adjacent layers are fused in a step-by-step manner. While this process is effective in many scenarios, it can result in the loss of fine details during the cross-level fusion process. Our feature refinement neck structure is illustrated in Figure 3a. In the first stage, intermediate features are generated using the Information Fusion Module (IF) and Information Injection Module (Inject) with the same settings as in the Gold-YOLO model [74], where the IF integrates features from four different scales to ensure sufficient and efficient information aggregation, and the Inject modules are responsible for further extraction and fusion. In the second stage, the neck network, information from P2, P3 and P4 is aggregated with a self-attention based module for future processing. In the refinement from high-level to low-level, the Inverse Discrete Wavelet Transform (IDWT) is used as an up-sampling process to reconstruct the high-resolution feature maps. IDWT can improve the fusion of multiscale features by combining both fine and coarse information, resulting in better detection of objects at different scales [75,76]. The structure of the Inverse Discrete Wavelet Alignment Fusion (IDWAF) is presented in the right part of Figure 3b. The fusion procedure employs Depthwise convolution (DWConv) [77] to process the input from adjacent layers concurrently, further extracting local details. This combination helps the model obtain complementary information. Finally, the results on three scales are used for object localization, boundary regression, classification, and loss computation.

3.4. Object Level Similarity Loss

To enhance feature consistency and improve detection robustness, we introduce an object level similarity loss (SimLoss) that operates on dual-branch feature representations. This mechanism maintains coherent feature learning across different processing paths while providing specialized supervision for detected objects. The model used target bounding boxes to perform Region of Interest align operation [78] of the feature maps from two parallel branches. The SimLoss are combined with a cosine similarity loss and a classification supervision loss. The cosine similarity loss is described as

$$\mathcal{L}_{\cos} = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{\mathbf{e}_A^i \cdot \mathbf{e}_B^i}{|\mathbf{e}_A^i||\mathbf{e}_B^i|}\right), \tag{11}$$

where $\mathbf{e}_j^i$ are obtained by the same encoder network from two separate streams. To ensure that the encoded feature representations capture semantically meaningful information and prevent the model from converging to non-discriminative feature spaces, classification supervision is applied to both processing branches

$$\mathcal{L}_{\text{cls}} = \frac{1}{2}\left(\mathcal{L}_{\text{CE}}(\mathbf{p}_A, \mathbf{y}) + \mathcal{L}_{\text{CE}}(\mathbf{p}_B, \mathbf{y})\right), \tag{12}$$

where $\mathcal{L}_{\text{CE}}$ represents the cross-entropy loss function, with $\mathbf{p}$ and $\mathbf{y}$ denoting the prediction and target, respectively. The final loss combines the original detection loss and the cosine similarity loss with the proposed object-level

constraints, which is defined as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{det}} + \lambda_1 \mathcal{L}_{\cos} + \lambda_2 \mathcal{L}_{\text{cls}}, \tag{13}$$

where λ_1 and λ_2 are balance weights for similarity and classification losses.

4. Results

To validate the effectiveness of our approach, we conducted extensive comparative experiments against state-of-the-art methods on two datasets.

4.1. Dataset and Setups

This study focuses on forest in several counties within Yichang City, Hubei Province, China, from which the data was collected. Yichang City is characterized by a vast forested area, with tree cover exceeding 55%, which presents significant potential for the application of intelligent detection algorithms. Our drones are equipped with Global Navigation Satellite receivers. During flight operations, a mobile receiver is deployed alongside a base station receiver to capture observations. The data were captured in 2022 by an image sensor at an altitude of more than 1000 m with a spatial resolution of better than 0.6 m. All images were preprocessed to a uniform resolution of 1000×1000 pixels. After annotation, we obtained 2360 images and annotated 2962 pine tree instances, which were divided into training, validation, and test datasets with 1321, 331, and 708 images, respectively, following an approximate ratio of 6:1:3. During training, several data augmentation techniques were employed, including saturation adjustment, vertical and horizontal flipping, mixup [79], and mosaic technique [40]. Three samples from our dataset and one augmented sample with bounding boxes are presented in Figure 4.

(a) (b) (c) (d)

Figure 4. Representative samples from the pine wilt disease dataset: (a–c) infected trees, (d) data augmentation preprocessing.

The object detection model is developed using the Ultralytics YOLO Vision platform [55], with the backbone structure and configurations based on the FasterNet architecture [71]. The detection heads are adopted from the YOLOv8 detector [80]. The AdamW optimizer [81] is used with an initial learning rate of 0.001. In addition, a warm-up strategy and cosine annealing scheduler are applied during training. The balance weights in our loss function λ_1 and λ_2 are set to 1.0 and 0.5. The entire project is implemented using PyTorch version 1.13.1, and all models are trained on an NVIDIA A100 GPU provided by the Digital Research Alliance of Canada [82]. During the training process, a hard negative mining strategy is used, and the model weights are updated using both SIoU [83] bounding box loss and our proposed similarity loss. All experiments are trained for 300 epochs with a batch size of 48.

4.2. Comparisons

Our focus is primarily on evaluating the performance of our models on our pine wilt trees dataset and to perform rapid detection of all infected trees in the test images. For the detection of pine wilt disease, it is crucial to identify all infected trees, as undetected infections can lead to rapid disease spread. Thus, we use Recall, Average Precision (AP), and AP at a threshold of 0.50 (AP@50) as the evaluation metrics, which are presented in the COCO dataset as the evaluation metrics [84]. Additionally, the number of Parameters (Params) of the model and Gigabyte Floating-point Operations (GFLOPS) of the model are calculated to assess computational cost. We compare representative models, including Cascade R-CNN [85], advanced YOLO algorithms, RT-DETR with ResNet50 as the backbone, and the proposed model, on our pine wilt dataset. As illustrated in Table 1, our model outperforms 3.2% and 2.6% compared to YOLOv8-s.

Table 1. Comparative analysis of various detectors was conducted on our pine trees dataset, with the top-ranked results highlighted in bold. The evaluation metrics used were Average Precision (AP) and Average Precision at an Intersection over Union (IoU) threshold of 0.5 (AP@50).

Model	Recall	AP@0.5	AP	Params	GFLOPs
Cascade R-CNN [85]	0.751	0.824	0.457	69.4 M	119.0
YOLOv5u-s [55]	0.756	0.851	0.491	9.1 M	24.0
YOLOv6-s [86]	0.746	0.850	0.490	16.3 M	44.2
YOLOv8-s [80]	0.801	0.860	0.504	11.1 M	28.6
YOLOv11-s [80]	0.776	0.874	0.518	9.4 M	21.6
RTDETR-r50 [47]	0.81	0.878	0.513	42.8 M	130.5
Ours	**0.824**	**0.892**	**0.53**	13.5 M	30.9

4.3. Ablation Study

To verify the effectiveness of the proposed modules, we separately examined the DWT down-sampling module, the CMCE module, and the similarity loss. The original YOLOv8-s model is selected as the baseline model. We integrated our proposed DWT backbone network, CMCE module and similarity loss into the baseline model, and tested their combinations to evaluate their synergistic effects. According to Table 2, the CMCE module demonstrates an improvement in both AP and AP@50 compared to simple concatenation fusion, indicating that the cross-modal feature enhancement mechanism effectively captures comprehensive information. The DWT down-sampling in the backbone also shows improvement compared to conventional convolutional down-sampling and the FasterNet backbone [71], demonstrating that frequency-domain decomposition preserves critical fine-grained features that are typically lost in down-sampling operations. The similarity loss contributes an additional improvement (from 0.881 to 0.887), which is crucial for maintaining consistency of the feature across the dual channels and reducing false positives. Although both streams employ the same architecture for feature extraction, the dual-channel fusion model achieves improved accuracy over single-channel approaches, validating our hypothesis that combining spatial and frequency-domain representations enhances detection capability. Furthermore, the combination of DWT, CMCE, and SimLoss results in a 2.7% enhancement in AP@50 compared to the baseline YOLOv8-s model.

Table 2. Ablation study on the proposed components: DWT-backbone, FasterNet, CMCE module, and SimLoss. ✓indicates the component is included in the configuration. The best scores are highlighted in bold.

DWT-backbone	FasterNet	CMCE	SimLoss	AP@50	AP
	✓			0.877	0.511
✓				0.880	0.514
✓		✓		0.881	0.519
✓			✓	0.884	0.523
✓		✓	✓	**0.887**	**0.524**

To investigate the performance of Haar wavelet transform, we compared models that replace the Haar transform with other wavelet transforms. We also tested learnable wavelets [87] to evaluate the adaptive model's performance. Results are shown in Table 3. Although other wavelet transforms achieve competitive results, Haar wavelet still achieves the highest mAP on the PWD dataset. We evaluated the Frame Per Second (FPS) performance on a laptop equipped with GTX 4070 GPU, and the Haar wavelet approach outperformed strided convolution. To further validate the effectiveness of our DWT-based approach, we conducted Grad-CAM [88] visualization analysis as shown in Figure 5. Comparing the attention heatmaps, the DWT channel focuses more on specific tree regions while the conventional layers focus on broader image contexts, which demonstrates their complementary strengths when combined in our dual-channel framework.

We also evaluated the hyperparameter λ_1 and λ_2 in our proposed similarity loss function. The experiments achieved best results when λ_1 and λ_2 were set to 1.0 and 0.5, respectively. Experimental results are shown in Table 4.

Table 3. Experimental results with different wavelet transforms and convolutional operations, with the top-ranked model scores highlighted in bold.

Transform	AP@50	AP	FPS
Conv	0.872	0.521	107.5
Daubechies (db2)	0.887	0.525	-
Learnable DWT	0.885	0.520	-
Haar	**0.892**	**0.53**	144.9

Table 4. Ablation study results for similarity loss hyperparameters λ_1 and λ_2, with the best scores highlighted in bold.

λ_1	λ_2	AP@50	AP
1.0	0.25	0.882	0.522
1.0	2.0	0.881	0.527
0.5	0.5	0.88	0.525
2.0	0.5	0.881	0.527
1.0	0.5	**0.892**	**0.53**

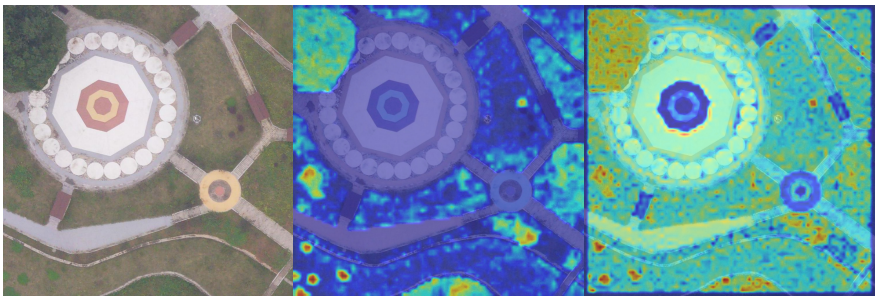


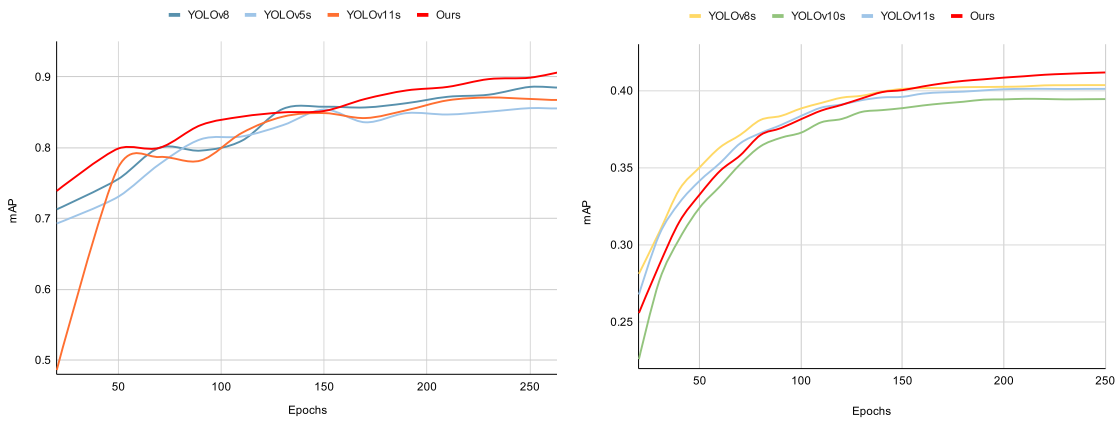
Figure 5. GradCAM visualization comparison showing original input image (**left**), attention heatmap with DWT-based feature extraction (**middle**), and attention heatmap with conventional convolution (**right**).

4.4. Comparisons on VisDrone Dataset

To further validate the generalizability and robustness of our proposed model, we conduct additional experiments on the VisDrone dataset [22], which presents similar challenges to our pine wilt disease recognition task. The VisDrone dataset has 6471 images for training and 1610 for validation, with 10 different classes being annotated. It is a large-scale benchmark dataset specifically designed for object detection in drone-captured images, containing dense annotations of small objects in complex aerial scenes. The experimental results in Table 5 show that our proposed model outperforms advanced YOLO algorithms on this publicly available challenging dataset, further validating the effectiveness of our approach for small object detection in aerial imagery. Figure 6 shows the smoothed training curves of different models on VisDrone, where our method (red line) demonstrates superior performance compared to baseline YOLO models.

Table 5. Experiments on VisDrone validation dataset with advanced object detection models, with the top-ranked scores highlighted in bold.

Models	Recall	AP@50	AP
Yolov8-s	0.393	0.404	0.238
Yolov10-s	0.387	0.395	0.236
Yolov11-s	0.391	0.401	0.238
Ours	**0.396**	**0.412**	**0.243**



(**a**) Training progress on pine wilt dataset

(**b**) Training progress on VisDrone dataset

Figure 6. Training curves comparison showing mAP evolution across epochs for different models.

5. Discussion

The experimental results demonstrate the effectiveness of our proposed approach for pine wilt disease detection in UAV imagery. The superior performance can be attributed to several key factors. First, the frequency-domain analysis through DWT decomposition enables the model to capture subtle texture changes that are often missed by conventional CNN-based methods. Second, the object-level similarity constraint prevents the dual streams from learning conflicting representations, ensuring consistent feature learning across processing paths.

As shown in Figure 7, our proposed model successfully identifies small infected objects that are often missed by YOLOv8-s algorithms, particularly for targets near image boundaries where traditional methods struggle with blurred features. The cross-dataset evaluation on VisDrone further validates the generalizability of our approach, though detecting extremely small objects remains challenging.

(**a**)

(**b**)

Figure 7. Comparison of detection of YOLOv8-s and our proposed model: (**a**) Results of our DWT-based model, (**b**) Results of YOLOv8-s model.

However, we observed limitations in detecting dead trees that lack distinctive visual features. Since these dead trees may still harbor vectors or pathogens, future work could explore multi-stage classification schemes to differentiate various infection stages and incorporate additional spectral information for comprehensive disease monitoring. Additionally, although learnable wavelets did not demonstrate superior results in this work, investigating adaptive wavelet selection mechanisms remains worthwhile and could optimize performance for different forest environments. Furthermore, extending the framework to real-time processing would enable autonomous UAV-based monitoring systems for practical deployment.

6. Conclusions

In this study, we propose a DWT-based dual-channel pine tree detection algorithm, aimed at enhancing the accuracy of detecting small, infected pine trees in UAV imagery. Our approach employs a two-stream architecture where input features are split into dual-channels for parallel processing. The DWT down-sampling module serves as an auxiliary feature extraction component that captures both spatial and frequency domain information while preserving fine-grained details often lost in conventional pooling operations. To ensure feature consistency between the two channels, we introduce an object-level similarity constraint that enforces cosine similarity between corresponding feature representations. In the decoder pathway, IDWT operations replace traditional upsampling methods, enabling principled reconstruction of spatial details from frequency components. The complete framework is integrated with YOLO detection heads to provide optimization for pine wilt disease recognition. Experimental results demonstrate that the proposed approach significantly enhances detection accuracy compared to baseline methods.

Supplementary Materials

The VisDrone2019 dataset is available at https://github.com/VisDrone/VisDrone-Dataset. The code will be available at https://github.com/zhoushengs/PWD-detect.

Author Contributions

All authors made significant contributions to the manuscript. Z.Z. conceived, designed, and conducted the research. Leadership responsibility, commentary, and critical review: S.X.Y. All authors discussed the basic structure of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

Data Availability Statement

The data are not publicly available due to its commercial nature.

Acknowledgments

We would like to express our gratitude for the generous support provided by Professor Dong Ren at the China Three Gorges University in supplying the data. Additionally, this research was made possible, in part, by support from the Digital Research Alliance of Canada (alliancecan.ca).

Conflicts of Interest

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

References

1. Zhao, J.; Huang, J.; Yan, J.; et al. Economic loss of pine wood nematode disease in mainland China from 1998 to 2017. *Forests* **2020**, *11*, 1042.
2. Shi, J.; Luo, Y.Q.; Song, J.Y.; et al. Traits of Masson pine affecting attack of pine wood nematode. *J. Integr. Plant Biol.* **2007**, *49*, 1763–1771.
3. Quirion, B.R.; Domke, G.M.; Walters, B.F.; et al. Insect and disease disturbances correlate with reduced carbon sequestration in forests of the contiguous United States. *Front. For. Glob. Chang.* **2021**, *4*, 716582.
4. Kiyohara, T.; Tokushige, Y. Inoculation experiments of a nematode, Bursaphelenchus sp., onto pine trees. *J. Jpn. For. Soc.* **1971**, *53*, 210–218.

5. Mamiya, Y.; Enda, N. Transmission of Bursaphelenchus Lignicolus (Nematoda: Aphelenchoididae) By Monochamus Alternatus (Coleoptera: Cerambycidae). *Nematologica* **1972**, *18*, 159–162.
6. Back, M.A.; Bonifácio, L.; Inácio, M.L.; et al. Pine wilt disease: A global threat to forestry. *Plant Pathol.* **2024**, *73*, 1026–1041.
7. Shimazu, M.; Katagiri, K. Pathogens of the pine sawyer, Monochamus alternatus Hope, and possible utilization of them in a control program. In Proceedings of the 17th IUFRO World Congress, Kyoto, Japan, 6–17 September 1981; Volume 504, pp. 291–295.
8. Yu, H.B.; Jung, Y.H.; Lee, S.M.; et al. Biological control of Japanese pine sawyer, Monochamus alternatus (Coleoptera: Cerambycidae) using Korean entomopathogenic nematode isolates. *Korean J. Pestic. Sci.* **2016**, *20*, 361–368.
9. Sousa, E.; Vale, F.; Abrantes, I. *Pine wilt Disease in Europe: Biological Interactions and Integrated Management*; FNAPF: Lisbon, Portugal, 2015.
10. Cybenko, G. Approximation by superpositions of a sigmoidal function. *Math. Control. Signals Syst.* **1989**, *2*, 303–314.
11. Qian, K.; Duan, Y.; Luo, C.; et al. Pixel-Level Domain Adaptation for Real-to-Sim Object Pose Estimation. *IEEE Trans. Cogn. Dev. Syst.* **2023**, *15*, 1618–1627.
12. Li, R.; Mo, T.; Yang, J.; et al. Bridge inspection named entity recognition via BERT and lexicon augmented machine reading comprehension neural model. *Adv. Eng. Inform.* **2021**, *50*, 101416.
13. Yan, C.; Meng, L.; Li, L.; et al. Age-invariant face recognition by multi-feature fusionand decomposition with self-attention. *ACM Trans. Multimed. Comput. Commun. Appl.* **2022**, *18*, 1–18.
14. Yan, C.; Li, Z.; Zhang, Y.; et al. Depth image denoising using nuclear norm and learning graph model. *ACM Trans. Multimed. Comput. Commun. Appl.* **2020**, *16*, 1–17.
15. Hopfield, J.J. Neurons with graded response have collective computational properties like those of two-state neurons. *Proc. Natl. Acad. Sci. USA* **1984**, *81*, 3088–3092.
16. Hochreiter, S. *Long Short-term Memory*; Neural Computation; MIT-Press: Cambridge, MA, USA, 1997.
17. Sutskever, I. Sequence to Sequence Learning with Neural Networks. *arXiv* **2014**, arXiv:1409.3215.
18. Vaswani, A. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1.
19. Egli, S.; Höpke, M. CNN-based tree species classification using high resolution RGB image data from automated UAV observations. *Remote. Sens.* **2020**, *12*, 3892.
20. Ke, C.; Ni, J.; Zhao, Y.; et al. Cross-Scale Feature Enhancement for Cotton Seedling Detection in UAV Images. *IEEE Geosci. Remote. Sens. Lett.* **2024**, *21*, 1–5.
21. Yu, R.; Luo, Y.; Zhou, Q.; et al. A machine learning algorithm to detect pine wilt disease using UAV-based hyperspectral imagery and LiDAR data at the tree level. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *101*, 102363.
22. Zhu, P.; Wen, L.; Du, D.; et al. Detection and tracking meet drones challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 7380–7399.
23. Yue, M.; Quan, L.; Cheng-Ming, Y.; et al. Study on early diagnosis technology of pine wilt disease. *J. Shandong Agric. Univ.* **2014**, *45*, 158–160.
24. Li, M.; Li, H.; Ding, X.; et al. The detection of pine wilt disease: A literature review. *Int. J. Mol. Sci.* **2022**, *23*, 10797.
25. Kong, Q.Q.; Ding, X.L.; Chen, Y.F.; et al. Comparison of morphological indexes and the pathogenicity of Bursaphelenchus xylophilus in northern and southern China. *Forests* **2021**, *12*, 310.
26. Hu, Y.; Kong, X.; Wang, X.; et al. Direct PCR-based method for detecting Bursaphelenchus xylophilus, the pine wood nematode in wood tissue of Pinus massoniana. *For. Pathol.* **2011**, *41*, 165–168.
27. Lecun, Y.; Bottou, L.; Bengio, Y.; et al. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324.
28. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1.
29. Sellers, T.; Lei, T.; Luo, C.; et al. A node selection algorithm to graph-based multi-waypoint optimization navigation and mapping. *Intell. Robot.* **2022**, *2*, 333–54.
30. Ni, J.; Zhang, Z.; Shen, K.; et al. An improved deep network-based RGB-D semantic segmentation method for indoor scenes. *Int. J. Mach. Learn. Cybern.* **2024**, *15*, 589–604.
31. Lei, T.; Luo, C.; Jan, G.E.; et al. Deep learning-based complete coverage path planning with re-joint and obstacle fusion paradigm. *Front. Robot. AI* **2022**, *9*, 843816.
32. Yan, C.; Hao, Y.; Li, L.; et al. Task-adaptive attention for image captioning. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 43–51.
33. LeCun, Y.; Boser, B.; Denker, J.S.; et al. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Comput.* **1989**, *1*, 541–551.
34. Ioffe, S. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv* **2015**, arXiv:1502.03167.
35. He, K.; Zhang, X.; Ren, S.; et al. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.

36. Girshick, R.; Donahue, J.; Darrell, T.; et al. Region-Based Convolutional Networks for Accurate Object Detection and Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 142–158.
37. Redmon, J. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
38. Redmon, J.; Farhadi, A. YOLO9000: better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
39. Farhadi, A.; Redmon, J. Yolov3: An incremental improvement. In *Computer Vision and Pattern Recognition*; Springer: Berlin/Heidelberg, Germany, 2018; Volume 1804, pp. 1–6.
40. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
41. Girshick, R. Fast r-cnn. *arXiv* **2015**, arXiv:1504.08083.
42. Ren, S.; He, K.; Girshick, R.; et al. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149.
43. Duan, K.; Bai, S.; Xie, L.; et al. Centernet: Keypoint triplets for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6569–6578.
44. Tian, Z.; Shen, C.; Chen, H.; et al. FCOS: A simple and strong anchor-free object detector. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 1922–1933.
45. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
46. Carion, N.; Massa, F.; Synnaeve, G.; et al. End-to-end object detection with transformers. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
47. Zhao, Y.; Lv, W.; Xu, S.; et al. Detrs beat yolos on real-time object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 16965–16974.
48. Iordache, M.D.; Mantas, V.; Baltazar, E.; et al. A machine learning approach to detecting pine wilt disease using airborne spectral imagery. *Remote. Sens.* **2020**, *12*, 2280.
49. Yu, R.; Huo, L.; Huang, H.; et al. Early detection of pine wilt disease tree candidates using time-series of spectral signatures. *Front. Plant Sci.* **2022**, *13*, 1000093.
50. Wu, W.; Zhang, Z.; Zheng, L.; et al. Research progress on the early monitoring of pine wilt disease using hyperspectral techniques. *Sensors* **2020**, *20*, 3729.
51. Zhou, Z.; Zhang, Y.; Gu, Z.; et al. Deep learning approaches for object recognition in plant diseases: A review. *Intell. Robot.* **2023**, *3*, 514–537.
52. Deng, X.; Tong, Z.; Lan, Y.; et al. Detection and location of dead trees with pine wilt disease based on deep learning and UAV remote sensing. *AgriEngineering* **2020**, *2*, 294–307.
53. Oide, A.H.; Nagasaka, Y.; Tanaka, K. Performance of machine learning algorithms for detecting pine wilt disease infection using visible color imagery by UAV remote sensing. *Remote. Sens. Appl. Soc. Environ.* **2022**, *28*, 100869.
54. Xie, W.; Wang, H.; Liu, W.; et al. Early-Stage Pine Wilt Disease Detection via Multi-Feature Fusion in UAV Imagery. *Forests* **2024**, *15*, 171.
55. Jocher, G.; Chaurasia, A.; Stoken, A.; et al. ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation. 2022. Available online: https://github.com/ultralytics/yolov5 (accessed on 10 May 2025).
56. Wu, Z.; Jiang, X. Extraction of pine wilt disease regions using UAV RGB imagery and improved mask R-CNN models fused with ConvNeXt. *Forests* **2023**, *14*, 1672.
57. Zhang, N.; Chai, X.; Li, N.; et al. Applicability of UAV-based optical imagery and classification algorithms for detecting pine wilt disease at different infection stages. *GIScience Remote. Sens.* **2023**, *60*, 2170479.
58. Zhou, Z.; Yang, X. Pine wilt disease detection in UAV-CAPTURED images. *Int. J. Robot. Autom* **2022**, *37*, 37–43.
59. Huang, X.; Gang, W.; Li, J.; et al. Extraction of pine wilt disease based on a two-stage unmanned aerial vehicle deep learning method. *J. Appl. Remote. Sens.* **2024**, *18*, 014503–014503.
60. Lin, T.Y.; Dollár, P.; Girshick, R.; et al. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
61. Sun, W.; Dai, L.; Zhang, X.; et al. RSOD: Real-time small object detection algorithm in UAV-based traffic monitoring. *Appl. Intell.* **2021**, *52*, 8448–8463.
62. Ye, T.; Qin, W.; Li, Y.; et al. Dense and Small Object Detection in UAV-Vision Based on a Global-Local Feature Enhanced Network. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–13.
63. Zhao, H.; Zhang, H.; Zhao, Y. Yolov7-sea: Object detection of maritime uav images based on improved yolov7. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2023, pp. 233–238.
64. Yang, L.; Zhang, R.Y.; Li, L.; et al. SimAM: A Simple, Parameter-Free Attention Module for Convolutional Neural Networks. In Proceedings of the 38th International Conference on Machine Learning, Virtual Event, 18–24 July 2021; Volume 139, pp. 11863–11874.

65. Ni, J.; Zhu, S.; Tang, G.; et al. A Small-Object Detection Model Based on Improved YOLOv8s for UAV Image Scenarios. *Remote. Sens.* **2024**, *16*, 2465.
66. Wang, F.; Wang, H.; Qin, Z.; et al. UAV target detection algorithm based on improved YOLOv8. *IEEE Access* **2023**, https://doi.org/10.1109/ACCESS.2023.3325677.
67. Zhang, Y.; Zuo, Z.; Xu, X.; et al. Road damage detection using UAV images based on multi-level attention mechanism. *Autom. Constr.* **2022**, *144*, 104613.
68. Xu, G.; Liao, W.; Zhang, X.; et al. Haar wavelet downsampling: A simple but effective downsampling module for semantic segmentation. *Pattern Recognit.* **2023**, *143*, 109819.
69. Finder, S.E.; Amoyal, R.; Treister, E.; et al. Wavelet Convolutions for Large Receptive Fields. *arXiv* **2024**, arXiv:2407.05848.
70. Porwik, P.; Lisowska, A. The Haar-wavelet transform in digital image processing: its status and achievements. *Mach. Graph. Vis.* **2004**, *13*, 79–98.
71. Chen, J.; Kao, S.h.; He, H.; et al. Run, don't walk: chasing higher FLOPS for faster neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 12021–12031.
72. Woo, S.; Park, J.; Lee, J.Y.; et al. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
73. Liu, S.; Qi, L.; Qin, H.; et al. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
74. Wang, C.; He, W.; Nie, Y.; et al. Gold-YOLO: Efficient object detector via gather-and-distribute mechanism. *Adv. Neural Inf. Process. Syst.* **2024**, *36*, 51094–51112.
75. Dharejo, F.A.; Deeba, F.; Zhou, Y.; et al. TWIST-GAN: Towards wavelet transform and transferred GAN for spatio-temporal single image super resolution. *ACM Trans. Intell. Syst. Technol.* **2021**, *12*, 1–20.
76. Yang, H.H.; Yang, C.H.H.; Wang, Y.C.F. Wavelet channel attention module with a fusion network for single image deraining. In Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP), Virtual, 25–28 October 2020; pp. 883–887.
77. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
78. He, K.; Gkioxari, G.; Dollár, P.; et al. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
79. Zhang, H.; Cisse, M.; Dauphin, Y.N.; et al. mixup: Beyond empirical risk minimization. *arXiv* **2017**, arXiv:1710.09412.
80. Jocher, G.; Qiu, J.; Chaurasia, A. Ultralytics YOLO (Version 8.0.0). Available online: https://github.com/ultralytics/ultralytics (accessed on 5 May 2025).
81. Loshchilov, I. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
82. Digital Research Alliance of Canada. Digital Research Alliance of Canada, n.d. Available online: https://alliancecan.ca/en (accessed on 3 October 2024).
83. Gevorgyan, Z. SIoU Loss: More Powerful Learning for Bounding Box Regression. *arXiv* **2022**, arXiv:2205.12740.
84. Lin, T.Y.; Maire, M.; Belongie, S.; et al. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part V 13, pp. 740–755.
85. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
86. Li, C.; Li, L.; Jiang, H.; et al. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976.
87. Xu, K.; Qin, M.; Sun, F.; et al. Learning in the frequency domain. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1740–1749.
88. Selvaraju, R.R.; Cogswell, M.; Das, A.; et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.