forests

MDPI

Article

A Detection Method for Individual Infected Pine Trees with Pine Wilt Disease Based on Deep Learning

Yan Zhou [1,2], Wenping Liu [1,2,*], Haojie Bi [3], Riqiang Chen [1,2], Shixiang Zong [3] and Youqing Luo [3]

[1] College of Information, Beijing Forestry University, Beijing 100083, China
[2] Engineering Research Center for Forestry-Oriented Intelligent Information Processing of National Forestry and Grassland Administration, Beijing, 100083, China
[3] College of Forestry, Beijing Forestry University, Beijing 100083, China
* Correspondence: wendyl@vip.163.com

Abstract: Pine wilt disease (PWD) can cause destructive death in many species of pine trees within a short period. The recognition of infected pine trees in unmanned aerial vehicle (UAV) forest images is a key technology for automatic monitoring and early warning of pests. This paper collected UAV visible and multispectral images of Korean pines (*Pinus koraiensis*) and Chinese pines (*P. tabulaeformis*) infected by PWD and divided the PWD infection into early, middle, and late stages. With the open-source annotation tool, LabelImg, we labeled the category of infected pine trees at each stage. After coordinate-correction preprocessing of the ground truth, the Korean pine and Chinese pine datasets were established. As a means of detecting infected pine trees of PWD and determining different infection stages, a multi-band image-fusion infected pine tree detector (MFTD) based on deep learning was proposed. Firstly, the Halfway Fusion mode was adopted to fuse the network based on four YOLOv5 variants. Simultaneously, the Backbone network was initially designed as a dual branching network that includes visible and multispectral subnets. Moreover, the features of visible and multispectral images were extracted. To fully utilize the features of visible and multispectral images, a multi-band feature fusion transformer (MFFT) with a multi-head attention mechanism and a feed-forward network was constructed to enhance the information correlation between visible and multispectral feature maps. Finally, following the MFFT module, the two feature maps were fused and input into Neck and Head to predict the categories and positions of infected pine trees. The best-performing MFTD model achieved the highest detection accuracy with mean average precision values (mAP@50) of 88.5% and 86.8% on Korean pine and Chinese pine datasets, respectively, which improved by 8.6% and 10.8% compared to the original YOLOv5 models trained only with visible images. In addition, the average precision values (AP@50) are 87.2%, 93.5%, and 84.8% for early, middle, and late stages on the KP dataset and 81.2%, 92.9%, and 86.2% on the CP dataset. Furthermore, the largest improvement is observed in the early stage with 14.3% and 11.6%, respectively. The results show that MFTD can accurately detect the infected pine trees, especially those at the early stage, and improve the early warning ability of PWD.

Keywords: unmanned aerial vehicle; pine wood nematode; convolutional neural network; object detection; YOLOv5

check for updates

Citation: Zhou, Y.; Liu, W.; Bi, H.; Chen, R.; Zong, S.; Luo, Y. A Detection Method for Individual Infected Pine Trees with Pine Wilt Disease Based on Deep Learning. *Forests* **2022**, *13*, 1880. https://doi.org/10.3390/f13111880

Academic Editor: Juan Antonio Martin

Received: 7 October 2022
Accepted: 8 November 2022
Published: 9 November 2022

1. Introduction

Pine wilt disease (PWD) caused by the pine wood nematode (PWN, *Bursaphelenchus xylophilus*) develops and spreads rapidly with an extremely high mortality rate, making it difficult to control [1]. Since its appearance in North America, it has expanded to many countries and regions in Europe and Asia, posing a severe threat to global ecological security [2]. China is the most seriously affected by PWD, among other nations. Since it was discovered in China in 1982, PWN, an exotic pest that significantly depletes Chinese forest resources, had covered an area of approximately 1.72 million hm2 by the end of 2021,

with over 14 million infected trees and economic losses reaching 300 billion yuan, and continues to spread, threatening the 60 million hm2 of pine forest resources in China [2,3]. Most of the pine trees that have contracted PWD progress from the initial infection to the serious infection stage within 5 weeks [4]. Therefore, it is necessary to continuously monitor suitable habitats for PWN and identify the infected pine trees in the early infection stage to avoid the further spread of the disease and effectively control PWD.

Currently, the general approach to managing PWD mainly involves burning, fumigation, and tree felling after an outbreak [5,6]. Thus, the identification of individual infected pine trees is a crucial step before PWD management. As a result of missing the early infected pine trees, PWD will continue to spread and find new sources of transmission. In recent years, unmanned aerial vehicles (UAVs) have been applied in forestry pest monitoring due to their advantages of easy operation, low cost, high efficiency, etc. [7,8]. In the practical application of UAV forestry pest monitoring, the extraction of infected pine trees from UAV images still relies heavily on manual screening, which primarily uses manual observation or common AI methods [9] to identify the infected area in UAV images and realize the manual identification and localization of infected pine trees in the forest. This method includes drawbacks such as low efficiency, a high misjudgment rate, and imprecise localization of infected pine trees. A pressing issue that has to be resolved presently is how to identify infected pine trees with PWD quickly and reliably in a complex natural environment and achieve accurate and efficient intelligent monitoring of PWD at the early stage.

With the continuous development of computer technology, UAVs combined with computer algorithms such as image analysis, machine learning, and pattern recognition have been widely used in forest protection to achieve infected tree identification [10–12]. These methods only consider low-level features such as color and texture and they are artificially designed, which greatly prolongs the identification cycle and prevents the timely determination of pine trees infected with early PWD. Thus, although these methods improved the efficiency of manual investigation to some extent, they cannot fully achieve automatic identification of infected trees at an early stage. In recent years, the object detection method based on deep learning and the utilization of a Convolutional Neural Network (CNN) independent of pre-processing has been able to automatically complete accurate object detection without the need to design features and classifiers manually. In UAV images, some approaches have contributed to progress in vegetation recognition and pest detection. By combining DCNN, deep Convolutional GAN, and the AdaBoost classifier, Hu et al. [13] efficiently identified pine trees whose color changed on the occasion of insufficient training samples. Deng et al. [14] proposed an improved Faster-RCNN detection algorithm to locate and identify withered and dead trees in UAV visible images. Tao et al. [15] introduced CNNs of ALexNet and GoogleNet to effectively detect dead pine trees in UAV visible images. However, the above-mentioned deep-learning-based object detection methods to distinguish healthy pine trees and PWD-infected trees can only rely on the features of visible images. It is tricky to differentiate infected pine trees from healthy pine trees in visible images since infected pine trees during the early stage are still yellow-green and have very similar color characteristics to healthy ones. In addition, as the aforementioned methods can only identify infected pine trees using a binary (yes or no) classification method, it is impossible to track the progression of PWD and determine the optimal time for disease prevention and control. Multispectral remote sensing data have always been a research hotspot in forestry applications, namely in the field of forest pest monitoring via remote sensing technology [16]. Yu [17] et al. adopted two deep learning methods, Faster R- CNN [18] and YOLOv4 [19], to detect infected pine trees at different stages in UAV multispectral pictures. However, the highest accuracy rate was only 66.7%. Although visible and multispectral images have been widely used to recognize infected trees, there is still no study that provides an accurate method that can detect each infection stage, especially the early stage, using UAV images.

Nowadays, many computer vision tasks apply multi-mode data fusion as the input form of the deep learning network. For example, in the field of automatic driving and

robotics, visible images are combined with deep information of images or LiDAR data to detect three-dimensional objects [20]. At the same time, they are mixed with infrared images to detect pedestrians or vehicles [21]. The data from the alternative mode can improve the model detection ability, further guide network learning, and offset visible image detection, which is easily affected by ambient brightness, target movement, and other shooting conditions. The most significant change in external characteristics of an infected pine tree is the start of the pine needles withering and discoloration [22]. There are clear distinctions of image features between the crowns of infected trees and healthy trees in the multispectral images, which is more obvious than in visible images. Since it is more convenient to discriminate between different plant species from visible images, the fusion of visible and multispectral image features can enhance the recognition ability of the detection algorithm in deep learning for infected pine trees during an early stage and increase the accuracy of classification at different infection stages.

Among all the current lightweight object detection networks, YOLO [23] directly predicts the category and location of objects by using the characteristics of the input image with a compact number of model parameters and high computational speed. With continuous improvement from V1 to V5 [19,24,25], it has gradually become the mainstream network of object detection, of which the performance grants a state-of-the-art network. Transformer [26], proposed in natural language processing, adopts the self-attitude mechanism, which aggregates information from each feature of the input sequence to realize global computations. The application of Transformer in object detection [27,28] can create the receptive fields of each feature map while realizing end-to-end training of object detectors. Thus, adding the Transformer structure to the YOLOv5 [29] network can more fully utilize the semantic information from visible and multispectral image features due to the enhancement of the correlation between features of different bands. While improving the accuracy of infected pine trees, the training and detection speeds of the models are maintained as much as possible.

Therefore, we propose a new method to fuse visible images and multispectral images in the convolution process focused on the problem of complex backgrounds in images acquired by multispectral UAVs and the difficulty of recognizing and locating the infected pine trees at an early stage. This method is called the multi-band image-fusion infected pine tree detector (MFTD). MFTD adopts the Halfway Fusion mode to modify the YOLOv5 to become a dual branching network to achieve image fusion, which combined a multi-band feature fusion transformer (MFFT) module mainly containing a multi-head attention (MHA) mechanism and feed-forward networks (FFN) to integrate the correlation information of visible and multispectral features.

2. Materials and Methods

2.1. Study Area

The experiment was conducted in the Dahuofang Experimental Forest Farm (124°1′~124°24′ E, 42°0′~41°16′ N), which is located around the Dahuofang Reservoir, Fushun City, Liaoning Province of Northeast China, covering an area of approximately 7330 km^2 (Figure 1). The forest is rich in vegetation types, most of which are mixed broadleaf–conifer forests. However, due to past forest exploitation, there are only a few natural secondary forests in the area; the majority are planted forests such as Korean pines (*Pinus koraiensis*), Chinese pines (*P. tabulaeformis*) and larches (*Larix* spp.). PWD is the main disease in the forest region. With a continental monsoon climate of medium latitudes and noticeable seasonal changes, the region is in the intermediate temperate zone. The annual average precipitation is approximately 935 mm, and the annual average temperature ranges from 0° to 14°.

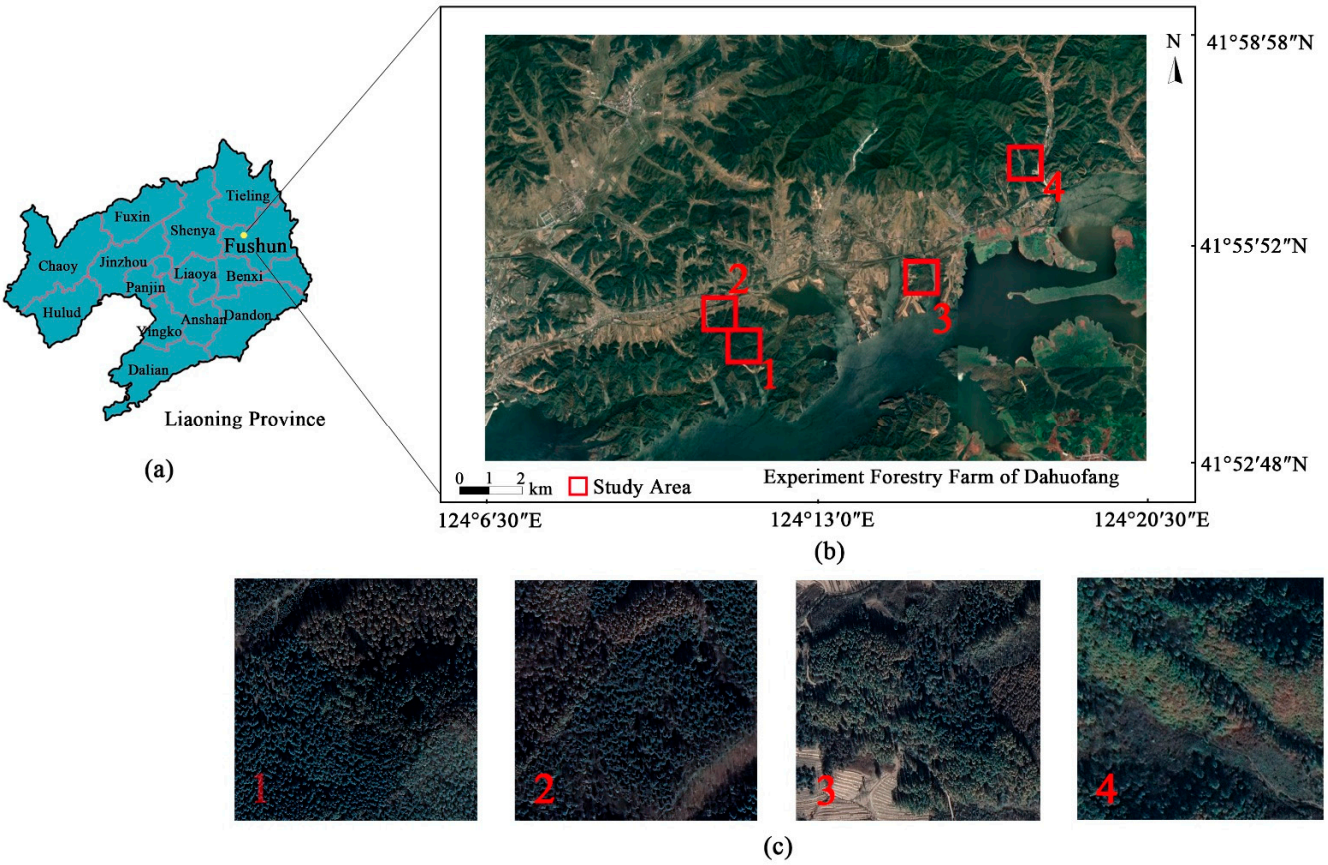


Figure 1. Location of the study areas: (**a**) The map of Liaoning Province; (**b**) the location of the study areas, where the red squares represent four study areas and the 1–4 is the numbers of the study areas; (**c**) the study areas as seen in Google Earth high-resolution images acquired on 16 October 2020.

2.2. UAV-Based Multispectral Data

The visible images and multispectral images were captured with the multispectral UAV (DJI Phantom 4 series, Shenzhen, China), which has a one-piece multispectral imaging system that integrates one visible light camera and five multispectral cameras, including a red edge, near-infrared (NIR), blue light, green light, and red light, to achieve visible imaging and multispectral imaging. A 35 mm F2.2 fixed focus lens was used for aerial photography. With the assistance of a multispectral UAV for the D-RTK2 mobile station, satellite signals from GPS, BeiDou, and other systems were received to provide real-time differential information for the UAV. With the signals, localization accuracy at the centimeter level and accurate location information of the images were obtained. Table 1 shows the key parameter data of the multispectral UAV. The sample plots of Korean pines and Chinese pines severely infected with PWD were chosen. The study areas shown in Figure 1 containing sample plots were scanned and photographed using a multispectral UAV. Table 2 shows the details of the UAV's flight mission. The images were acquired from 13:00 to 14:30 on 10 April, 13 May, 24 June, and 15 July 2021 in the study areas. The sample plots of Korean pines were located in areas 2 and 3, and the UAV flight height was 120 and 60 m, respectively. The sample plots of Chinese pines were in areas 1 and 4, and the flight height was 120 m. The relative height of the UAV and pine trees at the foot and top of the mountain changed continuously during the cruise, as the mountain terrain in areas 1, 2, and 4 changed greatly. Thus, the spatial resolution of the images acquired in the same flight mission ranged from 0.2 to 1.5. Furthermore, the images of area 2 had a spatial resolution of 0.75. An image set includes one visible image. There are 389 and 315 image sets of Korean pines and Chinese pines, respectively, and the images have a 1600 × 1300 pixels resolution. Moreover, the total surface area covered by the images is approximately 14.2 km^2.

Table 1. Key parameters of DJI multispectral UAV.

Items	Parameters	Values
DJI Phantom4 multispectral UAV	Length × Width × Height/(mm × mm × mm)	540 × 595 × 255
	Sensor	CMOS × 6, 1/2.9″
	Light filter	Blue: 450 nm ± 16 nm Green: 560 nm ± 16 nm Red: 650 nm ± 16 nm Red edge: 730 nm ± 16 nm NIR: 840 nm ± 26 nm
	Image resolution/pixels	1600 × 1300

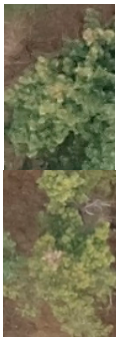
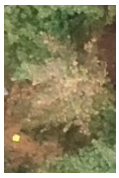
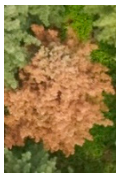
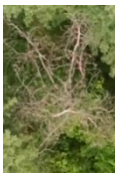
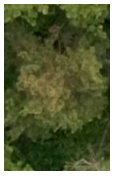
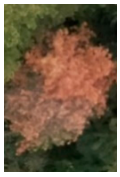
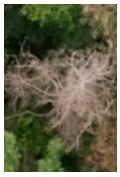
Table 2. Information on flight missions in study areas.

Study Areas	1	2	3	4
Tree species	Chinese pines	Korean pines	Korean pines	Chinese pines
Date	24 June 2021	13 May 2021	10 April 2021	15 July 2021
Flight height/m	120	120	60	120
Center coordinates	124°10′39″ E 41°54′53″ N	124°10′45″ E 41°54′48″ N	124°14′26″ E 41°55′48″ N	124°16′44″ E 41°57′33″ N
Spatial resolution/(cm/pixel)	0.2–1.5	0.2–1.5	0.75	0.2–1.5
Image sets amount	180	150	239	135
Surface area/hm^2	4.2	3.9	3.5	2.6

According to the external characteristics of pine needles such as the discoloration degree and the withering status of pine needles, we divided the infected pine trees into three infection stages: The early stage when pine trees begin to discolor or the crowns turn yellow–green or yellow–brown; the middle stage when most of the needles are becoming dry due to dehydration and their crowns turn reddish brown; and the late stage when the entire pine needles fall off and only branches and trunks remain. Table 3 displays the Korean pines and Chinese pines at each infection stage of PWD. The appearance of infected pine trees at the early stage is more complicated. There are three typical appearances of infection at the early stage: The tops of pine tree crowns wilt slightly and become light yellow or light red; part of the pine needles change color; and all pine needles wilt and become yellow–green or yellow–brown. Compared with Korean pines, the detection of Chinese pines appears to be more complicated with more diversification at the early stage.

The same parts of a visible image and a red-band image acquired by multispectral UAV are shown in Figure 2, where the white boxes represent the healthy pine trees and those in yellow, red, and blue are infected pine trees at early, middle, and late stages, respectively. The infected pine trees in the red-band image are white in color and become brighter as the discoloration increases. The difference in brightness is more pronounced in the yellow box for the early stage with slight discoloration at the crown top. The average grayscale values of the healthy and infection stages are shown in Figure 3. The average grayscale value of red-band images increased with the development of the infection stage, while the trend of the other band images is unstable. This stable change is more conducive for the model to learn the features of different stages. Moreover, the difference between healthy and early-stage-infected trees is largest in the red-band image compared to other bands. Therefore, the i red-band images show the distinction between early-stage and healthy trees more clearly, making it simpler to spot infected pine trees. Meanwhile, visible images contain texture features that can be used to distinguish infected pine trees from red or gray bare ground and other tree species. Therefore, this paper selects visible images and red-band images to be the input for the MFTD models.

Table 3. Infected pine trees in KP and CP datasets.

Infection Stage	Early Stage			Middle Stage	Late Stage
Description	Slightly wilted crown tops	Part of pine needles wilted and turning yellow or red	The whole crowns wilted and discoloration	Crowns turning reddish brown	Pine needles falling off
Korean pines					
Chinese pines					

(**a**) (**b**)

Figure 2. Sample images collected by the multispectral UAV: (**a**) Visible image, (**b**) red-band image.

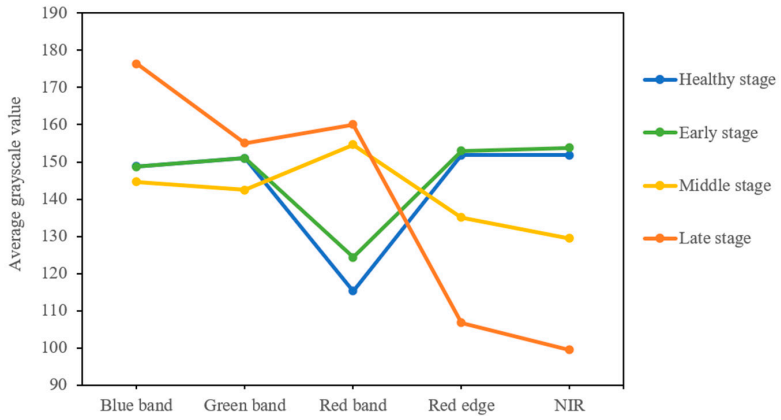


Figure 3. The average gray values of healthy and different PWD infection stages in multispectral images.

2.3. Dataset

After the collection of images, Korean pine (KP) and Chinese pine (CP) datasets were established according to tree species. The open-source annotation tool, LabelImage, was used to annotate the infected trees in the original visible and red-band images based on different infection stages. We divided infected pine trees into three categories according to the early stage, middle stage, and late stage and used the labels "yellow", "red", and "gray" to represent these stages during model training and testing, respectively.

The annotation contents include the coordinates of the rectangular bounding boxes surrounding the infected pine trees and the categories of infection stages, as shown in Figure 4a,c. For the training and testing of MFTD models, the annotation file was stored as a text file. Figure 4b,d depict the format of the annotation file, with the first column serving as the category, in which 0 represents the late stage, 1 represents the middle stage, and 2 represents the early stage. Columns 2 to 5 show the ratios of the bounding boxes' upper left and lower right horizontal and vertical coordinates to the length and width of the image. The annotation of a pair of one visible image and one red-band image was performed simultaneously to ensure the order of labeled infected pine trees in the two annotation files is the same so the accuracy of image fusion in the model training could be guaranteed. If an infected pine tree fell over several adjacent images, each part was marked as a different sample in the datasets for model training and testing. Each dataset contains visible images and red-band images and their annotation files. A 7:3 ratio was used to divide the two datasets into training sets and validation sets, and there was no overlap among all images. The number of samples of various infected pine trees in the two datasets can be seen in Figure 5. Infected pine trees at the early stage make up the majority, totaling approximately 1200 and 1400 in KP and CP datasets, respectively.

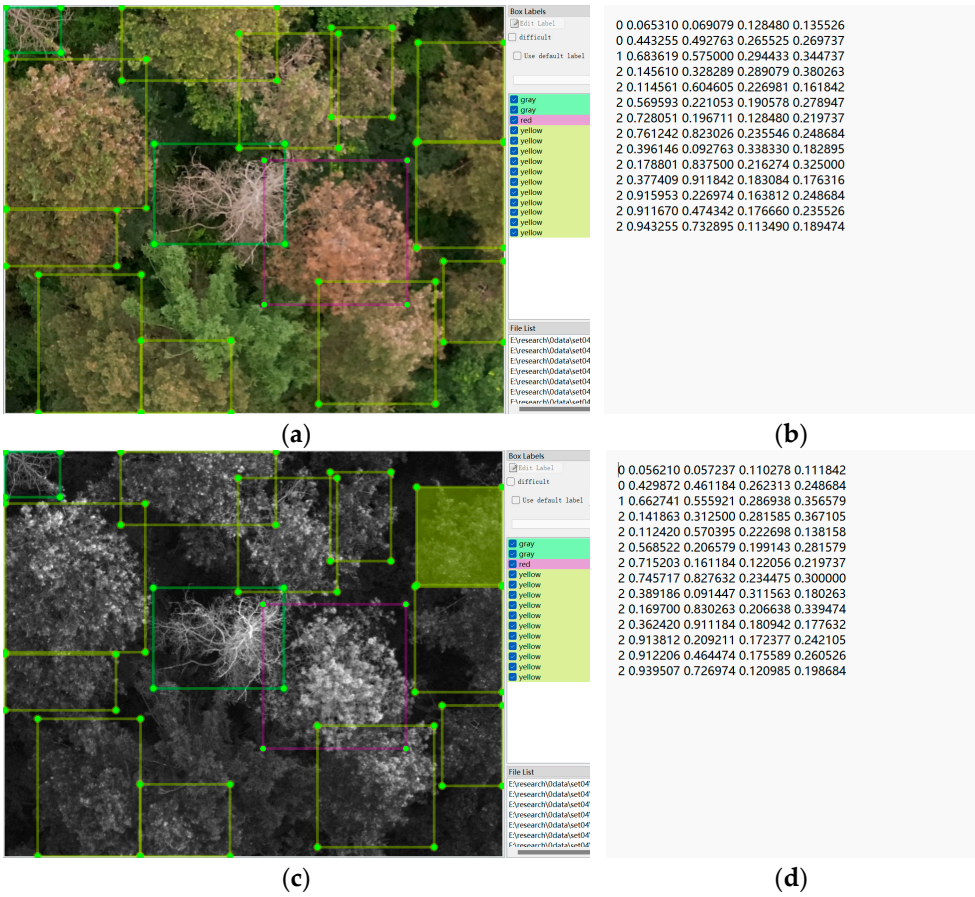
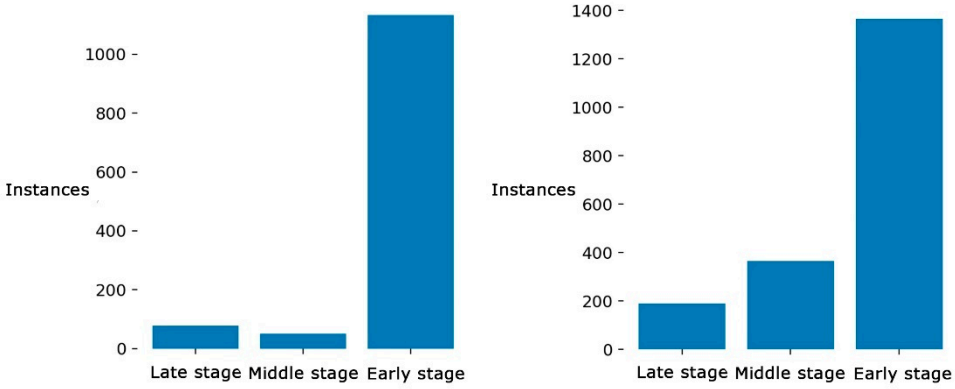


Figure 4. Data annotation, (**a**) labeling visible images, (**b**) annotation files of visible image, (**c**) labeling red-band images, and (**d**) annotation files of red-band image.

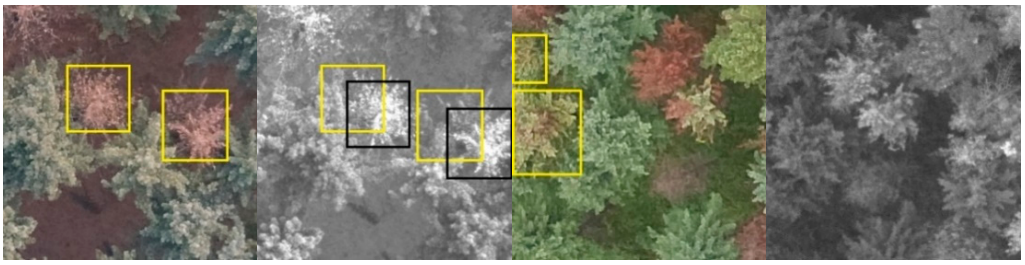


(**a**) (**b**)

Figure 5. The sample number of various categories: (**a**) KP dataset and (**b**) CP dataset.

2.4. Pre-Processing

There is a certain deviation in the installation positions of the visible and multi-spectral cameras on the pan-tilt of the UAV. The shot of each band in the same image set is not completely synchronized. Therefore, the coordinates of the same pine trees in different band images of a set will shift. The manually annotated box is marked as the ground truth in object detection. In Figure 6a, the left and right pictures represent the clipping parts of a pair of visible and red-band images with the same coordinate. The yellow box is the ground truth of two pine trees in the visible images, while the black box is the ground truth in the red-band images. This clearly shows that the positions of the two pine trees are shifted in the two pictures and cannot match each other. In Figure 6b, the pine tree at the edge of the visible image is absent in the red-band image, which also accounts for the two images' non-correspondence with the ground truth. The mismatch of the ground truth in visible and multispectral images will reduce the detection accuracy of infected pine trees.



(**a**) (**b**)

Figure 6. Mismatch of ground truth on visible and red-band images, where the yellow and black boxes are the ground truth in visible images and red-band images, respectively: (**a**) Position shifting of ground truth and (**b**) missing ground truth at the edge of red-band images.

Thus, after each pair of original visible images and red-band images was annotated, the images and annotation files needed to be preprocessed to align the coordinates of each ground truth to ensure the fusion accuracy of image features from different bands. The preprocessing process was as follows: Firstly, each pair of infected pine trees on the visible and red-band images were annotated. If the infected pine trees were missing on one of the images, it was marked as "unpaired" in the annotation files of the image, using the number 1 to represent this. Then, the offsets (Δx, Δy) of the two images were calculated. The calculation formula is as follows:

$$\Delta x = \frac{1}{N}\left(\sum_i \left(x_i^{rgb} - x_i^r\right)\right) \tag{1}$$

$$\Delta y = \frac{1}{N}\left(\sum_i \left(y_i^{rgb} - y_i^r\right)\right) \tag{2}$$

where N is the sum of ground truth pairs in this set. $\left(x_i^{rgb}, y_i^{rgb}\right)$ and $\left(x_i^r, y_i^r\right)$ are the coordinates of the center in each ground truth pair on visible and red-band images. If the ground truth was annotated as "unpaired", it was not included in the calculation. Finally, the image pair was cropped according to the offset to keep only the overlapping part of the visible images and red-band images. Furthermore, the coordinates of the ground truth in the annotation files were adjusted such that the same infected pine tree in the visible and red-band images had approximately the same coordinates. Next, the number representing "unpaired" in the annotation files was deleted. After pre-processing, the two types of images and their annotation files were used as the input of the models for training and testing.

2.5. YOLOv5

The lightweight object detection network YOLOv5 contains four variants: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. These variants have an increasing number of model parameters, which slows the detection speed while successively enhancing the detection accuracy. An example of YOLOv5l is presented in Figure 7, whose network structure is composed of three parts: Backbone, Neck, and Head. Backbone is a module for basic feature extraction that contains multiple convolution modules to extract features of the input images and output feature maps of each module with different scales of $W \times C \times H$. W and H represent the length and width of the feature map and C denotes the number of channels. The output feature maps F_{rgb1}, F_{rgb2}, and F_{rgb3} of the fifth, seventh, and tenth convolution modules were used as the input of Neck. The spatial pyramid pooling—fast (SPPF) module, which changes the parallel MaxPooling [30] layer in the spatial pyramid pooling (SSP) module [19] to serial to boost the training speed, was added at the end of Backbone. Neck is a multi-scale feature integration module that combines the feature pyramid network (FPN) [31] and the pyramid attention network (PAN) [32] to realize parameter aggregation of various output feature maps of Backbone. In addition, by referring to CSPNet [33], CSP1 and CSP2 modules with the residual structure [34] were incorporated into Backbone and Neck, respectively, thus further enhancing the feature information. Head is a prediction module that, after reducing the number of channels of the Neck output using the convolutional layer Conv with a 1×1 kernel, obtains feature maps P1~P3 to predict the category and location of infected pine trees.

2.6. MFTD Structure

In the object detection network, multiband image fusion means fusing the feature maps from images in different bands. The feature map generated by each layer of the network produces various information, with the high layers containing richer semantic information and the low layers containing better position information. The fusion of different layers of the network will lead to distinct detection effects [35]. Therefore, the network fusion mode needs to be specified before building the MFTD. We compared three network fusion modes on the YOLOv5 network, namely Input Fusion, Halfway Fusion, and Late Fusion, as shown in Figure 8, which fuse the feature maps of visible and multispectral subnets at different stages of the network. Input Fusion directly combines visible and multispectral image data through the matrix addition, using the results as the input of Backbone. Halfway Fusion involves Backbone consisting of visible and multispectral subnets. The feature maps of the two subnets are added to finish the fusion before Neck and Head. Late Fusion achieves fusion at the high layers of the network. The two subnets contain the complete Backbone, Neck, and Head. For prediction, the output feature maps of the Head are incorporated by addition.

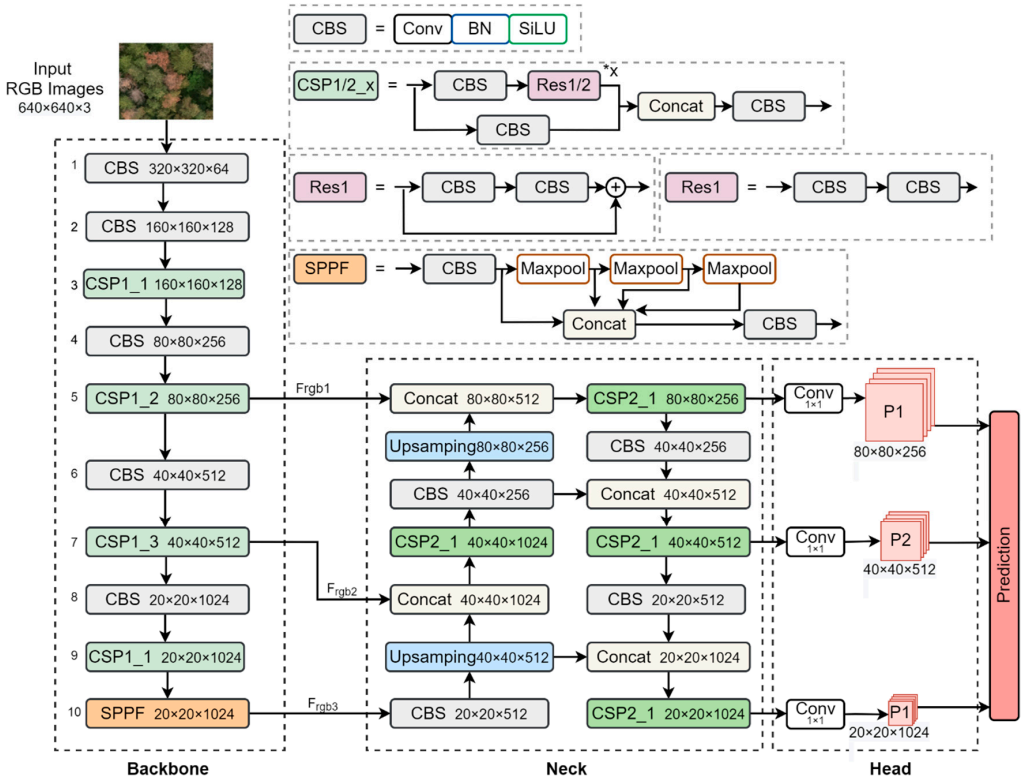


Figure 7. Structure of YOLOv5l network.

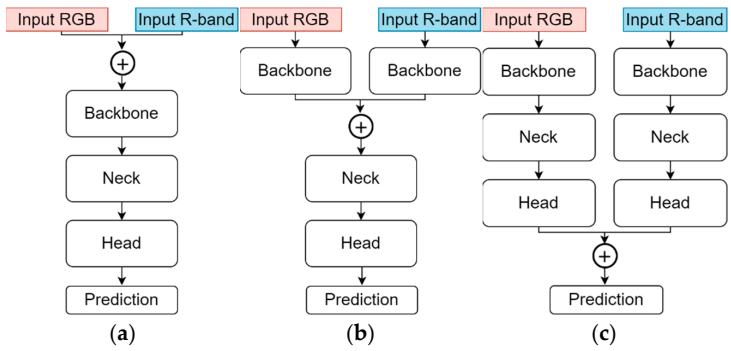


Figure 8. Network fusion modes: (**a**) Input Fusion, (**b**) Halfway Fusion, (**c**) Late Fusion.

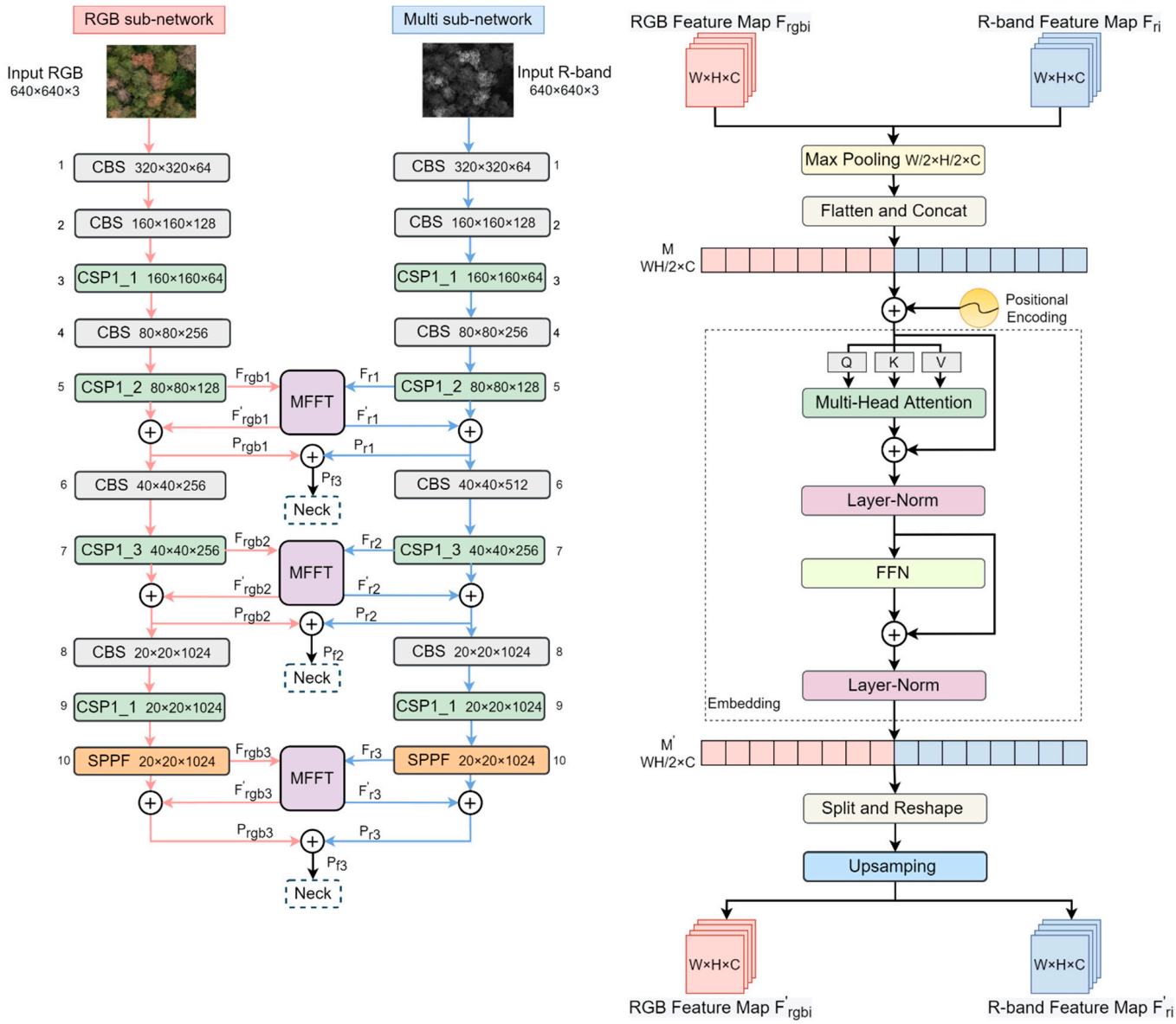
Halfway Fusion was adopted in MFTD, which expands Backbone of the original YOLOv5 into a dual branching network to build visible and multispectral subnets, inputting visible and red-band images, respectively, to extract image features. Furthermore, for the purpose of promoting the fusion of the two feature maps, the multi-band feature fusion transformer (MFFT) module was added to Backbone. Figure 9a shows the structure of MFTD Backbone, and the fusion process of the multiband feature map can be formulated as:

$$P_{fi} = P_{rgbi} + P_{ri} \tag{3}$$

$$P_{rgbi} = F_{rgbi} + MFFT\left(F_{rgbi}\right) \tag{4}$$

$$P_{ri} = F_{ri} + MFFT(F_{ri}) \tag{5}$$

where F_{rgbi} and F_{ri} represent the output feature maps pair from the fifth, seventh, and tenth modules of the visible and multispectral subnets, which are added to the output feature maps of MFFT to obtain P_{rgbi} and P_{ri}, respectively. P_{fi} as the input of Neck, which is the sum of P_{rgbi} and P_{ri}.



(**a**) (**b**)

Figure 9. Structure of proposed MFTD: (**a**) Backbone of MFTD, (**b**) structure of MFFT module.

2.7. MFFT Module

Apart from the network fusion mode, the calculation method of multiband feature map fusion also has a significant impact on detection accuracy. The correlation between the features of each band cannot be obtained via general fusion methods such as the direct concatenation, addition, or multiplication of feature maps. To integrate the overall feature information of visible and red-band feature maps, we proposed an MFFT module based on Transformer, which processes the two feature maps before fusion. The structure of MFFT is shown in Figure 9b. To reduce the calculation of the model, a structure without a decoder was used in MFFT to work as model embedding. The calculation process is as follows:

(1) A MaxPooling layer was used to reduce the dimensions of the input feature map F_{rgbi} and F_{ri}, from $W \times H \times C$ to $W/2 \times H/2 \times C$, which was concatenated after flattening each feature map to obtain M, a feature matrix with a dimension of $WH/2 \times C$.

(2) The position encoding (PE) method of Transformer was used to supplement the timing information for M. S, the input feature sequence of Embedding, can be expressed as:

$$S = M + PE \tag{6}$$

$$PE(p, 2c) = sin\left(\frac{p}{10000^{\frac{2c}{C}}}\right) \tag{7}$$

$$PE(p, 2c+1) = cos\left(\frac{p}{10000^{2c/C}}\right) \tag{8}$$

where $p \in (1, 2, 3, \ldots, WH/2)$, $c \in (1, 2, 3, \ldots, C/2)$.

(3) S was input into the Multi-Head Attention (MHA) module of Embedding. The MHA is the concatenating result after multiple Self-Attention parallel calculations, with its structure shown in Figure 10. Self-Attention can be described as a query mechanism, which applies vector Query (Q) to retrieve the target Key (K) and the corresponding Value (V) of the Key and takes the weighted sum of all Values as the output. Q, K, and V, as the input of MHA, are the matrixes of feature sequence S under the mapping of weights, W_q, W_k, and W_v, which are expressed as:

$$Q = SW_q \tag{9}$$

$$K = SW_k \tag{10}$$

$$V = SW_V \tag{11}$$

where the dimension of S is $WH/2 \times C$ and the weighted matrix is $C \times C$. Self-Attention uses the softmax function to calculate the weight of each value. The Self-Attention value of the nth Q is:

$$Attention(Q_n, K, V) = \sum_{i}^{C_K} softmax(e_i^n)V \tag{12}$$

$$softmax(e_i^n) = \frac{exp(e_i^n)}{\sum_j^{C_K} exp\left(e_j^n\right)} \tag{13}$$

$$e_i^n = \frac{Q_n K_i^T}{\sqrt{C_K}} \tag{14}$$

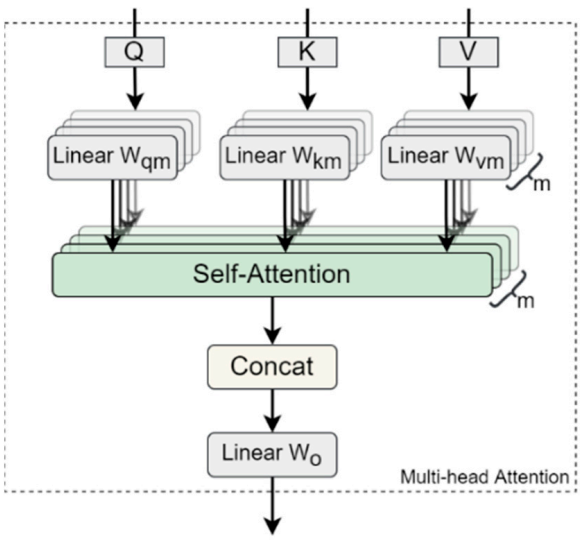


Figure 10. Structure of MHA.

To prevent an excessively large dot product of Q and K from leading to an extremely small softmax gradient, which further complicates the convergence of model training, the coefficient $\sqrt{C_K}$ is divided after the dot product, and $C_K = C$. MHA performs Self-Attention on the linear transformation of Q, K, and V, and the output M_{mha} can be shown as the following:

$$M_{mha} = MultiHeadAtten(Q, K, V) = Concat(h_1, h_2, \ldots, h_m)W_o \tag{15}$$

$$h_i = Attention(QW_{qi}, KW_{ki}, VW_{vi}) \tag{16}$$

where m is the number of parallel calculations of Self-Attention for which we employ $m = 8$. The dimensions of the weights W_{qi}, W_{ki}, and W_{vi}, are $C \times C$. The dimension of the weight W_o is $mC \times C$.

(4) After M_{mha} and S were added, they were normalized through the Normal Layer [36] and then input into the full-connection Feed-Forward Network (FFN), which contains two linear transformations and a Relu activation function. The output F is expressed as:

$$F = FFN(S') = Relu(S'W_{s1} + b_1)W_{s2} + b_2 \tag{17}$$

$$Relu(S'W_{s1} + b_1) = max(0, S'W_{s1} + b_1) \tag{18}$$

$$S' = M_{mha} + S \tag{19}$$

where W_{s1}, W_{s2}, and b_1 and b_2 are learnable parameters of the network, while the dimensions of the weights, W_{s1} and W_{s2} are $C \times C$.

(5) The output feature sequence M' of Embedding was obtained after adding F and M_{mha} through the Normal Layer. Two feature maps with the dimensions of $W/2 \times H/2 \times C$ were obtained by splitting M' and the reverse operation of flattening in step (1). After upsampling, the visible and multispectral enhancement feature maps, F'_{rgbi} and F'_{ri}, were output.

2.8. Evaluation Indicators

During the test stage, evaluation indicators, namely, Average Precision (AP) and mean Average Precision (mAP), proposed by the MSCOCO dataset [37], were used to evaluate the accuracy rate for the detection of infected trees. AP can evaluate the accuracy rate within the category, and mAP is the mean value of all AP. In addition to evaluating whether categories of infected pine trees can be accurately identified, the two indicators also judge whether the Intersection-over-Union (IoU) between the prediction boxes of the model and the ground truth meets the threshold, which is usually 0.5. At this time, AP and mAP are recorded as AP@50 and 50. AP is the area under the Recall–Precision curve, and recall and precision are defined as:

$$precision = \frac{TP}{(TP + FP)} \tag{20}$$

$$recall = \frac{TP}{(TP + FN)} \tag{21}$$

where TP is the number of True Positives predicted by the model, that is, the number of prediction boxes with the same category of ground truth and an IoU greater than the threshold value. FP is the number of False Positives, that is, the prediction boxes with the same category of ground truth but an IoU less than the threshold value. FN represents False Negatives, indicating the number of ground truths that do not match the prediction boxes. In addition, the number of parameters and frames per second (FPS) of the model were employed to evaluate its volume and detection speed. In general, the smaller the number of parameters amount, the faster the training speed of the model is, and the bigger FPS, the faster the detection speed is.

3. Results

3.1. Implementation Details

The Ubuntu 18.04 64-bit system was installed as the deep learning server for model training and testing. Furthermore, the Python (version 3.9) programming language created by Python software foundation in Delaware, USA and the PyTorch deep learning open-source framework were used. The server had an Intel CPU (64 GB) and NVIDIA RTX 3090 GPU (24 GB) installed. The stochastic gradient descent (SGD) algorithm with a momentum of 0.937 was selected to optimize the training process with the initial learning rate set at $1e-2$ with a batch size of 16. The weight decay was 0.0005. A total of 500 epochs were trained, and the learning rate decreased with the increase in epochs using the Cosine Annexing method [38]. To prevent model overfitting, two data augmentation methods were used to expand the data volume: (1) The mosaic method proposed in YOLOv4 was adopted to randomly cut, scale, and rotate four images, which were later combined into one for the detection of infected pine trees and to increase the number of small-scale targets; (2) and two patches, whose length–width ratio to the original image is 1:3, were randomly cut from the input image, with their sizes adjusted to the original YOLOv5 input size of 640×640 pixels. Then, they were randomly flipped in the horizontal or vertical direction. Random processing can improve the diversity of image data and strengthen the robustness of the model regarding infected pine trees in different scales. Moreover, all detection models for infected pine trees trained in this paper were pre-trained on the MS COCO dataset.

3.2. Evaluation of Network Fusion Modes

In order to validate the best network fusion mode of MFDT, we modified the YOLOv5l network and conducted training and tests on KP and CP datasets, according to Input Fusion, Halfway Fusion, and Late Fusion, respectively. The detection accuracy for infected pine trees, AP@50 and mAP@50, as well as the number of parameters of each model, are shown in Tables 4 and 5. The models on KP and CP datasets were assigned the letters K and C, respectively. RGB and R accordingly represent visible images and red-band images. K0 and K1 and C0 and C1 are original YOLOv5l models trained on independent visible images and red-band images. It can be seen that for models adopting any network fusion method, the AP@50 and mAP@50 were higher than those before fusion. Compared with Input Fusion and Late Fusion, the mAP@50 of K3 and C3 employing Half Fusion are the highest, reaching 83.2% and 82.1%, respectively. The AP@50 of K3's late stage is 1.7% lower than that of K4 but higher than K2 and K4 in the other two categories. Models C3 and C4 have nearly equal mAP@50 values, and except for the middle stage, the AP@50 values of the other two categories are greater than those of C2 and C4. Overall, there is minimal deviation in the detection accuracy between Half Fusion and Late Fusion on CP datasets, but Half Fusion performs better than Late Fusion on the KP dataset. Meanwhile, the model adopting Half Fusion is 39.2 MB smaller than the model adopting Late Fusion, and it trains and detects more rapidly. Therefore, Halfway Fusion was used for the MFDT network.

Table 4. Detection performance of different network fusion modes on KP dataset.

Model (YOLOv5l)	Network Fusion Mode	Image Type	mAP@50	AP@50			Parameter Amount/MB
				Early Stage	Middle Stage	Late Stage	
K0	\	RGB	79.9	72.9	87.3	78.6	93.7
K1	\	R	77.8	70.6	81.6	81.3	93.7
K2	Input Fusion	RGB + R	81.9	74.9	87.7	83.2	118.4
K3	Half Fusion	RGB + R	83.2	77.4	89.5	82.6	148.1
K4	Late Fusion	RGB + R	82.9	75.2	89.3	84.3	187.3

Table 5. Detection performance of different network fusion modes on CP dataset.

Model (YOLOv5l)	Network Fusion Mode	Image Type	mAP@50	AP@50			Parameter Amount/MB
				Early Stage	Middle Stage	Late Stage	
C0	\	RGB	76.0	69.6	85.5	72.8	93.7
C1	\	R	73.6	68.1	74.4	78.2	93.7
C2	Input Fusion	RGB + R	79.5	75.3	82.3	80.8	118.4
C3	Half Fusion	RGB + R	82.1	78.4	85.6	82.3	148.1
C4	Late Fusion	RGB + R	82.0	78.1	87.0	80.8	187.3

3.3. Ablation Experiments

There are outputs of three corresponding feature map pairs from the visible and multispectral subnets in Backbone of Halfway Fusion, [F_{rgb1}, F_{r1}], [F_{rgb2}, F_{r2}], and [F_{rgb3}, F_{r3}]. Scales vary between feature map pairs, with [F_{rgb1}, F_{r1}] having the largest and [F_{rgb3}, F_{r3}] having the lowest. On the feature map, different sizes of prediction boxes are generated to match target infected pine trees of various sizes. The larger the feature map, the smaller the generated prediction box is. Prior to the fusion of the feature maps, the MFFT module processes different feature map pairs, and the models' capacity to identify infected pine trees likewise varies. In order to validate the effectiveness of the MFFT module and the optimal combination for feature map pairs, we take YOLOv5l as an example to conduct ablation experiments on various combinations of feature map pairs. As shown in Tables 6 and 7, all models adopted half-fusion, and model K3 (C3) did not receive an MFFT module. For other models with the same serial number, feature map pairs processed by the MFFT module are identical.

Table 6. Ablation Experiments on KP Dataset.

Model (YOLOv5l)	Data Preprocessing	MFFT			mAP@50	AP@50		
		[F_{rgb1}, F_{r1}]	[F_{rgb2}, F_{r2}]	[F_{rgb3}, F_{r3}]		Early Stage	Middle Stage	Late Stage
K3	√				83.2	77.4	89.5	82.6
K5	√	√			83.3	77.7	89.0	83.1
K6	√		√		84.3	78.0	89.9	84.8
K7	√			√	83.5	76.1	92.1	82.2
K8	√	√	√		84.4	79.4	90.5	83.4
K9	√	√		√	85.4	78.0	93.6	84.7
K10	√		√	√	86.3	81.9	92.8	84.3
K11(MFTD)	√	√	√	√	88.5	87.2	93.5	84.8
K12		√	√	√	83.2	82.2	86.7	80.7

Table 7. Ablation Experiments on CP Dataset.

Model (YOLOv5l)	Data Preprocessing	MFFT			mAP@50	AP@50		
		[F_{rgb1}, F_{r1}]	[F_{rgb2}, F_{r2}]	[F_{rgb3}, F_{r3}]		Early Stage	Middle Stage	Late Stage
C3	√				82.1	78.4	85.6	82.3
C5	√	√			83.7	80.3	87.7	83.3
C6	√		√		82.6	81.1	86.3	80.5
C7	√			√	83.4	78.7	87.6	84.0
C8	√	√	√		84.9	82.4	88.8	83.5
C9	√	√		√	84.6	79.2	91.9	82.8
C10	√		√	√	85.3	81.0	90.1	84.7
C11(MFTD)	√	√	√	√	86.8	81.2	92.9	86.2
C12		√	√	√	84.9	79.2	90.4	85.0

MFFT was used separately for individual feature map pairs in K5~K7 (C5~C7). Compared with K3 (C3), each model's mAP@50 has been improved to a certain degree. While

the improvement of K6 is concentrated in the late stage, with AP@50 rising by 2.2%, K5's growth is relatively low. The greatest increase in AP@50 occurred at K7's middle stage, reaching 2.6%. The early stage of C5 and C6 were both enhanced greatly, and AP@50 was 1.9% and 2.7%, which were higher than that of C3, respectively. All three models' AP@50 at the middle stage showed significant improvement. However, AP@50 of K7's early stage and C6's late stage decreased by 1.3% and 1.8%, respectively, indicating that the size of the infected Korean pines at the early stage is mainly concentrated in the medium and small scale, while the number of infected Chinese pines at the late stage with a medium-scale size is relatively small. The ability of the model to identify infected pine trees is further enhanced, and the mAP@50 of each model is higher than that of K5~K7 (C5~C7) as the simultaneous enhancement of two different feature map pairs on the MFFT of K8~K10 (C8~C10) lead to more abundant feature information on the fusion feature maps of varying bands. The MFTD models K11 and C11 can achieve the best detection effect by employing MFFT concurrently for three feature map pairs. In K11, AP@50 of three categories of infected pine trees was 87.2%, 93.5, and 84.8%, while in C11, it was 81.2%, 92.9%, and 86.2%. The two models' mAP@50 were 88.5% and 86.8%, respectively. The MFFT can effectively improve the correlation between the features of different wavebands and increase the model's ability to detect individual infected pine trees. Furthermore, the MFTD was trained and tested on the non-preprocessed datasets to obtain models K12 and C12, and their respective mAP@50 values were 5.3% and 1.9% lower than those of K11 and C11. This suggests that the deviation of coordinates of infected pine trees in visible and multispectral images negatively impacts the detection accuracy of models, and the coordinate-correction preprocessing of the ground truth can further improve the detection accuracy of MFTD.

3.4. Experiments of YOLO Series

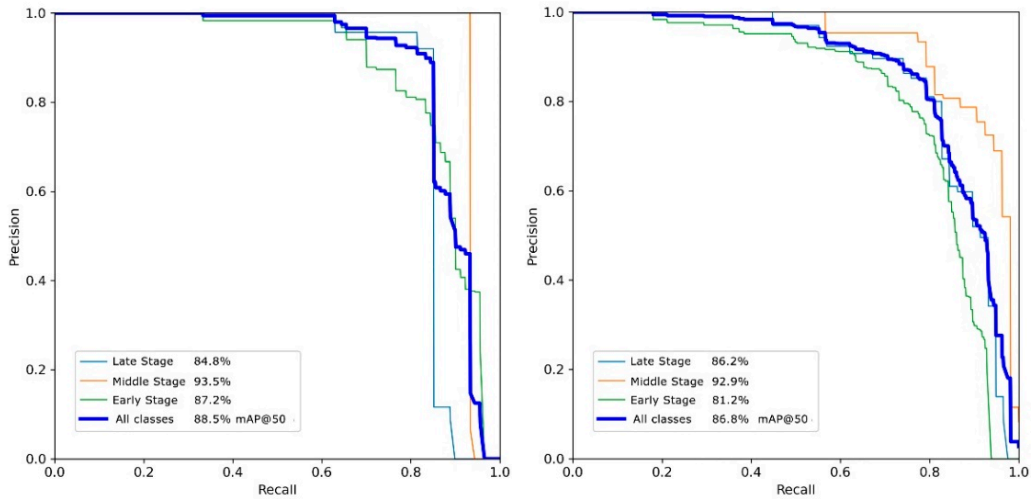
In KP and CP datasets, Tables 8 and 9 compare the experimental findings of the YOLO series and the method proposed in this paper. YOLOv5 variants trained on visible images outperform YOLOv3. Although mAP@50 values of YOLOv5s (K15 and C15) and YOLOv5m (C17) are slightly lower than that of YOLOv4 (K14 and C14), YOLOv5l (K0 and C0) and YOLOv5x (K19 and C19) perform much better in terms of accuracy, and the number of parameters and detection speed of YOLOv5 are both higher than YOLOv4. Thus, YOLOv5 is more suitable for the KP and CP datasets. The YOLOv5 variants were improved in the MFTD network structure for experimentation. The results demonstrated that four MFTD models based on YOLOv5 variants have increased detection ability for infected pine trees compared to their pre-improvement counterparts. Furthermore, YOLOv5l-based models K11 and C11 had the best detection effects, increasing by 8.6% and 10.8% in comparison to the original YOLOv5l models trained on the visible images. Although the identification of infected pine trees at the early stage is the most challenging, AP@50 of K11 and C11's early stage increased the most. K11's AP@50 was 14.3% and 16.6% higher than that of K0 and K1, respectively, before improvement, while C11's was 11.6% and 13.1% higher than that of C0 and C1. The Precision–Recall curves of K11 and C11 shown in Figure 11 reveal that the models are most sensitive to infected pine trees at the middle stage where AP@50 is the highest. Additionally, as the edge of individual Korean pines is clearer in the densely distributed area of pine trees in the UAV image, they are simpler to distinguish than Chinese pines, and the mAP@50 of the KP dataset is 1.7% higher. Moreover, the mAP@50 of YOLOv5x-based models K23 and C23 is lower than that of K11 and C11, indicating that simply increasing the network depth and width on KP and CP datasets will not further improve the model's detection accuracy. Under the experimental conditions in this paper, the detection speed of YOLOv5s-based models K21 and C21 is the fastest, with the FPS approaching 49.2 Hz. The FPS of K11 and C11 decreased to 35.6 Hz. Despite the fact that the detection speed was reduced following the addition of the MFFT module, the method in this study can still realize the real-time detection of individual infected pine trees on the server.

Table 8. Detection performances of YOLO series and MFDN on KP Dataset.

Model	YOLOx	Image Type	mAP@50	AP@50			Parameter Amount/MB	FPS/HZ
				Early Stage	Middle Stage	Late Stage		
K13	v3	RGB	73.9	70.1	80.7	71.1	248.2	80.3
K14	v4	RGB	78.7	72.7	86.1	77.2	257.9	65.7
K15	v5s	RGB	77.9	71.3	87.4	75.1	14.4	146.3
K16		R	77.8	69.5	82.8	81.1	14.4	146.3
K17	v5m	RGB	79.2	73.4	86.3	77.9	42.5	97.1
K18		R	77.6	72.7	76.1	84.0	42.5	97.1
K0	v5l	RGB	79.9	72.9	87.3	78.6	93.7	86.1
K1		R	77.8	70.6	81.6	81.3	93.7	86.1
K19	v5x	RGB	79.3	75.2	86.2	76.5	175.1	58.9
K20		R	78.4	71.0	82.4	81.8	175.1	58.9
K21(MFTD)	v5s	RGB + R	80.8	77.1	84.6	80.7	89.6	49.2
K22(MFTD)	v5m	RGB + R	85.3	84.3	90.1	81.5	216.8	40.2
K11(MFTD)	v5l	RGB + R	88.5	87.2	93.5	84.8	413.4	35.6
K23(MFTD)	v5x	RGB + R	85.9	84.7	93.1	80.0	690.1	30.4

Table 9. Detection performances of YOLO series and MFDN on CP Dataset.

Model	YOLOx	Image Type	mAP@50	AP@50			Parameter Amount/MB	FPS/HZ
				Early Stage	Middle Stage	Late Stage		
C13	v3	RGB	66.9	64.3	75.9	60.5	248.2	80.3
C14	v4	RGB	73.3	67.2	80.6	72.1	257.5	65.7
C15	v5s	RGB	68.6	65.0	75.5	65.2	14.4	146.3
C16		R	64.5	62.3	64.9	66.3	14.4	146.3
C17	v5m	RGB	72.6	68.8	82.2	66.7	42.5	97.1
C18		R	69.2	68.0	68.6	70.9	42.5	97.1
C0	v5l	RGB	76.0	69.6	85.5	72.8	93.7	86.1
C1		R	73.6	68.1	74.4	78.2	93.7	86.1
C19	v5x	RGB	75.6	70.6	86.0	70.3	175.1	58.9
C20		R	73.3	63.9	75.6	80.5	175.1	58.9
C21(MFTD)	v5s	RGB + R	80.2	79.7	81.6	79.3	89.6	49.2
C22(MFTD)	v5m	RGB + R	82.0	80.2	84.4	81.3	216.8	40.2
C11(MFTD)	v5l	RGB + R	86.8	81.2	92.9	86.2	413.4	35.6
C23(MFTD)	v5x	RGB + R	84.7	80.8	88.5	84.6	690.1	30.4



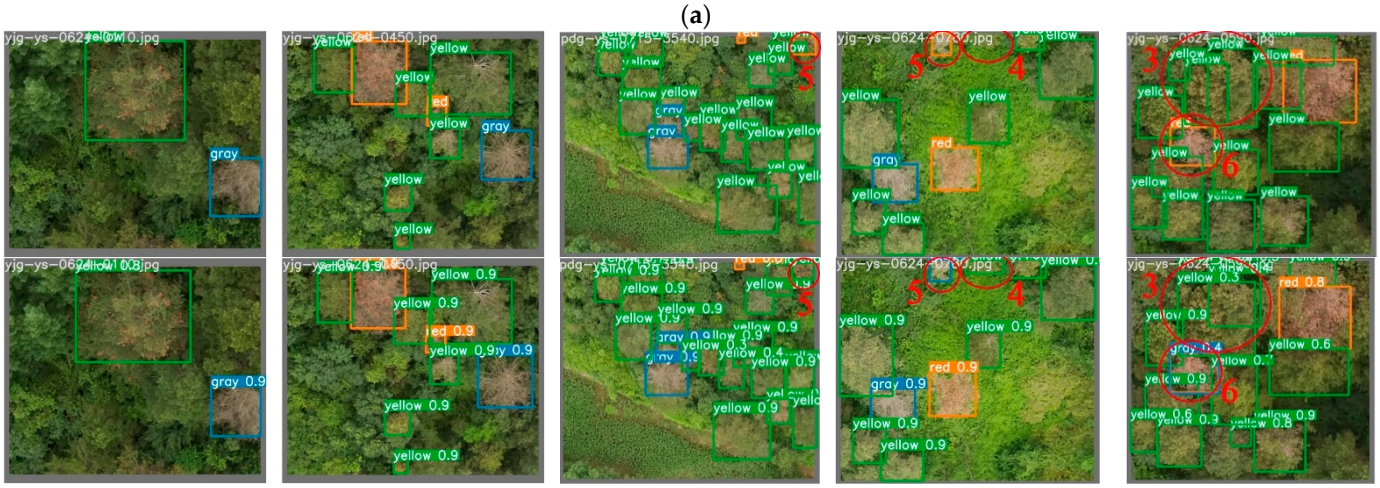
(**a**) (**b**)

Figure 11. The Precision–Recall curves of MFTD models based on YOLOv5l: (**a**) The Precision–Recall curve of model K11 on the validation set of KP dataset, (**b**) the Precision–Recall curve of model C11 on the validation set of CP dataset.

Figure 12 illustrates the detection results of models K11 and C11 on validation sets of KP and CP datasets. The upper line depicts the ground truth while the lower line is the models' prediction. The early stage, middle stage, and late stage are represented by green, orange, and blue boxes, respectively. The number on the box indicates the confidence coefficient of the category. The greater the confidence coefficient, the more certain the model is that infected pine trees in the box fall into the prediction category, and the value of the confidence coefficient is 0~1. The first and second columns indicate that MFTD can identify different infection stages while accurately recognizing and locating individual infected pine trees. Not only is MFTD capable of accurately recognizing medium-scale infected pine trees but it also has a relatively strong detection ability for infected pine trees on larger or smaller scales. It means that multi-scale infected pine trees in UAV images shot at a variety of heights can be simultaneously detected by MFTD.

i. ii. iii. iv. v.

(**a**)



i. ii. iii. iv. v.

(**b**)

Figure 12. Detection Results of Infected pine trees with PWD, where the 1–7 represent the numbers of the error areas and i.–v. are the numbers of the result images: (**a**) Detection results on KP dataset, (**b**) detection results on CP dataset.

4. Discussion

Given that single-band images cannot fully express the feature information of infected pine trees at different infection stages, which makes it difficult to accurately recognize those at the early stage, we fused visible and red-band images, and a coordinate-correction preprocessing method for ground truths was proposed when constructing datasets to modify the coordinate shifts of the same infected pine tree on visible and red-band images.

Moreover, regarding the low efficiency of existing traditional image analysis methods and the difficulty of current deep learning object detection methods to accurately recognize infected pine trees at the early stage and the inability of these methods to identify different infection stages of PWD, etc., an object detection network MFTD for infected pine trees with PWD was designed using deep learning combined with Transformer. Based on YOLOv5 and adopting the network fusion mode of Halfway Fusion, the MFTD established visible and multispectral subnets on Backbone of the network to extract multiband image features. Meanwhile, the MFFT module was then added to Backbone to integrate multiband image features and enhance the correlation between the feature maps of two subnets.

MFTD fuses the features of visible images and multispectral red-band images so it can remove the bare ground, green broad-leaved trees, and other interfering substances in the UAV image and detect individual infected pine trees accurately without the need to judge the stage of infected pine trees via manual visual inspection. This is in contrast to the existing detection methods for infected pine trees, which combine image analysis [10,11] and only segment the areas of infected pine trees, which makes it difficult to locate individual infected pine trees. Moreover, traditional machine learning technologies, such as random forest and support vector machine algorithms, were employed to identify the infected pine trees [39,40]. However, the need to design features for different experimental data adds time and extends the monitoring cycle for forest pests, and these studies do not have the ability to identify infected trees with PWD at the early stage.

By targeting massive data without the need for artificial feature design, MFTD can accelerate model training and object detection by using GPU. Thus, it has obvious advantages in the detection speed of infected pine trees and greatly decreases the monitoring process for forest pests. Moreover, among the methods using deep learning to identify infected pine trees in UAV images [13,15], the model proposed by Deng et al. [14] only used visible images, suggesting that it could not determine the infection stages, and some infected pine trees at the early stage were undetected. To segment infected pine trees in the visible image, Qin et al. [41] adopted a CNN. However, this method could only extract the area of infected pine trees at the middle stage and was unable to recognize the individual infected trees. Yu et al.'s [17] detection of infected pine trees at three stages on multispectral images had relatively low accuracy, as indicated by the mAP of the two models, which were 60.98% and 57.07%, respectively. The methods that only use visible image or spectral image data cannot utilize the correlation of feature information between images of different wavebands. Nonetheless, the MFTD can simultaneously extract the features of visible and multispectral images to achieve feature fusion. Multiple experiments show that MFTD is capable of fully utilizing the feature information of the two types of images when predicting the category and location of infected pine trees. With the addition of the MFFT module to Backbone, further improvement of the detection ability for infected pine trees at all stages can be obtained, especially the early stage, where the AP@50 on KP and CP datasets reach 87.2% and 81.2%, respectively, while mAP@50 reaches 88.5% and 86.8% accordingly. There are several primary factors that affect the detection accuracy of MFTD: (1) The visible images are easily affected by sunlight and excessive sunlight will make it harder to distinguish the features of infected pine trees at the early stage since they will resemble those of healthy pine trees. As shown in area 1 of Figure 12aiii, some infected pine trees at the early stage are undetected or misrecognized. (2) Several images in the KP dataset were collected after the leaves of green broad-leaved trees fell. Infected pine trees at the late stage are difficult to spot because the trunk of the green broad-leaved trees in area 2 of Figure 12av is very similar to those of the late stage in Figure 12aiv. (3) As illustrated in area 3 of Figure 12aiv,bv, multiple overlapping pine trees are easily mistaken as one in the densely distributed area as the edge of each tree is not sufficiently distinct. (4) Multispectral images collected at low sunlight angles have less evident features, which leads to the false detection of infected pine trees at the edge of the images. For example, the infected pine trees are undetected in Figure 12av,biv, and the categories are incorrectly recognized in area 5. (5) The sample of infected pine trees at the middle stage and the late stage is small in the two datasets,

which leads to false detection of these two categories, as shown in area 6 of Figure 12bv. In addition, some distractors can also cause false detection such as in area 7.

Three parts are necessary for the practical application of the MFTD method proposed in this paper to UAV monitoring for PWD: Image capture, the detection of infected pine trees, and the visualization of detection results. Figure 13 displays the monitoring process. First, the MFTD model was trained on the deep learning server in the laboratory, after which the optimal model was deployed to a laptop with an independent graphics card or an intelligent mobile device. Secondly, multispectral UAVs took images of the monitored area in the forest and transmitted visible and red-band images to local laptops or mobile devices. The MFTD model was then employed to detect infected pine trees at various stages in Windows or Android systems. Then, the detection results such as the categories and number of infected pine trees, GPS, and photographing time were uploaded to the cloud server database. Finally, to complete the remote monitoring of PWD, remote monitoring servers located in the Forest Pest Control and Quarantine Station and laboratories read the database information, visualize the infected areas on the 3D map, and display the detection results. The automation of UAV monitoring and monitoring efficiency of PWD can all be improved by using the MFTD network to monitor PWD.

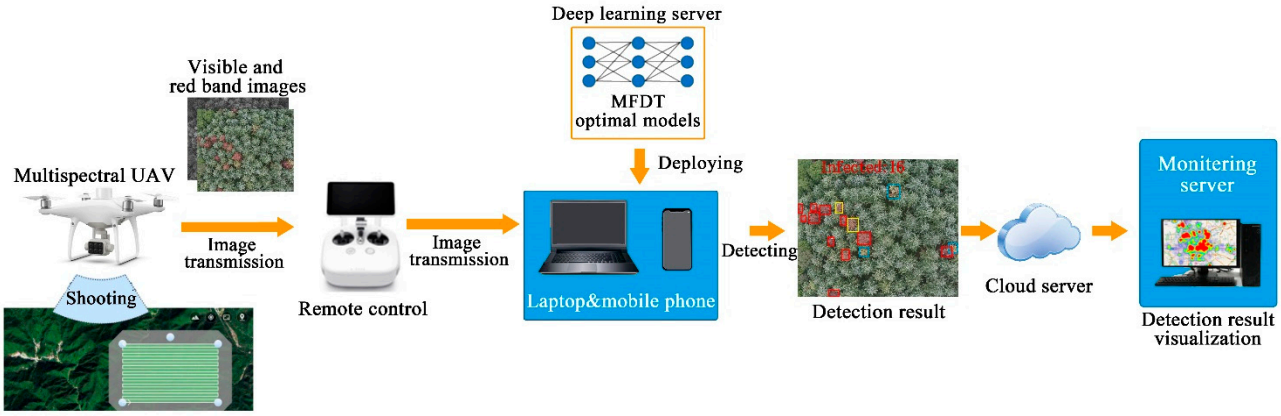


Figure 13. The UAV monitoring process with MFTD.

However, MFTD still has some drawbacks in detection regarding actual monitoring. The image collection environment will affect the image quality, and blurry or unclear image feature information will impair the accuracy of model detection. The detection of infected pine trees will also be influenced by the color change of pine crowns caused by drought and other pests. Hence, to reduce the interference of environmental factors on the detection accuracy, it is necessary to avoid too much direct sunlight and cloudy weather during the actual monitoring and specify the problems in the monitoring forest area. Additionally, MFTD increases the model's number of parameters to increase accuracy at the expense of a certain degree of detection speed. The MFTD can be implemented by using the four variants of YOLOv5. Due to its small number of parameters, the MFTD model based on YOLOv5s has a low detection accuracy but can reduce deployment costs as much as possible, which facilitates the detection of infected pine trees on mobile devices. In contrast, the MFTD model based on YOLOv5l has a relatively slow detection speed but the highest detection accuracy, making it more suitable to be deployed on laptops or cloud servers. Therefore, different MFTD models can be deployed according to different detection equipment.

In the follow-up study, to achieve the classification of different appearances of infected pine trees at the early stage and the recognition of different diseases of pine trees, we will attempt to combine hyperspectral data. Furthermore, we plan to design a lightweight detection network that can better balance accuracy and speed to further adapt to mobile devices.

5. Conclusions

In this study, a new detection network with a multi-band image-fusion infected pine tree detector (MFTD) for infected pine trees with pine wilt disease was developed using deep learning technology. In order to fuse the visible and multispectral image features, the Backbone network of MFTD was constructed with visible and multispectral subnetworks. A multi-band feature fusion transformer module was designed to integrate the correlation information between two types of images prior to fusion. Extensive experiments demonstrated that the detection accuracy of MFTD models based on YOLOv5 variants was significantly improved. The mean average precision (mAP@50) of the best-performing MFTD model reached 88.5% on the Korean pine dataset and 86.8% on the Chinese pine dataset. The average precision (AP@50) of pine trees at different infection stages was improved in all cases, and the early infected stage with the greatest improvement had AP@50 of 87.2% and 81.2% on the two datasets, respectively, which was an increase of 14.3% and 11.6% compared to the original YOLOv5.

MFTD with UAV visible and multispectral images can efficiently and accurately detect infected pine trees at an early stage and distinguish the various infection stages, which has obvious advantages in UAV early intelligent monitoring.

Author Contributions: Investigation, H.B.; methodology, Y.Z.; software, Y.Z.; supervision, W.L., S.Z. and Y.L.; writing—original draft, Y.Z.; writing—review and editing, W.L. and R.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Major emergency science and technology project of National Forestry and Grassland Administration, grant number ZD-202001 and the National Key R & D Program of China, grant number 2021YFD1400900.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Wu, W.; Zhang, Z.; Zheng, L.; Han, C.; Wang, X. Research Progress on the Early Monitoring of Pine Wilt Disease Using Hyperspectral Techniques. *Sensors* **2020**, *20*, 3729. [CrossRef] [PubMed]
2. Ye, J. Epidemic status of pine wilt disease in China and its prevention and control techniques and counter measures. *Sci. Silvae Sin.* **2019**, *55*, 10.
3. Announcement of the State Forestry and Grassland Administration (2022 No. 5) (Pinewood Nematode Epidemic Area in 2022). Available online: http://www.forestry.gov.cn/ (accessed on 6 April 2022).
4. Umebayashi, T.; Yamada, T.; Fukuhara, K.; Endo, R.; Kusumoto, D.; Fukuda, K. In situ observation of pinewood nematode in wood. *Eur. J. Plant Pathol.* **2017**, *147*, 463–467. [CrossRef]
5. 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]
6. Shin, S.-C. *Pine wilt Disease*; Zhao, B.G., Futai, K., Sutherland, J.R., Eds.; Springer: Tokyo, Japan, 2008; pp. 26–32.
7. Pajares, G. Overview and Current Status of Remote Sensing Applications Based on Unmanned Aerial Vehicles (UAVs). *Photogramm. Eng. Remote Sens.* **2015**, *81*, 281–329. [CrossRef]
8. Tang, L.; Shao, G. Drone remote sensing for forestry research and practices. *J. For. Res.* **2015**, *26*, 791–797. [CrossRef]
9. Li, W.; Shen, S.; He, P.; Hao, D.; Fang, Y.; Tao, L.; Zhang, S. A precisely positioning technique by remote sensing the dead trees in stands with inexpensive small UAV. *China For. Sci. Technol.* **2014**, *28*, 102–105.
10. Yuan, Y.; Hu, X. Random forest and objected-based classification for forest pest extraction from UAV aerial imagery. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *XLI-B1*, 1093–1098. [CrossRef]
11. Zhong, T.; Liu, W.; Luo, Y.; Hung, C.C. A New Type-2 Fuzzy Algorithm for Unmanned Aerial Vehicle Image Segmentation. *Int. J. Multimed. Ubiquitous Eng.* **2017**, *12*, 75–90. [CrossRef]
12. Takenaka, Y.; Katoh, M.; Deng, S.; Cheung, K. 25–27 October 2017 Detecting forests damaged by pine wilt disease at the individual tree level using airborne laser data and worldview-2/3 images over two seasons. In *The ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*; Shinshu University Library: Jyväskylä, Finland, 2017; pp. 181–184.
13. Hu, G.; Yin, C.; Wan, M. Recognition of diseased Pinus trees in UAV images using deep learning and AdaBoost classifier. *Biosyst. Eng.* **2020**, *194*, 138–151. [CrossRef]

14. 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]
15. 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]
16. Wu, H. A study of the potential of using worldview-2 of images for the detection of red attack pine tree. In Proceedings of the Eighth International Conference on Digital Image Processing (ICDIP 2016), Chengdu, China, 20–22 May 2016.
17. Run, Y.; Youqing, L.; Quan, Z.; Xz, A.; Dw, A.; Lra, B. Early detection of pine wilt disease using deep learning algorithms and UAV-based multispectral imagery. *For. Ecol. Manag.* **2021**, *497*, 119493.
18. Shaoqing, R.; Kaiming, H.; Ross, G.; Jian, S. Faster R-CNN: Towards real-time object detection with region proposal networks. In Proceedings of the International Conference on Neural Information Processing Systems (NIPS), Montreal, QC, Canada, 7 December 2015.
19. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
20. Shi, S.; Guo, C.; Jiang, L.; Wang, Z.; Li, H. PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14 June 2020.
21. Ha, Q.; Watanabe, K.; Karasawa, T.; Ushiku, Y.; Harada, T. MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In Proceedings of the International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 25 September 2017.
22. Vollenweider, P.; Günthardt-Goerg, M. Diagnosis of abiotic and biotic stress factors using the visible symptoms in foliage. *Environ. Pollut.* **2006**, *140*, 562–571. [CrossRef] [PubMed]
23. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27 June 2016.
24. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21 July 2017.
25. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18 June 2018.
26. 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 Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017.
27. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision (ECCV), Online, 23 August 2020.
28. Song, H.; Sun, D.; Chun, S.; Jampani, V.; Han, D.; Heo, B.; Kim, W.; Yang, M.-H. ViDT: An Efficient and Effective Fully Transformer-based Object Detector. In Proceedings of the International Conference on Learning Representations (ICLR), Online, 3–7 May 2021.
29. Jocher, G. Ultralytics-YOLOv5. Available online: https://github.com/ultralytics/yolov5 (accessed on 4 October 2022).
30. Tolias, G.; Sicre, R.; Jégou, H. Particular object retrieval with integral max-pooling of CNN activations. In Proceedings of the International Conference on Learning Representations (ICLR), San Juan, Puerto Rico, 2–4 May 2016.
31. Lin, T.Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21 July 2017.
32. Li, H.; Xiong, P.; An, J.; Wang, L. Pyramid Attention Network for Semantic Segmentation. In Proceedings of the British Machine Vision Conference, Newcastle, UK, 3–6 September 2018.
33. Wang, C.Y.; Liao, H.; Wu, Y.H.; Chen, P.Y.; Yeh, I.H. CSPNet: A New Backbone that can Enhance Learning Capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14 June 2020.
34. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27 June 2016.
35. Liu, J.; Zhang, S.; Wang, S.; Metaxas, D.N. Multispectral Deep Neural Networks for Pedestrian Detection. In Proceedings of the British Machine Vision Conference (BMVC), York, UK, 19–22 September 2016.
36. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *Stat* **2016**, *1050*, 21.
37. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014.
38. Loshchilov, I.; Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
39. 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]
40. Iordache, M.-D.; Mantas, V.; Baltazar, E.; Pauly, K.; Lewyckyj, N. A machine learning approach to detecting pine wilt disease using airborne spectral imagery. *Remote Sens.* **2020**, *12*, 2280. [CrossRef]
41. 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]