remote sensing

MDPI

Article

Detection Method of Infected Wood on Digital Orthophoto Map–Digital Surface Model Fusion Network

Guangbiao Wang [1,2], Hongbo Zhao [1,*], Qing Chang [1], Shuchang Lyu [1], Binghao Liu [1], Chunlei Wang [1] and Wenquan Feng [1]

[1] Department of Electronics and Information Engineering, Beihang University, Beijing 100191, China; wanggb@buaa.edu.cn (G.W.); changqing@buaa.edu.cn (Q.C.); lyushuchang@buaa.edu.cn (S.L.); liubinghao@buaa.edu.cn (B.L.); wcl_buaa@buaa.edu.cn (C.W.); buaafwq@buaa.edu.cn (W.F.)
[2] Qingdao Research Institute of Beihang University, Qingdao 266000, China
* Correspondence: bhzhb@buaa.edu.cn

Abstract: Pine wilt disease (PWD) is a worldwide affliction that poses a significant menace to forest ecosystems. The swift and precise identification of pine trees under infection holds paramount significance in the proficient administration of this ailment. The progression of remote sensing and deep learning methodologies has propelled the utilization of target detection and recognition techniques reliant on remote sensing imagery, emerging as the prevailing strategy for pinpointing affected trees. Although the existing object detection algorithms have achieved remarkable success, virtually all methods solely rely on a Digital Orthophoto Map (DOM), which is not suitable for diseased trees detection, leading to a large false detection rate in the detection of easily confused targets, such as bare land, houses, brown herbs and so on. In order to improve the ability of detecting diseased trees and preventing the spread of the epidemic, we construct a large-scale PWD detection dataset with both DOM and Digital Surface Model (DSM) images and propose a novel detection framework, DDNet, which makes full use of the spectral features and geomorphological spatial features of remote sensing targets. The experimental results show that the proposed joint network achieves an AP50 2.4% higher than the traditional deep learning network.

Keywords: pine wilt disease (PWD); digital orthophoto map (DOM); digital surface model (DSM); convolutional block attention module (CBAM)

Citation: Wang, G.; Zhao, H.; Chang, Q.; Lyu, S.; Liu, B.; Wang, C.; Feng, W. Detection Method of Infected Wood on Digital Orthophoto Map–Digital Surface Model Fusion Network. *Remote Sens.* **2023**, *15*, 4295. https://doi.org/10.3390/rs15174295

Academic Editors: Guangliang Cheng, Qi Zhao, Paolo Tripicchio and Hossein M. Rizeei

Received: 19 July 2023
Revised: 23 August 2023
Accepted: 28 August 2023
Published: 31 August 2023

1. Introduction

Pine wilt disease (PWD) caused by pine wood nematodes (PWN) poses a significant threat to pine forests due to its formidable destructive power and rapid propagation rate [1]. Nowadays, PWD is the most dangerous forest disease and is a major threat to our ecological security, biosecurity and economic development. In 1982, PWD spread to China; by the end of 2022, the epidemic has involved 701 county-level epidemic zones and 5250 township-level epidemic points nationwide and an area of 1.51 million hectares (https://www.forestry.gov.cn/search/501503 (accessed on 10 May 2023)). Pine wilt disease has the characteristics of fast spreading, short onset time and strong pathogenicity. Currently, efficient and accurate monitoring and treatment are effective means to control the spread of pine wood nematode disease.

The infected needles gradually lose their luster and change from green to yellow, before finally turning reddish brown without falling off the tree [2]. These characteristics provide the possibility to achieve the detection and localization of pine wilt-diseased trees based on spectral features. Spectrum-based methods for PWD detection, characterized by their simplicity and speed of operation, have become the primary approach for PWD monitoring. Traditional PWD monitoring is mainly based on artificial ground inspection. However, pine forest is often located in unfavorable working environments, such as high mountains, steep roads, dense forests, etc. The artificial ground survey has the disadvantages of low

efficiency, high cost and a high rate of missing detection. Over the past few years, the swift progress in geospatial information science and sensor technology has notably elevated the capacity for real-time and dynamic macro-scale Earth observation. The amalgamated Earth observation network, which integrates ground surveys, satellite remote sensing and unmanned aerial vehicles (UAVs), has found extensive application in the monitoring of geographical conditions. Furthermore, its utility has undergone extensive scrutiny and validation within the domain of PWD monitoring and detection [3].

With the development of deep learning technology, object detection is used to solve this problem at a lower cost. Since AlexNet [4] applied convolutional neural networks to the ImageNet classification task and achieved a breakthrough in performance, the research based on convolutional neural networks developed rapidly. With R-CNN [5] pioneering object detection in deep learning, Fast R-CNN [6] and Faster R-CNN [7] greatly improve the detection performance. Nowadays, deep learning object detection algorithms have evolved into two dominant directions, which are two-stage object detection algorithms based on region suggestion, such as R-CNN [5], SPP-Net [8], Fast R-CNN [6], Faster R-CNN [7], FPN [9],Cascade R-CNN [10,11],and Mask R-CNN [12], and single-stage object detection algorithms based on regression analysis, such as SSD [13], YOLO series [14–20], RetinaNet [21] and EfficientDet [22].

Employing deep learning techniques, the detection of PWD trees based on satellite and unmanned aerial vehicle (UAV) remote sensing imagery has emerged as a mainstream direction in recent studies [23,24]. Satellite remote sensing possesses characteristics such as wide monitoring coverage and multiple spectral bands, offering significant advantages in monitoring, precise localization and the assessment of discolored pine trees. This approach plays a crucial role in the field of forest pest-disease monitoring. Zhan et al. [25] compare the classification accuracies of Gaofen-2 (GF2) imagery and Sentinel-2 (S2) imagery at different spatial resolutions using pixel-based and object-based methods, providing a comprehensive analysis of the use of satellite remote sensing to detect tree mortality caused by the red turpentine beetle. Zhang et al. [26] propose a method based on multi-temporal remote sensing image comparison to solve the problem of the serious misjudgment of deciduous trees and dead grass. Li et al. [27] employ a medium-resolution satellite image analysis and simulations using an extended stochastic radiative transfer model to delineate areas affected by PWD. However, satellite remote sensing methods often achieve low PWD detection accuracy due to the constraints of the low spatial–temporal resolutions, weather complications and the challenge of capturing detailed changes, especially in cases where the number of infected trees in a forest is limited. In contrast to approaches reliant on satellite remote sensing, UAV remote sensing-based methods offer enhanced flexibility and efficiency, characterized by their ability to provide a low cost, a high spatial resolution and the accurate detection of PWD. Qin et al. [28] use UAV remote sensing images for pine nematode disease monitoring. This method proposes SCANet with a spatial information retention module to reduce the loss of spatial information and solve the problem of small targets and complex backgrounds in UAV images. Compared to DenseNet [29], HRNet [30] and other deep learning networks, Deng et al. [31] propose an improved model of Fast R-CNN, which can promote the detection of diseased trees by replacing the backbone and improving the anchor size. The above models have attained noteworthy accomplishments in PWD detection tasks. However, easily confused samples such as bare ground, houses and brown herbs are the most important constraint on detection performance and have not been intensively investigated. In view of this problem, Xu et al. [32] add dead trees to the sample database, proving that this method can effectively improve the detection rate of diseased trees.

UAV remote sensing images are typically acquired using UAVs equipped with RGB digital cameras or multi-spectral sensors. Despite the ability to capture a broader spectrum of information, multi-spectral cameras come with higher costs and exhibit a lower resolution in comparison to standard RGB digital cameras. Furthermore, their stability is diminished by the impact of various environmental factors on image quality [33]. Utilizing

RGB cameras in UAVs offers a notably more accessible option compared to subsequent approaches. Furthermore, the utilization of photogrammetry techniques and algorithms, such as Structure from Motion (SfM) and Multi-View Stereo (MVS), enables the acquisition of data beyond the confines of the imaged scenes, such as a Digital Orthophoto Map (DOM) and the Digital Surface Model (DSM) [34]. In view of this, the current study delves into PWD using remote sensing images obtained via UAVs equipped with RGB digital cameras.

In recent years, a multitude of deep learning methods for PWD detection using UAV RGB imagery have been investigated, yielding remarkable advancement [28,31,33,35–37]. However, the following challenges are still faced in practical application scenarios. (1) As shown in Figure 1a, some confusing samples are difficult to distinguish for detectors because of the similar RGB features, such as bare ground, houses, brown herbs, etc. (2) As shown in Figure 1b, the targets size distribution often spans a wide range; the targets in the image exhibit approximately a 50-fold difference in scale. (3) As shown in Figure 1c, the proportion of PWD samples in the image is very small and the unbalanced positive and negative samples make it difficult to learn valid information for the network. Furthermore, the lack of publicly available datasets specifically designed for PWD has constrained researchers from conducting extensive studies in this field, thereby affecting the introduction of more superior methodologies.

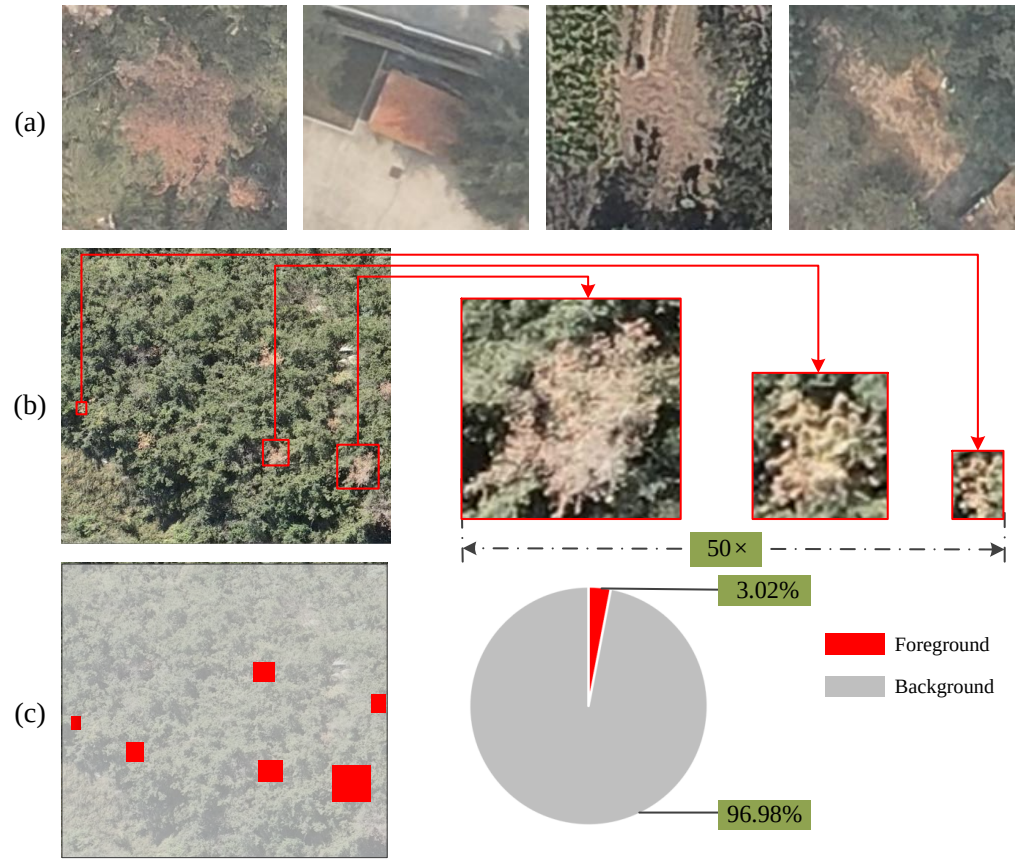


Figure 1. Challenges in PWD detection. (**a**) Easily confused targets: the first on the left is a PWD tree, and the three on the right are houses, landings and brown herbaceous plants. (**b**) Large range of target size span, with 50 times the size span. (**c**) Sparse distribution: target area accounts for only 3.02% of the overall regional area.

In this paper, we confront the challenges and problems mentioned above. Our motivation is as follows.

In response to the challenge of effectively distinguishing easily confused targets using traditional methods that solely rely on RGB-based DOM information, inspired by the research on multi-modal fusion [38–40], we endeavor to introduce a more diversified

DSM data source to reduce false detection rates through a multi-modal fusion approach. Furthermore, we investigate the "where" and "how" to fusion of multi-modal information [38,41]. Based on "where" to fuse DSM and DOM modalities, we conduct comparative analyses of fusion stages and ultimately determine the optimal DOM-DSM fusion stage for PWD detection. Regarding the aspect of "how" to fusion, the information contribution of data from different modalities varies across spatial and channel dimensions. Attention mechanisms [42–44] have the capability to autonomously allocate attention to the pertinent information. Therefore, to enhance the network's focus on valuable insights, we employ a cross-modal attention fusion mechanism to integrate DOM and DSM modality information. To enhance a better understanding of the rationale behind this study, it is imperative to elucidate the concept of the DSM. The DSM encompasses ground elevation models with the height of surface buildings, bridges and trees; can well represent the spatial features of landforms; and has been applied in the fields of ground feature classification [45–48], tree species classification [49], tree detection and delineation [50,51], ground cover change detection [26,52], etc. Through careful observation, we have observed significant distinctions in the features of buildings, bare land and dried grass on the DSM compared to those of trees on the same model, as illustrated in Figure 2. In this depiction, a pseudo-color map is utilized to represent the distribution of terrain heights, revealing that the variations in height are relatively minor for buildings, bare land and dried grass, whereas trees exhibit more substantial height changes, accompanied by circular contour lines. Based on the aforementioned analysis, we have resolved to incorporate the DSM information into the PWD detection task, with the expectation that the inclusion of the DSM data can mitigate the false detection rate for easily confused targets.

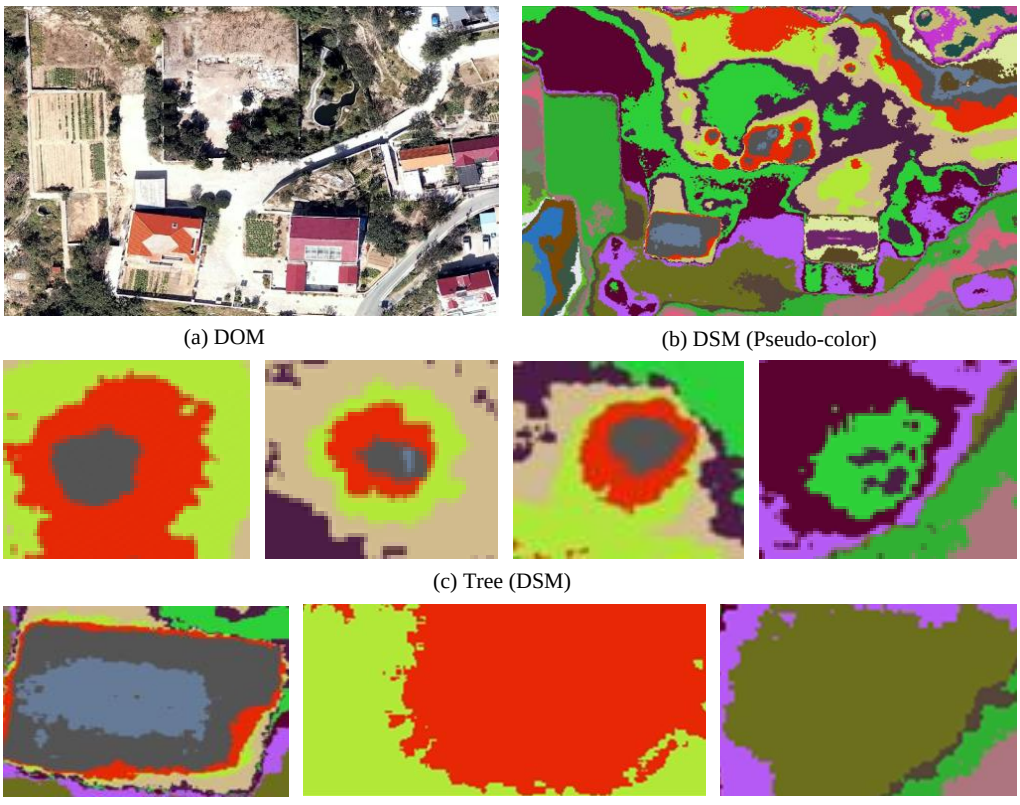


Figure 2. Comparison of features of different objects on the DSM image. (**a**) DOM in RGB, (**b**) DSM in pseudo-color house, (**c**) spatial morphology of tree canopies in DSM, (**d**–**f**) spatial morphology of house, brown herbs, bare land in DSM.

To address the significant variation in target scales, inspired by the feature pyramid network (FPN) [9,53], we introduce a feature pyramid structure, utilizing different-scale detection heads to adapt to varying object sizes in the detection tasks. In real-world scenarios, we capture pine canopy images from a top-down perspective using drones. Some trees have canopies of less than 1 square meter due to factors like occlusion, while larger trees can have canopies exceeding 50 square meters. As stated in the DSSD [53], the nodes in different layers have different receptive fields; it is natural to predict large objects from layers with large receptive fields (called higher or later layers within a ConvNet) and use layers with small receptive fields to predict small objects. Liu et al. [44] exploited the pyramid structure to excavate representative features for the buildings of various scales and shapes from local and global perspectives, respectively. Based on the above analysis, we adopt the FPN structure to address the issue of significant variations in target scales. Additionally, considering the elongated and morphological characteristics of diseased trees, we optimize the anchor radio.

To tackle the problems of sample imbalance, we design strategies from two perspectives: data augmentation and the loss function. In the context of PWD detection, the availability of positive samples is limited. One intuitive approach is data augmentation, which has been demonstrated to effectively enhance sample diversity and improve the performance of deep neural networks [17,54]. Furthermore, the PWD detection task is characterized by a substantial proportion of negative samples, resulting in an imbalance between positive and negative samples. Facing this challenge, some researchers have focused on the design of the loss function [21,55]. Their efforts have been aimed at reducing the dominance of simple negative samples in the loss weight, leading to significant improvements in performance.

Guided by the motivations mentioned above, we conduct a study on PWD detection based on DOM-DSM fusion. This paper proposes a joint DOM-DSM network, which not only introduces remote sensing target spectral features but also fuses geomorphological spatial features to improve the detection effect. The main contributions of this paper are as follows:

- We construct a large-scale PWD dataset acquired by UAVs equipped with RGB digital cameras, which contains a total of 7379 DOM-DSM image pairs (600 pixels $\times$ 600 pixels, 0.05 m resolution) and 23,235 PWD targets. To the best of our knowledge, this is the first publicly accessible dataset for PWD detection tasks.
- We propose a flexible and embedded branching network for DSM feature extraction. Alongside this, we intricately design a novel DOM-DSM multi-modal fusion approach, introducing innovative ideas for both the fusion stage and the fusion method. Building upon these foundations, we propose a novel detection framework named DDNet.
- Extensive experiments demonstrate the effectiveness of our network and achieve SOTA results on our proposed dataset. In addition, we conduct numerous ablation experiments to validate the effectiveness of our design choices in aspects such as the incorporation of DSM data, the DOM-DSM cross-modality attention module and varifocal loss.

2. Materials and Methods

2.1. PWD Dataset

Datasets are the basis for the development of deep neural networks. In recent years, thanks to the availability of large-scale datasets [56–58], deep neural object detection networks have shown increasing performance. However, in the field of PWD detection, to the best of our knowledge, there are still no large-scale publicly available datasets, which seriously restricts the improvement in PWD detection methods. Based on the above problems, this paper proposes a large-scale PWD detection dataset, which contains a total of 7379 image pairs (including one DOM and one DSM map) and 24,235 PWD targets.

2.1.1. Data Acquisition Area

The study area encompasses the entire Laoshan District (120°24′33″E∼120°43′E, 36°03′10″N∼36°20′23″N), which is situated in Qingdao City, Shandong Province, China, as indicated by the orange area in Figure 3. The topography of Laoshan District is in the order of low and middle mountains, hills, coastal plains and inter-mountain valleys in a stepped distribution, with a forested area of about 22,600 hectares. The forested area is mostly a mixed forest containing pine, chestnut, Quercus and other tree species. The study area covers a variety of features, such as forest areas, houses, bare land, mixed forests and brown herbaceous plants.

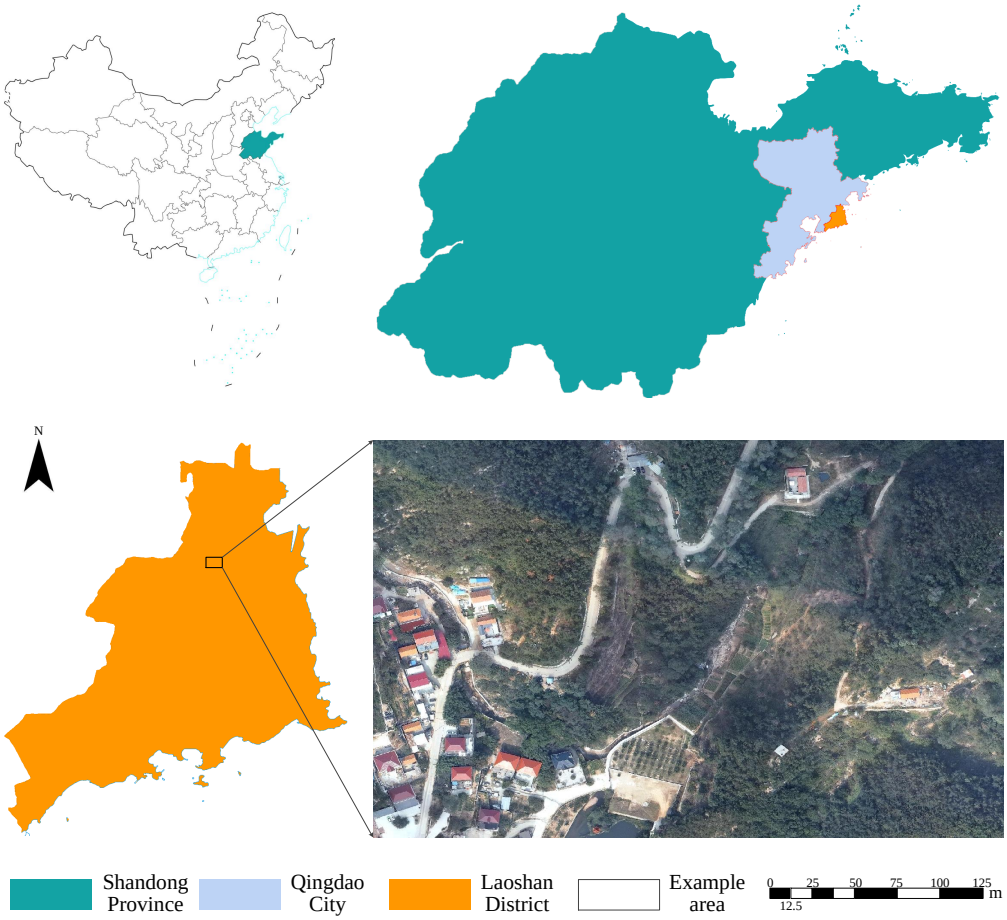


Figure 3. The geographical location of the data collection area. The research area is Laoshan District, which is the orange area in the figure. The lower right corner of the figure shows the data collected from one of the areas.

2.1.2. Dataset Production Process

According to the workflow shown in Figure 4, the process of producing a PWD detection dataset can be divided into three stages: data collection, data processing and data annotation.

(a) Data collection: We use the DJI M300RTK multi-rotor UAV platform (DJI, Shenzhen, China) with a ZenmuseP1 RGB digital camera (fixed focus lens 35 mm; 45 Megapixel CMOS sensor; an 8192 × 5460 resolution; and no need for Ground Control Points (GCPs)) to conduct data acquisition of about 20,000 hectares of forest area in Laoshan District from 8 October to 20 October 2022, collecting 61,256 images in JPG format with a resolution of approximately 0.05 m, which includes GPS information (longitude, latitude and altitude) and camera details (focal length and aperture value). To ensure the quality of the imagery, we establish a front overlap rate of 80% and a side overlap rate of 65%. Simultaneously, we ensure that the elevation variation within each flight mission remains below one-fourth of

the flying altitude, where the flying altitude is defined as the unmanned aerial vehicle's height relative to the lowest point within the specific area covered by the respective flight mission. It is crucial to underscore that spectral data exhibit distinct characteristics under varying illumination conditions, topography and atmospheric factors [59]. Such disparities can significantly undermine the efficacy of algorithms designed for target detection relying on spectral features. To ensure the uniformity of spectral data, our approach involves data collection within a suitable lighting environment, typically scheduled between 8:00 am to 10:00 am and 3:00 pm to 5:00 pm. This time frame is chosen to coincide with periods of moderate light intensity, avoiding extremes of strong or weak illumination.

(b) Data processing: We use DJI Terra software (version: 3.5.5, DJI, Shenzhen, China) to splice the collected images in batches based on the flight missions. First, we create a visible light reconstruction task. Next, we import the aerial images, select 2D reconstruction, choose a high reconstruction resolution, select the rural construction scene and opt for WGS84 as the output coordinate system. Finally, we click on "Start Reconstruction" to initiate the process. After image stitching, the software outputs DOM and DSM files, with a DOM resolution of 0.05 m and a DSM resolution of 0.1 m. In order to make the DSM resolution the same as the DOM resolution, we adjust the DSM resolution to 0.05 m by using the bilinear interpolation upsampling method. The stitched image suffers from distortion, poor resolution and warp deviation at the edges, which makes it necessary to excise the image edges.

(c) Data annotation: We design and develop our own human–machine tagging system for remote sensing images called "Tianxing", which is used for tagging PWD targets. The data annotation is performed by professional experts; furthermore, we also conduct on-site verification of certain annotated data within the forest area, which ensures the reliability of the annotated data.

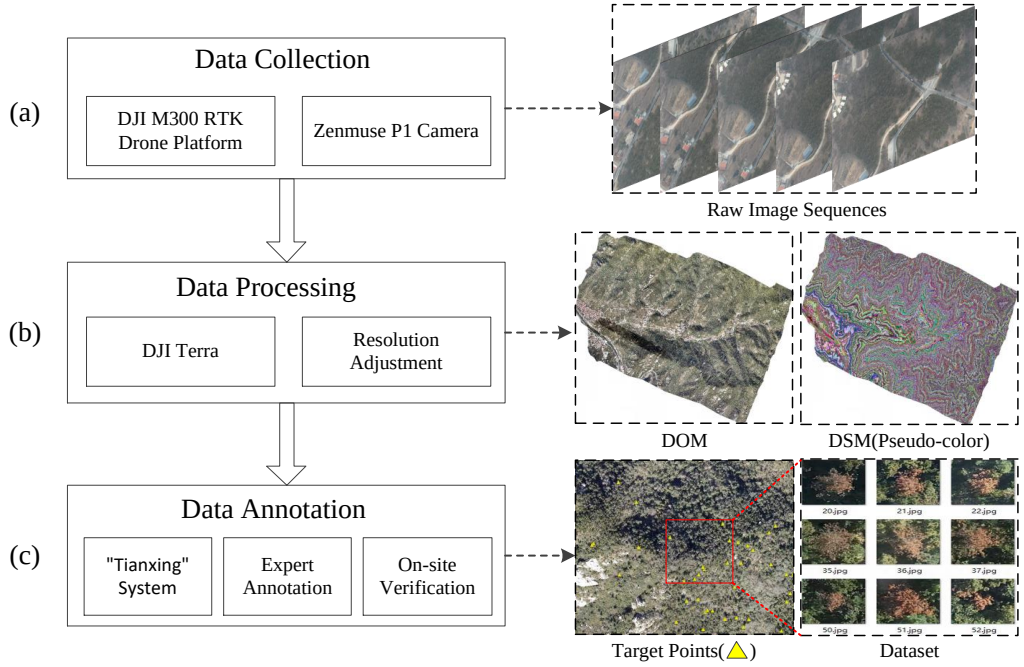


Figure 4. Flow chart of PWD dataset production. (**a**) Raw data collection, (**b**) data processing and (**c**) data annotation.

2.1.3. Introduction of PWD Dataset

Using the method described above, we construct a large-scale PWD detection dataset, as shown in Figure 5, which contains a total of 7379 pairs of sliced images (containing one DOM and one DSM map, respectively, with dimensions of 600 pixels × 600 pixels) and 24,235 PWD targets, as shown in Figure 6. Figure 6a shows the size distribution of the

targets in width and height, with a range of 10 pixels to 211 pixels in width and 9 pixels to 233 pixels in height; red dots represent target boxes with aspect ratios less than 0.5, green dots represent target boxes with aspect ratios greater than 2 and blue dots represent target boxes with aspect ratios between 0.5 and 2. Figure 6b shows the size distribution of the targets in area, with a range of 156 pixels to 42,180 pixels in the pixel values contained in the targets. Figure 6a,b illustrate that the target scales in this dataset have a great range of distribution, which reflects the true size distribution of PWD to the greatest extent, and provide a good basis for subsequent PWD object detection studies. Figure 6c shows the aspect ratio of the target frame of a PWD tree, which is concentrated around 1, which reflects the close square shape of PWD tree frames in terms of length and width, and also provides a basis for the optimal design of the anchor. The dataset is openly available at (PWD dataset, https://pan.baidu.com/s/1TTdx_pINE2sds1t-J04JCg?pwd=1e74, accessed on 10 May 2023, (pw:1e74)). Based on the information available to us, this is the first publicly accessible dataset for PWD detection tasks.

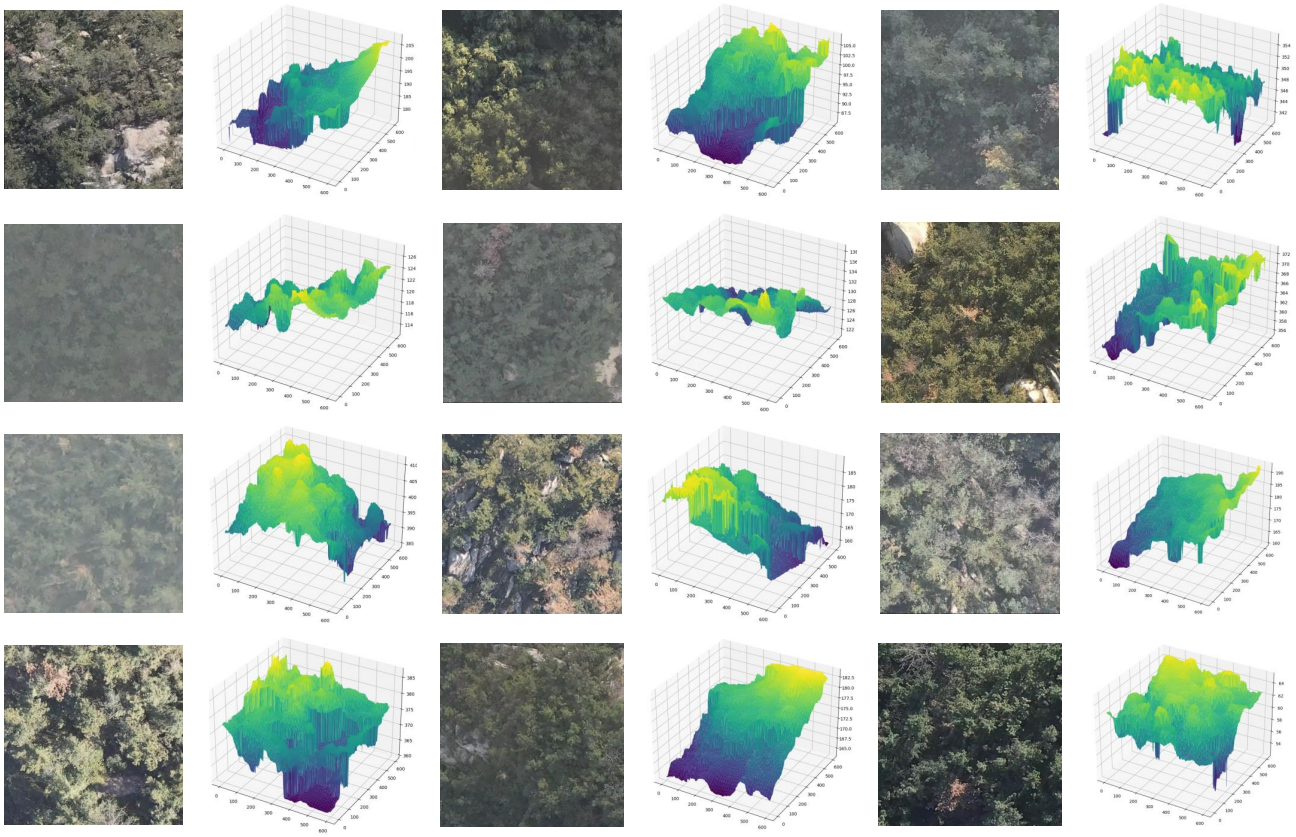


Figure 5. Some cases in the large-scale remote sensing dataset for PWD, which contains a total of 7379 image pairs (including one DOM map (**left**) and one DSM map (**right**)) and 23,235 PWD targets.

2.2. Network Structure

2.2.1. DDNet: DOM-DSM Fusion Network

Because traditional networks only use spectral features and inefficiently distinguish confusing targets such as bare ground, houses and brown herbs in the scene, we introduce significant difference features in the DSM between confusing targets and PWD trees into the detection network and design a flexible and embeddable network branch for DSM feature extraction, which can be easily adapted to a variety of networks, such as Faster RCNN [7], TPH-YOLOv5 [20], RetinaNet [21] and so on [60,61]. All that is needed is to duplicate the backbone part of the network for DSM feature extraction. The network with the addition of the DSM feature extraction branch can extract both infected tree spectral and

geomorphological spatial features; Figure 7 shows our proposed novel network structure named DDNet.

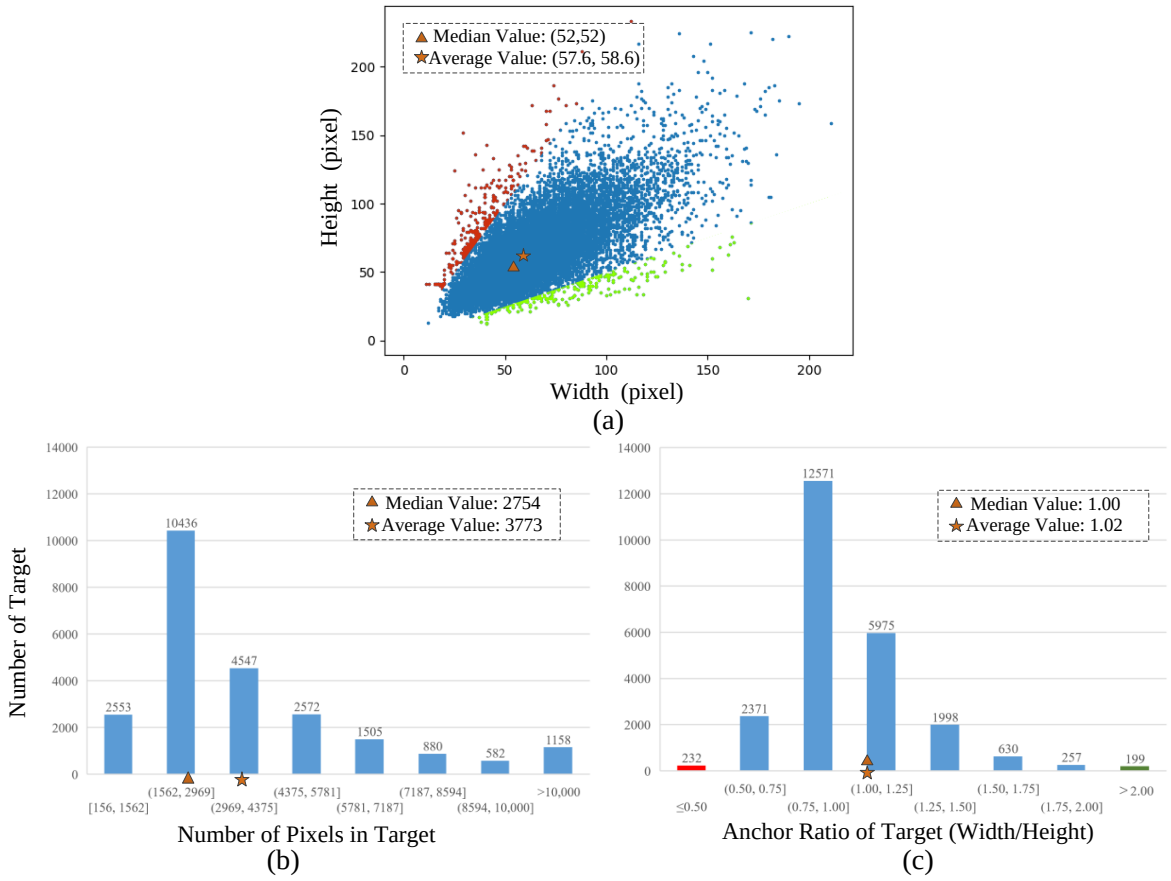


Figure 6. Distribution of PWD dataset. (**a**) shows the size distribution of the targets in width and height, with a range of 10–211 pixels in width and 9–233 pixels in height; red dots represent target boxes with aspect ratios less than 0.5, green dots represent target boxes with aspect ratios greater than 2 and blue dots represent target boxes with aspect ratios between 0.5 and 2. (**b**) shows the size distribution of the targets in area, with a range of 156–42,180 pixels in the pixel values contained in the targets. (**c**) shows that the aspect ratio of the target frame of the PWD tree is concentrated around 1, which reflects the tight square shape of the PWD trees in terms of length and width, and provides a basis for the optimal design of the anchor.

As shown in Figure 7, the DDNet network consists of three components: the backbone, FPNNeck and head. The network has two data inputs, DSM (top) and DOM (bottom), defined as $\boldsymbol{X_{DSM}}$ and $\boldsymbol{X_{DOM}}$, respectively, which are then fed into the corresponding backbone network branches (DSM branch (top) and DOM branch (bottom)) for feature extraction. To solve the problem of object detection at different scales as illustrated in Figure 1b, we select a residual network structure [21,62] to extract the feature information at different scales. Note that the DSM branch is a copy of the DOM branch with modifications only in the input channels. The DSM branch and the DOM branch output 4 feature layers, respectively, denoted as H_i and C_i, which correspond to $conv_i$ in the literature [62], where $i = 2, \cdots, 5$.

The backbone extracts multiscale features and then fuses the features in the FPNNeck module, with the two branch features passing through a structure called the "CAC", where the features of the two branches are concatenated. Then, we introduce an attention network to improve the correlated feature extracting. Finally, we use a 1×1 convolution or 3×3 convolution operation to resize the feature data. The "CAC" module can integrate the cross-modality features of the DOM and DSM, which represent the spectral features and

the feature height distribution features, respectively. The output features are denoted as $\boldsymbol{G_i}$, where $i = 3, \cdots, 6$. The mathematical formulation can be denoted as in Equation (1).

$$\boldsymbol{G_i} = \begin{cases} Conv_1(Att(Concat(C_i, H_i))), i = 3, 4, 5 \\ Conv_3(Att(Concat(C_{i-1}, H_{i-1}))), i = 6 \end{cases} \tag{1}$$

where $Conv_1(\cdot)$ and $Conv_3(\cdot)$ are convolution operations with 1×1 and 3×3 convolutional kernels, respectively. $Att(\cdot)$ is an attention operation. $Concat(\cdot)$ is a concatenation operation.

The feature pyramid network can fuse cross-scale features ($\boldsymbol{P_i}, i = 3, \cdots, 7$), which can be expressed as in Equation (2).

$$\boldsymbol{P_i} = \begin{cases} Add(\boldsymbol{G_i}, Up(\boldsymbol{P_{i+1}})), i = 3, 4 \\ \boldsymbol{G_i}, i = 5, 6 \\ MaxPooling(\boldsymbol{P_{i-1}}), i = 7 \end{cases} \tag{2}$$

where $Add(\cdot)$ is a summation operation. $Up(\cdot)$ is a 2-times upsampling operation. $MaxPooling$ is a pooling operation with 2×2 filters with stride 2.

The features fused by FPNNeck are then sent to the head network and used for target categorization and border prediction. The head network is consistent with the structures proposed in the literature [21].

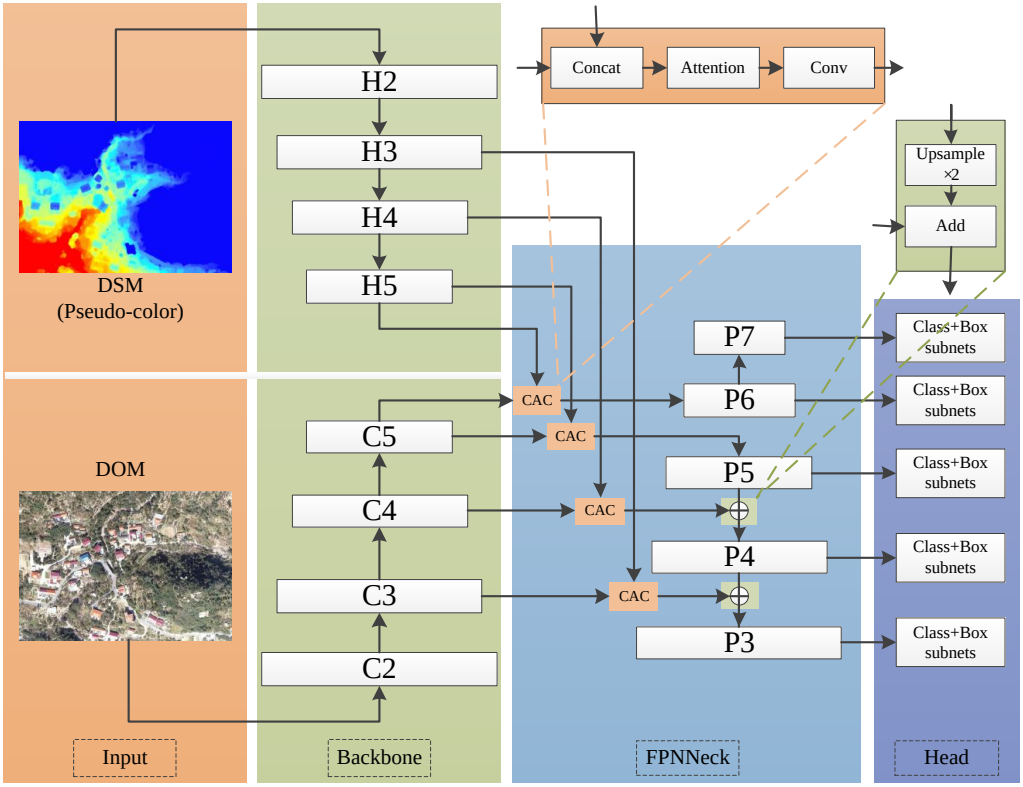


Figure 7. DDNet framework. Features of the DOM and DSM extracted from the backbone are fused in the FPNNeck. FPNNeck contains two modules, where the first is the "CAC" module for cross-modality feature fusion and the second is the FPN module for multiscale fusion. The features fused by FPNNeck are sent to the head and used for target category and border prediction.

2.2.2. DOM-DSM Cross-Modality Attention Module

A DOM is the feature of the visible light spectrum of the ground object, while the DSM is the feature of the height distribution. These two features of the object from different dimensions are highly complementary. However, the data of different modes also make it

difficult to extract effective information. To solve this problem, we use the cross-modality attention calculation, which is the "CAC" module shown in Figure 7. The "CAC" module serves the features of the DSM and DOM as inputs and then extracts effective information through the expanded attention module, which can make the network focus on important information. We embed the spatial attention and channel attention modules into the residual feature extraction network, where the structure diagram is shown in Figure 8.

Here, Figure 8a contains a channel attention module (Equation (3)) and spatial attention module (Equation (4)).

$$\boldsymbol{M_C(F)} = \sigma(MLP(AvgPool(\boldsymbol{F})) + MLP(MaxPool(\boldsymbol{F}))) \tag{3}$$

$$\boldsymbol{M_S(F')} = \sigma(f^{7\times7}(Concat(AvgPool(\boldsymbol{F'}), MaxPool(\boldsymbol{F'})))) \tag{4}$$

where $\sigma(\cdot)$ denotes a sigmoid function. $MLP(\cdot)$ consists of a reduced-dimensional convolution, a ReLU activation function and a raised-dimensional convolution. $f^{7\times7}$ represents a convolution operation with the filter size of 7×7.

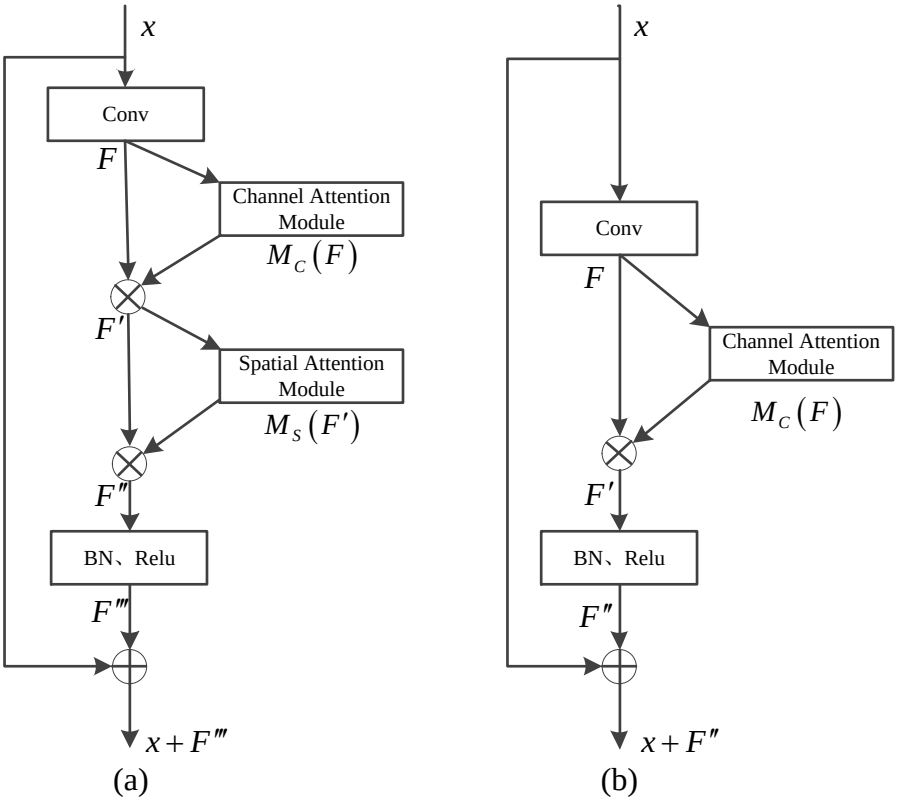


Figure 8. Cross-Modality Attention Calculation (CAC). (**a**) The structure contains channel attention and spatial attention modules, and (**b**) the structure only contains channel attention modules.

2.2.3. Optimize DDNet with Varifocal Loss

As shown in Figure 1a,c, the PWD detection scenario has the problems of confusing difficult targets and unbalanced positive and negative samples. To solve these problems, we use varifocal loss [55] to optimize the DDNet network. Compared with focal loss [21], varifocal loss (Equation (5)) only reduces the loss contribution from negative samples by scaling their losses with a factor of p^{γ} and does not reduce the positive samples in the same way. Because positive samples are extremely rare compared to negative samples, we should retain their valuable learning signal.

$$VFL(p,q) = \begin{cases} -q(qlog(p) + (1-q)log(1-p)), q > 0 \\ -\alpha p^{\gamma} log(1-p), q = 0 \end{cases} \quad (5)$$

where p is the predicted IACS (IoU-aware classification score) and q is the target score. For a foreground pixel point, q is set as the IoU between the generated bounding box and ground truth. For a background pixel point, the target q for all classes is set as 0.

2.3. Evaluation Metrics

In the task of PWD detection, our objective is to reduce both the false positive rate and the false negative rate. Therefore, we choose precision (P) and recall (R) as the evaluation metrics. Additionally, we utilize the average precision (AP), $AP50$ and $AP75$ as comprehensive evaluation indicators. Here, AP50 and AP75 represent the average precision computed at IoU thresholds of 0.5 and 0.75, respectively, where IoU corresponds to the ratio between the areas of the intersection of the corresponding pair (ground truth and inference) by union. The higher the conformity of the tracings of the two masks, the closer the IoU value is to 1. AP represents the average precision computed over a range of IoU thresholds (0.50 to 0.95 with a step size of 0.05). Furthermore, the computational time is crucial for the detection task, so we employ frames per second (fps) as a metric for assessing the detection speed. The calculation of precision, recall and AP can be formulated in Equation (6).

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN}, \quad AP = \sum_{k=1}^{N} (p(k)r(k)) \quad (6)$$

where TP represents correctly detected PWD trees; FP represents background areas mistakenly identified as PWD trees; and FN represents PWD trees missed during detection. P represents the proportion of positive samples in the identification sample. R represents the proportion of positive samples in the identification sample to all positive samples. AP is an important metric to evaluate the overall accuracy of the model, which is calculated from P and R.

3. Results

3.1. Implementation Details

3.1.1. Dataset

We divide our proposed dataset into training and testing sets in a 9:1 ratio. Specifically, the training set contains 6641 pairs of images with 21,642 epidemic targets, while the testing set contains 728 pairs of images with 2593 epidemic targets. Data enhancement methods such as RandomFlip, Pad and Mosaic are applied.

3.1.2. Experimental Setting

We design DDNet based on Pytorch and the MMDetection toolbox. All experiments are conducted on a single NVIDIA RTX2080TI GPU with 11 GB memory. We select ResNet101 and FPN as the backbone and the feature fusion neck, respectively. We also adopt the SGD optimizer, where the initial learning, momentum and weight decay are, respectively, set as 0.001, 0.9 and 0.0001. The training batch size is 5. All the experiments are run on an ubuntu 18.04 system with Pytorch version 1.8.0 and CUDA version 11.1. In the anchor design, we analyze the epidemic size distribution and set the anchor ratio to 0.7, 1 and 1.3, which experimentally prove that such a design can make the network converge faster.

3.2. Experimental Results

We conduct comparison experiments on the PWD dataset. First, we select the Faster RCNN, RetinaNet and YOLOV5 networks as the baseline. Then, we integrate the DSM branch and use the "CAC" module for feature fusion to construct a series of novel networks named Faster RCNN-DSM, RetinaNet-DSM and YOLOV5-DSM. The above networks are

compared with our proposed DDNet. Moreover, the networks all use ResNet101 as the backbone. The experimental results are shown in Table 1.

Table 1. The experiment results of infected wood using different networks. The best scores are highlighted in bold.

Models	AP50	AP75	AP	FPS
Faster RCNN	0.872	0.734	0.624	6.582
RetinaNet	0.891	0.735	0.613	12.336
YOLOv5	0.884	0.703	0.628	**14.354**
Faster RCNN-DSM	0.889	0.741	0.629	4.359
RetinaNet-DSM	0.907	0.736	0.628	10.455
YOLOv5-DSM	0.893	0.715	0.631	12.119
DDNet	**0.915**	**0.751**	**0.632**	8.759

From Table 1, we can observe that Faster RCNN-DSM, RetinaNet-DSM and YOLOV5-DSM with the DSM branch have performance improvements in the AP50 metric compared to the baseline by 1.7, 1.6 and 0.9, respectively. In AP75 and AP, the improvements are also significant. Obviously, the DSM branch can effectively improve the network's performance. Among all the methods, DDNet achieves SOTA results, which proves the effectiveness of our method. At the same time, it should also be noted that the introduction of the DSM has led to a reduction in the inference speed.

3.3. Fusion Stage Experiment

Based on "Where" to fusion multi-modal features, it is generally categorized into three types based on the different fusion stage: early fusion (data level), middle fusion (feature level) and late fusion (decision level) [41]. In this paper, the feature-level fusion approach is adopted, and the experiments are conducted at various stages to validate the superiority of the proposed fusion method.

In the early fusion experiments, similar to the approach in [50], we concatenate the RGB bands data (Red, Green and Blue) in a DOM with the single-channel data in the DSM to form four-channel data. In the late fusion experiments, independent backbone feature extraction and FPN feature fusion are performed on the DOM and DSM channels. Unlike the DDNet network, where fusion occurs before the FPNNeck, we utilize the "CAC" module for fusion just before the prediction head.

Table 2 displays the results of the different fusion stages. The experiments demonstrate that the middle fusion approach in DDNet achieves the optimal outcome in APs (AP50, AP75 and AP), and the early fusion approach exhibits the most optimal inference speed.

Table 2. The experiment results of different fusion stages. The best scores are highlighted in bold.

Fusion Stage	AP50	AP75	AP	FPS
Early fusion (data level)	0.893	0.736	0.618	**11.054**
Middle fusion (feature level)	**0.915**	**0.751**	**0.632**	8.759
Late fusion (decision level)	0.896	0.737	0.626	6.254

3.4. Visualization Experiment

In this section, we conduct a visual analysis to provide an intuitive demonstration of the method proposed in this paper. Two visual analysis experiments are presented as follows. Experiment 1: To present the overall performance of the network proposed in this paper, we perform a comparative analysis of the network discussed in Section 3.2. The results are depicted in Figure 9. Experiment 2: To further elucidate the reasons

behind the improved results achieved by incorporating the DSM data branch, without loss of generality, we choose the RetinaNet and RetinaNet-DSM networks for a Grad-CAM analysis. The outcomes of this analysis are illustrated in Figure 10.

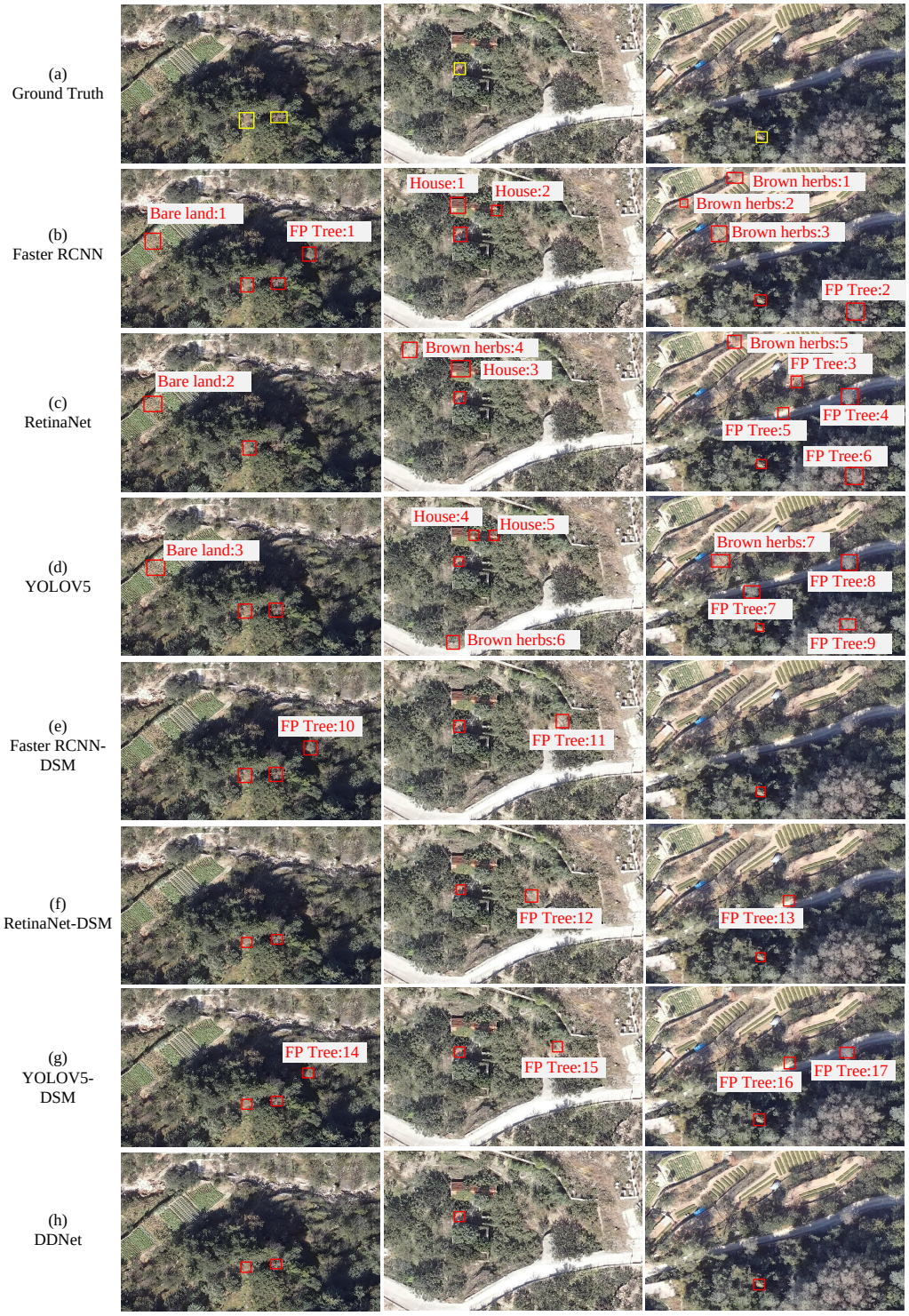
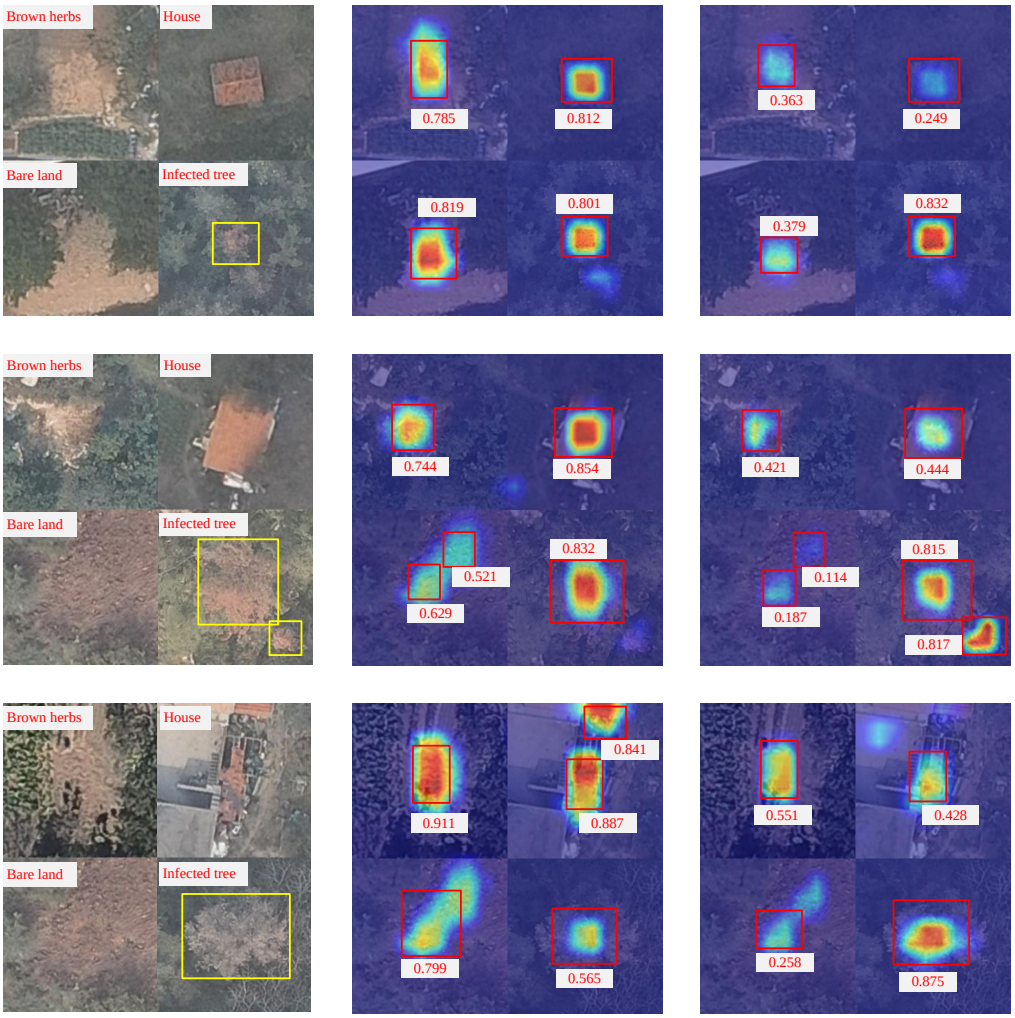


Figure 9. Visualization results of different networks. (**a**) Ground Truth, (**b**) Faster RCNN, (**c**) RetinaNet, (**d**) YOLOV5, (**e**) Faster RCNN-DSM, (**f**) RetinaNet-DSM, (**g**) YOLOV5-DSM and (**h**) DDNet.



(a) Ground Truth

(b) Grad-CAM and Results without DSM

(c) Grad-CAM and Results with DSM

Figure 10. Visualization of the results of Grad-CAM. (**a**) Ground truth, (**b**) Grad-CAM without DSM branch, (**c**) result without DSM branch.

To intuitively show the superiority of DDNet, we conduct a visualization analysis shown in Figure 9. In rows (b), (c) and (d), there exist some confusing targets, such as bare ground 1–3, houses 1–5 and brown herbs 1–7 marked in the figure. Compared to networks (b), (c) and (d), networks (e), (f) and (g) with the DSM involved intuitively filter the above-mentioned confusing targets. Compared to networks (a)–(g), our proposed DDNet shows excellent performance on both the confusing target and the "FP Tree" (a tree misidentified as a PWD tree).

As shown in Figure 10, we select three sets of data (as depicted in the three rows of data in Figure 10). Each set includes four images of brown herbs, a house, bare land and an infected tree. To provide a more intuitive analysis, we concatenate each set into a single plot (the first plot in each row). We analyze the recognition performance of these datasets using the RetinaNet (without DSM branch) and RetinaNet-DSM (with DSM branch) networks and conduct a Grad-CAM [63] analysis. Each row in the figure represents the (a) ground truth, (b) Grad-CAM without a DSM branch, (c) result without a DSM branch, (d) Grad-CAM with a DSM branch and (e) result with a DSM branch. Based on the results shown in Figure 10, the following conclusions can be drawn: (1) The Grad-CAM and prediction results in the figure maintain perfect consistency. Higher Grad-CAM values correspond to higher confidence in target detection. (2) For confusing targets, the introduction of the DSM

branch significantly reduces misidentified targets. The Grad-CAM values for confusing targets decrease substantially, while the overall Grad-CAM values for PWD targets show improvement. (3) For small-scale targets, the introduction of the DSM branch enhances the recognition capability. As shown in the second set of data, after incorporating the DSM branch, the network effectively identifies the PWD tree in the lower right corner.

3.5. Ablation Study

In Table 1 and Figure 9, we prove the effectiveness of involving the DSM module in DDNet. To further verify the effectiveness of "CAC" and varifocal loss, we conduct ablation experiments as shown in Table 3. It is clear that applying a combination of "Structure A" and varifocal loss in DDNet can obtain the best prediction results.

Table 3. Ablation study for cross-modality attention mechanism and the loss function. " Structure A" corresponds to the structure in Figure 8a; "Structure B" corresponds to the structure in Figure 8b. The best scores are highlighted in bold.

Structure A	Structure B	Focal Loss	Varifocal Loss	AP50	AP75	AP
✓	✘	✓	✘	0.912	0.746	0.628
✘	✓	✓	✘	0.909	0.741	0.626
✓	✘	✘	✓	**0.915**	**0.751**	**0.632**
✘	✓	✘	✓	0.911	0.748	0.629

Table 4 displays the results of various data augmentation techniques. The experiments demonstrate that the RandomFlip, Pad and Mosaic methods substantially enhance the model's performance, leading to a significant 1.9% improvement in AP.

Table 4. Ablation study for data augmentation. The best scores are highlighted in bold.

RandomFlip	Pad	Mosaic	AP50	AP75	AP
✘	✘	✘	0.901	0.727	0.613
✓	✘	✘	0.908	0.731	0.621
✘	✓	✘	0.907	0.728	0.619
✘	✘	✓	0.910	0.732	0.622
✓	✓	✘	0.913	0.745	0.627
✓	✘	✓	0.912	0.741	0.625
✘	✓	✓	0.913	0.747	0.629
✓	✓	✓	**0.915**	**0.751**	**0.632**

3.6. Discussion

As shown in Table 1, the networks with the DSM branch show performance improvements compared to the original network, with AP50 improvements of 1.7%, 1.6% and 0.9%, respectively. These results indicate that the DSM branch is beneficial to the network performance improvement. We believe that this improvement is due to the complementary effects of the DOM and DSM data on the representation of the target. The DSM data contain precise and dense 3D spatial coordinate information of the target but cannot reflect intuitive information, such as the chromaticity and luminance of the target. The DOM data contain rich surface texture information, semantic information of the target, etc. The fusion of the two types of remote sensing data has obvious superiority than previous methods with only single-modality data available. Simultaneously, it is imperative to acknowledge that the incorporation of the DSM has resulted in a decrease in the pace of inference. The cause for the decrease in inference speed is intuitive: the introduction of the DSM branch in the

network results in an increase in both the model's parameter count and the computational workload during inference.

Table 2 displays the results of the different fusion stages. The experimental results reveal that the middle fusion approach employed in DDNet yields the most favorable outcomes in terms of the APs (AP50, AP75 and AP). The DSM and DOM provide descriptions of the Earth's surface features from the perspectives of spectral and spatial morphology, exhibiting both modal disparities and modal correlations. The early fusion approach ignores the differences between modalities, while the late fusion approach falls short in adequately integrating cross-modal features.

Table 3 shows the ablation experiments comparison results on the "CAC" and the varifocal loss function. The results show that the performance of DDNet has been further improved by the "CAC" module and the varifocal loss function. Toward this result, we believe that the cross-modality attention calculation can enable the network to adjust different modality information and extract more effective features. Moreover, the varifocal loss can allow the network to retain more information of rare positive samples, which is beneficial to the network.

Table 4 exhibits the outcomes achieved through different data augmentation techniques. The experimental results clearly indicate that employing the RandomFlip, Pad and Mosaic methods leads to notable improvements in the model's performance, with a substantial increase of 1.9% in AP. Data augmentation can effectively enhance the diversity of data and to some extent alleviate the issue of imbalanced positive and negative samples, which is beneficial for improving network performance.

4. Conclusions

To reduce the false detection rate of easily confused targets such as bare land, houses and brown herbs in PWD detection, we propose a flexible and embeddable DOM-DSM detection network called DDNet. Our approach incorporates DSM data to enhance the detection accuracy. By utilizing a feature pyramid structure, we can effectively handle targets of various scales. The adoption of varifocal loss helps address the issue of sample imbalance. Additionally, we introduce cross-modal attention to effectively fuse spectral and spatial features of landforms. The results demonstrate that our proposed algorithm achieves state-of-the-art results, with a notable improvement of up to 2.4% in AP50. Furthermore, we substantiate the effectiveness of our approaches in aspects including the incorporation of DSM data, the DOM-DSM cross-modality attention module and varifocal loss through an extensive array of experiments. In our future research, we plan to explore the integration of DSM data into object detection, object classification and change detection tasks. Moreover, to achieve real-time PWD detection, we recognize the need for further lightweight research on our network. Simultaneously, research on algorithms robust to spectral variations is of significant value.

Author Contributions: Conceptualization, G.W. and H.Z.; methodology, G.W., C.W. and B.L.; software, G.W.; validation, S.L., Q.C. and G.W.; formal analysis, G.W. and Q.C.; writing—original draft preparation, G.W. and H.Z; writing—review and editing, W.F. and S.L.; visualization, G.W., W.F. and B.L.; supervision, H.Z. and Q.C.; funding acquisition, H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China under Grant 61901015.

Data Availability Statement: The dataset of this paper is openly available at (PWD dataset, https://pan.baidu.com/s/1TTdx_pINE2sds1t-J04JCg?pwd=1e74, accessed on 10 May 2023, (pw:1e74)).

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

References

1. Futai, K. Pine wood nematode, Bursaphelenchus xylophilus. *Annu. Rev. Phytopathol.* **2013**, *51*, 61–83. [CrossRef] [PubMed]
2. Li, M.; Li, H.; Ding, X.; Wang, L.; Wang, X.; Chen, F. The detection of pine wilt disease: A literature review. *Int. J. Mol. Sci.* **2022**, *23*, 10797. [CrossRef] [PubMed]
3. Wu, W.; Zhang, Z.; Zheng, L.; Han, C.; Wang, X.; Xu, J.; Wang, X. Research progress on the early monitoring of pine wilt disease using hyperspectral techniques. *Sensors* **2020**, *20*, 3729. [CrossRef]
4. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1106–1114. [CrossRef]
5. Girshick, R.B.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
6. Girshick, R.B. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
7. Ren, S.; He, K.; Girshick, R.B.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
8. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
9. Lin, T.; Dollár, P.; Girshick, R.B.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
10. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
11. Cheng, G.; Wang, Y.; Xu, S.; Wang, H.; Xiang, S.; Pan, C. Automatic Road Detection and Centerline Extraction via Cascaded End-to-End Convolutional Neural Network. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3322–3337. [CrossRef]
12. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
13. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the European Conference Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Volume 9905, pp. 21–37.
14. Redmon, J.; Divvala, S.K.; Girshick, R.B.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 July 2016; pp. 779–788.
15. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
16. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
17. Bochkovskiy, A.; Wang, C.; Liao, H.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
18. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO Series in 2021. *arXiv* **2021**, arXiv:2107.08430.
19. Wang, C.; Bochkovskiy, A.; Liao, H.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:2207.02696.
20. 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 IEEE International Conference on Computer Vision Workshops, Montreal, BC, Canada, 11–17 October 2021; pp. 2778–2788.
21. Lin, T.; Goyal, P.; Girshick, R.B.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2999–3007.
22. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10778–10787.
23. Lee, S.H.; Cho, H.K.; Lee, W.K. Detection of the pine trees damaged by pine wilt disease using high resolution satellite and airborne optical imagery. *대한원격탐사학회지* **2007**, *23*, 409–420.
24. 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]
25. Zhan, Z.; Yu, L.; Li, Z.; Ren, L.; Gao, B.; Wang, L.; Luo, Y. Combining GF-2 and Sentinel-2 images to detect tree mortality caused by red turpentine beetle during the early outbreak stage in North China. *Forests* **2020**, *11*, 172. [CrossRef]
26. 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]
27. Li, X.; Tong, T.; Luo, T.; Wang, J.; Rao, Y.; Li, L.; Jin, D.; Wu, D.; Huang, H. Retrieving the Infected Area of Pine Wilt Disease-Disturbed Pine Forests from Medium-Resolution Satellite Images Using the Stochastic Radiative Transfer Theory. *Remote Sens.* **2022**, *14*, 1526. [CrossRef]
28. Qin, J.; Wang, B.; Wu, Y.; Lu, Q.; Zhu, H. Identifying Pine Wood Nematode Disease Using UAV Images and Deep Learning Algorithms. *Remote Sens.* **2021**, *13*, 162. [CrossRef]
29. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.

30. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep High-Resolution Representation Learning for Human Pose Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5693–5703.
31. Deng, X.; Zejing, T.; Yubin, L.; 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]
32. Xu, X.; Tao, H.; Li, C.; Cheng, C.; Guo, H.; Zhou, J. Detection and location of pine wilt disease induced dead pine trees based on Faster R-CNN. *Trans. Chin. Soc. Agric. Mach.* **2020**, *51*, 228–236.
33. You, J.; Zhang, R.; Lee, J. A deep learning-based generalized system for detecting pine wilt disease using RGB-based UAV images. *Remote Sens.* **2021**, *14*, 150. [CrossRef]
34. Pádua, L.; Vanko, J.; Hruška, J.; Adão, T.; Sousa, J.J.; Peres, E.; Morais, R. UAS, sensors, and data processing in agroforestry: A review towards practical applications. *Int. J. Remote Sens.* **2017**, *38*, 2349–2391. [CrossRef]
35. Sun, Z.; Wang, Y.; Pan, L.; Xie, Y.; Zhang, B.; Liang, R.; Sun, Y. Pine wilt disease detection in high-resolution UAV images using object-oriented classification. *J. For. Res.* **2022**, *33*, 1377–13893. [CrossRef]
36. Wu, B.; Liang, A.; Zhang, H.; Zhu, T.; Zou, Z.; Yang, D.; Tang, W.; Li, J.; Su, J. Application of conventional UAV-based high-throughput object detection to the early diagnosis of pine wilt disease by deep learning. *For. Ecol. Manag.* **2021**, *486*, 118986. [CrossRef]
37. Li, F.; Liu, Z.; Shen, W.; Wang, Y.; Wang, Y.; Ge, C.; Sun, F.; Lan, P. A remote sensing and airborne edge-computing based detection system for pine wilt disease. *IEEE Access* **2021**, *9*, 66346–66360. [CrossRef]
38. Hong, D.; Gao, L.; Yokoya, N.; Yao, J.; Chanussot, J.; Du, Q.; Zhang, B. More Diverse Means Better: Multimodal Deep Learning Meets Remote Sensing Imagery Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 4340–4354. [CrossRef]
39. Huang, Y.; Du, C.; Xue, Z.; Chen, X.; Zhao, H.; Huang, L. What Makes Multimodal Learning Better than Single (Provably). *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 10944–10956.
40. Hong, D.; Gao, L.; Yao, J.; Zhang, B.; Plaza, A.; Chanussot, J. Graph Convolutional Networks for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 5966–5978. [CrossRef]
41. Zhang, P.; Du, P.; Lin, C.; Wang, X.; Li, E.; Xue, Z.; Bai, X. A Hybrid Attention-Aware Fusion Network (HAFNet) for Building Extraction from High-Resolution Imagery and LiDAR Data. *Remote Sens.* **2020**, *12*, 3764. [CrossRef]
42. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef]
43. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the Computer Vision—ECCV 2018, Lecture Notes in Computer Science, Munich, Germany, 8–14 September 2018; pp. 3–19. [CrossRef]
44. Liu, T.; Gong, M.; Lu, D.; Zhang, Q.; Zheng, H.; Jiang, F.; Zhang, M. Building Change Detection for VHR Remote Sensing Images via Local-Global Pyramid Network and Cross-Task Transfer Learning Strategy. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4704817. [CrossRef]
45. Lewis, Q.W.; Edmonds, D.A.; Yanites, B.J. Integrated UAS and LiDAR reveals the importance of land cover and flood magnitude on the formation of incipient chute holes and chute cutoff development. *Earth Surf. Process. Landf.* **2020**, *45*, 1441–1455. [CrossRef]
46. Olmanson, L.G.; Bauer, M.E. Land cover classification of the Lake of the Woods/Rainy River Basin by object-based image analysis of Landsat and lidar data. *Lake Reserv. Manag.* **2017**, *33*, 335–346. [CrossRef]
47. Zhao, Q.; Ma, Y.; Lyu, S.; Chen, L. Embedded Self-Distillation in Compact Multibranch Ensemble Network for Remote Sensing Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
48. Zhao, Q.; Lyu, S.; Li, Y.; Ma, Y.; Chen, L. MGML: Multigranularity Multilevel Feature Ensemble Network for Remote Sensing Scene Classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *34*, 2308–2322. [CrossRef]
49. Wu, Y.; Zhang, X. Object-Based Tree Species Classification Using Airborne Hyperspectral Images and LiDAR Data. *Forests* **2019**, *11*, 32. [CrossRef]
50. Lucena, F.; Breunig, F.M.; Kux, H. The Combined Use of UAV-Based RGB and DEM Images for the Detection and Delineation of Orange Tree Crowns with Mask R-CNN: An Approach of Labeling and Unified Framework. *Future Internet* **2022**, *14*, 275. [CrossRef]
51. Hao, Z.; Lin, L.; Post, C.J.; Mikhailova, E.A.; Li, M.; Chen, Y.; Yu, K.; Liu, J. Automated tree-crown and height detection in a young forest plantation using mask region-based convolutional neural network (Mask R-CNN). *ISPRS J. Photogramm. Remote Sens.* **2021**, *178*, 112–123. [CrossRef]
52. Cheng, G.; Huang, Y.; Li, Y.; Lyu, S.; Xu, Z.; Zhao, Q.; Xiang, S. Change Detection Methods for Remote Sensing in the Last Decade: A Comprehensive Review. *arXiv* **2023**, arXiv:2305.05813.
53. Fu, C.Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. Dssd: Deconvolutional single shot detector. *arXiv* **2017**, arXiv:1701.06659.
54. Lv, Z.; Huan, H.; Jia, M.; Benediktsson, J.; Chen, F. Iterative Training Sample Augmentation for Enhancing Land Cover Change Detection Performance With Deep Learning Neural Network. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, 1–14. [CrossRef]
55. Zhang, H.; Wang, Y.; Dayoub, F.; Sunderhauf, N. VarifocalNet: An IoU-Aware Dense Object Detector. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8514–8523.
56. Lin, T.; Maire, M.; Belongie, S.J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the European Conference Computer Vision, Zurich, Switzerland, 6–12 September 2014; Volume 8693, pp. 740–755.

57. Everingham, M.; Gool, L.V.; Williams, C.K.I.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
58. Xia, G.; 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 Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3974–3983.
59. Hong, D.; Yokoya, N.; Chanussot, J.; Zhu, X.X. An augmented linear mixing model to address spectral variability for hyperspectral unmixing. *IEEE Trans. Image Process.* **2018**, *28*, 1923–1938. [CrossRef]
60. Zhao, Q.; Liu, B.; Lyu, S.; Wang, C.; Zhang, H. TPH-YOLOv5++: Boosting Object Detection on Drone-Captured Scenarios with Cross-Layer Asymmetric Transformer. *Remote Sens.* **2023**, *15*, 1687. [CrossRef]
61. Wang, Q.; Feng, W.; Yao, L.; Zhuang, C.; Liu, B.; Chen, L. TPH-YOLOv5-Air: Airport Confusing Object Detection via Adaptively Spatial Feature Fusion. *Remote Sens.* **2023**, *15*, 3883. [CrossRef]
62. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
63. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.