remote sensing

MDPI

Article

DROMAL-Net: A Forest Larch Casebearer Detection Model Focusing on Global Feature Enhancement and Positional Dynamic Clustering

Jiaxuan Wang [1], Lu Liu [2], Yuyang Tang [1], Tao Ma [3], Xiangyu Song [4], Pan Qiao [5] and Zhen Ding [6,*]

[1] Aulin College, Northeast Forestry University, Harbin 150040, China; 2023224713@nefu.edu.cn (J.W.); 2023214664@nefu.edu.cn (Y.T.)
[2] Heilongjiang Natural Resources Rights and Interests Investigation and Monitoring Institute, Harbin 150040, China; lkdl_999@126.com
[3] School of Geomatics, Liaoning Technical University, Fuxin 123000, China; 472510052@stu.lntu.edu.cn
[4] Key Laboratory of Polar Environment Monitoring and Public Governance, Wuhan University, Ministry of Education, Wuhan 430079, China; songxiangyu@stdu.edu.cn
[5] Hebei Provincial Institute of Cartography, Shijiazhuang 050031, China; hbjgydqb@163.com
[6] School of Computer Science and Artificial Intelligence, Northeast Forestry University, Harbin 150040, China
* Correspondence: dingzhen@nefu.edu.cn

Highlights

What are the main findings?

- The proposed DROMAL-Net model achieves 74.12% precision, 72.67% recall, and 77.98% mAP on the SLFPD test set, improving by 2.75%, 1.72%, and 2.46%, respectively, over the baseline YOLOv11n. The small standard deviations across five random seed experiments confirm the effectiveness and stability of the proposed improvement strategies.
- Ablation studies demonstrate that the MALAPSA module improves mAP to 76.04%. Introducing DCCC3k further yields the largest performance gain, raising mAP to 77.36%. The complete model with 2D-RoPE achieves the performance of 77.98%. The three modules provide complementary contributions.

What are the implications of the main findings?

- The proposed MALAPSA linear attention module leverages discriminative information embedded in feature magnitudes to enhance background noise suppression while maintaining linear computational complexity and global feature extraction. This offers an efficient solution for global context modeling in remote sensing object detection.
- The combination of DCCC3k dynamic clustering convolution and 2D-RoPE effectively addresses the lack of inductive bias in self-attention mechanisms and the spatial uncertainty inherent in dynamic clustering. This approach is generalizable to other remote sensing detection tasks characterized by large scale variations and densely distributed small targets.

Academic Editor: Michael Sprintsin

Received: 9 April 2026
Revised: 4 August 2026
Accepted: 25 August 2026
Published: 31 August 2026

Abstract

Larch casebearer (*Coleophora laricella*) poses a serious threat to forest ecological security, and remote sensing object detection is essential for early warning and targeted intervention. However, existing detectors are hindered by insufficient feature extraction, limited self-attention inductive bias, and weak spatial localization, which collectively constrain their practical deployment. To address these issues, we propose DROMAL-Net, which integrates global feature enhancement and positional dynamic clustering for forest pest detection. Built upon the YOLOv11n baseline, our model incorporates three key innovations. First, we design the MALAPSA module by embedding a magnitude-aware linear

attention (MALA) mechanism into the feature extraction pipeline, achieving linear computational complexity while preserving global contextual modeling. Second, to address the inductive bias deficiency of C2PSA and better preserve spatial topology, we introduce C3kDR (DCCC3k), a positional dynamic clustering module that combines dynamic clustering convolution with complex-domain positional compensation to enhance multi-scale localization accuracy. Third, we integrate 2D Rotary Position Embedding (2D-RoPE) to further mitigate spatial ambiguity in dynamic clustering.Extensive experiments on the public SLFPD dataset validate the effectiveness of each component. Under the standard random partition, DROMAL-Net achieves 77.98% mAP@0.5, outperforming YOLOv11n by 2.46 percentage points. Furthermore, under a strict leave-one-block-out protocol to prevent spatial leakage, DROMAL-Net consistently outperforms YOLOv11n across all five held-out blocks, achieving a macro-average mAP@0.5 of 69.12% vs. 65.94%, corresponding to a 3.18 percentage point improvement, which confirms its strong generalization capability to unseen geographic regions. Results across five random seeds yield consistently low standard deviations, confirming the stability of these gains. We further evaluate generalization on the TreeFinder dataset, which encompasses diverse tree species and complex canopy structures, where DROMAL-Net maintains robust performance, underscoring the effectiveness of MALAPSA and C3kDR for global feature enhancement and accurate localization in cross-regional applications.

Keywords: larch casebearer; object detection; positional dynamic clustering; global feature enhancement

1. Introduction

Among the numerous pests and diseases of the forest, the larch casebearer (*Coleophora laricella*) is recognized as one of the most destructive defoliators of larch forests [1]. Its larvae cause persistent defoliation by feeding on needles, leading to reduced tree growth, crown structure degradation, and declining forest health [2]. This pest has experienced long-term outbreaks in North America and Eurasia, posing serious threats to forest productivity, carbon sink functions, and ecosystem stability [3]. Therefore, timely and accurate monitoring of larch casebearer damage is of great importance for forest resource protection and pest prevention and control.

However, in the face of increasingly severe pest and disease outbreaks, the limitations of traditional monitoring methods have become progressively more pronounced. Conventional forest pest monitoring has been largely based on manual patrols and field sampling [4–6]. Although this approach enables intuitive assessments of tree health through observation of symptoms [7], it is constrained by complex terrain, subjective judgment, and high labor costs, making it applicable only to small-scale areas [6,8]. Consequently, it does not meet the requirements for timely responses to pests and diseases characterized by rapid spread and strong concealment. To address the issues of high costs and limited coverage associated with manual methods, a target detection technology based on remote sensing imagery and computer vision has emerged. This technology utilizes remote sensing platforms to achieve efficient data acquisition [9] and, when integrated with intelligent analytical methods, significantly enhances the spatiotemporal coverage of forest monitoring [10]. It enables the rapid identification of large-scale forest damage [11], providing critical support for scientific prevention and control strategies. In recent years, remote sensing technology, leveraging its distinct advantages in large-scale, multi-scale, and periodic monitoring capabilities [4,12], has been widely adopted, driving the evolution of forest health monitoring towards greater efficiency and precision.

Consequently, the academic community has begun to explore more intelligent approaches to data processing. Starting with conventional image processing techniques and shallow machine learning models, researchers have attempted to achieve automated pest and disease identification through spectral analysis and feature engineering. The optimized Normalized Difference Vegetation Index (NDVI) was combined with object-based image analysis (OBIA) classification techniques to classify five infestation levels caused by oak splendour beetle [13]. Minařík et al. [14] integrated Individual Tree Crown Delineation (ITCD) with Support Vector Machines (SVM) to extract spectral and three-dimensional features from high-density point clouds, successfully distinguishing between healthy, green-attacked, and damaged trees, thus providing an automated solution for precise forest health monitoring. However, although shallow machine learning methods perform well in specific datasets, their limitations become obvious when applied to large-scale applications in complex forest scenarios. Traditional methods rely on handcrafted features, which refer to the process of manually extracting low-level features (e.g., color histograms, SIFT, SURF, Gabor) and combining them with traditional classifiers for classification [15]. This process is time-consuming and subjective, making it difficult to obtain the optimal feature combination [16]. Furthermore, handcrafted features struggle to comprehensively capture pest-related information, handle noise in real-world images, or distinguish between closely related species with similar appearances [15], and they also lack robustness against common image disturbances such as rotation, scaling, translation, and viewpoint changes [17]. Therefore, deep learning techniques, which possess the capacity for automatic feature learning and high-level abstract representation, demonstrate greater potential in addressing the complexities of multi-source forest pest and disease monitoring tasks involving heterogeneous data from different sensors such as optical imagery [18], multispectral [19], LiDAR [19], and others.

With the rapid development of deep learning technology, deep learning methods based on computer vision have become a major research direction in the field of forest pest and disease detection. According to their detection paradigms, existing object detection methods can generally be divided into two-stage detectors and one-stage detectors. In two-stage detection frameworks, the first stage generates candidate object regions through a region proposal network, while the second stage performs refined classification and bounding-box regression on the proposed regions. Ma et al. [20] improved Faster R-CNN for the identification of the crown of pine wilt disease by introducing Inception-ResNet-V2 and an optimized attention mechanism. To address the problems of high model complexity and insufficient detection accuracy, Li et al. [17] and Guan et al. [21] further improved Faster R-CNN. Li et al. proposed a lightweight LLA-RCNN model that integrates a MobileNetV3 backbone, a coordinate attention module, GIoU loss, and RoI Align to reduce computational complexity and improve feature extraction. Guan et al. proposed the GC-Faster-RCNN model, which incorporates a hybrid attention mechanism that combines GCT and CBAM into the ResNet50+FPN backbone and employs EIoU loss together with a cosine annealing learning-rate strategy.

In contrast, one-stage detectors directly predict object categories and locations without generating region proposals, thus significantly improving detection efficiency while maintaining competitive detection accuracy [22]. Among them, the YOLO series [23,24] has become the most representative lightweight detection framework for forest pest and disease detection due to its favorable balance between detection accuracy and inference speed. With the continuous evolution of the YOLO architecture, representative models have progressively advanced the development of lightweight detectors. Early models, such as YOLOv3 [25] and YOLOv7-Tiny [26], laid the foundation for forest pest and disease detection on resource-constrained platforms due to their efficient network structures and

real-time inference capabilities. Subsequently, YOLOv8, YOLOv9, YOLOv10, and newly released YOLOv11n further optimized network architectures, feature representations, and training strategies, consistently improving detection performance while maintaining a lightweight design. Among these, YOLOv11n, as the latest lightweight version released by Ultralytics in 2024, achieves an excellent balance between detection accuracy, inference speed, and model complexity, and is therefore selected as the baseline model for this study.

To adapt to complex forest remote sensing scenarios, researchers have made extensive improvements to YOLO in terms of optimizing the bounding box regression, attention mechanisms, and dynamic receptive field design. For example, introducing the CIoU loss function improves the object localization capability and improves the recognition accuracy of pine wilt disease, yet missed detections and localization deviations still occur frequently in complex forest canopy backgrounds [9]. To strengthen feature representation in complex backgrounds, modules such as SE, CBAM, BiFPN, and DynamicConv have been integrated into the YOLO framework, effectively improving the discriminability of pest and disease targets, as well as the performance of the deployment on edge devices [27]. Dynamic convolution and adaptive receptive field mechanisms have also been adopted to adjust sampling positions according to the spatial distribution of disease regions, thus improving the detailed representation of small-scale disease targets [28]. In recent years, improved task-oriented models, including YOLO-DP [29], YOLO-lychee [30], and CPD-YOLO [31], have further validated the effectiveness of targeted feature enhancement strategies for specific pest and disease detection tasks.

Nevertheless, most existing lightweight detection methods primarily improve individual network components, such as convolutional blocks, attention mechanisms, and feature fusion strategies, to enhance feature extraction efficiency while reducing computational complexity [32–35]. Although these strategies have achieved promising results in generic object detection, they generally focus on optimizing a single aspect of feature representation, with limited consideration of the coordinated optimization of global semantic modeling, local detail representation, and spatial localization capability within a unified lightweight framework. This limitation becomes particularly evident in UAV-based forest remote sensing imagery, where pest and disease symptoms are usually manifested as subtle spectral variations, slight canopy discoloration, local texture degradation, or structural anomalies rather than objects with explicit boundaries [36,37]. In addition, forest scenes are inherently challenging due to dense canopy occlusion, high spectral and textural similarity between vegetation, illumination variations, seasonal phenological changes, and substantial target scale variation, all of which significantly increase the difficulty of accurate object detection and localization in optical remote sensing imagery [38–40]. Consequently, existing lightweight detectors still struggle to simultaneously preserve fine-grained disease characteristics, capture long-range contextual dependencies, and maintain robust spatial localization under complex forest conditions, leading to suboptimal performance in forest pest and disease detection tasks.

Therefore, improving discriminative feature representation in complex forest backgrounds, enhancing the detection of multi-scale small objects, and increasing localization robustness remain key research challenges in forest pest and disease image detection.

1.1. Research Question

RQ1: How can we simultaneously achieve local detail extraction and global context modeling to suppress complex background noise?

RQ2: How can we overcome the lack of inductive bias inherent in self-attention mechanisms to handle the challenges of drastic scale variation and dense distribution of targets in UAV imagery?

RQ3: How can we eliminate the spatial position ambiguity introduced by dynamic clustering mechanisms while leveraging their benefits for multi-scale representation enhancement?

1.2. Contributions

To address the above issues and limitations, this study proposes a forest pest detection model named DROMAL-Net based on a specific dataset (SLFPD) for the larch casebearer, and to fully leverage the discriminative information embedded in feature magnitudes to enhance background noise suppression, we construct a lightweight module named MALAPSA by embedding the MALA linear attention mechanism. Based on the YOLOv11n baseline, DROMAL-Net integrates the linear attention module (MALAPSA) and the Positional Dynamic Clustering convolution module (C3kDR). The C3kDR module is a novel module built on top of our proposed DCCC3k with 2D-RoPE embedded. The main contributions of this paper are summarized as follows:

1. To address the challenges of insufficient feature extraction capability and limited global context modeling, and to fully leverage the discriminative information embedded in feature magnitudes to enhance background noise suppression, we construct a lightweight module named MALAPSA by embedding the MALA linear attention mechanism.
2. To tackle the challenges of large scale variation and dense distribution of small targets, and to further address the lack of inductive bias in the self-attention mechanism, we design the DCCC3k module inspired by Dynamic Clustering Convolution, which enhances the feature representation capability for multi-scale small objects.
3. We introduce two-dimensional Rotary Position Embedding (2D-RoPE) into the designed DCCC3k module to mitigate the inherent spatial uncertainty in the dynamic clustering mechanism.
4. Extensive experiments on the UAV remote sensing dataset (SLFPD) and the TreeFinder multispectral dataset validate the effectiveness of DROMAL-Net.

2. Materials and Methods

2.1. Experimental Environment

The experiments were conducted on a 64-bit Windows 10 operating system, utilizing Python (v3.8.0) and the PyTorch (v1.10.0) deep learning framework for model training and evaluation. The hardware configuration includes an NVIDIA GeForce RTX 4090 GPU with 24 GB of memory (NVIDIA Corporation, Santa Clara, CA, USA), provided through the AutoDL cloud computing platform (operated by SeetaCloud Technology Co., Ltd., Nanjing, China), ensuring efficient computation and the capacity to handle large-scale data. All experiments were performed under this unified setup to maintain consistency and ensure the comparability of results. Table 1 lists the key hyperparameters used in model training and evaluation. All models are trained from scratch without pretrained weights. Data augmentation strategies, including mosaic, color jittering, and horizontal flip, are applied during training to improve generalization. Evaluation settings, including IoU threshold, confidence threshold, and NMS IoU threshold, are also specified.

Table 1. Hyperparameter settings for the model training phase.

Parameter	Value
Image Size	640×640
Training Epochs	100
Batch Size	8

Table 1. *Cont.*

Parameter	Value
Optimizer	AdamW
Initial Learning Rate	2×10^{-4}
Weight Decay	5×10^{-4}
Learning Rate Schedule	Cosine Annealing
Gradient Clipping	Max norm clipped to 10.0
Mosaic Augmentation	Probability $p = 0.2$ (training only)
Color Augmentation	YOLO-style (hue, saturation, brightness)
Horizontal Flip	Enabled (default probability)
Mixed Precision	Automatic Mixed Precision (AMP) on GPU
Pretrained Weights	None (trained from scratch)
IoU Threshold for Matching	0.5 (evaluation)
Confidence Threshold	0.001 (evaluation)
NMS IoU Threshold	0.5 (evaluation)

2.2. Evaluation Metrics

To comprehensively evaluate the proposed forest pest and disease detection model, four metrics were employed: Recall (R), Precision (P), mean Average Precision (mAP), and $F1$ score. The corresponding calculation formulas are presented in Equations (1)–(5).

$$P = \frac{TP}{TP + FP} \tag{1}$$

$$R = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = \frac{2 \cdot P \cdot R}{P + R} \tag{3}$$

$$AP = \int_0^1 P(R)\, dR, \quad mAP@0.5 = \frac{1}{Q} \sum_{q=1}^{Q} AP_q^{\text{IoU}=0.5} \tag{4}$$

$$mAP@0.5:0.95 = \frac{1}{10} \sum_{\substack{\text{IoU}=0.50 \\ \text{step } 0.05}}^{0.95} mAP@\text{IoU} \tag{5}$$

In detection tasks, TP denotes the number of correctly detected positive samples, FP represents the number of negative samples incorrectly identified as positive, and FN indicates the number of positive samples that were missed. Precision measures the proportion of correctly predicted positive instances among all instances classified as positive, while Recall reflects the proportion of actual positive instances that were correctly identified. The $F1$ score, as the harmonic mean of Precision and Recall, provides a comprehensive assessment of model performance. mAP, calculated as the mean of Average Precision (AP) across all categories, evaluates the overall detection effectiveness of the model across the entire dataset.

2.3. Datasets Introduction

2.3.1. SLFPD Datasets

The SLFPD dataset (Swedish Larch Forest Pest Dataset) utilized in this study comprises standard RGB aerial images acquired from larch forests in Västra Götaland County, Sweden, intended for the identification of tree species and the assessment of the severity of infestation caused by the larch casebearer (*Coleophora laricella*). The images were taken on two days, 27 May and 19 August 2019, in five regions of Västergötland (Bebehojd,

Ekbacka, Jallasvag, Kampe, and Nordkap). The dataset consists of 835 images that span three categories, with a subset of larch samples annotated according to fine-grained health levels reflecting the degree of infestation by larch casebearers, namely healthy, lightly infested, and severely infested. Captured in various scenarios, the dataset exhibits a spatial resolution of 1500 × 1500 pixels. The dataset was partitioned into training, validation, and test sets according to the distribution provided with the public YOLO-format release: 501 training images, 167 validation images, and 167 test images, corresponding to a ratio of approximately 60%/20%/20%. All reported results are evaluated on the test set.

VDVI [41] simulates the logic of NDVI and highlights canopy color anomalies caused by chlorophyll loss. The Laplacian texture map is sensitive to grayscale abrupt changes (edges, corners) in the image, which is crucial for delineating tree crown contours and identifying structural degradation caused by disease. NGBDI [42] leverages the characteristics of healthy vegetation, namely strong absorption of blue light and strong reflection of green light, to assess the health status of vegetation through the difference between the green and blue bands. The corresponding calculation formulas are presented in Equations (6)–(8).

$$VDVI = \frac{2G - R - B}{2G + R + B} \tag{6}$$

$$T(x, y) = \left|\nabla^2 I(x, y)\right| = \left|\frac{\partial^2 I}{\partial x^2} + \frac{\partial^2 I}{\partial y^2}\right| \tag{7}$$

$$NGBDI = \frac{G - B}{G + B} \tag{8}$$

As shown in Figure 1, VDVI and NGBDI can explicitly extract disease-sensitive signals, effectively distinguishing infected trees from healthy ones, and differentiating background elements such as shadows or grasslands from true tree crowns. In contrast, the Laplacian texture map focuses more on local canopy sharpness and structural integrity, highlighting structural deterioration phenomena such as sparse foliage or branch dieback.

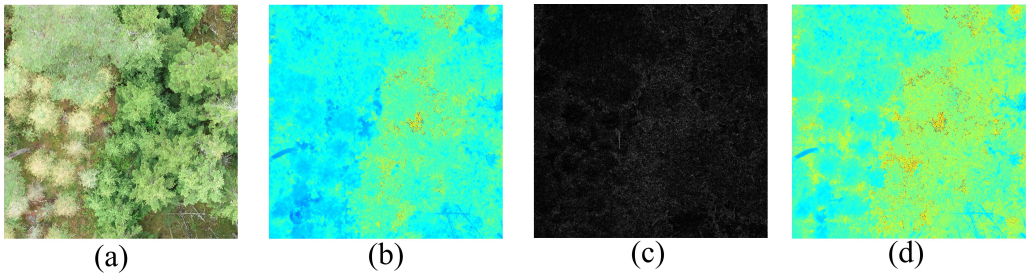

Figure 1. Multi-dimensional feature visualization for SLFPD dataset: (**a**) original RGB image; (**b**) VDVI output; (**c**) texture feature map; (**d**) NGBDI output.

2.3.2. TreeFinder Datasets

TreeFinder [43] is constructed based on aerial imagery from the National Agriculture Imagery Program (NAIP), achieving a spatial resolution of 0.6 m and covering four spectral bands: red, green, blue, and near-infrared. The dataset spans 48 states across the Contiguous United States (CONUS), comprising 1000 sampling sites with a total annotated area exceeding 23,000 hectares. It includes 15,489 image patches of 224 × 224 pixels each, along with pixel-level segmentation masks of dead trees. All reported results are evaluated on the test set. Figure 2 illustrates the multispectral visualization examples of the TreeFinder dataset.

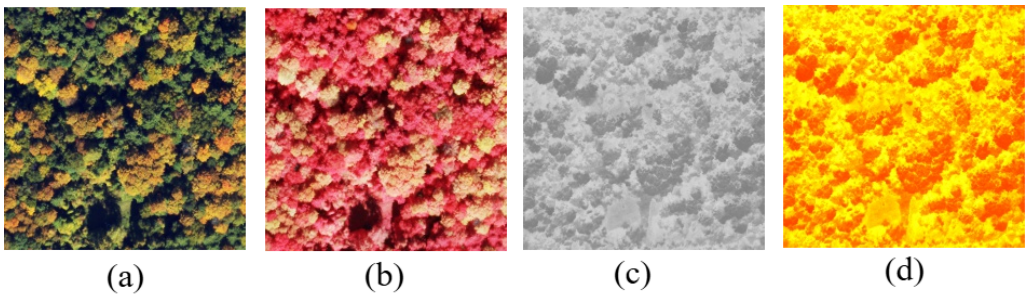

Figure 2. Multispectral visualization examples of scattered individual tree mortality from the TreeFinder dataset, using four spectral representations: (**a**) true-color (RGB); (**b**) false-color (NIR-R-G), a common remote sensing approach that highlights healthy vegetation in red; (**c**) grayscale NDVI map; (**d**) colored NDVI map, where dark red regions indicate dead trees and bright yellow regions indicate healthy vegetation.

For detection training with DROMAL-Net, each TreeFinder binary segmentation mask was converted to object-level bounding boxes using a deterministic procedure. Specifically, each mask was normalized to [0,1] and binarized at a threshold of 0.5, followed by 8-connected-component labeling. For every retained connected component, the tight axis-aligned bounding box was generated from its minimum and maximum row and column coordinates. Components smaller than 4 pixels, or boxes with width or height below 2 pixels, were removed as annotation noise to filter out trivial fragments. Spatially disconnected fragments were treated as separate objects, while adjacent dead-tree regions that remained disconnected in the binary mask were retained as separate boxes; touching regions forming a single connected component were represented by one enclosing box. No manual splitting or merging was performed. This preprocessing pipeline ensures that the detection training is reproducible and consistent with the object detection paradigm of DROMAL-Net.

DROMAL-Net was trained separately on TreeFinder from random initialization using only the TreeFinder training partition. No SLFPD-trained weights or SLFPD images were used for TreeFinder training, validation, or testing. The TreeFinder results therefore measure within-dataset learning and the region-held-out robustness defined for TreeFinder, rather than zero-shot transfer from SLFPD.

2.3.3. Development of DROMAL-Net

YOLOv11 represents a recent advancement in the YOLO series of object detection models. Based on variations in network depth and width, it encompasses five distinct architectures: YOLOv11n, YOLOv11l, YOLOv11s, YOLOv11x and YOLOv11m. Considering the real-time requirements of forest pest and disease detection, this study selects the lightweight YOLOv11n model as the baseline and subsequently introduces optimizations. The architecture of DROMAL-Net is shown in Figure 3. DROMAL-Net is primarily composed of four components: the Input Layer, the backbone layer, the neck layer, and the output layer.

The primary function of the input layer is to receive raw remote sensing pest and disease images and process them through a series of data augmentation operations to meet the requirements of model training. These operations include hue adjustment, image resizing, and the application of Mosaic data augmentation. The Mosaic augmentation technique operates by randomly selecting regions from four distinct images, which are then subjected to random cropping and scaling before being combined to form a new composite image.

The backbone layer is primarily responsible for extracting salient features from the input image and comprises four modules: Conv, C3kDR, MALAPSA, and SPPF. The Conv

module performs feature transformation through convolutional operations, followed by batch normalization (BN) and the SiLU activation function. The C3kDR module, built upon C3k2 to reconfigure its original units within the baseline model, constructs its core C3k as the Dynamic Clustering Convolution module (DCCC3k) to achieve a dynamic receptive field and multi-scale feature representation. This module inherits the channel splitting and variable kernel size bottleneck units from the C2f structure, enhancing multi-scale modeling capability by dynamically aggregating similar features. Meanwhile, two-dimensional Rotary Position Embedding (2D-RoPE) is introduced to explicitly encode spatial position information, effectively improving the model's localization robustness for local details. The MALAPSA module integrates MALA linear attention into the C2PSA structure. This module inherits the cross-stage feature fusion paradigm from the C2f architecture and, through channel splitting, convolutional transformations, and a parallel self-attention mechanism, achieves near-Softmax global modeling capability while maintaining linear complexity, effectively promoting the joint extraction of local and global features. MALAPSA is deployed after the SPPF module at the end of the backbone, where the features already possess rich semantic information and a large receptive field. Therefore, MALAPSA explicitly models the feature magnitude and applies it to high-level semantic features, enabling the response strength to more accurately discriminate disease-affected regions. In contrast to the SPP module employed in earlier YOLO versions, the SPPF module applies three sequential pooling operations to reduce computational overhead while preserving multi-scale feature fusion and expanding the effective receptive field.

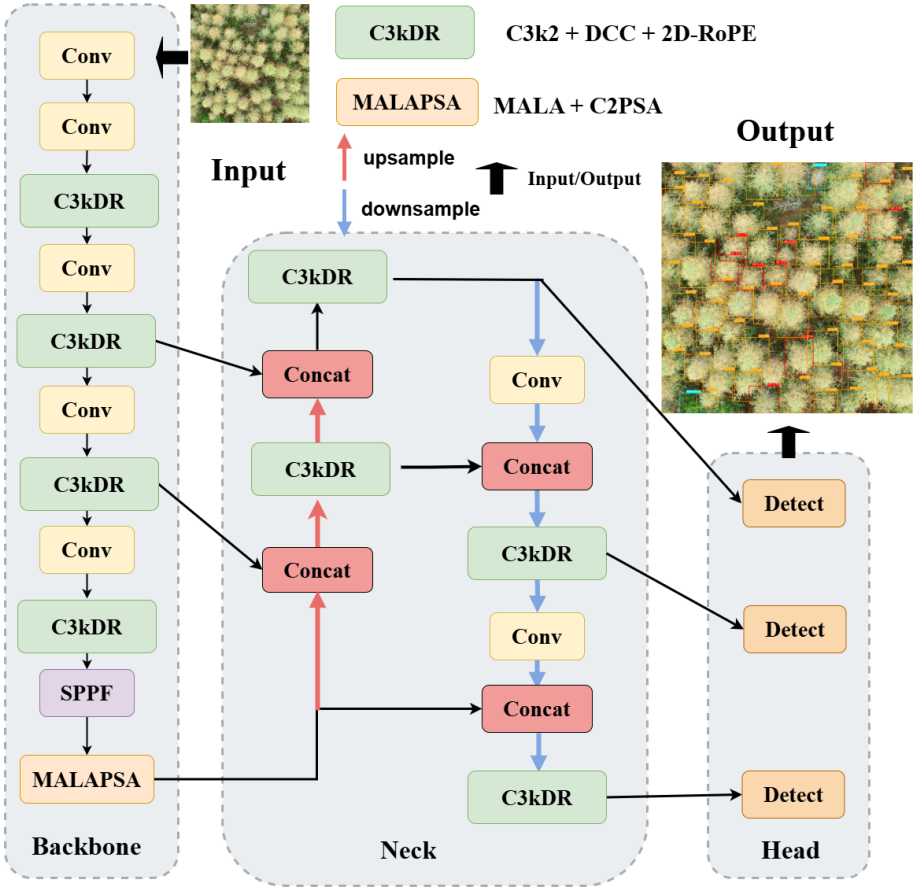

Figure 3. Overall workflow based on the YOLOv11 architecture. MALAPSA integrates MALA linear attention into C2PSA for near-Softmax global modeling at linear complexity, while C3kDR incorporates Dynamic Clustering Convolution (DCC) for dynamic receptive fields and multi-scale representation, further enhanced by 2D-RoPE for robust spatial localization.

The core function of the Neck layer is to enable cross-dimensional integration of feature representations. Through the utilization of Feature Pyramid Network (FPN) and Path Aggregation Network (PAN) structures, it efficiently merges feature maps from different levels, ensuring the precise preservation of spatial information. This layer directs the model's focus toward target-specific features, thereby substantially enhancing detection performance. The C3kDR module is deployed after every feature fusion node to facilitate the deep refinement and optimized propagation of cross-stage features.

The output layer is responsible for generating the final object detection results. It utilizes the refined feature maps produced by the neck layer to predict bounding box coordinates, class probabilities, and other requisite information for each feature map. Additionally, the output layer applies Non-Maximum Suppression (NMS) to eliminate redundant detections, thereby preserving accurate predictions.

2.4. The Proposed DROMAL-Net Network

2.4.1. MALAPSA (C2PSA with Magnitude-Aware Linear Attention)

To address the issues of insufficient feature extraction capability and limited global context modeling in remote sensing images, the traditional linear attention mechanism, although capable of approximating a global receptive field through kernel functions, suffers from uniformly distributed attention weights due to the absence of a Softmax-like exponential scaling mechanism. This limitation prevents it from effectively capturing the detailed features of objects.

To address this efficiency bottleneck, this study proposes an efficient feature extraction module based on a linear attention mechanism—MALAPSA. The overall structure of the module is shown in Figure 4. The MALA [44] mechanism employed in this module introduces query magnitude information while maintaining linear complexity, enabling linear attention to more closely approximate the weight distribution of Softmax, thereby enhancing attention quality without increasing computational cost. Specifically, this module integrates the linear attention mechanism MALA into the C2PSA structure of the backbone network by modifying the computation of linear attention to fully incorporate the magnitude information of the Query. This adjustment enables MALA to generate an attention score distribution that closely resembles Softmax attention while exhibiting a more balanced structure. As a result, it achieves global modeling capability close to that of Softmax attention with low computational overhead, based on linear complexity.

Magnitude-Aware Linear Attention (MALA) builds upon the original linear attention by introducing a scaling factor and an offset term, while replacing division-based normalization with addition-based normalization.

To address this issue, MALA introduces a scaling factor β and an offset term γ based on the original linear attention, explicitly incorporating the query magnitude information into the attention weight computation. Specifically, the MALA score is defined as:

$$\mathrm{Attn}(Q_i, K_j) = \beta\phi(Q_i)\phi(K_j)^T - \gamma \tag{9}$$

where the parameters β and γ are dynamically determined based on the interaction between the query and all keys:

$$\beta = 1 + \frac{1}{\phi(Q_i)\sum_{m=1}^{N}\phi(K_m)^T} \tag{10}$$

$$\gamma = \frac{\phi(Q_i)\sum_{m=1}^{N}\phi(K_m)^T}{N} \tag{11}$$

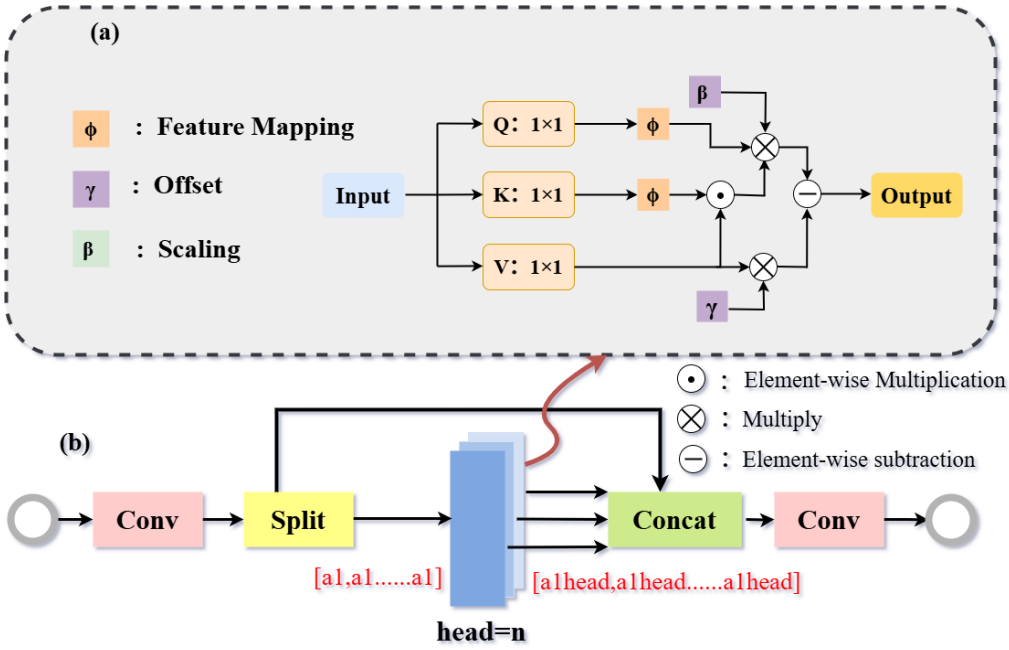

Figure 4. The structure of MALAPSA and Magnitude-Aware Linear Attention: (**a**) the structure of Magnitude-Aware Linear Attention; (**b**) the structure of MALAPSA.

This design achieves normalization through addition rather than traditional division, ensuring that the attention scores satisfy $\sum_{j=1}^{N} \mathrm{Attn}(Q_i, K_j) = 1$.

MALA retains the core advantages of linear attention. Its complete output formulation can be rewritten as:

$$Y_i = \beta\phi(Q_i) \sum_{j=1}^{N} \phi(K_j)^T V_j - \gamma \sum_{j=1}^{N} V_j \tag{12}$$

When considering all attention scores as positive values, the ratio of Q_i's attention scores for K_m and K_n is given by:

$$\frac{\beta\phi(Q_i)\phi(K_m)^T - \gamma}{\beta\phi(Q_i)\phi(K_n)^T - \gamma} = p;$$

We assume that Q_i assigns a higher attention score to K_m, i.e., $\beta\phi(Q_i)\phi(K_m) > \beta\phi(Q_i)\phi(K_n)$ and $p > 1$. When the direction of $\phi(Q_i)$ remains unchanged and its magnitude is scaled by a factor of $a > 1$, it is straightforward to derive that the new β and γ can be written as:

$$\beta_{new} = \frac{\beta + a - 1}{a}, \tag{13}$$

$$\gamma_{new} = a\gamma; \tag{14}$$

At this point, the ratio of the attention scores of $a\phi(Q_i)$ for K_m and K_n becomes:

$$\frac{\beta_{new} a\phi(Q_i)\phi(K_m)^T - \gamma_{new}}{\beta_{new} a\phi(Q_i)\phi(K_n)^T - \gamma_{new}} = \frac{\beta\phi(Q_i)\phi(K_m)^T - \frac{a\beta}{\beta+a-1}\gamma}{\beta\phi(Q_i)\phi(K_n)^T - \frac{a\beta}{\beta+a-1}\gamma} = p_m; \tag{15}$$

Since $\beta > 1$ and $a > 1$, it is straightforward to prove that $\frac{a\beta}{a+\beta-1} > 1$. From this, we can further easily prove that when considering all attention scores as positive values, $p_m > p$. Moreover, since $\sum_{j=1}^{N} \mathrm{Attn}(Q_i, K_j) = 1$, as the magnitude of $\phi(Q_i)$ increases, MALA concentrates more attention on the keys that originally received higher attention, while

allocating less attention to the keys that originally had lower attention. This behavior is similar to Softmax Attention.

In summary, by introducing a magnitude-aware mechanism, MALA successfully reintegrates query value information into linear attention, enabling it to exhibit more reasonable dynamic distribution characteristics while maintaining linear complexity. This provides a feasible solution for pest detection in resource-constrained scenarios.

2.4.2. C3kDR (C3k2 with DCCC3k and 2D-RoPE)

To enhance multi-scale feature aggregation while preserving spatial topology in complex canopy scenes, we propose C3kDR, a C3k2-style module that integrates two complementary components: dynamic clustering convergence (DCCC3k) and 2-dimensional rotational position encoding (2D-RoPE). The overall design philosophy is to adaptively group semantically similar pixels to improve local feature representation, while simultaneously compensating for the positional ambiguity introduced by such dynamic grouping.

- DCCC3k (C3k with Dynamic Clustering Convolution)

While the self-attention mechanism is capable of capturing global semantic correlations, it often requires training on large-scale datasets due to the lack of inductive biases inherent. Therefore, the dynamic convolutional clustering method is adopted to combine the advantages of both approaches.

To effectively address this challenge, this paper proposes an improved feature extraction module, DCCC3k, whose structure is shown in Figure 5. Inspired by the idea of Dynamic Clustering Convolution [45], this module is integrated into the C3k2 structure of the backbone network to enhance the extraction of small target features through dynamic clustering convolution. Its workflow is as follows: after the input feature map undergoes channel compression through two parallel branches, Branch 1 passes through n stacked Bottleneck_DCC modules, each of which employs DClusConv to globally aggregate semantically similar feature points. Branch 2 serves as a direct residual connection. Finally, the outputs of the two branches are concatenated along the channel dimension and integrated through a DClusConv fusion head to generate enhanced features.

Specifically, Dynamic Clustering Convolution dynamically builds clustering neighborhoods in the feature space to aggregate features from semantically relevant regions, thereby enabling dynamic expansion of the receptive field and enhancement of fine-grained contextual features. This allows DCCC3k to effectively capture multi-scale features and model contextual semantic relationships, consequently improving the representation capability for small objects and their semantic associations.

The DCCC3k module enhances the model's capability for small target detection through a configurable kernel size parameter k. Given an input feature map $\mathbf{X} \in \mathbb{R}^{C_1 \times H \times W}$, the output $\mathbf{Y} \in \mathbb{R}^{C_2 \times H \times W}$ is computed as:

$$\mathbf{Y} = \mathcal{F}_{\text{DClusConv}}\left(\left[\mathcal{F}_{\text{Bottleneck\_DCC}^{(n)}}(\text{Conv}_{1\times1}(\mathbf{X})), \text{Conv}_{1\times1}(\mathbf{X})\right]\right), \tag{16}$$

where $\text{Conv}_{1\times1}(\cdot)$ reduces the channel dimension by a compression factor $e = 0.5$, $\mathcal{F}_{\text{Bottleneck\_DCC}^{(n)}}(\cdot)$ denotes n stacked Bottleneck_DCC modules (each with $k \times k$ convolutions), $[\cdot,\cdot]$ denotes channel-wise concatenation, and $\mathcal{F}_{\text{DClusConv}}(\cdot)$ is the dynamic clustering convolution used as the fusion head with a 3×3 kernel. We set $k = 3$ to provide a standard receptive field suitable for small target detection in aerial imagery.

The Bottleneck_DCC adopts a residual structure consisting of two consecutive DClusConv layers. Given an input feature map $\mathbf{X} \in \mathbb{R}^{C_{\text{in}} \times H \times W}$, the output $\mathbf{Y} \in \mathbb{R}^{C_{\text{out}} \times H \times W}$ is computed as:

$$\mathbf{Y} = \mathcal{F}_{\text{DClusConv}}(\sigma(\mathcal{F}_{\text{DClusConv}}(\mathbf{X}))) + \alpha\mathbf{X}, \tag{17}$$

where $\mathcal{F}_{\text{DClusConv}}(\cdot)$ denotes the dynamic clustering convolution, $\sigma(\cdot)$ is the StarReLU activation, and $\alpha = 1$ when shortcut is enabled and $C_{\text{in}} = C_{\text{out}}$; otherwise $\alpha = 0$. The channel compression factor $e = 0.5$ is applied within each DClusConv to reduce computational cost.

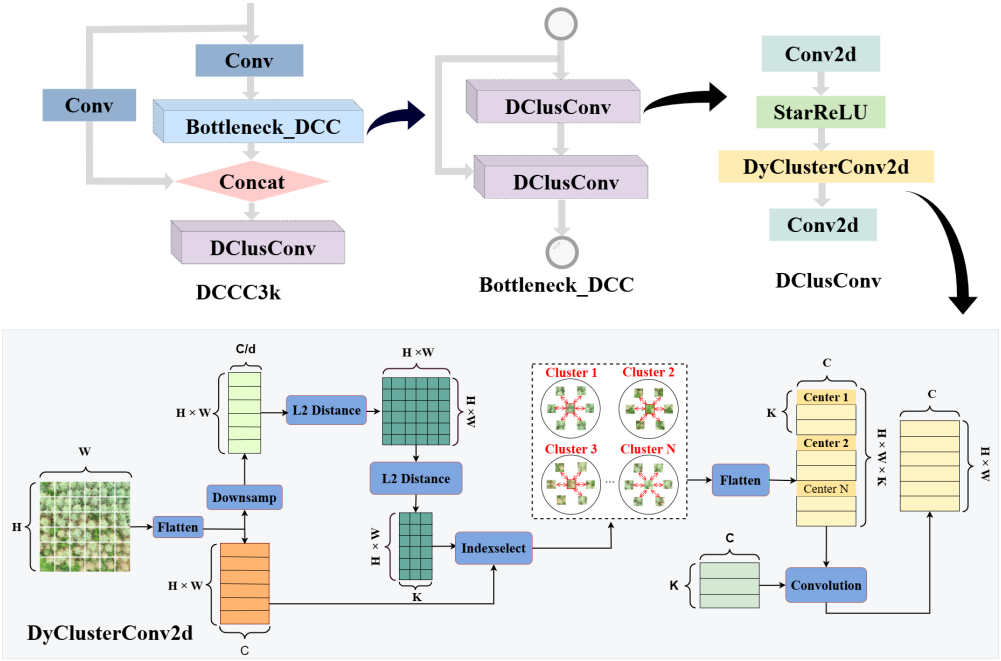

Figure 5. Illustration of DCCC3k. H: height of the feature map; W: width of the feature map; C: feature dimension; d: downsampling interval; K: convolution kernel size of dynamic clustering convolution; Flatten: flattening the feature vectors along the spatial dimensions; Downsampling: extracting a set of sub-vectors for each patch; IndexSelect: selecting feature vectors using the index of clusters.

The specific implementation process of DClusConv is as follows: First, the input features are projected into the feature space, and the Euclidean distance between cluster centers and other feature points is computed, yielding a distance matrix $\mathbf{M}$. This process can be formulated as:

$$distance(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^2, \quad \mathbf{M} \in \mathbb{R}^{n \times n} \tag{18}$$

where $\mathbf{x}$ is the cluster center feature, $\mathbf{y}$ is other patch features, and n is the number of cluster centers. Based on this matrix, the model selects the $K - 1$ feature points closest to each cluster center through a Top-K operation, constructing dynamic clustering neighborhoods:

$$idx = TopkIndex(\mathbf{M}), \quad idx \in \mathbb{R}^{n \times K} \tag{19}$$

where idx represents the indices of patches within each cluster. Using these indices, we rearrange the patch features to obtain a feature tensor convenient for subsequent convolution:

$$\mathbf{X}' = IndexSelect(idx, \mathbf{X}), \quad \mathbf{X}' \in \mathbb{R}^{n \times K \times D} \tag{20}$$

where $\mathbf{X}$ is the original patch features and D is the feature dimension. To address the computational complexity $\Omega(DC) = h^2w^2C + 2hwC$ associated with high-resolution images, we adopt an efficient dynamic clustering strategy using sub-vectors:

$$\Omega(EDC) = \frac{h^2w^2}{d}C + \frac{2hw}{d}C \tag{21}$$

where h and w are the height and width of the feature map, C is the feature dimension, and d is the sampling interval for sub-vectors (set to 8 in our model). Finally, we perform convolution on the obtained clusters using shared depthwise separable convolution kernels:

$$Y_{i,j,c} = \sum_{k=0}^{K-1} \mathbf{X}'_{i*w+j,k,c} \cdot W_{k,c} + b_c \tag{22}$$

where Y is the output feature, W and b are the weight and bias of the convolution kernels, i and j are spatial coordinates, and k is the coordinate within the cluster. To further enhance local feature expression, we incorporate a 3×3 depthwise separable convolution into the feed-forward network:

$$\mathbf{Y} = DWConv(Linear^{C \to 4C}(\mathbf{Y})) \tag{23}$$

$$\mathbf{Y} = Linear^{4C \to C}\{\mathrm{GELU}(\mathbf{Y})\} \tag{24}$$

where $Linear^{4C \to C}$ denotes a linear layer with input channels of $4C$ and output channels of C, and GELU is the activation function.

- 2D-RoPE (2-Dimensional Rotary Position Embedding)

Although the DCCC3k module implicitly captures local spatial neighborhood relationships through a clustering mechanism, the inherent spatial stochasticity of cluster assignment makes it difficult to explicitly preserve topological continuity and geometric associations between adjacent regions. In other words, the local organization induced solely by clustering is insufficient to reliably characterize the relative spatial dependencies of pest and disease lesions.

To alleviate this limitation, we further introduce two-dimensional Rotary Position Embedding (2D-RoPE) to explicitly encode relative positional information in the two-dimensional spatial domain and inject geometric priors into feature representations (as illustrated in Figure 6. Due to the insufficient localization accuracy of the original model, we explicitly encode spatial relative relationships by incorporating positional information and leveraging Euler's formula to add magnitude information in the two-dimensional coordinate space. By doing so, the model is encouraged to perceive structured spatial distribution patterns of lesions along both horizontal and vertical directions, thereby enhancing its ability to represent local morphological variations, boundary shifts, and complex spatial layouts. This design improves both localization accuracy and robustness, while enriching the spatial semantics of structured event representations.

Specifically, given an input voxel feature tensor $\mathbf{V} \in \mathbb{R}^{2N \times H \times W}$, where $2N$ denotes the number of channels and H and W denote the spatial height and width, respectively, we first partition the channel dimension into a set of two-dimensional rotational subspaces and map the real-valued representation into the complex domain. This formulation enables spatial positional information to be encoded explicitly through phase rotations on the complex plane. For the k-th frequency component, the corresponding frequency factor is defined as

$$\omega_k = \mathrm{base} \cdot \frac{2k}{C}, \quad k = 0, 1, \ldots, \frac{C}{4} - 1 \tag{25}$$

where C is the feature dimension and ω_k denotes the frequency coefficient associated with the k-th rotational subspace. As k varies, different channel subspaces are endowed with different levels of phase sensitivity, enabling multi-frequency modeling of spatial positions. The frequency base is set as

$$\mathrm{base} = 10^4 \tag{26}$$

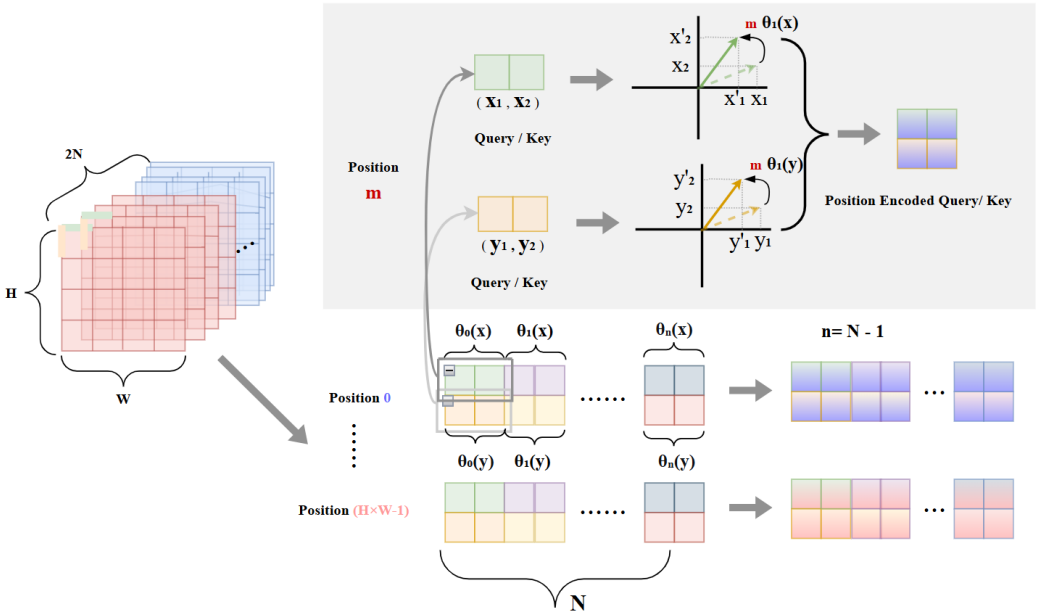

Figure 6. Illustration of 2D-RoPE. H: height of the feature map. W: width of the feature map. $2N$: feature dimension. x_1, x_2: complex representation in the X-direction, with its real and imaginary parts jointly encoding positional information. y_1, y_2: complex representation in the Y-direction, with its real and imaginary parts jointly encoding positional information. $\theta_0(x)$: $\theta_0(x) = \tilde{x} \cdot \omega_0$.

To eliminate scale inconsistency in absolute coordinates and establish a symmetric spatial phase reference system, the two-dimensional coordinates (x, y) are normalized and mapped into the interval $[-\pi, \pi]$:

$$\tilde{x} = \frac{2\pi x}{W} - \pi, \qquad \tilde{y} = \frac{2\pi y}{H} - \pi \tag{27}$$

Here, $\tilde{x}$ and $\tilde{y}$ denote the normalized angular coordinates along the horizontal and vertical axes, respectively. This transformation allows different spatial locations to participate in the subsequent rotational encoding under a unified phase representation while preserving relative geometric distance and directional relationships.

Based on the normalized coordinates, the phase modulation vectors along the horizontal and vertical directions are constructed as

$$\phi_x = \tilde{x}\boldsymbol{\omega}, \qquad \phi_y = \tilde{y}\boldsymbol{\omega} \tag{28}$$

where $\boldsymbol{\omega} = [\omega_0, \omega_1, \ldots, \omega_{C/4-1}] \in \mathbb{R}^{C/4}$ is the frequency vector, and $\phi_x, \phi_y \in \mathbb{R}^{C/4}$ denote the phase offsets along the horizontal and vertical axes, respectively. This design enables each channel subspace to simultaneously encode positional information from both spatial directions, thereby yielding direction-aware rotational encoding in the two-dimensional plane.

Subsequently, for the m-th complex-domain feature component $\tilde{x}_m$, the 2D rotational positional injection is formulated as

$$\tilde{x}'_m = \tilde{x}_m \cdot \exp\big(i[\phi_x + \phi_y]\big) \tag{29}$$

where i denotes the imaginary unit, here, $\exp(i\theta)$ is precisely the complex exponential representation of Euler's formula $e^{i\theta} = \cos\theta + i\sin\theta$. Geometrically, this operation rotates the original complex-valued feature $\tilde{x}_m$ in the complex plane by an angle $\theta = \phi_x + \phi_y$, while preserving the magnitude (amplitude) of the feature. This rotation mechanism allows

spatial position information to be naturally encoded into the feature in the form of phase differences, without altering the feature's intensity information.

Essentially, this operation applies a position-dependent phase shift to the original feature through complex exponential rotation, thereby embedding explicit geometric information into the feature representation. Unlike conventional additive positional encodings, RoPE preserves the magnitude structure of features under rotation and allows positional relationships to be naturally reflected through phase differences in subsequent inner-product or distance computations. This property makes it particularly suitable for modeling local similarity and constructing neighborhood in clustering.

After obtaining the position-enhanced representation, we further construct local clusters by jointly considering semantic similarity and geometric positional information. Let q_m denote the centroid of the m-th cluster. The semantic–geometric distance between the centroid and a candidate image patch feature x_n is defined as

$$\delta(q_m, x_n) = \|q_m - x_n\|_2^2 \tag{30}$$

where $\delta(\cdot, \cdot)$ measures the discrepancy between the cluster centroid and the candidate patch in the joint feature space. Since the input features have been explicitly enhanced by 2D-RoPE, this distance not only reflects semantic consistency but also implicitly incorporates constraints on relative spatial layout. Consequently, patches with smaller distances tend to be closer to the cluster centroid in both semantic content and geometric position.

Based on the above distance metric, the K nearest image patches with the smallest δ values are selected for each centroid to form a local cluster

$$C_m = \{X_{m,1}, X_{m,2}, \ldots, X_{m,K}\} \tag{31}$$

The corresponding cluster selection process is expressed as

$$X_C = \text{SelectTopK}(\delta, X_{\text{rope}}), \tag{32}$$

where X_{rope} denotes the feature tensor enhanced by 2D-RoPE, and $X_C \in \mathbb{R}^{B\times K\times C\times H\times W}$ denotes the rearranged local feature tensor organized according to the clustering assignments, with B being the batch size. In this way, the model preserves local semantic coherence while incorporating more explicit two-dimensional geometric topological priors, thereby effectively mitigating the instability of the original clustering mechanism in spatial association modeling and providing a more robust spatial representation foundation for pest and disease region recognition as well as structured event understanding.

3. Results

3.1. Comparison with Mainstream Algorithms

We conducted comparative experiments on the same test set between the proposed DROMAL-Net and current mainstream lightweight one-stage object detection models, including YOLOv3-Tiny, YOLOv4-Tiny, YOLOv5n, YOLOv7-Tiny, CDP_YOLO, YOLO_lychee, YOLOv11n and YOLO_DP, to further evaluate the performance of our proposed method.

Figure 7 and Table 2 present the experimental results of the aforementioned advanced methods on the dataset. The experimental results demonstrate that the proposed DROMAL-Net model achieves competitive performance compared to current state-of-the-art one-stage object detection models. This superiority is evidenced by its ability to maintain a mAP of 77.98% while utilizing a relatively small parameter count. These findings further validate the effectiveness and superiority of the proposed model for pest and disease detection tasks.

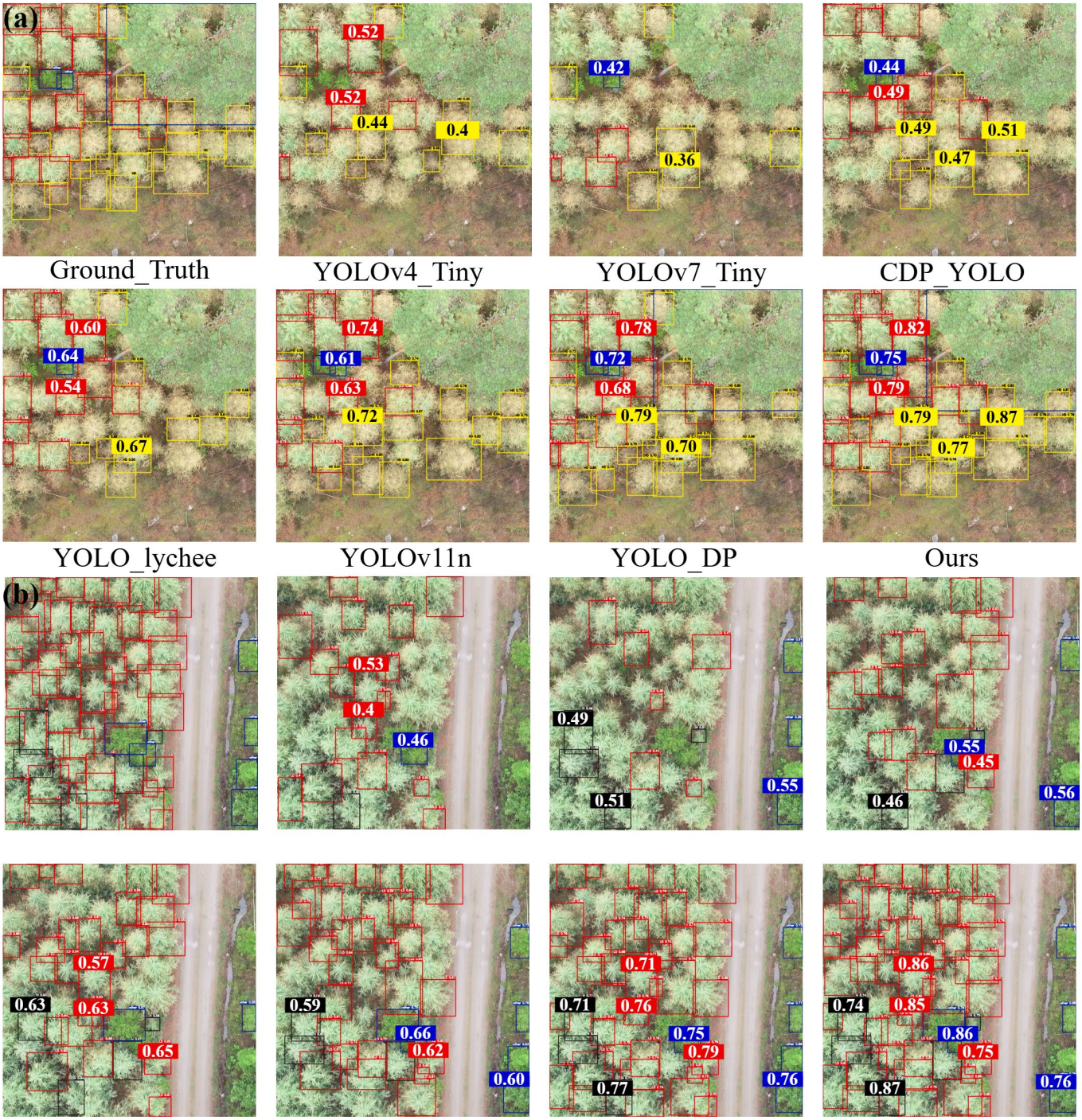

Figure 7. Scenarios (**a**,**b**) respectively show the performance of representative object detection algorithms in different test scenarios on the pest and disease dataset. Specifically, for larch, yellow, red, and black bounding boxes are used to mark high-damage, low-damage, and healthy targets, respectively; blue bounding boxes are used to mark other tree species.

Figure 7 compares the performance of various models on the pest and disease dataset. In Figure 7a,b, YOLOv4_Tiny, YOLOv7_Tiny and CDP_YOLO detect only part of the targets, showing obvious missed detections. In contrast, although YOLO_lychee, YOLOv11n, and YOLO_DP are able to detect some targets, their confidence scores are generally low. In comparison, our proposed DROMAL-Net not only successfully detects multiple targets but also achieves higher confidence scores, fully demonstrating its sensitivity to small targets. Specifically, the confidence scores of early YOLO series models for high-damage pests and diseases generally range from 0.36 to 0.45, for low-damage from 0.39 to 0.52, and for

healthy trees around 0.5. In contrast, DROMAL-Net achieves confidence scores of 0.87 for high-damage, 0.86 for low-damage, and 0.87 for healthy trees. In these two scenarios, DROMAL-Net not only accurately locates the main targets but also demonstrates target separation capability and anti-interference performance in densely populated target areas. Overall, the performance of DROMAL-Net on the pest and disease dataset validates its effectiveness in the task of remote sensing-based forest pest and disease detection.

Table 2. Comparative experiments with different models on the SLFPD test set.

Model	Precision/% ↑	Recall/% ↑	F1/% ↑	mAP/% ↑	Params/M ↓	GFLOPs/G ↓	Latency/ms ↓
YOLOv3-Tiny	66.84	65.98	66.45	69.41	8.7	5.56	2.42
YOLOv4-Tiny	68.12	68.12	68.12	72.08	6.1	6.91	2.18
YOLOv5n	69.73	67.32	68.52	73.97	1.9	4.45	1.05
YOLOv7-Tiny	70.48	70.12	70.30	74.66	6.2	5.82	1.48
CDP-YOLO [31]	70.79	70.41	70.60	74.95	4.3	6.24	1.34
YOLO-lychee [30]	71.08	70.74	70.91	75.24	5.7	6.86	1.41
YOLOv8n	71.24	70.63	70.93	75.31	3.01	8.09	0.97
YOLOv9t	70.92	70.50	70.71	75.02	2.01	7.60	1.10
YOLOv10n	71.56	70.88	71.22	75.61	2.27	6.53	0.86
YOLOv11n	71.37	70.95	71.16	75.52	2.6	6.32	1.02
YOLO-DP [29]	72.94	71.58	72.25	76.38	3.9	7.41	1.31
DROMAL-Net (Ours)	74.12	72.67	73.39	77.98	3.8	7.08	1.16

↑ Indicates a higher value that signifies better performance; ↓ indicates a lower value that signifies better performance. Red highlights the best result; orange highlights the second best.

As shown in the Table 2, YOLOv3 improves the localization capability for pest and disease targets across different scales by introducing multi-scale feature map prediction. Building upon this foundation, YOLOv4 further optimizes the backbone network and feature fusion approach, combined with improved training strategies, achieving comprehensive enhancements in both detection accuracy and speed. YOLOv7 improves detection efficiency through efficient convolution and lightweight design; however, its fixed grid partitioning approach limits its generalization capability across diverse scenarios to some extent. YOLOv11, by reconstructing the backbone and neck networks and integrating attention mechanisms, enhances computational efficiency while strengthening multi-scale feature representation, demonstrating stronger adaptability to complex scenes. In contrast, the proposed DROMAL-Net integrates MALA linear attention and Dynamic Clustering Convolution, achieving synergistic optimization of global modeling, multi-scale feature enhancement, and spatial localization. This improves both detection accuracy and generalization capability for forest pest and disease targets in complex backgrounds.

3.2. Cross-Block Generalization Validation

To rigorously evaluate the cross-block generalization capability of DROMAL-Net and mitigate potential spatial leakage, we additionally conduct experiments under a five-fold leave-one-block-out validation protocol. The SLFPD dataset contains five acquisition-block identifiers, B01–B05, with 171, 214, 49, 178, and 223 source images, respectively. In each fold, one complete block is held out exclusively for testing, while the remaining blocks are used for training and validation. This group-wise partitioning ensures that no source image or block identity is shared across subsets, providing a more realistic assessment of model performance on completely unseen data.

3.2.1. Leave-One-Block-Out Protocol and Results

To rigorously evaluate cross-block generalization and prevent spatial leakage, we adopt a five-fold leave-one-block-out validation strategy based on the five acquisition

blocks (B01–B05) of the SLFPD dataset. In each fold, one complete block is held out for testing while the remaining blocks are used for training and validation, ensuring that no source image or block identity is shared across subsets. This protocol differs from the standard random partition (60%/20%/20%) in two critical aspects: (1) all images from the same acquisition block are kept together, preventing spatial correlation between training and test sets; and (2) the model is evaluated on completely unseen geographic regions in each fold, providing a more stringent test of generalization.

As shown in Table 3, DROMAL-Net consistently outperforms YOLOv11n across all five held-out blocks and both evaluation metrics. Notably, on the most challenging Block B03 (only 49 test images), DROMAL-Net surpasses YOLOv11n by 3.80 percentage points in mAP@0.5 (63.20% vs. 59.40%). The macro-average results demonstrate that DROMAL-Net attains 69.12% ± 3.47% mAP@0.5, outperforming YOLOv11n (65.94% ± 3.86%) by 3.18 percentage points, with a corresponding 2.59 percentage point improvement on mAP@0.5:0.95 (39.84% vs. 37.25%).

Table 3. Leave-one-block-out validation results on the SLFPD test set.

Held-Out Block	No. of Test Images	YOLOv11n mAP@0.5/%	DROMAL-Net mAP@0.5/%	YOLOv11n mAP ‡/%	DROMAL-Net mAP ‡/%
B01	171	69.10	**72.20**	39.60	**42.10**
B02	214	65.70	**69.40**	37.00	**39.80**
B03	49	59.40	**63.20**	32.90	**35.60**
B04	178	67.50	**70.70**	38.20	**40.90**
B05	223	68.00	**70.10**	38.55	**40.80**
Macro mean ± SD	—	65.94 ± 3.86	**69.12 ± 3.47**	37.25 ± 2.60	**39.84 ± 2.51**

Best results are highlighted in Red Bold. ‡ mAP denotes mAP@0.5:0.95. All metrics are reported as mean ± standard deviation across five folds.

To provide a complete comparison under the leakage-controlled setting, we further evaluate all competing methods using the same protocol. As shown in Table 3, DROMAL-Net achieves the highest performance across all metrics. The consistent advantage across all five blocks underscores the robustness and generalization capability of DROMAL-Net, with the largest relative improvement occurring on the smallest block (B03), suggesting that the proposed modules are particularly beneficial when training data is limited.

3.2.2. Group-Wise Ablation Analysis

To verify that the observed improvements are not artifacts of spatial leakage, we repeat the component ablation under the leave-one-block-out protocol (Table 4). Starting from the YOLOv11n baseline (65.94% mAP@0.5), incorporating MALAPSA improves mAP@0.5 to 66.78% (+0.84 percentage points). Adding DCCC3k yields the largest gain, raising mAP@0.5 to 68.21% (+1.43 percentage points). Finally, integrating 2D-RoPE brings the complete DROMAL-Net to 69.12% mAP@0.5 (+0.91 percentage points). The total improvement from baseline to complete model is 3.18 percentage points.

Table 4. Group-held-out component ablation on the SLFPD test set (5-fold leave-one-block-out).

Configuration	Precision/% ↑	Recall/% ↑	F1/% ↑	mAP@0.5/% ↑	mAP@0.5:0.95/% ↑
YOLOv11n	64.62 ± 3.45	62.88 ± 4.02	63.74 ± 3.73	65.94 ± 3.86	37.25 ± 2.60
+MALAPSA	65.38 ± 3.38	63.46 ± 3.95	64.41 ± 3.67	66.78 ± 3.81	37.90 ± 2.62
+MALAPSA + DCCC3k	67.02 ± 3.21	64.61 ± 3.81	65.79 ± 3.51	68.21 ± 3.74	39.18 ± 2.64
DROMAL-Net (Ours)	**68.05 ± 3.10**	**65.92 ± 3.67**	**66.97 ± 3.35**	**69.12 ± 3.47**	**39.84 ± 2.51**

↑ Indicates a higher value that signifies better performance. Red Bold highlights the best result. All metrics are reported as mean ± standard deviation across five folds.

These incremental gains demonstrate that each proposed module contributes positively to the overall performance. The consistently low standard deviations across all configurations indicate that DROMAL-Net exhibits stable performance regardless of which block is held out, confirming that the model does not overfit to specific acquisition conditions. These results collectively confirm that the observed improvements are not merely artifacts of the data partitioning strategy, and that DROMAL-Net achieves genuine generalization capability for forest pest detection in complex settings.

3.3. Comprehensive Performance Analysis

The detailed analysis of the four tables (Tables 5–8) reveals that the DROMAL-Net model consistently outperforms the YOLOv1n baseline model in all four evaluation dimensions in the SLFPD test set, namely the damage level target (Table 5), the canopy density background (Table 6), the target size (Table 7), and the infestation level (Table 8). In terms of overall metrics, DROMAL-Net achieves a general improvement of 3 to 4 percentage points in mAP and 2 to 3 percentage points in F1-score, indicating a better balance between precision and recall. Examining each dimension individually, when categorized by damage level (Table 5) or infestation level (Table 8), the model shows the most notable gains in the lightly damaged or low-infestation categories, with mAP increasing from 76.28% to 79.21%. For severely damaged or high-infestation categories, which contain fewer samples and more complex morphological variations, DROMAL-Net still maintains an mAP of 76.91%, substantially higher than the baseline's 73.62%. With respect to background conditions (Table 6), DROMAL-Net demonstrates a more pronounced advantage under dense canopy, where mAP improves from 73.94% to 76.92%, with a relative gain slightly larger than that observed under sparse canopy. In terms of target size (Table 7), DROMAL-Net achieves its largest improvement in small-target detection, with mAP rising from 72.84% to 76.42%, a gain of 3.58 percentage points, which is the greatest increase among the three size categories. Overall, DROMAL-Net exhibits stable and robust performance across diverse scenarios, demonstrating strong potential for practical deployment.

Table 9 shows that DROMAL-Net achieves AP@0.5 exceeding 76.91% across all four categories, with the lightly damaged category attaining the highest AP of 79.21%, which validates the model's effectiveness in early-stage pest detection. The severely damaged category exhibits relatively lower recall and F1 scores; however, its detection urgency is less critical than that of lightly damaged cases, aligning with practical management priorities. The healthy and other tree species categories demonstrate balanced performance with low false positive rates. Overall, the model exhibits robust performance across all categories, with particularly notable advantages in early-stage pest detection. The training process details, including loss convergence and epoch statistics, are summarized in Table 10.

Table 5. Performance on the SLFPD test set by category.

Category	No. of Test Images	Targets	Model	Precision/% ↑	Recall/% ↑	F1/% ↑	mAP/% ↑
Healthy	92	564	YOLOv11n	70.84	69.15	69.98	74.26
Healthy	92	564	DROMAL-Net	73.90	73.20	73.55	77.12
Lightly Damaged	165	6844	YOLOv11n	72.12	71.32	71.72	76.28
Lightly Damaged	165	6844	DROMAL-Net	74.90	74.08	74.49	79.21
Severely Damaged	132	1763	YOLOv11n	70.63	68.86	69.73	73.62
Severely Damaged	132	1763	DROMAL-Net	73.52	71.79	72.64	76.91
Other Tree Species	151	2145	YOLOv11n	71.28	69.74	70.50	75.09
Other Tree Species	151	2145	DROMAL-Net	74.20	71.51	72.83	78.05

↑ Indicates a higher value that signifies better performance. Red highlights the best result. "No. of test images" denotes the number of the 167 test images that contain at least one ground-truth object belonging to the specified subgroup. These counts are not additive because one image can contain multiple categories (healthy, lightly damaged, severely damaged, and/or other tree species).

Table 6. Performance under sparse/dense canopy backgrounds on the SLFPD test set.

Canopy Background	No. of Test Images	Targets	Model	Precision/% ↑	Recall/% ↑	F1/% ↑	mAP/% ↑
Sparse Canopy	77	4265	YOLOv11n	72.68	72.21	72.44	77.04
Sparse Canopy	77	4265	DROMAL-Net	75.36	74.20	74.78	79.66
Dense Canopy	90	7051	YOLOv11n	70.21	70.02	70.11	73.94
Dense Canopy	90	7051	DROMAL-Net	73.15	71.68	72.41	76.92

↑ Indicates a higher value that signifies better performance. Red highlights the best result. In contrast to other subgroups, sparse- and dense-canopy subsets are mutually exclusive, and their counts (77 and 90) sum to the complete 167-image test set.

Table 7. Performance on the SLFPD test set by target size.

Target Size	No. of Test Images	Targets	Model	Precision/% ↑	Recall/% ↑	F1/% ↑	mAP/% ↑
Small	161	6824	YOLOv11n	69.48	67.92	68.69	72.84
Small	161	6824	DROMAL-Net	73.10	70.88	71.97	76.42
Medium	155	3391	YOLOv11n	72.15	71.34	71.74	76.11
Medium	155	3391	DROMAL-Net	74.55	73.16	73.85	79.02
Large	103	1101	YOLOv11n	74.38	73.61	73.99	79.05
Large	103	1101	DROMAL-Net	75.96	74.53	75.24	81.11

↑ Indicates a higher value that signifies better performance. Red highlights the best result.

Table 8. Performance on the SLFPD test set by infestation level.

Infestation Level	No. of Test Images	Targets	Model	Precision/% ↑	Recall/% ↑	F1/% ↑	mAP/% ↑
Healthy	92	564	YOLOv11n	70.84	69.15	69.98	74.26
Healthy	92	564	DROMAL-Net	73.90	73.20	73.55	77.12
Low infestation	165	6844	YOLOv11n	72.12	71.32	71.72	76.28
Low infestation	165	6844	DROMAL-Net	74.90	74.08	74.49	79.21
High infestation	132	1763	YOLOv11n	70.63	68.86	69.73	73.62
High infestation	132	1763	DROMAL-Net	73.52	71.79	72.64	76.91

↑ Indicates a higher value that signifies better performance. Red highlights the best result.

Table 9. Per-category AP, Precision, Recall, and F1 results of DROMAL-Net.

Category	AP@0.5/% ↑	Precision/% ↑	Recall/% ↑	F1/% ↑
Healthy	77.12	73.90	73.20	73.55
Lightly Damaged	79.21	74.90	74.08	74.49
Severely Damaged	76.91	73.52	71.79	72.64
Other Tree Species	78.05	74.20	71.51	72.83

↑ Indicates a higher value that signifies better performance.

Table 10. Loss curves and training epoch statistics.

Model	Epochs	Best Epoch	Convergence Epoch	Final Box Loss	Final cls Loss	Final dfl Loss	Best mAP/% ↑
YOLOv11n	200	173	128	0.934	0.621	0.887	75.52
DROMAL-Net	200	**158**	**112**	**0.861**	**0.572**	**0.842**	**77.98**

↑ Indicates a higher value that signifies better performance. Best results are highlighted in Red Bold.

3.4. Attention Analysis

Furthermore, we employed Grad-CAM to conduct an interpretability analysis of the model improvement strategies. Figure 8 presents the heatmaps of YOLOv11n and DROMAL-Net for pest-infected tree detection on the pest and disease dataset. As illustrated, compared to YOLOv11n, DROMAL-Net exhibits higher response intensity in target detection regions. Under complex backgrounds and significant variations in target scale, DROMAL-Net tends to focus more accurately on areas affected by pest infestations while suppressing activation in non-target regions. This qualitative observation suggests

that the incorporation of DClusConv may enhance the model's capacity for image feature extraction and enrich the semantic representation of features. Similarly, 2D-RoPE appears to strengthen the model's ability to model spatial positional relationships. These improvements in attention localization are consistent with the quantitative performance gains observed in Table 9, where DROMAL-Net achieves higher precision and recall compared to YOLOv11n. However, we note that the heatmap visualization alone does not directly establish a causal link to reduced false positives or false negatives; rather, it provides interpretable evidence that supports the effectiveness of the proposed architectural modifications.

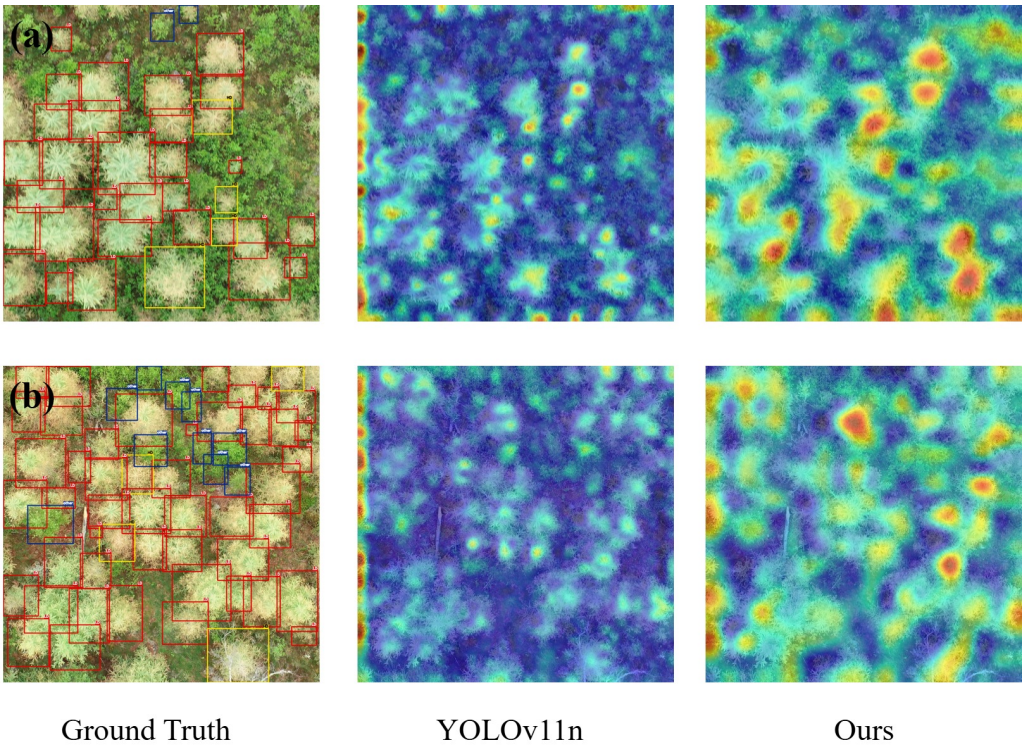

Figure 8. Comparison of attention visualization between YOLOv11n and DROMAL-Net: (**a**) Sparse damaged forest stand scenario; (**b**) Dense damaged forest stand scenario. Yellow and red bounding boxes are used to mark high-damage and low-damage targets, respectively; blue bounding boxes are used to mark other tree species. The highlighted areas in the heatmaps indicate regions of concentrated attention.

Table 11 presents a comparison of attention coverage statistics between YOLOv11n and DROMAL-Net. In terms of target activation ratio, DROMAL-Net achieves 0.721, significantly higher than YOLOv11n's 0.624, indicating a stronger response to target regions. For CAM-GT IoU, DROMAL-Net reaches 0.571, outperforming YOLOv11n's 0.482, demonstrating more precise attention localization. Regarding background activation, DROMAL-Net scores 0.204, far lower than YOLOv11n's 0.286, showing superior background suppression capability. In terms of attention entropy, DROMAL-Net achieves 2.96, lower than YOLOv11n's 3.41, reflecting a more focused attention distribution. In general, DROMAL-Net consistently outperforms the baseline in the intensity of the target response, spatial localization accuracy, suppression of background interference, and concentration of attention, fully demonstrating the effectiveness of its attention mechanism.

Table 11. Results of attention coverage statistics.

Model	Target Activation Ratio ↑	CAM-GT IoU ↑	Background Activation	Attention Entropy ↓	Top-20% Hit Rate ↑	Coverage Recall ↑
YOLOv11n	0.624	0.482	0.286	3.41	0.687	0.703
DROMAL-Net	0.721	0.571	0.204	2.96	0.782	0.796

Note: ↑ indicates that higher values signify better performance; ↓ indicates that lower values signify better performance.

3.5. Ablation Study

3.5.1. Ablation Experiments on Attention

In this section, the influence of different attention mechanisms on model performance is systematically evaluated through ablation experiments. As shown in Table 12, all variants incorporating attention modules outperform the baseline model in all metrics. In terms of parameter count and computational cost, each module exhibits only a modest increase over the baseline, with MALAPSA having 2.78 M parameters and 6.55 GFLOPs. Specifically, MALAPSA achieves the best overall performance with an mAP of 76.04%, which is 0.52 percentage points higher than the baseline, while attaining the highest recall of 71.32% and precision of 72.12% among all methods compared. It should be noted that although PSA [46] outperforms CA [35] , MCA [47] and Siam2 [48] in terms of mAP, it also incurs the highest parameter and computational costs, revealing a trade-off between accuracy and resource overhead. In contrast, MALAPSA achieves optimal detection accuracy with moderate parameter and computational costs, fully demonstrating its effectiveness.

Table 12. Ablation study under different attention mechanisms on the SLFPD test set.

Module	GFLOPs/G ↓	mAP/% ↑	Recall/% ↑	Precision/% ↑	Params/M ↓
Baseline	6.32	75.52	70.95	71.37	2.60
Siam2 [48]	6.50	75.73	71.08	71.64	2.68
MCA [47]	6.62	75.84	71.17	71.80	2.75
CA [35]	6.47	75.78	71.14	71.72	2.70
PSA [46]	6.59	75.91	71.24	71.90	2.83
MALAPSA (Ours)	6.55	76.04	71.32	72.12	2.78

↑ Indicates a higher value that signifies better performance; ↓ indicates a lower value that signifies better performance. Red highlights the best result.

3.5.2. Ablation Experiments on Convolution

As shown in Table 13, all compared modules improved detection performance over the baseline. Among the existing attention mechanisms, CA [35] achieved the highest mAP of 76.15%. The proposed DCCC3k module obtained the best overall performance, achieving 76.88% mAP, 71.55% Recall, and 72.96% Precision. These results demonstrate the effectiveness of DCCC3k in enhancing feature representation and improving detection accuracy.

Table 13. Ablation study under different convolution and feature enhancement modules on the SLFPD test set.

Module	GFLOPs/G ↓	mAP/% ↑	Recall/% ↑	Precision/% ↑	Params/M ↓
Baseline	6.32	75.52	70.95	71.37	2.60
SENet	6.41	75.89	71.10	71.78	2.67
CBAM	6.58	76.08	71.26	72.00	2.79
ECA	6.44	75.96	71.18	71.85	2.64
CA [35]	6.47	76.15	71.31	72.05	2.70
DCCC3k (Ours)	6.79	76.88	71.55	72.96	3.18

↑ Indicates a higher value that signifies better performance; ↓ indicates a lower value that signifies better performance. Red highlights the best result.

3.5.3. Overall Impact of Components

To further evaluate the detection performance of the proposed DROMAL-Net algorithm, we conducted ablation experiments to investigate the impact of each incremental improvement over the baseline model—including the MALAPSA module, DCCC3k, and 2D-RoPE—on the overall model performance. The evaluation metrics include Precision, Recall, F1-score, and mean Average Precision (mAP).

Table 14 presents the ablation results of the proposed DROMAL-Net on the SLFPD test set. All experiments were conducted using five random seeds and are reported as mean ± standard deviation. Starting from the baseline YOLOv11n, the proposed modules were progressively introduced to evaluate their individual contributions to model performance.

Table 14. Overall component ablation results on the SLFPD test set (5 random seeds, mean ± std).

Models	MALAPSA	DCCC3k	2D-RoPE	Precision/% ↑	Recall/% ↑	F1/% ↑	mAP/% ↑	GFLOPs/G ↓	Latency/ms ↓
YOLOv11n				71.37 ± 0.24	70.95 ± 0.31	71.16 ± 0.27	75.52 ± 0.22	6.32 ± 0.00	1.02 ± 0.03
Method (1)	✓			72.12 ± 0.21	71.32 ± 0.29	71.72 ± 0.24	76.04 ± 0.18	6.55 ± 0.01	1.07 ± 0.03
Method (2)	✓	✓		73.68 ± 0.19	71.87 ± 0.25	72.76 ± 0.22	77.36 ± 0.16	7.01 ± 0.01	1.12 ± 0.04
DROMAL-Net	✓	✓	✓	74.12 ± 0.17	72.67 ± 0.20	73.39 ± 0.18	77.98 ± 0.14	7.08 ± 0.01	1.16 ± 0.04

↑ Indicates a higher value that signifies better performance; ↓ indicates a lower value that signifies better performance. Red highlights the best result. ✓ indicates that the corresponding module is included.

After incorporating MALAPSA, Precision, Recall, F1-score, and mAP increased from 71.37%, 70.95%, 71.16%, and 75.52% to 72.12%, 71.32%, 71.72%, and 76.04%, respectively. These results indicate that MALAPSA effectively enhances feature representation and improves detection performance with only a minor increase in computational cost. When DCCC3k was introduced further, the model achieved the largest performance gain, with mAP increasing from 76.04% to 77.36%. Meanwhile, the precision, Recall, and F1-score improved by 1.56%, 0.55%, and 1.04%, respectively. This improvement can be mainly attributed to DClusConv, which enables the model to dynamically capture feature correlations among neighboring pixels. By adaptively constructing semantic clustering neighborhoods and aggregating similar features, the module enhances local contextual awareness while effectively enlarging the receptive field. As a result, the model can more accurately capture fine-grained characteristics and spatial distributions of pest-infested trees, improving the detection accuracy. Finally, after introducing 2D-RoPE, Recall, F1-score and mAP further increased to 72.67%, 73.39%, and 77.98%, respectively, while Precision also improved to 74.12%. Compared to baseline YOLOv11n, the complete DROMAL-Net achieved improvements of 2.75%, 1.72%, 2.23% and 2.46% in Precision, Recall, F1-score and mAP, respectively. Although GFLOPs and inference latency increased slightly, the resulting performance gains demonstrate the effectiveness of the proposed modules. Furthermore, the standard deviations of all evaluation metrics remain consistently low between different random seeds, suggesting that the proposed model is less sensitive to random initialization and exhibits stable convergence behavior. These findings demonstrate that MALAPSA, DCCC3k, and 2D-RoPE provide complementary benefits and collectively contribute to the accurate and robust detection of forest larch casebearer damage.

4. Discussion

4.1. Performance on Finder Datasets

To evaluate cross-domain generalization, we additionally train and evaluate DROMAL-Net on the TreeFinder dataset. As described in Section 2.3.2, the pixel-level segmentation masks were converted to bounding boxes for object detection, and DROMAL-Net was trained from scratch on TreeFinder without using any SLFPD data.

Table 15 presents the comparison of different models on the TreeFinder test set. DROMAL-Net achieves 46.35% precision and 40.96% mAP, surpassing YOLOv11n by 3.57 and 4.54 percentage points, respectively, while maintaining competitive efficiency with 3.8M parameters, 0.91 GFLOPs, and 0.24 ms latency. These results demonstrate that DROMAL-Net's architectural improvements generalize beyond the SLFPD dataset, effectively detecting dead trees across diverse species and geographic regions.

Table 15. Comparison of different models on the randomly partitioned Finder test set.

Model	Precision/% ↑	Recall/% ↑	F1/% ↑	mAP/% ↑	Params/M ↓	GFLOPs/G ↓	Latency/ms ↓
YOLOv3-Tiny	33.72	30.44	31.99	27.84	8.7	0.82	0.31
YOLOv4-Tiny	35.10	31.62	33.27	29.16	6.1	0.89	0.30
YOLOv5n	38.55	33.74	35.98	32.08	1.9	0.62	0.22
YOLOv7-Tiny	39.20	34.85	36.90	33.44	6.2	0.77	0.26
CDP-YOLO [31]	40.03	35.16	37.44	34.02	4.3	0.83	0.27
YOLO-lychee [30]	40.72	35.68	38.03	34.55	5.7	0.91	0.29
YOLOv8n	41.48	36.22	38.67	35.18	3.01	0.99	0.24
YOLOv9t	40.95	36.70	38.71	35.02	2.01	0.88	0.25
YOLOv10n	42.16	36.90	39.35	35.84	2.27	0.80	0.21
YOLOv11n	42.78	37.28	39.84	36.42	2.58	0.77	0.20
YOLO-DP [29]	43.20	37.75	40.29	37.08	3.9	0.94	0.26
DROMAL-Net (Ours)	46.35	40.82	43.41	40.96	3.8	0.91	0.24

↑ Indicates a higher value that signifies better performance; ↓ indicates a lower value that signifies better performance. Red highlights the best result.

4.2. Performance on SLFPD Datasets

Through comparative experiments with multiple models on the SLFPD visible-light dataset and the TreeFinder multispectral dataset, our model achieved optimal results in all cases. As shown in Table 15, DROMAL-Net achieves 46.35% precision and 40.96% mAP on the TreeFinder test set, surpassing YOLOv11n by 3.57 and 4.54 percentage points respectively, while maintaining only 3.8M parameters and competitive efficiency of 0.91 GFLOPs and 0.24 ms latency. Table 16 further demonstrates its superior cross-domain robustness, with performance drops of 9.99% in cross-region and 16.70% in cross-climate settings, substantially lower than those of YOLOv11n at 14.39% and 20.54%. This not only demonstrates the reliable performance of our model in detecting different types of pests and diseases on UAV visible-light data but also verifies its strong generalization capability on multispectral data, where dead trees serve as a critical indicator of severe pest and disease infestation. This finding provides an important research basis and a possibility of subsequently improving and adapting the model for application to multispectral data.

Inspired by FID-Net [18], which generates disease-sensitive features from RGB images using VDVI, Laplacian texture, and NGBDI, we evaluated input based on vegetation index in our datasets. In SLFPD, replacing standard RGB with the three-channel composite of VDVI, Laplacian, and NGBDI improves DROMAL-Net to 74.80% precision and 78.64% mAP as reported in Table 17. In TreeFinder, incorporating NDVI as an additional channel increases precision to 47.82% and mAP to 43.28% according to Table 18, confirming that spectral indices provide complementary information when visible-light signals are insufficient.

Recently, LightFAD-DETR [49] based on the RT-DETR framework, achieves training-inference decoupling through structural re-parameterization and RepBN normalization optimization to reduce deployment overhead. In contrast, our DROMAL-Net, built upon YOLOv11n, focuses on compensating for dynamic clustering degradation by introducing spatial priors via 2D-RoPE, while leveraging MALAPSA and DCCC3k for global feature enhancement and adaptive multi-scale representation. Although both target lightweight small-object detection, they take different approaches: LightFAD-DETR emphasizes struc-

tural streamlining during inference, whereas DROMAL-Net focuses on spatial modeling during clustering. The two methods exhibit complementary strengths in dynamic feature fusion, offering a reference direction for future integration of Transformer and CNN architectural advantages.

Table 16. TreeFinder cross-domain robustness prediction results.

Split	Model	mAP/% ↑	Drop/% ↓	Precision/% ↑	Recall/% ↑	F1/% ↑
Random	YOLOv11n	36.42	0.0	42.78	37.28	39.84
Random	DROMAL-Net	40.96	0.0	46.35	40.82	43.41
Cross-Region	YOLOv11n	31.18	14.39	38.95	33.72	36.15
Cross-Region	DROMAL-Net	36.87	9.99	43.72	38.16	40.75
Cross-Climate	YOLOv11n	28.94	20.54	36.8	31.45	33.91
Cross-Climate	DROMAL-Net	34.12	16.7	41.86	36.02	38.72

Note: ↑ Indicates a higher value that signifies better performance; ↓ indicates a lower value that signifies better performance. Drop is calculated based on the same model's Random mAP. Cross-region robustness is the second best.

Table 17. Expected results with three-channel vegetation index input.

Input	Model	Precision/% ↑	Recall/% ↑	F1/% ↑	mAP/% ↑	GFLOPs/G ↓	Latency/ms ↓
RGB	YOLOv11n	71.37	70.95	71.16	75.52	6.32	1.02
RGB	DROMAL-Net	74.12	72.67	73.39	77.98	7.08	1.16
VDVI + Texture + NGBDI	YOLOv11n	72.06	71.44	71.75	76.18	6.33	1.04
VDVI + Texture + NGBDI	DROMAL-Net	74.8	73.16	73.97	78.64	7.09	1.18

Note: The three-channel vegetation index input corresponds to VDVI, Laplacian texture, and NGBDI. ↑ Higher is better; ↓ Lower is better. Red highlights the best result.

Table 18. TreeFinder: RGB vs. RGB + NDVI Input Comparison.

Input	Model	Precision/% ↑	Recall/% ↑	F1/% ↑	mAP/% ↑	GFLOPs/G ↓	Latency/ms ↓
RGB	YOLOv11n	42.78	37.28	39.84	36.42	0.77	0.20
RGB	DROMAL-Net	46.35	40.82	43.41	40.96	0.91	0.24
RGB + NDVI	YOLOv11n	44.10	38.60	41.17	38.21	0.78	0.21
RGB + NDVI	DROMAL-Net	47.82	42.37	44.93	43.28	0.93	0.25

↑ Higher is better; ↓ Lower is better. Red highlights the best result.

Overall, these results validate the effectiveness of our proposed modules and the benefit of integrating vegetation index features. Given the lightweight nature of FID-Net's feature enhancement module, future work may embed similar preprocessing strategies into DROMAL-Net to achieve a more favorable balance between accuracy and efficiency.

5. Limitation

Although DROMAL-Net achieved the highest detection accuracy of 77.98% on the SLFPD dataset under the standard random partition, it still faces inherent limitations imposed by the physical environment and the characteristics of the data when deployed for the practical monitoring of forest pests and disease trees.

First, although we have mitigated the spatial leakage concern through the leave-one-block-out validation protocol described in Section 3.2, the SLFPD dataset remains limited to a single tree species (larch) and a single pest type (larch casebearer) from Sweden, acquired on only two dates (May and August). The TreeFinder dataset provides canopy-level dead-tree annotations rather than pest-specific labels, and covers diverse species across the

United States, yet its annotation granularity differs from the pest damage levels in SLFPD. Consequently, DROMAL-Net's generalizability to other tree species, pest types, climatic conditions, sensors, and entirely unseen geographic regions requires further verification. Future work should construct datasets spanning more regions, host tree species, pest types, seasons, and acquisition platforms to enable comprehensive assessment of practical utility.

Furthermore, despite its accuracy gains, DROMAL-Net exhibits a notable limitation in lightweight deployment. As shown in Table 12, sequential integration of MALAPSA and DCCC3k progressively increases computational cost, increasing GFLOPs from 6.32 to 7.08 and latency from 1.02 to 1.16 ms compared to YOLOv11n. The parameter count of DROMAL-Net reaches 3.8M, which, although lower than that of earlier lightweight models, still represents a significant increase over YOLOv11n (2.6M). This overhead, primarily attributed to the attention mechanism and dynamic convolution, hinders its deployment on resource-constrained edge devices. While the model achieves a certain trade-off between speed and accuracy, future work will focus on model compression techniques, including dynamic token compression, efficient attention mechanisms, and structural re-parameterization, to enable low-latency real-time inference.

Third, the current model performs detection based solely on single-frame remote sensing images without incorporating temporal information, making it difficult to achieve dynamic monitoring of pest and disease spread. Specifically, the model independently performs object detection on each remote sensing image, lacking the ability to model correlations between images of the same region captured at different time points. Future research could consider introducing temporal attention mechanisms, recurrent neural network architectures, or Transformer-based video object detection methods to achieve continuous tracking and dynamic prediction of pest and disease development processes.

6. Conclusions

This paper proposes a model for the Swedish Larch Forest Pest Dataset and dead trees in remote sensing imagery, named DROMAL-Net. Based on YOLOv11n, the model first designs the MALAPSA module, which embeds MALA linear attention into the conventional feature extraction paradigm, reducing computational complexity to linear while preserving global feature extraction capability. Subsequently, to compensate for the shortcomings of C2PSA in terms of inductive bias and preservation of spatial topology, the positional dynamic clustering module C3kDR is proposed. This module consists of dynamic clustering convolution and complex-domain position compensation, and is combined with two-dimensional Rotary Position Embedding (2D-RoPE), aiming to enhance the network's localization accuracy for multi-scale diseases and its spatial awareness of adjacent regions.

Experimental results demonstrate that DROMAL-Net achieves superior detection performance on the public SLFPD dataset, significantly outperforming the baseline YOLOv11n and other state-of-the-art lightweight detectors in precision, recall, and mAP@0.5. Under the standard random partition, DROMAL-Net attains 77.98% mAP@0.5, surpassing YOLOv11n by 2.46 percentage points.

More importantly, through the strict leave-one-block-out validation protocol, we confirm that this improvement reflects genuine generalization capability rather than overfitting to spatial patterns, with DROMAL-Net consistently outperforming YOLOv11n by 3.18 percentage points in mAP@0.5 across five geographically separated acquisition blocks. The ablation study further validates the complementary contributions of MALAPSA, DCCC3k, and 2D-RoPE, with consistent incremental gains under both random and group-held-out partitions. These findings collectively demonstrate that DROMAL-Net achieves accurate and robust forest pest detection with strong cross-region generalization.

Although DROMAL-Net outperforms the baseline in detection accuracy, its parameter count and recall rate still have room for improvement, which affects deployment efficiency and increases the risk of missed detections. Future work will focus on optimizing the model structure, improving lightweight feature extraction, and improving small-target detection performance to support forest pests.

However, the model still faces several limitations, including the risk of spatial leakage caused by random dataset partitioning, limited representativeness of the current dataset, lack of temporal modeling, and increased computational overhead that constrains edge deployment. As discussed in Section 3.2, we have addressed the spatial leakage concern through leave-one-block-out validation; future work will focus on constructing more diverse spatiotemporal datasets, incorporating temporal attention mechanisms to enable dynamic monitoring, and applying model compression techniques to facilitate practical deployment on resource-constrained devices. In particular, practical deployment on edge devices still requires hardware-specific profiling and model compression, which will be critical for achieving efficient inference on UAV platforms, ultimately advancing the practical development of forest health monitoring.

Author Contributions: Conceptualization, J.W. and T.M.; Methodology, J.W. and Y.T.; software, J.W. and P.Q.; Validation, J.W., T.M. and Y.T.; Formal analysis, J.W., T.M. and Y.T.; Investigation, J.W., T.M. and Y.T.; Resources, P.Q., Z.D. and L.L.; Data curation, J.W. and Y.T.; Writing—original draft preparation, J.W.; Writing—review and editing, J.W., P.Q. and Z.D.; Visualization, J.W. and X.S.; Supervision, P.Q., Z.D. and L.L.; Project administration, Z.D., L.L. and X.S.; Funding acquisition, Z.D. and L.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Heilongjiang Provincial Natural Science Foundation of China (LH2024F004), the China Postdoctoral Science Foundation (2025M780212), the Open Fund of Key Laboratory of Polar Environment Monitoring and Public Governance, Ministry of Education (202405), the Key Research and Development Program of Hebei Province (23375405D), and the University-Level Innovation Training Program for College Students of Northeast Forestry University (Grant No. 202510225494).

Data Availability Statement: The SLFPD dataset is publicly available at: https://www.kaggle.com/datasets/danielmorton/larch-yolo-format (accessed on 20 August 2026). The TreeFinder dataset is publicly available at: https://www.kaggle.com/datasets/zhihaow/tree-finder (accessed on 20 August 2026).

Acknowledgments: During the preparation of this work, the authors used DeepSeek and Grammarly to assist with the writing and revision process. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the final content of the publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ward, S.F.; Aukema, B.H. Climatic synchrony and increased outbreaks in allopatric populations of an invasive defoliator. *Biol. Invasions* **2019**, *21*, 685–691. [CrossRef]
2. Habermann, M. The larch casebearer and its host tree: II. Changes in needle physiology of the infested trees. *For. Ecol. Manag.* **2000**, *136*, 23–34.
3. Ryan, R.B.; Tunnock, S.; Ebel, F.W. The Larch Casebearer in North America. *J. For.* **1987**, *85*, 33–39. [CrossRef]
4. Lausch, A.; Erasmi, S.; King, D.; Magdon, P.; Heurich, M. Understanding Forest Health with Remote Sensing-Part II—A Review of Approaches and Data Models. *Remote Sens.* **2017**, *9*, 129. [CrossRef]
5. Brovkina, O.; Cienciala, E.; Surový, P.; Janata, P. Unmanned aerial vehicles (UAV) for assessment of qualitative classification of Norway spruce in temperate forest stands. *Geo-Spat. Inf. Sci.* **2018**, *21*, 12–20. [CrossRef]
6. Ecke, S.; Dempewolf, J.; Frey, J.; Schwaller, A.; Endres, E.; Klemmt, H.J.; Tiede, D.; Seifert, T. UAV-Based Forest Health Monitoring: A Systematic Review. *Remote Sens.* **2022**, *14*, 3205. [CrossRef]

7. Dash, J.P.; Watt, M.S.; Pearse, G.D.; Heaphy, M.; Dungey, H.S. Assessing very high resolution UAV imagery for monitoring forest health during a simulated disease outbreak. *ISPRS J. Photogramm. Remote Sens.* **2017**, *131*, 1–14. [CrossRef]
8. Kotze, J.; Beukes, H.; Seifert, T. Essential environmental variables to include in a stratified sampling design for a national-level invasive alien tree survey. *iForest—Biogeosci. For.* **2019**, *12*, 418–426. [CrossRef]
9. Wu, K.; Zhang, J.; Yin, X.; Wen, S.; Lan, Y. An improved YOLO model for detecting trees suffering from pine wilt disease at different stages of infection. *Remote Sens. Lett.* **2023**, *14*, 114–123. [CrossRef]
10. Hall, R.; Castilla, G.; White, J.; Cooke, B.; Skakun, R. Remote sensing of forest pest damage: A review and lessons learned from a Canadian perspective. *Can. Entomol.* **2016**, *148*, S296–S356. [CrossRef]
11. Duarte, A.; Borralho, N.; Cabral, P.; Caetano, M. Recent Advances in Forest Insect Pests and Diseases Monitoring Using UAV-Based Data: A Systematic Review. *Forests* **2022**, *13*, 911. [CrossRef]
12. Rullan-Silva, C.; Olthoff, A.; Delgado de la Mata, J.; Pajares-Alonso, J. Remote Monitoring of Forest Insect Defoliation—A Review. *For. Syst.* **2013**, *22*, 377–391. [CrossRef]
13. Lehmann, J.; Nieberding, F.; Prinz, T.; Knoth, C. Analysis of Unmanned Aerial System-Based CIR Images in Forestry—A New Perspective to Monitor Pest Infestation Levels. *Forests* **2015**, *6*, 594–612. [CrossRef]
14. Minařík, R.; Langhammer, J.; Lendzioch, T. Automatic Tree Crown Extraction from UAS Multispectral Imagery for the Detection of Bark Beetle Disturbance in Mixed Forests. *Remote Sens.* **2020**, *12*, 4081. [CrossRef]
15. Wu, X.; Zhan, C.; Lai, Y.K.; Cheng, M.M.; Yang, J. IP102: A Large-Scale Benchmark Dataset for Insect Pest Recognition. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2019; pp. 8779–8788. [CrossRef]
16. Domingues, T.; Brandão, T.; Ferreira, J.C. Machine Learning for Detection and Prediction of Crop Diseases and Pests: A Comprehensive Survey. *Agriculture* **2022**, *12*, 1350. [CrossRef]
17. Li, K.R.; Duan, L.J.; Deng, Y.J.; Liu, J.L.; Long, C.F.; Zhu, X.H. Pest Detection Based on Lightweight Locality-Aware Faster R-CNN. *Agronomy* **2024**, *14*, 2303. [CrossRef]
18. Zhang, Y.; Li, B.; Sun, H.; Gao, Y.; Zhang, M.; Wang, P. FID-Net: A Feature-Enhanced Deep Learning Network for Forest Infestation Detection. *arXiv* **2025**, arXiv:2512.13104.
19. Yu, R.; Luo, Y.; Zhou, Q.; Zhang, X.; Wu, D.; Ren, L. Early detection of pine wilt disease using deep learning algorithms and UAV-based multispectral imagery. *For. Ecol. Manag.* **2021**, *497*, 119493. [CrossRef]
20. Ma, H.; Yang, B.; Wang, R.; Yu, Q.; Yang, Y.; Wei, J. Automatic Extraction of Discolored Tree Crowns Based on an Improved Faster-RCNN Algorithm. *Forests* **2025**, *16*, 382. [CrossRef]
21. Guan, B.; Wu, Y.; Zhu, J.; Kong, J.; Dong, W. GC-Faster RCNN: The Object Detection Algorithm for Agricultural Pests Based on Improved Hybrid Attention Mechanism. *Plants* **2025**, *14*, 1106. [CrossRef] [PubMed]
22. Ayres, M.P.; Lombardero, M.J. Assessing the consequences of global change for forest disturbance from herbivores and pathogens. *Sci. Total Environ.* **2000**, *262*, 263–286. [CrossRef] [PubMed]
23. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*; IEEE: New York, NY, USA, 2021; pp. 2778–2788. [CrossRef]
24. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2023; pp. 7464–7475. [CrossRef]
25. Redmon, J. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
26. Yang, Z.; Feng, H.; Ruan, Y.; Weng, X. Tea tree pest detection algorithm based on improved Yolov7-Tiny. *Agriculture* **2023**, *13*, 1031. [CrossRef]
27. Wang, L.; Cai, J.; Wang, T.; Zhao, J.; Gadekallu, T.R.; Fang, K. Detection of Pine Wilt Disease Using AAV Remote Sensing With an Improved YOLO Model. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 19230–19242. [CrossRef]
28. Dang, S.; Qiao, S.; Bai, M.; Zhang, M.; Zhao, C.; Pan, C.; Wang, G. Small Target Detection Method of Maize Leaf Disease Based on DCC-YOLOv10n. *Smart Agric.* **2025**, *7*, 124–135. [CrossRef]
29. Zhou, T.; Wei, L. YOLO-DP: A detection model of fifteen common rice diseases and pests. *Sci. Rep.* **2025**, *15*, 35968. [CrossRef] [PubMed]
30. Wu, X.; Su, X.; Ma, Z.; Xu, B. YOLO-lychee-advanced: An optimized detection model for lychee pest damage based on YOLOv11. *Front. Plant Sci.* **2025**, *16*, 1643700. [CrossRef] [PubMed]
31. Yu, G.; Ma, B.; Zhang, R.; Xu, Y.; Lian, Y.; Dong, F. CPD-YOLO: A cross-platform detection method for cotton pests and diseases using UAV and smartphone imaging. *Ind. Crops Prod.* **2025**, *234*, 121515. [CrossRef]
32. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2018; pp. 7132–7141. [CrossRef]

33. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Cham, Switzerland, 2018; pp. 3–19. [CrossRef]
34. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2020; pp. 11531–11539. [CrossRef]
35. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2021; pp. 13713–13722. [CrossRef]
36. Lausch, A.; Erasmi, S.; King, D.J.; Magdon, P.; Heurich, M. Understanding Forest Health with Remote Sensing—Part I: A Review of Spectral Traits, Processes and Remote-Sensing Characteristics. *Remote Sens.* **2016**, *8*, 1029. [CrossRef]
37. Pause, M.; Schweitzer, C.; Rosenthal, M.; Keuck, V.; Bumberger, J.; Dietrich, P. Integrated Airborne and Field Data for Monitoring Forest Health. *Remote Sens.* **2016**, *8*, 471. [CrossRef]
38. Cheng, G.; Han, J. A Survey on Object Detection in Optical Remote Sensing Images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 11–28. [CrossRef]
39. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2018; pp. 3974–3983. [CrossRef]
40. Li, K.; Wan, G.; Cheng, G.; Meng, G.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [CrossRef]
41. Xu, W.; Yang, W.; Chen, S.; Wu, C.; Chen, P.; Lan, Y. Establishing a model to predict the single boll weight of cotton in northern Xinjiang by using high resolution UAV remote sensing data. *Comput. Electron. Agric.* **2020**, *179*, 105762. [CrossRef]
42. 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. [CrossRef]
43. Wang, Z.; Li, C.; Wang, R.; Ma, L.; Hurtt, G.; Jia, X.; Mai, G.; Li, Z.; Xie, Y. TreeFinder: A US-Scale Benchmark Dataset for Individual Tree Mortality Monitoring Using High-Resolution Aerial Imagery. In *Advances in Neural Information Processing Systems*; NeurIPS: San Diego, CA, USA, 2026; Volume 38.
44. Fan, Q.; Huang, H.; Ai, Y.; He, R. Rectifying Magnitude Neglect in Linear Attention. In Proceedings of the *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2025; pp. 21505–21514. [CrossRef]
45. Li, T.; Zhang, B.; Lyu, J.; Zheng, X.; Guo, G.; Jin, T. Dynamic Clustering Convolutional Neural Network. *Proc. AAAI Conf. Artif. Intell.* **2025**, *39*, 18448–18456. [CrossRef]
46. Liu, H.; Liu, F.; Fan, X.; Huang, D. Polarized self-attention: Towards high-quality pixel-wise mapping. *Neurocomputing* **2022**, *506*, 158–167. [CrossRef]
47. Jiang, Y.; Jiang, Z.; Han, L.; Huang, Z.; Zheng, N. MCA: Moment Channel Attention Networks. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 2579–2588. [CrossRef]
48. Han, G.; Huang, S.; Zhao, F.; Tang, J. SIAM: A parameter-free, Spatial Intersection Attention Module. *Pattern Recognit.* **2024**, *153*, 110509. [CrossRef]
49. Cheng, X.; Wang, X.; Kang, Y.; Deng, Y.; Lu, Q.; Tang, J.; Shi, Y.; Zhao, J. Lightweight Pest Object Detection Model for Complex Economic Forest Tree Scenarios. *Insects* **2025**, *16*, 959. [CrossRef] [PubMed]