remote sensing

MDPI

Article

Monitoring Pine Wilt Disease Using High-Resolution Satellite Remote Sensing at the Single-Tree Scale with Integrated Self-Attention

Wenhao Lv, Junhao Zhao and Jixia Huang *

Key Laboratory for Silviculture and Conservation of Ministry of Education, Beijing Forestry University, Beijing 100083, China; lvwenhao24@mails.ucas.ac.cn (W.L.); zhaojunhao@bjfu.edu.cn (J.Z.)
* Correspondence: huangjx@bjfu.edu.cn

Abstract

Pine wilt disease has caused severe damage to China's forest ecosystems. Utilizing the rich information from very-high-resolution (VHR) satellite imagery for large-scale and accurate monitoring of pine wilt disease is a crucial approach to curbing its spread. However, current research on identifying infected trees using VHR satellite imagery and deep learning remains extremely limited. This study introduces several advanced self-attention algorithms into the task of satellite-based monitoring of pine wilt disease to enhance detection performance. We constructed a dataset of discolored pine trees affected by pine wilt disease using imagery from the Gaofen-2 and Gaofen-7 satellites. Within the unified semantic segmentation framework MMSegmentation, we implemented four single-head attention models—NLNet, CCNet, DANet, and GCNet—and two multi-head attention models—Swin Transformer and SegFormer—for the accurate semantic segmentation of infected trees. The model predictions were further analyzed through visualization. The results demonstrate that introducing appropriate self-attention algorithms significantly improves detection accuracy for pine wilt disease. Among the single-head attention models, DANet achieved the highest accuracy, reaching 73.35%. The multi-head attention models exhibited an excellent performance, with SegFormer-b2 achieving an accuracy of 76.39%, learning the features of discolored pine trees at the earliest stage and converging faster. The visualization of model inference results indicates that DANet, which integrates convolutional neural networks (CNNs) with self-attention mechanisms, achieved the highest overall accuracy at 94.43%. The use of self-attention algorithms enables models to extract more precise morphological features of discolored pine trees, enhancing user accuracy while potentially reducing production accuracy.

Keywords: pine wilt disease; high-resolution satellite imagery; single tree identification; semantic segmentation; self-attention

check for updates

Academic Editor: Carlos Alberto Silva

Received: 10 May 2025
Revised: 21 June 2025
Accepted: 22 June 2025
Published: 26 June 2025

Citation: Lv, W.; Zhao, J.; Huang, J. Monitoring Pine Wilt Disease Using High-Resolution Satellite Remote Sensing at the Single-Tree Scale with Integrated Self-Attention. *Remote Sens.* **2025**, *17*, 2197. https://doi.org/10.3390/rs17132197

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Pine wilt disease (PWD), caused by the pine wood nematode (Bursaphelenchus xylophilus), is a destructive forest pathogen originating from North America. It has become one of the most severe pests and diseases affecting coniferous forests worldwide [1]. Due to the unpredictable nature of PWD outbreaks and the lack of effective monitoring methods, the disease has continuously spread in China since its initial invasion in 1982. By the end of 2022, PWD had affected 701 county-level regions and 5250 township-level locations in

China, with a total impacted area of approximately 22.67 million mu. The epidemic is expected to further intensify in the coming years [2].

Many researchers have conducted detailed analyses of the biochemical composition and spectral changes in tree leaves following infection with pine wilt disease. Huang et al. found that the spectral reflectance of infected pine needles was characterized by a decrease in the green peak, a blue shift in the red edge, and reduced reflectance in the near-infrared spectrum [3]. Kim et al. determined through experiments that green and red spectral bands were the optimal combination for monitoring pine wilt disease [4]. Li et al. performed biochemical composition analysis and hyperspectral diagnosis of infected pine needles, discovering that the chlorophyll, carotenoid, and water content in the needles decreased with the progression of the disease, resulting in variations in spectral reflectance. They identified the green, red-edge, and shortwave infrared bands as diagnostic spectral bands for pine wilt disease [5]. Yu et al. analyzed the temporal dynamics of diseased trees throughout the growing season using UAV-based hyperspectral imagery and found that in the early stages of infection, classification based on absolute spectral reflectance values yielded better results. As the disease progressed, classification using vegetation indices such as NDVI became more effective [6]. These studies provide a solid foundation for utilizing remote-sensing technology to monitor pine wilt disease.

Monitoring forest health is essential for managing threats such as pests, diseases, and climate change [7,8]. Compared with traditional field-based surveys, remote sensing provides an efficient means of large-scale, long-term forest disturbance monitoring. Many studies have explored the use of remote sensing for detecting PWD with promising results. For example, Takenaka et al. combined airborne LiDAR and WorldView imagery with SVM and vegetation indices for accurate classification [9]. Syifa et al. used UAV visible-light imagery to detect dead trees, achieving 86–94% accuracy [10]. Tao et al. applied deep learning to UAV data in pure forests with 65–80% accuracy [11]. Zhang et al. employed high-resolution satellite imagery for individual-tree-level detection, reaching 81.2% user accuracy [12]. Zhou et al. used CNNs with satellite data to identify infected areas [13]. UAVs offer high spatial and spectral resolution, enabling precise localization and early detection, but are limited by cost and coverage. In contrast, satellite imagery is more scalable and cost-effective but often lacks sufficient resolution. Enhancing disease localization using high-resolution satellite data is thus crucial for effective large-scale PWD management.

Convolutional neural networks (CNNs) have become widely used in computer vision for their strong spatial feature extraction capabilities and often outperform traditional methods in remote-sensing tasks [14,15]. Several studies have applied deep learning to the satellite-based monitoring of pine wilt disease. Huang et al. used an enhanced SqueezeNet with high-resolution imagery for outbreak region detection [16]. Ye et al. integrated multi-scale attention into U-Net with Landsat 8 and UAV data for regional mapping [17]. Wang et al. proposed a semi-supervised GAN_HRNet_semi model for individual tree detection, achieving 72.07% accuracy [18]. However, most current studies still rely on conventional CNNs, leading to coarse outputs, high noise levels, and poor generalizability. In addition, performance evaluations often focus solely on accuracy, with limited attention to model interpretability, which limits real-world applicability.

Self-attention algorithms have emerged as one of the most innovative techniques in deep learning in recent years. By focusing solely on its information to quantify the interdependencies between input features, self-attention allows each element to access global context information. With its powerful feature extraction capability, it has been widely applied in fields such as natural language processing [19], computer vision [20], and speech recognition [21]. In the field of remote-sensing-based pest and disease monitoring, Yang et al. used a three-branch Swin Transformer model to identify plant diseases and their

severity [22]. Li et al. proposed a novel model that integrates 3D convolutional layers with a Transformer architecture, achieving pixel-level segmentation of PWD with an F1-score of 0.9 [23]. Feng et al. introduced an innovative data fusion framework that leverages the Real-Time Detection Transformer architecture (RTDETR). Applied to a pine wilt disease dataset, their method yielded an average detection accuracy improvement of around 10% compared to baseline approaches [24]. These studies demonstrate the excellent performance of self-attention models in various remote-sensing applications. Self-attention possesses strong feature extraction capabilities, enabling the efficient utilization of global semantic information in remote-sensing imagery. However, to date, there has been no research using self-attention algorithms for the satellite-based monitoring of pine wilt disease.

Currently, the performance of high-resolution satellite remote sensing for monitoring pine wilt disease remains unsatisfactory. The existing identification algorithms are relatively outdated, and few studies have achieved the efficient and precise identification of infected trees at the single-tree scale. Moreover, models often exhibit limited generalization capabilities and produce results with significant noise. To address these issues, this study introduces self-attention algorithms capable of extracting global features for the satellite-based remote-sensing monitoring of pine wilt disease. The self-attention mechanism overcomes the receptive field limitations of convolutional neural networks, enabling the model to capture spectral variations and morphological texture features of PWD at a global scale. This reduces the interference of local features on the recognition results and enhances the model's ability to distinguish diseased trees more accurately. By leveraging a unified framework, multiple self-attention models were implemented through transfer learning, and their performance characteristics in this task were systematically explored through the visualization of the inference results. This study marks the first application of self-attention algorithms in the satellite-based remote-sensing monitoring of pine wilt disease. It provides a fair evaluation of the performance of various self-attention algorithms in this context. The findings offer novel technical references for achieving the efficient and precise monitoring of pine wilt disease.

The main contributions of this article lie in the following three aspects.

(1) As far as we know, this study is the first to introduce the self-attention algorithm into the task of high-resolution satellite remote-sensing monitoring of pine wilt disease. We constructed a semantic segmentation dataset for pine wilt disease and, based on this dataset, achieved intelligent, precise, and efficient monitoring using high-resolution satellite imagery.
(2) Within a unified model implementation framework, we demonstrated that self-attention algorithms perform exceptionally well in the satellite remote-sensing monitoring of pine wilt disease. A detailed comparison of the performance of various self-attention algorithms was also provided.
(3) We found that self-attention algorithms can reduce the number of wrong detections of infected trees in this task, resulting in a smoother inferred diseased tree morphology. However, they may increase missed detections of diseased trees with less salient features. Combining convolutional neural networks with self-attention algorithms can achieve the best overall performance.

2. Materials and Methods

2.1. Study Area

The study area encompasses Yichun and Fuzhou in Jiangxi Province, and Nanping in Fujian Province, located in southern China (Figure 1). This region is characterized by hilly terrain with diverse landforms and abundant vegetation resources, with mixed coniferous and broad-leaved forests as the dominant forest type. The area experiences a mild annual

average temperature and ample precipitation, creating favorable conditions for the spread of pine wilt disease. It is considered one of the representative outbreak regions of pine wilt disease in China. Due to monitoring and survey delays, pine wilt disease has continuously spread within Jiangxi and Fujian Provinces. The disease was first detected in Yichun City in 2010, and between 2018 and 2021, a total of 1.26 million pine wilt disease-affected trees were removed in Yichun. Fuzhou City, like Yichun, is heavily impacted by pine wilt disease in Jiangxi Province. The disease was also discovered in Nanping City, Fujian Province, in 2008, where 5360 dead pine wilt disease-infected trees were cleared in 2018, and 4663 dead pines were removed in Shunchang County in 2021. The study area contains a substantial number of pine wilt disease-affected trees, and high-resolution satellite imagery provides numerous clear and standardized samples of pine wilt disease-affected trees, making it highly suitable for researching the satellite-based intelligent monitoring of pine wilt disease.

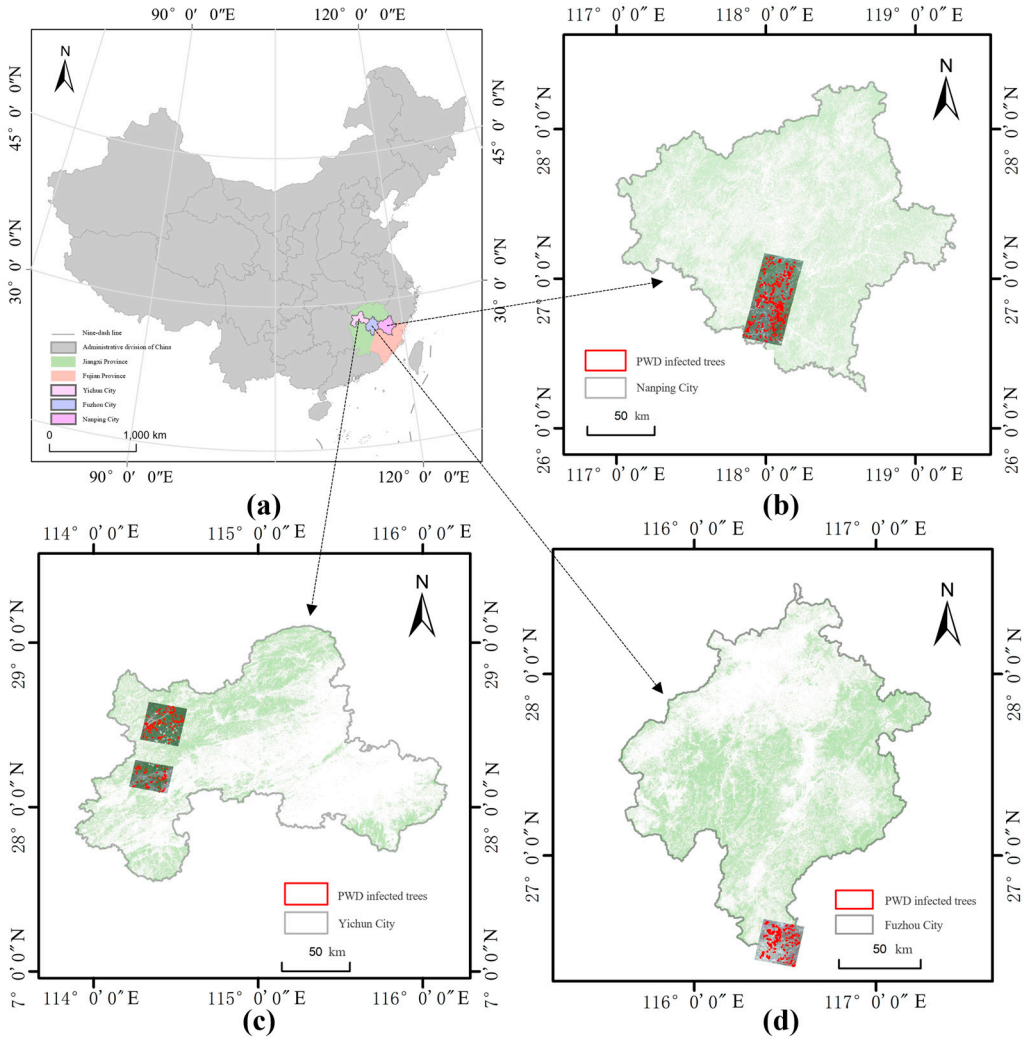


Figure 1. Study area and data. (**a**) Jiangxi Province and Fujian Province. (**b**) Fuzhou City. (**c**) Nanping City. (**d**) Yichun City.

2.2. Very-High-Resolution Remote-Sensing Image Data

This study utilizes high-resolution imagery products generated by the Gaofen-2 (GF-2) and Gaofen-7 (GF-7) satellites of China. The GF-2 satellite is equipped with two push-broom PMS cameras that have a swath width of 45 km, with a spatial resolution of 4 m for multispectral bands, which include red, green, blue, and near-infrared bands. The panchromatic band has a spatial resolution of up to 0.8 m. The GF-7 satellite is equipped

with a two-line array stereo camera with a swath width of 20 km, offering multispectral band resolutions better than 3.2 m, also including red, green, blue, and near-infrared bands, with a panchromatic band resolution of 0.8 m.

The study selected two scenes of GF-7 imagery covering Yichun City in Jiangxi Province, acquired in August 2022, and one scene of GF-7 remote-sensing imagery covering Fuzhou City, acquired in September 2022. Additionally, three scenes of GF-2 remote-sensing imagery covering Nanping City in Fujian Province were obtained, all acquired in September 2019, totaling six scenes. The research used 1A-level high-resolution satellite remote-sensing imagery products provided by the China Centre for Resources Satellite Data and Application (CRESDA), Beijing, China. These data underwent standard preprocessing steps, including radiometric calibration, atmospheric correction, orthorectification, and image fusion. The imagery consists of four spectral bands—red, green, blue, and near-infrared—all with a spatial resolution of 0.8 meters. The images were acquired between August and October, during which pine trees infected with PWD in southern China typically exhibit prominent wilting symptoms. Furthermore, during this period, the leaves of broad-leaved trees in southern China have not yet undergone natural color changes, and the humid climate of the study area minimizes the occurrence of spectral variations in tree crowns caused by drought stress. These factors collectively reduce seasonal and drought-related interference, thereby enhancing the accuracy of detecting PWD-affected trees.

2.3. Construction of the Semantic Segmentation Dataset for Pine Wilt Disease-Affected Trees

The construction of the semantic segmentation dataset for pine wilt disease involved the following steps: image enhancement, vegetation index construction, visual interpretation, and sample cropping.

The original true-color composite imagery appeared dim, making it difficult to visually distinguish diseased trees from healthy ones. A stretching operation was applied to enhance the imagery's vibrancy, facilitating the visual interpretation of diseased trees. This enhancement was achieved using a Python (version: 3.8.18) script that truncated pixel values based on specified percentages. The spectral reflectance of trees infected with pine wilt disease undergoes significant changes in the green, red, and near-infrared bands. To assist in visual interpretation, the Normalized Difference Vegetation Index (NDVI) and the Normalized Green–Red Difference Index (NGRDI) were constructed using the spectral reflectance from these bands, and the latter calculated as the difference between green and red band reflectance divided by their sum. The results of image enhancement and vegetation index construction are shown in Figure 2.

Based on data from 119 ground survey points in Nanping City, Fujian Province, provided by the Forest and Grassland Pest Control Center of the National Forestry and Grassland Administration, visual interpretation standards for trees infected with pine wilt disease were established. Diseased tree crowns typically appear orange-red, bright red, or reddish-brown, with shapes resembling circles or ellipses (distinguishable from bare land based on shape characteristics). They are relatively scattered in distribution, with only occasional clusters of multiple diseased trees. Their texture is distinct, and due to lighting effects, areas of brightness and shadow can be observed. In contrast, healthy tree crowns appear bright green or dark green, with dense and aggregated foliage. The visual interpretation of diseased trees was conducted in ArcGIS Pro. To improve the model's generalization ability, additional typical object samples were included, such as water bodies, farmland, building complexes, healthy forests, grasslands, red-roofed houses within forests, and forest trails. Examples of the dataset samples are shown in Figure 3.

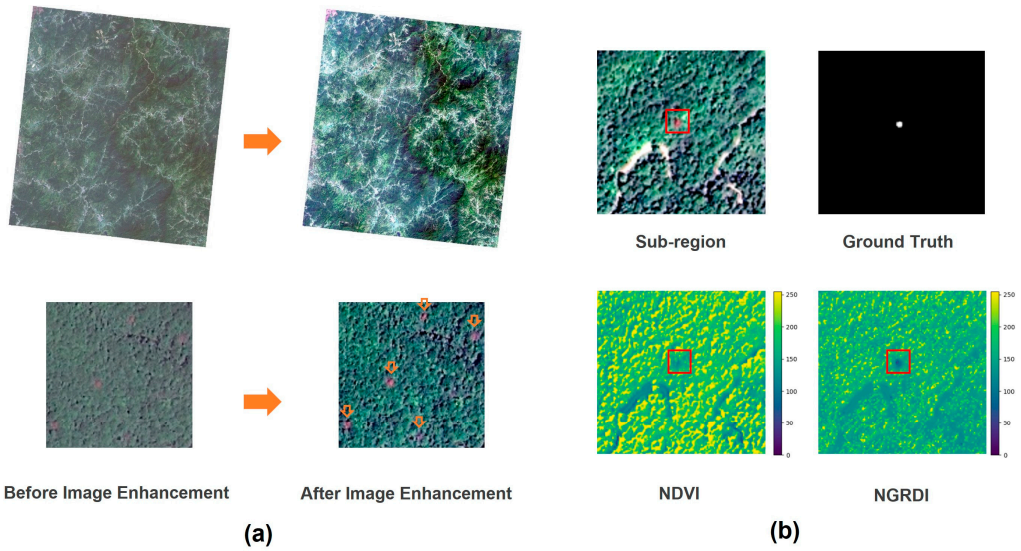


Figure 2. (**a**) Image enhancement. (**b**) Vegetation indices visual interpretation. (The red arrows and squares indicate the locations of trees infected by pine wilt disease).

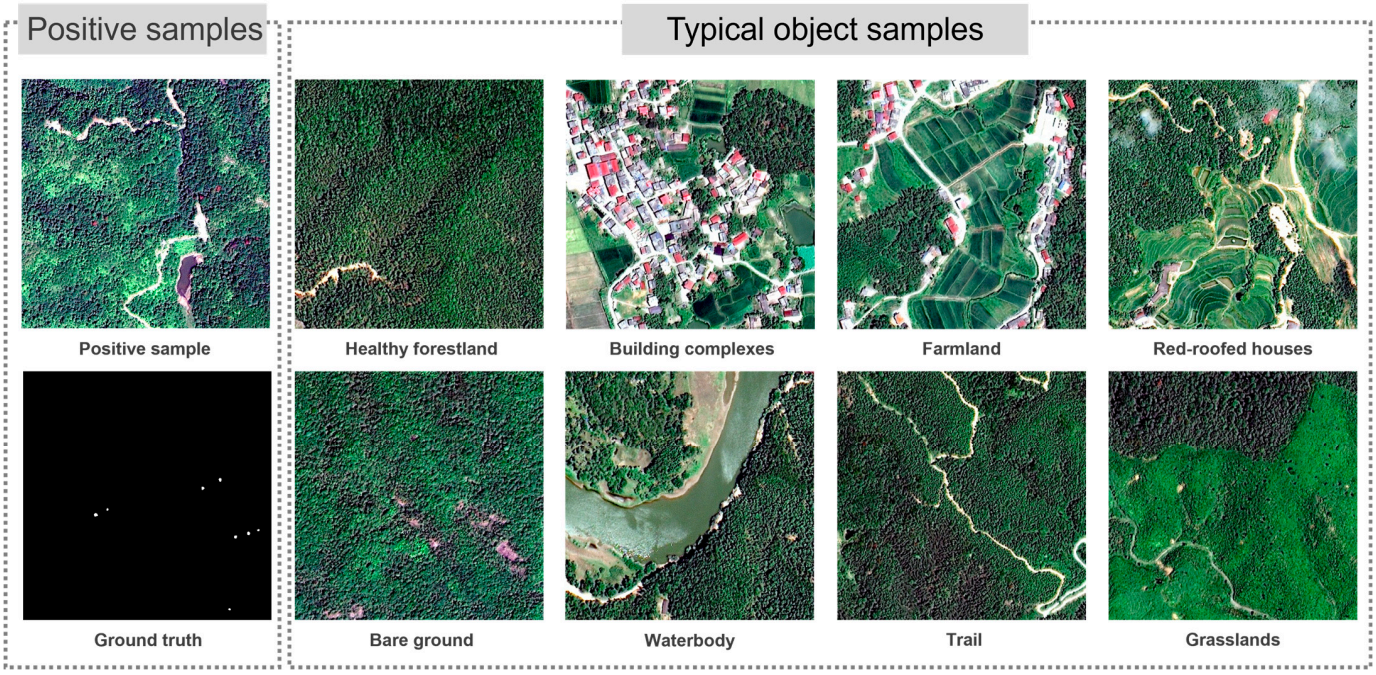


Figure 3. Example samples from the dataset. (The typical object samples do not contain any positive cases; therefore, the corresponding label is composed entirely of zeros).

Finally, sample labels drawn in ArcGIS Pro (version: 3.0.1) were used to clip the images, and a Python (version: 3.8.18) script was employed to standardize the data types of image slices and labels, facilitating their input into the model for training. This process resulted in the construction of a pine wilt disease discolored tree semantic segmentation dataset comprising 4352 sample slices. Among them, 404 slices were typical object samples, while 3948 slices contained discolored tree samples. The number of diseased and typical samples was roughly balanced across GF-2 and GF-7 imagery. Each slice measured 512 pixels in width and height and included four bands—red, green, blue, and near-infrared—along with corresponding binary labels. The dataset was randomly split into training, validation, and testing sets in a 6:2:2 ratio. The training and validation sets were used to train and evaluate the models, while the testing set was utilized for visual analysis of the model predictions.

2.4. Selection and Improvement of Visual Self-Attention Models

The principle of the self-attention algorithm is to compute the degree of association between each element in the input sequence and other elements through query and key vectors, then perform a weighted sum of value vectors based on these associations. This generates a new representation of the sequence, capturing both short-term and long-term dependencies within the sequence. In the field of image processing, self-attention algorithms can be divided into single-head attention and multi-head attention. The core idea of single-head attention is to establish the dependency relationship between each element in the feature map along a specific dimension. Multi-head attention, on the other hand, uses multiple single-head attention mechanisms to perform self-attention calculations on the input tensor separately. The results of these multiple attention heads are then fused and output (Figure 4). Compared to single-head attention, multi-head attention outputs a richer and more complex result, as it integrates information from multiple subspaces, enhancing the feature extraction ability. In this study, various forms of self-attention algorithms are systematically introduced into the field of pine wilt disease satellite monitoring, exploring their effectiveness and performance characteristics in this specific application.

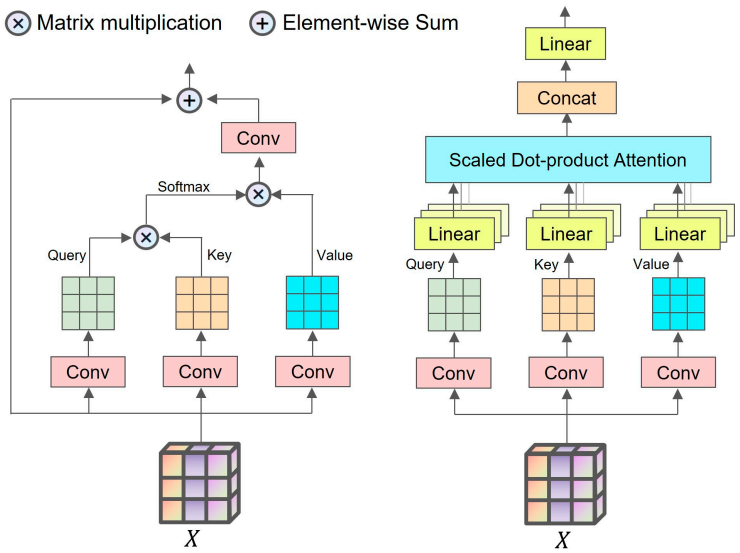


Figure 4. The structure of the single-head attention and multi-head attention module.

Considering that convolutional neural networks (CNNs) are still the most widely used deep learning method in the remote-sensing monitoring of pine wilt disease, we attempted to integrate self-attention algorithms into this task from two aspects. First, we selected the fully convolutional neural network (FCN) [25] as the baseline model for our experiment. FCN is a classical deep learning architecture widely used in semantic segmentation tasks due to its end-to-end pixel-level prediction capability and computational efficiency. Unlike traditional convolutional neural networks that rely on fully connected layers, FCN replaces them with convolutional layers, enabling it to process input images of arbitrary size while preserving spatial information throughout the network. This characteristic is particularly advantageous for remote-sensing applications, where accurate delineation of spatial features such as tree crowns is essential. We replaced the decoder head of the FCN with a single-head attention mechanism to assist in extracting features related to pine wilt disease. By observing the increase or decrease in model recognition accuracy, we could demonstrate the effectiveness of the self-attention algorithm in this task. Second, we completely discarded the traditional convolutional neural network and instead utilized the powerful

multi-head attention mechanism to form Transformer models for the feature extraction of pine wilt disease. Through a comprehensive comparison of multiple evaluation metrics, we deeply analyzed the performance of the self-attention algorithm in this task.

In the context of single-head attention models, we selected several advanced attention modules to enhance feature extraction for pine wilt disease remote-sensing monitoring. These included the non-local module [26], which utilizes a self-attention mechanism to establish dependencies between arbitrary pairs of pixels, thereby overcoming the receptive field limitations of traditional convolutional neural networks and enabling global feature extraction; the criss-cross attention module [27], which reduces computational complexity by computing attention along the horizontal and vertical directions in two passes but may suffer from information loss during the attention propagation process; the dual attention module [28], which performs self-attention in both the spatial and channel dimensions, allowing the model to simultaneously capture relationships among spatial locations and spectral channels; and the global attention module [29], which simplifies the non-local operation to efficiently model spatial dependencies and employs a Squeeze-and-Excitation (SE) mechanism to reweight the importance of each channel. For these models, ResNet50v1c [30] served as the backbone network, with the attention modules (NLNet, CCNet, DANet, and GCNet) integrated as decoder heads to refine features related to pine wilt disease.

In terms of multi-head attention models, Swin Transformer [31] and SegFormer [32] were chosen for their superior feature extraction abilities. The Swin Transformer backbone employs a hierarchical multi-head self-attention mechanism to capture pine wilt disease features at multiple scales. By restricting self-attention computation within a shifting window, it significantly reduces computational complexity. The final feature map from the backbone is passed to the decoder head to produce the PWD detection result. Similarly, the SegFormer model also utilizes an efficient multi-scale attention mechanism to focus on features at various resolutions. However, unlike Swin Transformer, SegFormer does not constrain attention to local windows, resulting in a higher computational cost but a larger effective receptive field. Additionally, SegFormer feeds features from all backbone stages into the decoder head, enabling more effective multi-scale feature fusion and improving recognition performance. To ensure parameter sizes remained comparable to single-head attention models, we employed Swin Transformer-T, SegFormer-B0, and SegFormer-B2 for PWD detection.

The model structure is shown in Figure 5. In constructing the single-head attention models, we used ResNet50v1c as the backbone network for the single-head attention models. In constructing the single-head attention models, considering that the results of self-attention computation are significantly influenced by the sizes of the query matrix (Query), key matrix (Key), and value matrix (Value), larger sizes could fully capture image features but would also increase algorithm complexity. On the other hand, smaller sizes helped to avoid excessive computational load but might blur some image features. To address this, we first adjusted the 'redu' parameter in each single-head attention module to ensure that the feature map sizes used for self-attention computation were consistent across all modules. Given the availability of sufficient computational resources, we aimed to deeply explore the capabilities of each single-head attention module in the pine wilt disease satellite monitoring task. We chose to construct an ultra-large similarity matrix with dimensions [4096, 4096] to fully establish the relationships between elements in the feature map, thereby taking advantage of the rich contextual information in high-resolution satellite imagery. After feature extraction by the backbone network, the sample generated the input for the single-head attention modules. The single-head attention modules were responsible for the secondary mining of pine wilt disease features from the backbone network output, ultimately outputting the fused result of the backbone network output

and the self-attention algorithm output, which was then passed through a classification layer to produce the final model prediction.

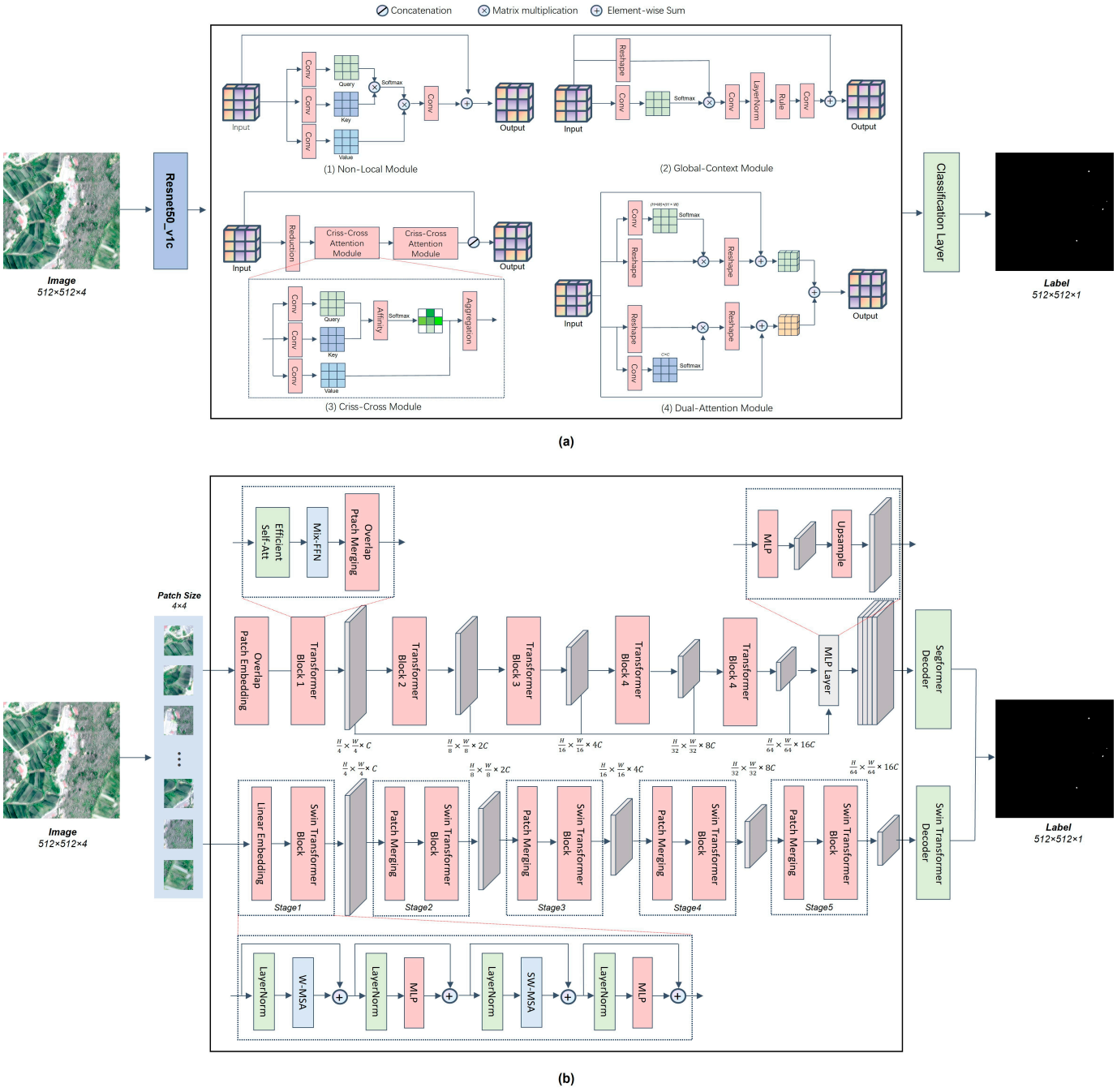


Figure 5. Model structures. (**a**) The structure of single-head attention models. (**b**) The structure of multi-head attention models.

In constructing the multi-head attention models, we first used small-scale windows to divide the sample slices into image patches, then performed linear projection on the sliced image patches, using the projection results as inputs for the two Transformer backbone networks. The resolution of the satellite imagery used in this study was 0.8 m, with the average crown width of pine wilt disease-infected trees being approximately 8 pixels. Therefore, infected trees in the image patches are classified as small targets. To enable the model to focus on fine-scale pine wilt disease features, two layers of 64-times downsample multi-head attention modules were added to the final parts of the two backbone networks. Additionally, the number of attention heads in these modules was doubled compared to the

previous stage. The multi-level feature extraction results from the backbone networks were then used as input for the decoder head, where feature fusion and final output generation took place. Additionally, to provide an objective model evaluation baseline, we constructed a fully convolutional network (FCN) as the baseline model. The backbone network used ResNet50v1c, and the decoder head did not use any special strategies, simply relying on several convolutional layers to restore the output size of the backbone network.

2.5. Model Accuracy Evaluation Metrics

During the model transfer learning phase, the effectiveness of transfer learning for each pretrained network was evaluated using the following metrics: the number of epochs required for the model to first learn diseased tree features (learning speed), the number of epochs required for the model to converge (convergence speed), and the mean Intersection over Union (IoU) of the validation set over the final ten epochs (recognition accuracy). The calculation formula is shown in Equation (1).

$$mIoU = \left(\frac{TP_{pixel}}{TP_{pixel} + FP_{pixel} + FN_{pixel}} + \frac{TN_{pixel}}{TN_{pixel} + FP_{pixel} + FN_{pixel}} \right) \quad (1)$$

where TP_{pixel} refers to the number of pixels predicted as positive that are indeed positive in the ground truth, FP_{pixel} refers to the number of pixels predicted as positive that are actually negative, TN_{pixel} refers to the number of pixels predicted as negative that are indeed negative, and FN_{pixel} refers to the number of pixels predicted as negative that are actually positive.

In the model inference result visualization analysis phase, the inference results were visualized to intuitively demonstrate the models' detection performance. The numbers of missed detections, wrong detections, and correct detections of diseased trees were statistically analyzed. Additionally, user's accuracy, producer's accuracy, and mean accuracy were calculated using the formulas provided in Equations (2)–(4).

$$Producer's\ accuracy = \frac{TP_{count}}{TP_{count} + FN_{count}} \quad (2)$$

$$User's\ accuracy = \frac{TP_{count}}{TP_{count} + FP_{count}} \quad (3)$$

$$mean\ accuracy = \frac{Producer's\ accuracy + User's\ accuracy}{2} \quad (4)$$

where TP_{count} represents the number of correctly classified diseased trees, FP_{count} represents the number of trees predicted as diseased but are in fact healthy, and FN_{count} represents the number of trees predicted as healthy but are in fact diseased

2.6. Implementation Detail

This study employed the open-source model training framework MMSegmentation [33] for transfer learning. MMSegmentation is an open-source toolbox that provides a unified framework for model implementation and evaluation in semantic segmentation tasks. The open-source scripts were first modified to support the training and inference of multi-band imagery. Within the customized MMSegmentation framework, we implemented single-head attention models, including NLNet, CCNet, DANet, and GCNet, as well as multi-head attention models, such as Swin Transformer-T, SegFormer-b0, and SegFormer-b2. Additionally, the classical fully convolutional network (FCN) was implemented as the baseline to evaluate the performance of the attention-based models.

The training environment consisted of an Ubuntu 9.4.0 operating system, with training accelerated using four NVIDIA A100-PCIE-40GB GPUs. The AdamW optimizer was employed, with a learning rate set to 0.001, a batch size of 96, and a total of 100 training epochs. The selected models have large parameter counts, in the tens of millions. Direct training from scratch would result in excessive computational overhead and poor generalization capabilities. To address these issues, the ADE20K dataset pretrained weights officially provided by MMSegmentation were loaded for transfer learning. The models trained with these transferred weights were used for inference and validation, with the results visualized for analysis.

3. Results

3.1. Transfer Learning Results for Self-Attention Models

The baseline model, FCN, along with single-head attention models (NLNet, CCNet, DANet, and GCNet), and multi-head attention models (Swin Transformer-T, SegFormer-b0, SegFormer-b2) were trained using the training set from the constructed semantic segmentation dataset for pine wilt disease. Recognition accuracy was evaluated on the validation set, and the mean Intersection over Union (mIoU) averaged over the final ten training epochs was used as the final performance metric to reduce the uncertainty introduced by the randomness of deep learning training and ensure more robust accuracy assessment. The experimental results are shown in Table 1. The study found that self-attention algorithms demonstrated an excellent performance in the fine-grained satellite-based identification of pine wilt disease. Appropriate single-head attention modules effectively assisted convolutional neural networks in extracting more meaningful features of pine wilt diseased trees. Compared to the baseline model, both the identification accuracy and performance improved significantly. Furthermore, multi-head attention models achieved a notably higher identification accuracy and superior performance compared to single-head attention models.

Table 1. Results of model transfer learning recognition.

Models	Backbone	Params (Millions)	Learning Speed (Epochs)	Convergence Speed (Epochs)	mIoU (%)
FCN (baseline)	ResNet50v1c	49.7	18	54	71.95
NLNet	ResNet50v1c	40.4	15	57	72.79 ↑
CCNet	ResNet50v1c	36.2	15	55	69.05 ↓
DANet	ResNet50v1c	46.7	10	49	**73.35** ↑
GCNet	ResNet50v1c	44.5	15	54	73.12 ↑
SegFormer-b0	Mit-b0	3.8	10	22	75.34 ↑
SegFormer-b2	Mit-b2	27.5	7	17	**76.39** ↑
Swin Transformer-T	Swin-Tiny	28.3	11	28	74.77 ↑

↑: indicates an improvement in accuracy compared to the FCN baseline; ↓: indicates a decrease in accuracy compared to the FCN baseline; Bold: highlights the highest accuracy achieved among single-head or multi-head attention models.

After transfer learning, the baseline model FCN achieved an accuracy of 71.95%. Among the four selected single-head attention modules, all but the criss-cross attention module significantly improved the pine wilt disease segmentation accuracy of the backbone network. Additionally, all single-head attention models were able to learn the color-changing tree features more quickly compared to the baseline model. Among the single-head attention models, DANet was the best-performing model, achieving the highest recognition accuracy of 73.35%, 1.4% higher than the baseline model, and was the fastest to learn the color-changing tree features and complete convergence. Furthermore, we conducted an ablation study on DANet (Table 2) and found that both the spatial self-

attention module and the channel self-attention module played a positive role in this task. Notably, the channel attention module had a more significant impact on improving recognition accuracy than the spatial module. Overall, by modeling correlations between pixels and spectral channels, DANet was able to extract a richer set of features related to PWD, making it the most effective single-head attention model. GCNet achieved the second-highest accuracy among the single-head attention models, outperforming NLNet. This can be attributed to its incorporation of a channel reweighting mechanism on top of the spatial attention design of NLNet, which proved beneficial for extracting PWD-related features. However, the approach used by DANet to model inter-channel relationships was more effective than GCNet's channel-wise importance ranking. NLNet, by leveraging a global spatial self-attention mechanism, overcomes the receptive field limitations of convolutional neural networks. It enables the model to capture the spectral, morphological, and textural characteristics of PWD at a global scale. Compared with the fully convolutional network (FCN), NLNet achieved a noticeable improvement in detection accuracy, with a gain of 0.84%.

Table 2. Ablation experiment of DANet.

Models	Backbone	PAM [1]	CAM [2]	mIoU (%)
FCN (baseline)	ResNet50v1c			71.95
DANet	ResNet50v1c	√		72.46 ↑
DANet	ResNet50v1c		√	73.03 ↑
DANet	ResNet50v1c	√	√	73.35 ↑
Models	Backbone	PAM	CAM	mIoU (%)

[1] PAM: Position Attention Module; [2] CAM: Channel Attention Module. √: indicates that the module was included in the current experiment; ↑: indicates an improvement in accuracy compared to the FCN baseline.

We also observed that not all single-head self-attention mechanisms are well-suited for the remote-sensing-based monitoring of PWD. The criss-cross attention (CCNet) model exhibited a nearly 3% decrease in recognition accuracy compared to the baseline. This performance degradation can be attributed to the structural characteristics of diseased pine trees, whose crowns typically appear circular in satellite imagery. In such cases, the pixels within a single crown tend to have a low correlation with most other pixels along the same row or column. However, the criss-cross attention mechanism enforces dependencies between each pixel and others in its row and column, and propagates attention globally through two stages of criss-cross attention. During this indirect attention transmission process, a significant portion of the spectral and morphological features of discolored trees were lost in the backbone network, ultimately resulting in a substantial drop in detection accuracy. For the remote-sensing monitoring of pine wilt disease based on convolutional neural networks, it is important to select the appropriate single-head attention algorithm to assist the network in extracting features. An inappropriate self-attention algorithm can negatively impact tree detection performance.

All multi-head attention models achieved a significantly higher identification accuracy than single-head attention models and were able to learn the features of discolored pine trees faster, converging more rapidly. The Swin Transformer-T model, with its multi-scale self-attention backbone, showed superior feature-extraction capabilities for pine wilt disease compared to CNN, achieving an accuracy of 74.77%, which is 2.82% higher than the baseline. The SegFormer model further improved performance by integrating multi-scale features directly within its decoder head, allowing it to exploit PWD features more comprehensively than the Swin Transformer. As a result, SegFormer-B2 achieved the highest recognition accuracy of 76.39%, surpassing the baseline by 4.44%. Remarkably, this model learned the features of discolored trees within only 7 epochs and converged in

just 17 epochs, demonstrating an efficient and effective training process. These findings underscore the importance of fully utilizing multi-scale features for accurate PWD detection. Furthermore, the efficiency of the SegFormer architecture is evident from the SegFormer-B0 model, which, with only about one-ninth of the parameters of Swin Transformer, achieved a higher accuracy and faster training and convergence. This surprising result suggests that future work should consider developing PWD-specific recognition models based on the SegFormer architecture to maximize performance and efficiency.

A performance scatter plot was generated based on model parameter size and recognition accuracy to analyze the efficiency of the different models (Figure 6). In this plot, points closer to the upper-left corner represent better-performing models, as they achieve a higher recognition accuracy with fewer parameters. To facilitate comparison across different model types, a vertical and a horizontal reference line were drawn through the scatter points of the baseline model FCN and the least accurate multi-head attention model, Swin Transformer-T, respectively. Models located in the upper-left region relative to this reference point exhibit a superior performance. It was observed that all multi-head attention models were positioned above and to the left of the single-head attention models and the baseline, indicating that multi-head attention models are generally more efficient. Among the single-head attention models, all except CCNet outperformed the baseline FCN.

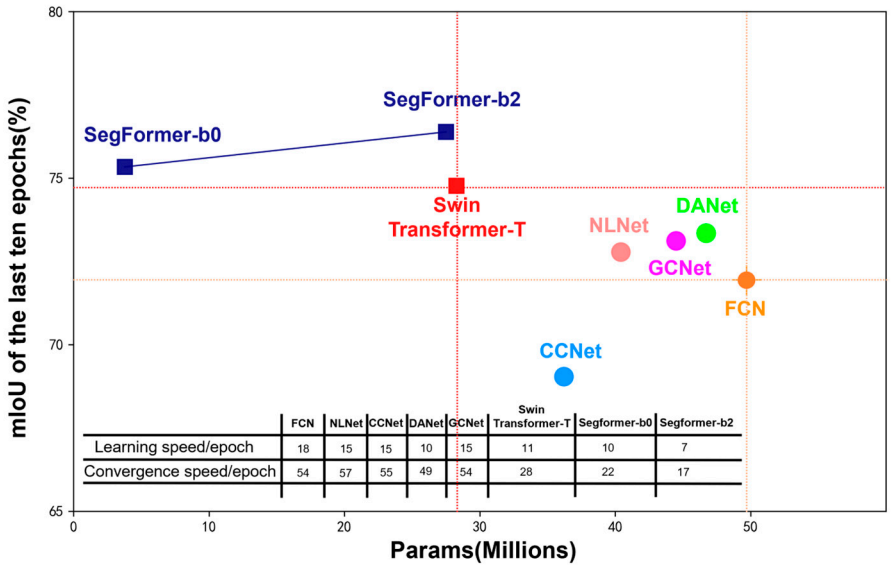


Figure 6. Scatter plot of model performance.

3.2. Visualization and Analysis of Self-Attention Model Inference Results

Evaluating models solely based on the overall recognition accuracy (mIoU) can be misleading, as global average metrics may obscure deficiencies in local recognition details, which are inconsistent with real-world application requirements. Therefore, from a practical application perspective, we performed inference on the test set using the best-performing single-head and multi-head attention models after training. The inference results were visualized and subjected to statistical analysis to further investigate the behavioral characteristics of self-attention mechanisms in the context of this task. The model identification results are shown in Table 3. It can be observed that the baseline model, FCN, based on a fully convolutional neural network, detected the highest number of discolored pine trees correctly, with 1901 trees identified, 136 missed, and 96 falsely identified. Compared to FCN, the DANet model had a slightly lower number of correctly identified trees (1889), fewer wrong detections (89), but more missed detections (148). The SegFormer-b2 model identified the fewest trees correctly (1858) and had the lowest number of wrong detec-

tions (68) but the highest number of missed detections (179). FCN exhibited the highest producer accuracy (93.32%), SegFormer-b2 had the highest user accuracy (96.51%), and DANet achieved the highest overall accuracy (94.43%). The results indicate that using self-attention algorithms reduces the number of wrong detections in identifying infected trees, thereby improving user accuracy. However, this comes at the cost of increased missed detections, leading to a noticeable decline in producer accuracy and the total number of correctly identified trees. Compared with the SegFormer model that is entirely based on the self-attention mechanism, the optimal single-head attention model DANet, which integrates convolutional neural networks with self-attention, achieves the best balance between missed and wrong detections, resulting in the highest mean accuracy.

Table 3. Results of the identification of discolored standing trees for pine wilt disease in the test set.

Models	Correct Detections	Missed Detections	Wrong Detections	Producer Accuracy (%)	User Accuracy (%)	Mean Accuracy (%)
FCN (baseline)	1901	136	96	93.32	95.19	94.25%
DANet	1889	148	89	92.73	96.13	94.43%
SegFormer-b2	1858	179	68	91.21	96.51	93.86%

Visualizations of the model inference results further clarified these findings. It was observed that the infected trees identified by the baseline FCN model appeared sharp and deformed, inconsistent with the actual shapes of infected trees and ground truth labels. In contrast, the DANet and SegFormer-b2 models, which incorporate self-attention algorithms, significantly improved this issue (Figure 7). The use of self-attention algorithms resulted in the smoother identification of infected trees, aligning better with the ground truth labels and reducing noise, thus markedly lowering the number of wrong detections compared to FCN. However, self-attention mechanisms tend to overlook infected trees with weaker feature saliency within the same image patch, leading to a high rate of missed detections (Figure 8). This issue is more pronounced in SegFormer, which relies entirely on self-attention, compared to DANet, which integrates convolutional neural networks with self-attention.

In high-resolution satellite imagery, forest scenes are highly complex, and the visual characteristics of different infected trees within the same image patch often vary. For instance, differences in the severity of infection lead to varying degrees of reflectance changes: in the early stage of pine wilt disease, the changes in canopy reflectance are subtle and difficult to detect; dead trees appear with dim coloration and are also challenging to identify; trees in the mid-to-late stages of infection exhibit the most pronounced reflectance variations and are therefore the easiest to recognize. In addition, infected trees vary in height and crown width—trees with a lower height and smaller crowns are more susceptible to shadow effects from neighboring trees, which results in reduced reflectance. Moreover, bare ground within forests tends to exhibit spectral and textural characteristics in the visible bands that are similar to those of trees in the early or mid-infection stages. Due to these interfering factors and the limited receptive field, CNNs are easily influenced by local features, leading to missed or wrong detections. In contrast, the self-attention mechanism can capture infected tree features at a global level, making it less susceptible to local interference. As a result, self-attention exhibits superior capability in modeling infected tree features. This is evident in the smoother morphology and significantly fewer wrong detections observed in masks extracted by DANet and SegFormer, compared to those produced by FCN.

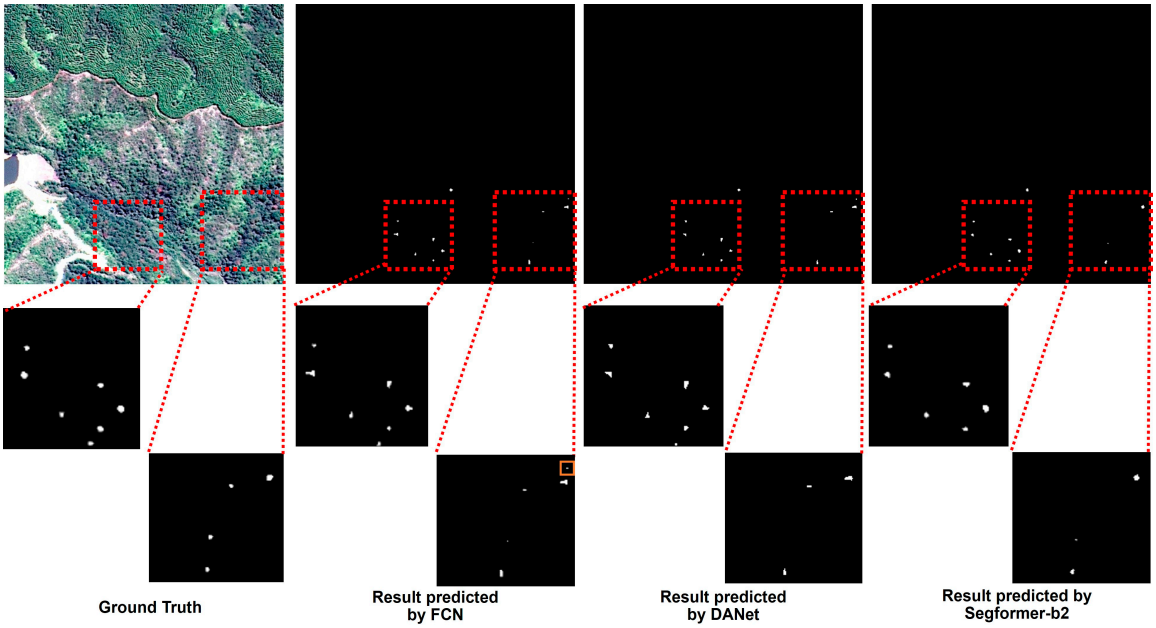


Figure 7. The self-attention algorithms make the morphology of the identified infected trees smoother. (The red boxes indicate sub-regions within the image slice where pine wilt disease outbreaks occur, while the orange box indicates wrong detection).

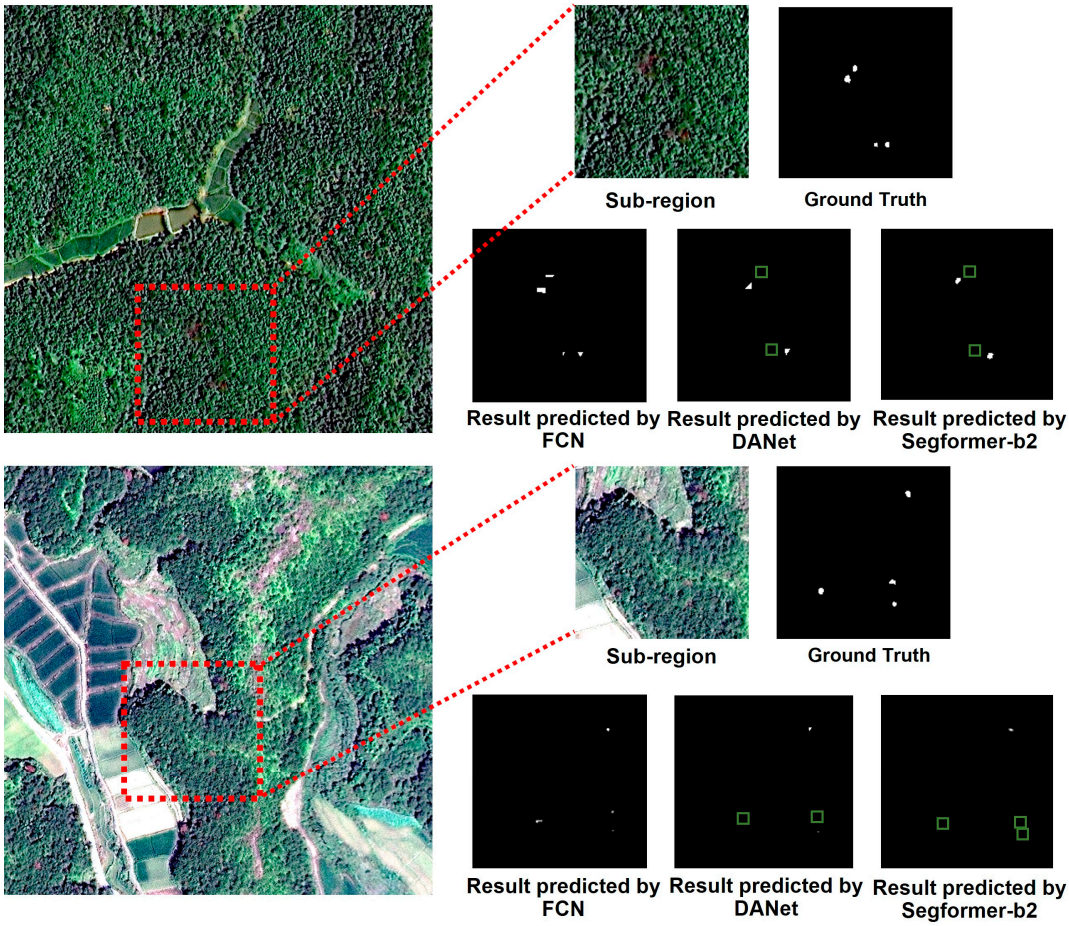


Figure 8. Self-attention algorithms are susceptible to missed detections due to differences in feature saliency among infected trees. (The red boxes indicate sub-regions within the image slice where pine wilt disease outbreaks occur, while the green boxes indicate missed detections.)

In satellite imagery, the crown width of trees infected by pine wilt disease typically spans fewer than ten pixels (approximately 8 m), which qualifies them as small targets relative to the 512 × 512 image patch size. Affected by the infection severity, tree height, crown size, and surrounding environmental conditions such as shadow and bare ground, the feature saliency of different diseased trees within the same patch varies considerably. In general, infected trees with more pronounced spectral changes, larger crown areas, and less interference from surrounding features exhibit stronger feature saliency. The study finds that self-attention-based models perform exceptionally well in detecting such highly salient infected trees. However, they struggle to allocate sufficient attention to trees with weaker feature saliency, resulting in missed detections (Figure 8). This explains why the SegFormer model, which relies entirely on self-attention, produces significantly more missed detections than FCN. DANet, although hybrid in design, also exhibits varying degrees of omission. To overcome this limitation, we suggest utilizing more extensive and higher-quality datasets to enhance the training of self-attention models, or alternatively, designing dedicated detection networks specifically for pine wilt disease.

4. Discussion

4.1. Comparison with Previous Research

Previous research on pine wilt disease monitoring based on satellite remote sensing has primarily focused on image classification, which typically only allows for detection at the scale of epidemic-prone areas [13,16,17]. This limitation has made it difficult for previous research to contribute significantly to the precise monitoring of pine wilt disease epidemics, and few researchers have attempted to conduct a detailed analysis of model performance at the inference result level to evaluate the strengths and weaknesses of the models. In contrast, this study introduces advanced self-attention techniques from the deep learning field into high-resolution satellite-based pine wilt disease monitoring, achieving efficient and accurate identification of diseased trees. This approach provides more precise recognition results at the individual tree level, which can directly support government efforts in diseased tree removal operations.

In the field of precise monitoring of pine wilt disease, Zhang et al. proposed a spatiotemporal detection method for single-tree pine wilt disease detection using high-resolution satellite imagery [12]. However, this method is cumbersome, cannot differentiate interference from land cover changes, and requires relatively short time intervals between images, which limits its applicability and makes it difficult to scale for the efficient monitoring of pine wilt disease. Wang et al. achieved single-tree recognition of pine wilt disease based on a semi-supervised model and high-resolution imagery, achieving the segmentation accuracy of 68.63% [18]. However, the semi-supervised network training process in their study was highly unstable, and the small model size led to an insufficient generalization ability, resulting in poor tree shape recognition and significant noise in the output. In contrast, this study develops self-attention models with tens of millions of parameters based on a unified framework, leveraging pretrained weights for transfer learning. In our experiments, even the weakest-performing model (CCNet) outperformed Wang et al.'s best result by 0.43%, while the best-performing model (SegFormer-B2) exceeded their accuracy by a substantial 7.76%, demonstrating a decisive advantage. Furthermore, the proposed method features a simple and stable training process, strong generalization ability, and nearly noise-free recognition results. The trained model can be directly applied to pine wilt disease diagnosis using high-resolution satellite imagery across different locations and periods, showing strong scalability and overcoming the limitations of Zhang et al.'s cumbersome method.

Compared to satellite imagery, drone imagery offers a higher spatial resolution and more detailed spectral information, making it easier to achieve precise identification of pine wilt disease. Many researchers have utilized drone technology and deep learning for the accurate identification of dead trees killed by pine wilt disease [11,34,35]. Inspired by the significant achievements of self-attention algorithms in the field of artificial intelligence, researchers have recently introduced self-attention algorithms into the drone-based remote-sensing monitoring task for pine wilt disease, achieving excellent results [23,24]. However, drone remote sensing is limited by its coverage area and high flight costs, making it difficult to scale for the global monitoring of pine wilt disease. In contrast, satellite remote sensing offers the advantages of wide coverage, low cost, and repeated observations. Furthermore, sub-meter resolution high-resolution satellite imagery is already well-suited for identifying diseased trees infected with pine wilt disease, making it highly significant for global efforts to prevent and control pine wilt disease. This study fully utilizes the rich contextual information in high-resolution satellite imagery to achieve the large-scale, accurate identification of diseased trees caused by pine wilt disease.

4.2. Research Limitations and Future Perspectives

This study has certain limitations. First, the spatial resolution of the GF-2 and GF-7 imagery used is 0.8 m, which is relatively coarse for the precise detection of PWD via satellite remote sensing. In areas of natural forest where tree crowns are densely overlapped, it is challenging to distinguish individual infected trees. The use of satellite imagery with a higher spatial resolution could help to alleviate this issue, but it may also lead to increased economic costs. Second, due to the limited spectral resolution of high-resolution satellite imagery and the unavailability of ground sampling data, the model developed in this study may be insensitive to the subtle spectral changes associated with early-stage infection, making it difficult to ensure high accuracy in identifying trees at the onset of disease. Moreover, some trees affected by other pest infestations may exhibit spectral variations similar to those of PWD-infected trees. Due to the lack of corresponding ground survey data, we were unable to eliminate the influence of such cases on the study's results. By integrating ground-based surveys, these sources of error could be constrained within a controllable range. Finally, in Section 3.2 of this study, we provided only a qualitative analysis of the causes behind the missed and wrong detections of the self-attention model, without quantifying the feature saliency of diseased trees or the attention weights allocated by the self-attention mechanism to trees with low feature saliency. Future research could consider developing a feature saliency metric for diseased trees and investigating its relationship with the attention weights assigned by self-attention models.

This study relied solely on a single high-resolution satellite imagery data source and did not explore the potential benefits of integrating multiple data sources. Technologies such as satellite imagery, UAVs, and LiDAR have proven effective in identifying changes in canopy structure, defoliation, and tree mortality. Future research could make progress in this area by leveraging data fusion approaches. Moreover, this study systematically investigated the performance of self-attention algorithms in the context of high-resolution satellite remote sensing for monitoring pine wilt disease. The findings can serve as a reference for future studies aiming to develop specialized models for PWD detection, thereby achieving the more accurate identification of infected trees.

5. Conclusions

This study utilizes self-attention to capture the rich contextual information between pine wilt disease-infected trees and the forest scene in very-high-resolution satellite imagery, significantly improving detection accuracy. Additionally, it thoroughly explores the

performance characteristics of self-attention algorithms within a unified model framework. The findings indicate that self-attention performs remarkably well in the task of pine wilt disease monitoring from very-high-resolution remote-sensing data. Single-head attention modules enhance convolutional neural networks by enabling the extraction of more disease-relevant features. Multi-head attention models exhibit significantly higher accuracy and performance than CNNs, and the diseased tree masks they produce are smoother and more closely resemble the actual crown morphology. However, their sensitivity to variations in feature saliency may lead to a high number of missed detections. In contrast, DANet—an architecture that integrates CNNs with a dual-attention module—achieves the highest mean detection accuracy by effectively reducing wrong detections without significantly increasing missed detections.

This study is the first to implement self-attention mechanisms for the fine-grained detection of pine wilt disease at the single-tree level using very-high-resolution satellite imagery. The proposed method achieves not only a high detection accuracy but also a precise spatial localization of diseased trees, addressing a key challenge in remote-sensing-based forest health monitoring. By enabling the early and accurate identification of infected individuals within complex forest environments, the approach significantly enhances the operational value of remote-sensing technologies. With strong scalability and adaptability, it holds great potential for integration into large-scale, intelligent forest disease surveillance systems, contributing to more effective ecological management and mitigation of the economic losses caused by pine wilt disease.

Author Contributions: Conceptualization, W.L.; methodology, W.L.; software, W.L.; validation, W.L.; formal analysis, W.L.; investigation, W.L.; resources, J.H.; data curation, W.L. and J.Z.; writing—original draft preparation, W.L.; writing—review and editing, J.H.; visualization, W.L.; supervision, J.H.; project administration, J.H.; funding acquisition, J.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Central Guidance on Local Science and Technology Development Fund of 2024ZY0171, the National Forestry and Grassland Administration under the Best Candidates Project, "Research and Development of the Key Technologies for Three-North Shelterbelt Forest Program" (202401-10), and the Beijing Natural Science Foundation of China "Research on Early Monitoring and Early Warning of Pine Wood Nematode Disease in Discolored Stumps Based on Deep Reinforcement Learning".

Data Availability Statement: Gaofen images can be requested from China Centre for Resources Satellite Data and Application (CRESDA, https://data.cresda.cn/#/2dMap, accessed on 27 February 2024), and ground survey site data can be obtained from the Center for Biological Disaster Prevention and Control of the State Forestry and Grassland Administration of China (BDPC, https://www.bdpc.org.cn/).

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

References

1. Hao, Z.; Huang, J.; Li, X.; Sun, H.; Fang, G. A Multi-Point Aggregation Trend of the Outbreak of Pine Wilt Disease in China over the Past 20 Years. *For. Ecol. Manag.* **2022**, *505*, 119890. [CrossRef]
2. Wang, W.; Zhu, Q.; He, G.; Liu, X.; Peng, W.; Cai, Y. Impacts of Climate Change on Pine Wilt Disease Outbreaks and Associated Carbon Stock Losses. *Agric. For. Meteorol.* **2023**, *334*, 109426. [CrossRef]
3. Huang, M.; Gong, J.; Li, S. Study on Pine Wilt Disease Hyper-Spectral Time Series and Sensitive Features. *Remote Sens. Technol. Appl.* **2012**, *27*, 954–960. [CrossRef]
4. Kim, S.-R.; Lee, W.-K.; Lim, C.-H.; Kim, M.; Kafatos, M.C.; Lee, S.-H.; Lee, S.-S. Hyperspectral Analysis of Pine Wilt Disease to Determine an Optimal Detection Index. *Forests* **2018**, *9*, 115. [CrossRef]
5. Li, N.; Huo, L.; Zhang, X. Classification of Pine Wilt Disease at Different Infection Stages by Diagnostic Hyperspectral Bands. *Ecol. Indic.* **2022**, *142*, 109198. [CrossRef]

6. Yu, R.; Luo, Y.; Ren, L. Detection of Pine Wood Nematode Infestation Using Hyperspectral Drone Images. *Ecol. Indic.* **2024**, *162*, 112034. [CrossRef]
7. Torres, P.; Rodes-Blanco, M.; Viana-Soto, A.; Nieto, H.; García, M. The Role of Remote Sensing for the Assessment and Monitoring of Forest Health: A Systematic Evidence Synthesis. *Forests* **2021**, *12*, 1134. [CrossRef]
8. Abdollahnejad, A.; Panagiotidis, D.; Surový, P.; Modlinger, R. Investigating the Correlation between Multisource Remote Sensing Data for Predicting Potential Spread of Ips Typographus L. Spots in Healthy Trees. *Remote Sens.* **2021**, *13*, 4953. [CrossRef]
9. Takenaka, Y.; Katoh, M.; Deng, S.; Cheung, K. Detecting forests damaged by pine wilt disease at the individual tree level using airborne laser data and worldview-2/3 images over two seasons. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, *XLII-3-W3*, 181–184. [CrossRef]
10. Syifa, M.; Park, S.-J.; Lee, C.-W. Detection of the Pine Wilt Disease Tree Candidates for Drone Remote Sensing Using Artificial Intelligence Techniques. *Engineering* **2020**, *6*, 919–926. [CrossRef]
11. Tao, H.; Li, C.; Zhao, D.; Deng, S.; Hu, H.; Xu, X.; Jing, W. Deep Learning-Based Dead Pine Tree Detection from Unmanned Aerial Vehicle Images. *Int. J. Remote Sens.* **2020**, *41*, 8238–8255. [CrossRef]
12. Zhang, B.; Ye, H.; Lu, W.; Huang, W.; Wu, B.; Hao, Z.; Sun, H. A Spatiotemporal Change Detection Method for Monitoring Pine Wilt Disease in a Complex Landscape Using High-Resolution Remote Sensing Imagery. *Remote Sens.* **2021**, *13*, 2083. [CrossRef]
13. Zhou, H.; Yuan, X.; Zhou, H.; Shen, H.; Ma, L.; Sun, L.; Fang, G.; Sun, H. Surveillance of Pine Wilt Disease by High Resolution Satellite. *J. For. Res.* **2022**, *33*, 1401–1408. [CrossRef]
14. Ha, J.; Moon, H.; Kwak, J.T.; Hassan, S.; Dang, L.M.; Lee, O.; Park, H. Deep Convolutional Neural Network for Classifying Fusarium Wilt of Radish from Unmanned Aerial Vehicles. *J. Appl. Remote Sens.* **2017**, *11*, 042621. [CrossRef]
15. Rançon, F.; Bombrun, L.; Keresztes, B.; Germain, C. Comparison of SIFT Encoded and Deep Learning Features for the Classification and Detection of Esca Disease in Bordeaux Vineyards. *Remote Sens.* **2018**, *11*, 1. [CrossRef]
16. Huang, J.; Lu, X.; Chen, L.; Sun, H.; Wang, S.; Fang, G. Accurate Identification of Pine Wood Nematode Disease with a Deep Convolution Neural Network. *Remote Sens.* **2022**, *14*, 913. [CrossRef]
17. Ye, W.; Lao, J.; Liu, Y.; Chang, C.-C.; Zhang, Z.; Li, H.; Zhou, H. Pine Pest Detection Using Remote Sensing Satellite Images Combined with a Multi-Scale Attention-UNet Model. *Ecol. Inform.* **2022**, *72*, 101906. [CrossRef]
18. Wang, J.; Zhao, J.; Sun, H.; Lu, X.; Huang, J.; Wang, S.; Fang, G. Satellite Remote Sensing Identification of Discolored Standing Trees for Pine Wilt Disease Based on Semi-Supervised Deep Learning. *Remote Sens.* **2022**, *14*, 5936. [CrossRef]
19. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
20. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations, Online, 2 October 2020.
21. Dong, L.; Xu, S.; Xu, B. Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 5884–5888.
22. Yang, B.; Wang, Z.; Guo, J.; Guo, L.; Liang, Q.; Zeng, Q.; Zhao, R.; Wang, J.; Li, C. Identifying Plant Disease and Severity from Leaves: A Deep Multitask Learning Framework Using Triple-Branch Swin Transformer and Deep Supervision. *Comput. Electron. Agric.* **2023**, *209*, 107809. [CrossRef]
23. Li, J. Detecting Pine Wilt Disease at the Pixel Level from High Spatial and Spectral Resolution UAV-Borne Imagery in Complex Forest Landscapes Using Deep One-Class Classification. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102947. [CrossRef]
24. Feng, H.; Li, Q.; Wang, W.; Bashir, A.K.; Singh, A.K.; Xu, J.; Fang, K. Security of Target Recognition for UAV Forestry Remote Sensing Based on Multi-Source Data Fusion Transformer Framework. *Inf. Fusion* **2024**, *112*, 102555. [CrossRef]
25. Shelhamer, E.; Long, J.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [CrossRef] [PubMed]
26. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-Local Neural Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 7794–7803.
27. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. CCNet: Criss-Cross Attention for Semantic Segmentation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 603–612.
28. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 3141–3149.

29. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Republic of Korea, 27 October–2 November 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1971–1980.
30. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 770–778.
31. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 9992–10002.
32. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In Proceedings of the Advances in Neural Information Processing Systems; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 12077–12090.
33. MMSegmentation Contributors OpenMMLab Semantic Segmentation Toolbox and Benchmark 2020. Available online: https://github.com/open-mmlab/mmsegmentation (accessed on 27 February 2024).
34. Deng, X.; Tong, Z.; Lan, Y.; Huang, Z. Detection and Location of Dead Trees with Pine Wilt Disease Based on Deep Learning and UAV Remote Sensing. *AgriEngineering* **2020**, *2*, 294–307. [CrossRef]
35. Ye, X.; Pan, J.; Shao, F.; Liu, G.; Lin, J.; Xu, D.; Liu, J. Exploring the Potential of Visual Tracking and Counting for Trees Infected with Pine Wilt Disease Based on Improved YOLOv5 and StrongSORT Algorithm. *Comput. Electron. Agric.* **2024**, *218*, 108671. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.