remote sensing

MDPI

Article

LLAM-MDCNet for Detecting Remote Sensing Images of Dead Tree Clusters

Zongchen Li [1,†], Ruoli Yang [1,†], Weiwei Cai [2], Yongfei Xue [1,*], Yaowen Hu [1] and Liujun Li [3]

[1] College of Computer & Information Engineering, Central South University of Forestry and Technology, Changsha 410004, China; 20193085@csuft.edu.cn (Z.L.); 20193062@csuft.edu.cn (R.Y.); 20192764@csuft.edu.cn (Y.H.)

[2] Graduate College, Northern Arizona University, P.O. Box 4084, Flagstaff, AZ 86011, USA; vivitsai@csuft.edu.cn

[3] Department of Civil, Architectural and Environmental Engineering, University of Missouri-Rolla, Rolla, MO 65401, USA; llpwc@umsystem.edu

* Correspondence: author: xueyongfei@csuft.edu.cn

† These authors contributed equally to this work.

check for updates

Citation: Li, Z.; Yang, R.; Cai, W.; Xue, Y.; Hu, Y.; Li, L. LLAM-MDCNet for Detecting Remote Sensing Images of Dead Tree Clusters. *Remote Sens.* **2022**, *14*, 3684. https://doi.org/10.3390/rs14153684

Academic Editor: Jan Altman

Received: 4 July 2022
Accepted: 29 July 2022
Published: 1 August 2022

Abstract: Clusters of dead trees are forest fires-prone. To maintain ecological balance and realize its protection, timely detection of dead trees in forest remote sensing images using existing computer vision methods is of great significance. Remote sensing images captured by Unmanned aerial vehicles (UAVs) typically have several issues, e.g., mixed distribution of adjacent but different tree classes, interference of redundant information, and high differences in scales of dead tree clusters, making the detection of dead tree clusters much more challenging. Therefore, based on the Multipath dense composite network (MDCN), an object detection method called LLAM-MDCNet is proposed in this paper. First, a feature extraction network called Multipath dense composite network is designed. The network's multipath structure can substantially increase the extraction of underlying and semantic features to enhance its extraction capability for rich-information regions. Following that, in the row, column, and diagonal directions, the Longitude Latitude Attention Mechanism (LLAM) is presented and incorporated into the feature extraction network. The multi-directional LLAM facilitates the suppression of irrelevant and redundant information and improves the representation of high-level semantic feature information. Lastly, an AugFPN is employed for down-sampling, yielding a more comprehensive representation of image features with the combination of low-level texture features and high-level semantic information. Consequently, the network's detection effect for dead tree cluster targets with high-scale differences is improved. Furthermore, we make the collected high-quality aerial dead tree cluster dataset containing 19,517 images shot by drones publicly available for other researchers to improve the work in this paper. Our proposed method achieved 87.25% mAP with an FPS of 66 on our dataset, demonstrating the effectiveness of the LLAM-MDCNet for detecting dead tree cluster targets in forest remote sensing images.

Keywords: object detection; forest fires prevention; attention mechanism; remote sensing images

1. Introduction

The forest is an ecosystem capable of nourishing water and purifying the atmosphere [1,2]. Concurrently, the frequent forest fires in recent years have posed a significant threat to human life and property. Due to the characteristics of inevitable harmful gases released, their prevention and control have drawn attention progressively [3]. People have realized that three conditions must be presented to form a forest fire: (1) Combustible materials (including trees, grasses, shrubs, and other plants); (2) Fire-danger weather is an important condition; (3) The fire source is the dominant factor. Forest fires will not occur without even one of them. Various facts show that forest fires can be prevented, combustible materials and fire sources can be controlled by human beings, and fire-danger weather can be forecasted [4]. As a consequence, the establishment of a firebreak or barrier to prevent the spread of the fire has been deemed the primary choice for forest fire defense to effectively limit the frequency of forest fires.

For a long period, firebreaks were established without considering the possible fire sites. Therefore, they typically cut down trees arbitrarily to create firebreaks [5]. Unsuitable selection of firebreaks not only leads to the cutting down of quality trees and the waste of forest resources but also leads to the fire-prone dead and old plants being neglected, which can lay a potential hazard for forest fires. When establishing firebreaks, it is becoming common practice to cut dead and old plants with low water content selectively. In most countries, hiring rangers to monitor is the primary method of locating dead and old plants. However, this method is labor-intensive, and occasionally omissions will be unavoidable [6].

The advancement of UAV remote sensing technology in recent years has brought new ideas for the prevention and control of forest fires. Because of the UAV's overhead view, it can monitor stand types, search for dead and old plants more comprehensively, and provide a reference for constructing fire isolation zones [7]. Before employing remote sensing images, people typically need to parse the massive quantity of data; otherwise, it is difficult to extract valuable information directly. The initial stage in parsing is image classification, and its accuracy has a direct impact on the quality of parsing in the subsequent process. The introduction of neural networks with the development of machine learning and deep learning approaches allows for the classification of remote sensing images with high accuracy and automation. Many researchers try to apply it to the field of target detection in remote sensing images and have achieved good results. When completing the scene classification assignment, Yan used Googlenet with subnetworks, which can handle low and high-complexity remote sensing images separately. The main network, on the other hand, is weakly coupled with the auxiliary branches and necessitates a significant number of training sets [8]. Acquiring adequate remote sensing image features is the key to achieving remote sensing image classification. These methods, in general, improve feature extraction capabilities by developing network structures or attention mechanisms. Xu [9] selected the YOLO-V3 as the backbone for remote sensing object detection at different scales and employed DenseNet to enhance the network feature extraction capability. Guo used Generative Adversarial Network (GAN) constraint as domain adaptive constraint in the adaptive module based on Faster R-CNN to achieve instance-level domain adaptation and learn migratable features [10]. The final experimental results show that the method detects better than the base network in the small SAR object detection task.

However, the abovementioned methods simply utilize the features output by each convolutional layer in a crude manner without considering high-level semantic features. This is often sufficient when applied to object detection for natural and remote sensing scenes closer to the ground. Nevertheless, in remote sensing image object detection tasks that are far from the ground and have high information entropy, detailed feature information is needed to discriminate the differences between similar classes. When the target size in the remote sensing image is small, it is challenging to precisely locate the tiny area, leading to detection inaccuracy. Figure 1 shows some remote sensing images of the forests we collected. As seen in the Figure, there are still several issues to be addressed when performing the task of detecting targets in clusters of dead and old trees, as follows.

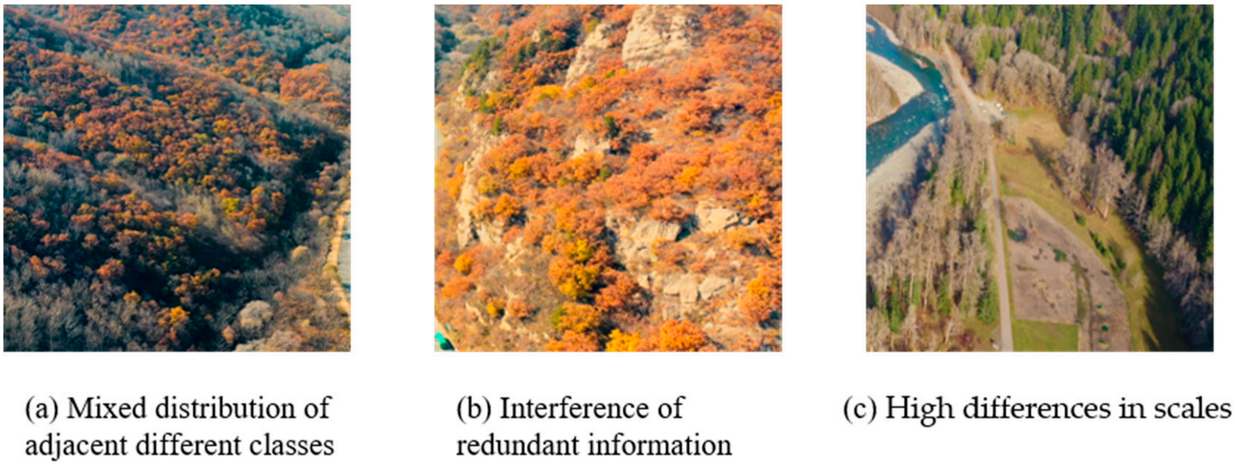

(a) Mixed distribution of adjacent different classes
(b) Interference of redundant information
(c) High differences in scales

Figure 1. Showcase of partial forest remote sensing images in the datasets.

- Mixed distribution of adjacent different classes: Because of the transformation of healthy trees to dead trees under adverse environments such as water shortage and soil salinization, remote sensing images in this area typically show interspersed rows of healthy trees and dead trees near each other.
- Interference of redundant information: Due to the complexity of the forest environment, remote sensing images captured by UAVs typically have irrelevant semantics or noise, e.g., exposed rocks, shaking water streams, vignettes due to lens shake, etc. The abovementioned redundant information may trigger the gradient explosion problem during the neural network training.
- High differences in scales: Some of the dead trees are too scattered in the forest cluster, showing characteristics such as small target features and high differences in scales from the clustered dead trees in the remote sensing images. In addition, when collecting remote sensing images, the scale of the objects will change due to the height of the sampling UAV and its shooting angle towards the ground.

To address the difficulty of locating detailed features of dead trees cluster contours due to the mixed distribution of adjacent different classes, some scholars have focused on locating important areas using bounding boxes and additional annotations [11]. Huang et al. suggested Part-Stacked CNN architecture to locate multiple object parts based on strong part annotation with manual labeling [12]. Di et al. proposed a detail feature recognition detection network Deep LAC, using a backpropagation linked valve linkage function to form a system for depth and fine-grained localization and alignment [13]. Zhang et al. presented a network for detecting multiple semantic parts or the whole object based on shared convolutional filter computation [14]. Xiang et al. proposed a CCA-ResNet where a CCA mechanism was introduced that equipped the feature map with hierarchical semantic features by unsupervised reconstruction of local and global features to retain detailed features from shallower layers [15]. The above methods focus only on the independent solution of differentiated local localization and local semantic feature-based learning when extracting detail features. However, the joint role of local detail localization and semantic feature learning is ignored.

To address the issue of miss and false detection due to interference of redundant information, He et al. presented an object detection method based on weighted image entropy. The method weights local entropy measurements by multiscale grayscale differences and adaptive thresholding operations to improve the SNR of small targets under circumstances where the interfering objects in the scene have similar thermal intensity measurements relative to the background [16]. Finally, it effectively enhanced the critical information and suppressed the redundant features. Han et al. suggested a multiscale detection algorithm using relative local contrast measurement (RLCM), which can effectively suppress the interference of all types of redundant information by computing multiscale RLCM for each pixel of the original IR image [17]. Huang proposed a Ship-YOLOv3 based on YOLOv3. The method reduces the interference of redundant features for the to-be-detected target by using k-means++ to cluster the dimensions of the bounding box, reducing some of the convolution operations, and adding the jump join mechanism [18].

To overcome the challenge of high-scale differences in clusters of dead trees due to naturally occurring and the camera angle of the UAV, Wang proposed an effective method for multiscale small target detection under natural scenes. The method improves the accuracy of small target detection by inferring a criterion for detecting small multiscale objects from the nature of the average grayscale absolute difference maximum map (AGADMM) of a natural scene [19]. Wu et al. suggested an improved small target detection network based on the YOLO-v5, in which the multiscale anchor mechanism of Faster R-CNN was utilized to make the network highly adaptable to different scenes, thus improving the accuracy of the network for small target detection [20].

Considering the above problems, in this paper, we present a new Longitude latitude cross attention-multipath dense composite network (LLAM-MDCNet) to overcome the difficulties of low detection accuracy due to the mixed distribution of adjacent different classes, redundant feature interference, and high differences in object scales. Our contributions are summarized as follows:

1. An object detection framework is presented based on a feature extraction network called Multipath dense composite network (MDCN), which can substantially increase the extraction of underlying and semantic features to enhance its extraction capability for information-rich regions. Eventually, the object detection accuracy of the LLAM-MDCNet is improved for remote sensing images.
2. In the row, column, and diagonal directions, the Longitude Latitude Attention Mechanism (LLAM) is presented and incorporated into the feature extraction network. First, it is capable of suppressing irrelevant and redundant information and improving the representation of high-level semantic feature information. Following that, irrelevant sematic features in different directions are also considered, which helps to improve the detection accuracy of irregular clusters of dead trees and the ability to detect small clusters of interspersed rows.
3. An AugFPN is added in MDCN, yielding a more comprehensive representation of image features with the combination of low-level texture features and high-level semantic information. Consequently, the network's detection effect for dead tree cluster targets with high-scale differences is improved.

The workflow diagram of LLAM-MDCNet is shown in Figure 2. First, the remote sensing images are preprocessed. Image preprocessing is divided into two parts: image annotation and image augmentation by rotation, flip, random cropping, and brightness transformation. Following that, the feature extraction network MDCN is utilized to extract the features in the remote sensing images. Then, the target bounding box is generated by RPN for the output features. The detection network consists of an ROI pooling layer and a classification and regression layer. The ROI Pooling layer outputs its pooled features and feeds them to two cascaded, fully connected layers for feature mapping before outputting them to the classification and regression layers. The fully-connected layer is decomposed into two sub-fully connected layers using singular value decomposition (SVD), which accelerates the computation of the fully-connected layer and significantly reduces the computations. The classification layer and regression layer adopt the same structure and loss function calculation method as the correlation layer in the RPN, and further classify and regress the candidate regions through the classification and regression network. Finally, remove the redundant detection frames using the non-maximum suppression method to obtain the final object detection results of forest remote sensing images for dead and old tree clusters.

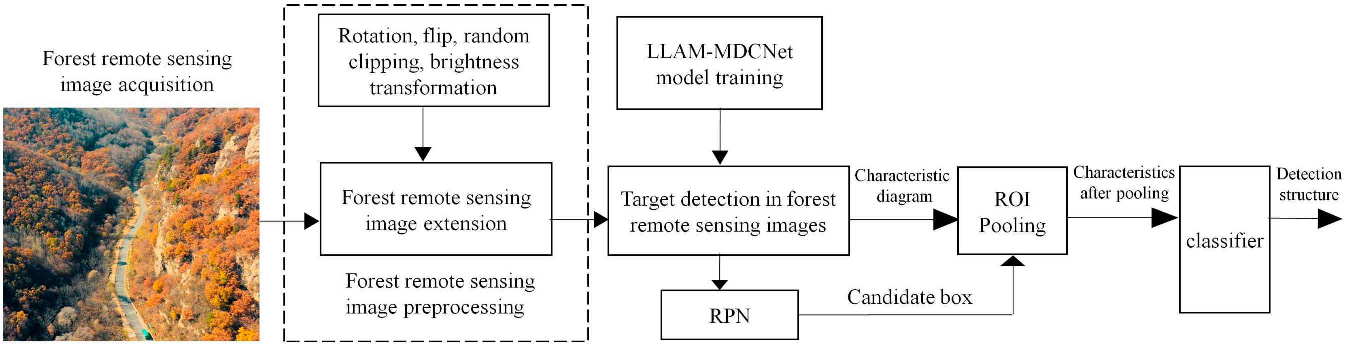

Forest remote sensing image acquisition
Rotation, flip, random clipping, brightness transformation
Forest remote sensing image extension
Forest remote sensing image preprocessing
LLAM-MDCNet model training
Target detection in forest remote sensing images
Characteristic diagram
ROI Pooling
Characteristics after pooling
classifier
Detection structure
RPN
Candidate box

Figure 2. The workflow diagram of LLAM-MDCNet.

2. Materials and Method

2.1. Data Acquisition

Due to the limited number of researchers working on ex-ante prevention and control of forest fires through remote sensing image object detection, publicly available datasets of dead tree cluster images in forests cannot be found yet. Consequently, we need to collect forest image data with dead tree clusters before conducting the research in this paper. First, we used DTI Mavic 3 (1080p, 60FPS) to capture 9766 images in different states at an altitude of about 90 m. Almost every image included both healthy and dead trees. Next, to fit the real scenarios and obtain images with smaller object targets, we raised the drone to an altitude of 200 m to capture 9751 images. Further, to cover as many forests states as possible, we collected forest stand types, including coniferous, deciduous broad-leaved, and evergreen broad-leaved forests from the Xiaoxing&apos; an Ling in northeastern China, the Qin Ling in central China, the Zepu Jinhuyang National Forest Park in northwestern China, and the Hengduan Mountain Range in southwestern China, mainly using climatic zones as divisions. These data are first cropped to an appropriate size and then precisely labeled by experts in the field of forestry. Then these images are converted into VOC format (the same as our open-source format for a more convenient comparison with other target detection networks.) Before input to LLAM-DRNet, our VOC format data is converted into YOLO format for easier reading.

To avoid the negative impact of height-width ratio mismatch on model training, we cropped all images in the dataset to a uniform size. To facilitate the study, by filtering, cropping, and normalizing the images to unify the pixel standards for depicting the real forest environment, we ended up with the aerial dead tree clusters dataset, which includes 19,517 images (see Figure 3). Since healthy trees are sometimes present in patches in real situations, only healthy trees are presented in some images in the dataset. In this paper, we have realized that the presence of only healthy trees in a small number of images may lead to an inter-class imbalance effect, and the detection of this issue will be discussed in the Discussion section.

2.2. Longitude Latitude Cross Attention Multipath Dense Composite Network (LLAM-MDCNet)

The distribution of some dead trees in the forest is too scattered, and tiny dead trees require the ability of small feature detection of the network. Healthy trees and dead trees are interspersed, which poses a great challenge to the network’s ability to extract details of dead trees. Tree withering is typically due to the lack of water or pest invasion, with an inevitable withering stagnation. Timely detection of the forest dead trees percentage within a certain range can alert the relevant forestry bureau to remedy the incompletely dead vegetation. The timely felling of dead trees can largely reduce the risk of forest fires. Therefore, the detection of dead tree clusters is of great practical importance.

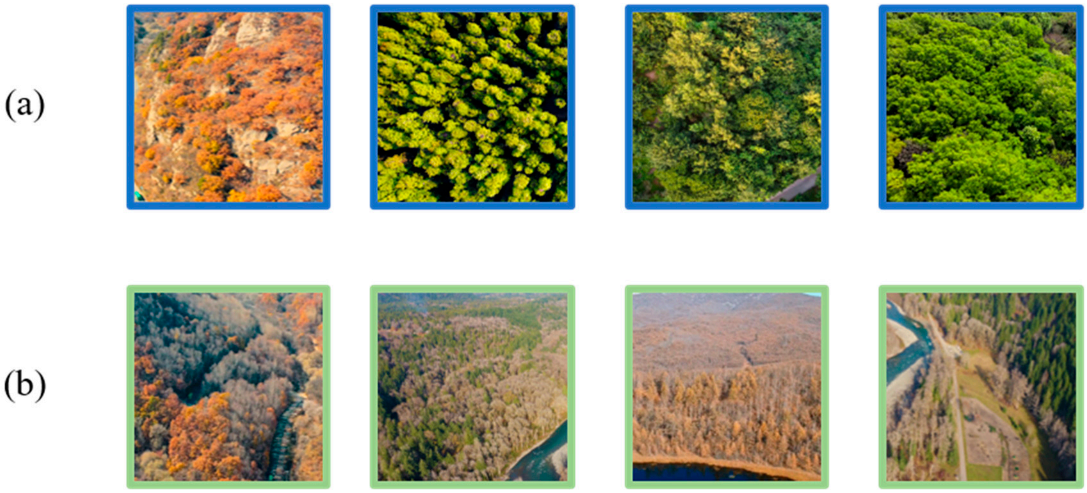


Figure 3. Example images from the aerial dead tree clusters dataset: (**a**) images without dead and old trees; (**b**) images including dead and old trees.

To address the above problems, in this paper we propose the Multipath Dense Network (MDCN) is proposed in this paper to extract remote sensing image features, and the overall structure is shown in Figure 4. The MDCN's baseline is DenseNet, whose densely connected module has powerful feature extraction and reuse capability, providing a solid foundation for subsequent detailed feature extraction. Firstly, the same training set images are simultaneously input into the DenseNet-based framework of the row-column direction convolution and the diagonal direction convolution for feature extraction. Secondly, LLAM is introduced in the row-column and diagonal directions after the second dense block of the base network. When the baseline network module incorporates an attention mechanism, some irrelevant, redundant information can be suppressed, and the representation of information with high-level semantic features can be improved. The LLAM direction in the row-column-dense path mainly considers the semantic features shown in the row and column directions which helped the realization of the detection of regular dead tree clusters in remote sensing images. The LLAM direction in the diagonal path mainly considers the semantic features characterized on the left and right diagonals to detect irregular clusters of dead trees in remote sensing images.

The combination of different multi-directional attention mechanisms in the two different paths not only enhances the network's ability to detect small clusters of dead trees scattered among healthy tree clusters but also to capture global features of the image. Thus, it can effectively improve detection accuracy when clusters of dead trees are detected. Further, low-level texture features and high-level semantic information are better combined using AugFPN for up-sampling to fuse their features. The low-level texture features can significantly characterize the contrast of the color features of healthy trees, and the high-level semantic features can express the overall trend of irregular dead tree clusters. Combining images at different levels improves the network's detection effect of dead tree clusters with high-scale differences. Then, the target bounding box is generated by RPN for the output layer.

Finally, the final classification structure and location information of the target is obtained. The object detection network consists of an ROI pooling layer and a classification and regression layer. The ROI pooling layer outputs its pooled features and feeds them to two cascaded fully connected layers for feature mapping before outputting them to the classification and regression layers. Simultaneously, the fully connected layer is decomposed into two sub-fully connected layers using singular value decomposition to speed up the computation of the fully connected layer and significantly reduce the computations. The classification layer and regression layer adopt the same structure and loss function calculation method in the correlation layer of the RPN, and further classify and regress the candidate regions through the classification regression network. And lastly, the redundant detection frames are removed by using the non-maximal so as to obtain the final object detection results of forest remote sensing images for dead and old tree clusters.

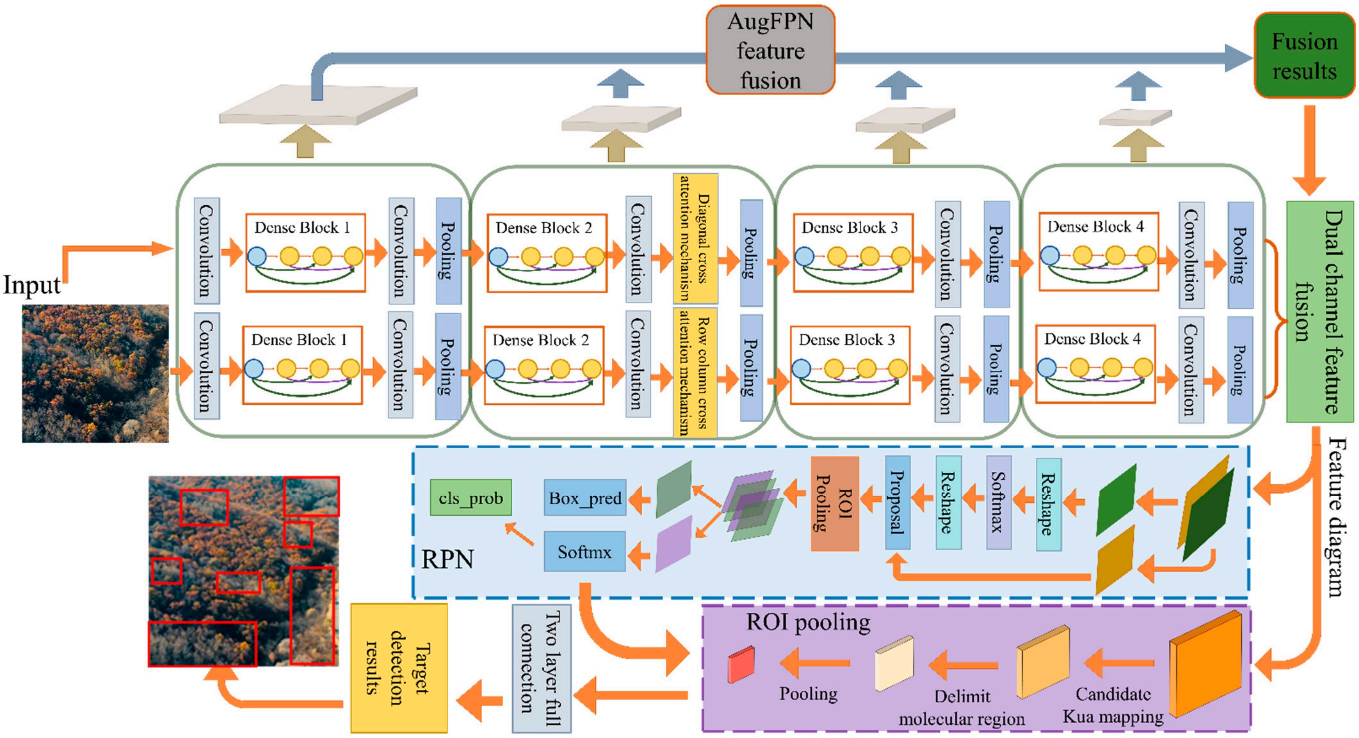


Figure 4. Structure of the LLAM-MDCNet.

As shown in Figure 4, LLAM-MDCNet mainly consists of MDCN, LLAM, and AugFPN. The specific implementation of each method is described below.

2.2.1. Multipath Dense Composite Network (MDCN)

MDCN consists of two identical DenseNet skeletons in parallel, and feature extraction is performed on the input image in each independent dense composite path. The structure of MDCN mainly consists of Dense Blocks and Transitions. There are 4 Dense Blocks in a DenseNet and multiple Bottleneck layers in a Dense Block. Normal convolutional networks generate L connections in the layer L of the dense composite path. Still, the densely connected module passes the feature overlay mapping from the previous layer to the other subsequent layers using dense connections. Each layer is connected with all previous layers in the channel dimension by concat, enabling the combination of multiple features, reducing the number of parameters, and enhancing feature reuse. Therefore, in the network with L layers, our network has $L * (L+1)/2$ connections. The expression for the dense connection module is as follows, where X_L represents the output feature map of layer L.

$$X_L = H_L([X_0, X_1, X_2, \cdots, X_{L-1}]) \tag{1}$$

$[X_0, X_1, X_2, \cdots, X_{L-1}]$ represent the feature mapping connections of all previous layers before layer L, H_L denotes the non-linear transformation function, including the batch normalization layer, the activation function ReLU, the dropout layer, and the 3 × 3 convolutional layer. MDCN employs a concat operation, which will make it likely to produce a large number of output channels. To control the model complexity, the transition block is introduced to not only halve the height and width of the input but also to vary the number of channels using 1 × 1 convolution. MDCN connects each layer directly to all subsequent layers to achieve feature reuse. During training, MDCN connects the feature maps at different levels, which can effectively improve the gradient propagation, enhance the number of features extracted, and increase the available semantic information for classification. The structure of mdcn is shown in Figure 5.

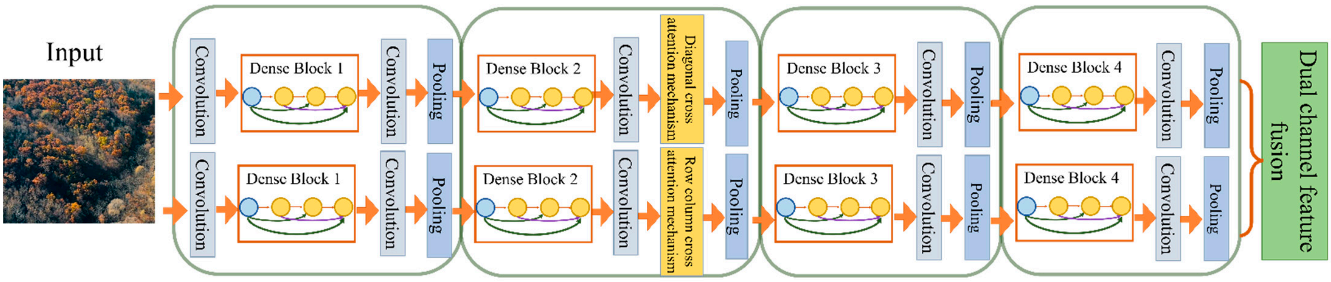


Figure 5. Structure of the MDCN.

2.2.2. Longitude Latitude Attention Mechanism (LLAM)

With the development of computer vision, various attention mechanisms have been proposed to address the lack of attention of neural networks to high-level semantics, which serves to ignore redundant features on the image that are not of interest, thus making the computer pay more attention to useful information on the image [21–23]. This selective attention mechanism is consistent with some of the mechanisms in discriminative detail features. Therefore, for more complete extraction of the underlying features and semantic features in the forest images, we introduce the Longitude latitude attention mechanism (LLAM) with different attention directions in two paths of the Double path dense composite path. The following figure shows the LLAM algorithm in row-column and diagonal directions. The working principle of Llam is shown in Figure 6.

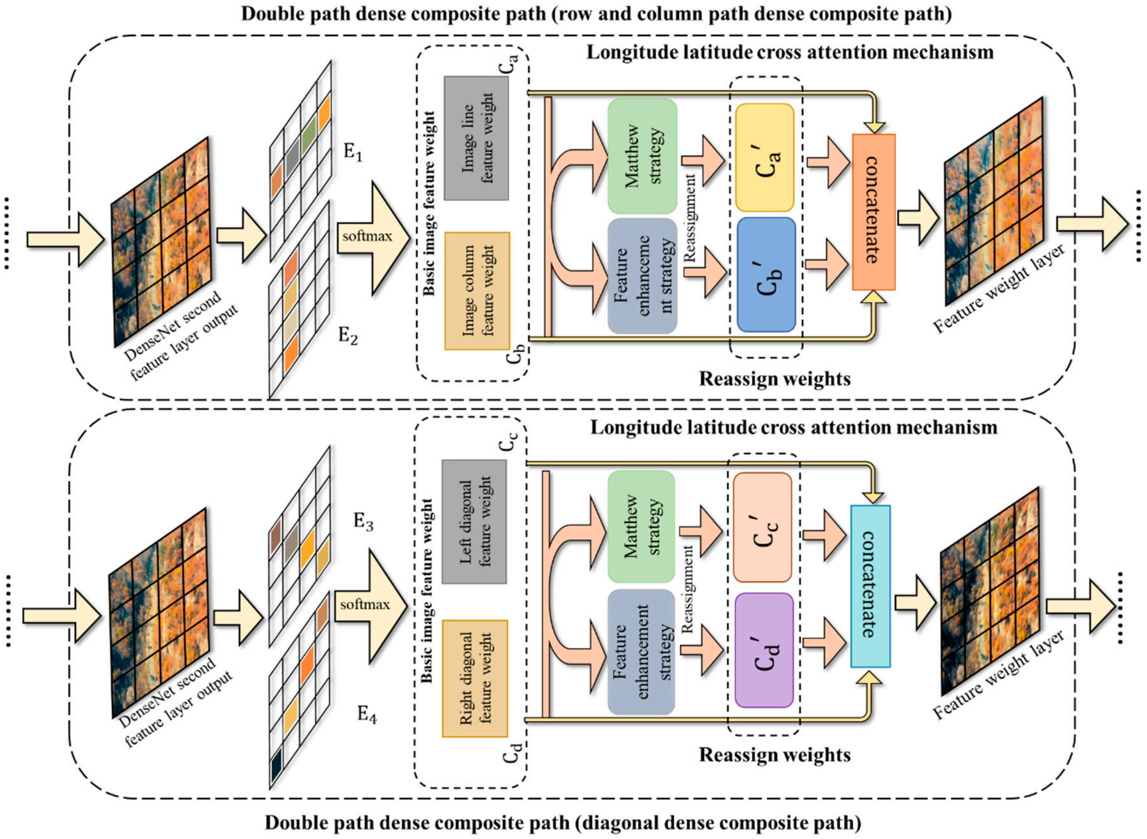

Double path dense composite path (row and column path dense composite path)
Longitude latitude cross attention mechanism
DenseNet second feature layer output
E_1
E_2
softmax
Basic image feature weight
Image line feature weight
Image column feature weight
C_a
C_b
Matthew strategy
Feature enhanceme nt strategy
Reassignment
C_a'
C_b'
concatenate
Feature weight layer
Reassign weights
Longitude latitude cross attention mechanism
DenseNet second feature layer output
E_3
E_4
softmax
Basic image feature weight
Left diagonal feature weight
Right diagonal feature weight
C_c
C_d
Matthew strategy
Feature enhancement strategy
Reassignment
C_c'
C_d'
concatenate
Feature weight layer
Reassign weights
Double path dense composite path (diagonal dense composite path)

Figure 6. Diagram of the LLAM.

Because the LLAM in MDCN only has directional differences, we introduce the principle of our LLAM algorithm in row-column directions.

- Step 1: LLAM operates on the upper layer of the input image to mine the shallow texture features in the image. Then, LLAM assigns weight coefficients to the feature from each row and column. Specifically, the input row-column feature weights C_a, C_b are generated in the row and column directions first. The set of eigenvectors of the vertices of the lth layer $h^l = [h_1, h_2, \ldots, h_N]^T$, N is the number of vertices in the image. The weight matrix is W. The corresponding attention coefficient can be obtained for each pixel, and the attention coefficient is $e_{i,j} = a * (W^T h_i, W^T h_j)$. Then the weights assigned to each pixel in each direction of the feature sequence j are obtained. Finally, the softmax activation function is introduced to regularize the attention coefficients. $e_{i,j}$ denotes the weight coefficient of each pixel assigned by the attention mechanism, and j denotes the feature sequence. i denotes a vertex node pixel, and h_j denotes the hidden layer information of the feature sequence j.

$$C_i = \sum_{j=1}^{n} \frac{exp(e_{i,j})}{\sum_{k=1}^{n} \exp(e_{i,k})} h_j \tag{2}$$

- Step 2: The assigned weight coefficients are reassigned using the Matthew strategy and feature reinforcement strategy, and deep features with more prominent semantic features are generated by convolutional layers and average pooling. Specifically, the extracted row-column weight features obtain the first reassigned weight $C_a{}'$ by the Matthew strategy. The formula for Matthew's strategy is shown in 4. which will deepen the impact of the high weight coefficients by multiplying by the minimum term penalty. The base weights can be multiplied to further enlarge the larger weights and further reduce the smaller weights. Thus, the redundant information with smaller diameters remaining after weight assignment is further filtered out, and weights with stronger information in multiple directions are extracted. The second part of the reassignment weights is to take the row-column basic weight features and obtain a second reassignment weight $C_b{}'$ by a feature enhancement strategy. The formula for the maximum weighting strategy is shown in Equation (4), where the maximum feature is considered a valid feature and is added to the α multiplier of the minimum value feature. In the equation, $0 \leq a \leq 1$ is the condition. The method takes the maximum value as the main feature and considers other subtle features, and fuses the main features with the subtle features to obtain the integrated features. These integrated features not only enhance the influence of the main factor on the results but also reduce the loss of subtle features.

$$C_a{}' = C_a * C_b - min(C_a, C_b) \tag{3}$$

$$C_b{}' = max(C_a, C_b) + a * min(C_a, C_b) \tag{4}$$

- Step 3: The softmax function is used to map the corresponding adjacent matrix for the reassigned deep features. LLAM better constructs the relationship between the features by deep mining the features and reassigning the original feature weights. Thus, the accuracy of dead tree cluster detection is improved. Specifically, the maximum of the four weight features $[C_a, C_b, C_a{}', C_b{}']$ is matched by concatenate and used to complement the results of the base image weight coefficients $[C_a, C_b]$. The four image weight coefficients in *LLAM* integrate the processed feature information in a tandem manner by the concatenate function. *LLAM* refers to the final attention mechanism in the row-column direction.

$$LLAM = concatenate([C_a, C_b, C_a{}', C_b{}']) \tag{5}$$

2.2.3. AugFPN

To preserve as much as possible the detailed features of dead tree clusters in images so that the dead tree cluster features, regardless of size, can be effectively detected by the MDCN and yield the correct detection results, we added AugFPN [24] to the network to fuse the features of different scales. The reason is that although the forest cluster images input into MDCN's two paths extract a large number of forest cluster features by convolution, the perceptual field of image features becomes larger and inevitably accompanied by the loss of small dead tree cluster features with the fragmented distribution. Therefore, in this paper, we first concatenate the convolutional paths in row-column and diagonal directions and add the AugFPN after the fused layers. This structure groups the layers that do not change the size of the feature map into a stage in the forward propagation of the network, and up-samples the feature map, then fuses the two and performs the convolution through the lateral connection. The advantage of this is to use both the high-resolution information of the lower layer features and the high semantic information of the higher layer features to predict by fusing them. The schematic diagram is shown in Figure 7, where B1~4 represent the four layers of MDCN with an added attention mechanism. M1~4 layers represent the auxiliary losses of the four layers, and P represents the main loss. The same supervisory signal is added to the features of each layer. In the following, the algorithm principle of AugFPN is shown specifically.

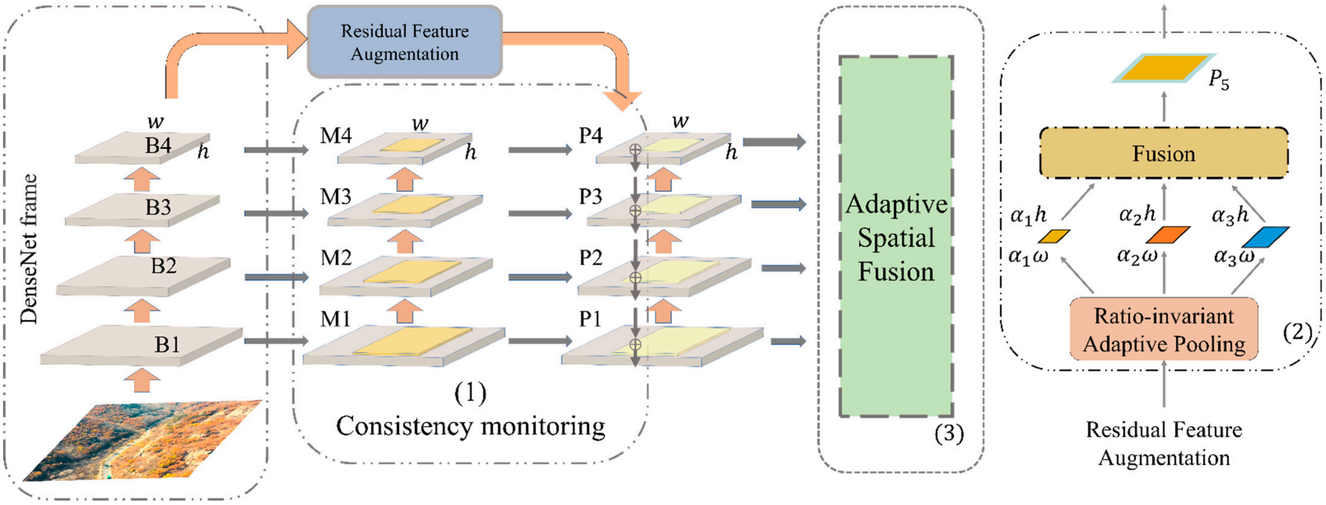


Figure 7. Schematic diagram of AugFPN fusion framework.

1. Bottom-up

First, the bottom-up feature extraction is performed by feeding the images into the MDCN for feature extraction. The outputs of the convolution blocks conv2, conv3, conv4, and conv5 are defined as $\{B_1, B_2, B_3, B_4\}$, and these are the outputs of the last residual block in each stage. Next, the features learned by $\{B_1, B_2, B_3, B_4\}$ are performed dimensionality reduction by 1×1 convolution, and the downscaling is done to get $\{M_1, M_2, M_3, M_4\}$ respectively, until they have the same number of channels, and then the features are summed. When the features are summed up, the expression ability of multiscale features will be weakened because of the different semantic information contained. Therefore, the corresponding feature map is obtained separately for each candidate region in $\{M_1, M_2, M_3, M_4\}$, and classification and regression operations are made. Then a weighted sum is made between the obtained loss and the loss of the network itself. The same supervision signal is implemented on these feature maps by the consistent supervision mechanism so that the laterally connected feature maps contain similar semantic information, which solves the problem of weak multiscale feature expression arising from different semantic information.

2. Top-down feature fusion

Second, the more abstract and semantic high-level feature map is up-sampled, and the result of the up-sampling is fused with the same size feature map generated from the bottom-up. Since the two horizontally connected layers of features are the same size, the bottom-level localization detail information is better utilized. Although the low-level features are enhanced by the high-level features from the top layer during feature fusion because the top-level features have performed 1×1 dimensionality reduction, it will inevitably result in the loss of M_4 layer information. However, since the information of C_4 is more comprehensive, the structure of residual feature enhancement is used to perform adaptive pooling of C_4. Then the feature maps at each scale are performed 1×1 dimensionality reduction before up-sampling. The features after up-sampling are summed according to the learned weights, and the obtaining M_4 and M_5 are fused. Different contextual information is extracted using ratio-invariant adaptive pooling to reduce the information loss of the highest-level features in AugFPN in terms of residuals during top-down feature fusion.

3. After feature fusion

Finally, soft ROI selection is introduced to use ROI features in different pyramid levels better to provide better ROI features for subsequent location refinement and classification. Since the features of each region of interest (ROI) are selected based on the scale of the proposal to determine the corresponding feature map, the feature information in the ignored layers is then lost, which directly affects the final detection results. So similar to the adaptive summation used in Residual Feature Augmentation, the features corresponding to an arbitrary ROI feature map $\{P_1, P_2, P_3, P_4\}$ are extracted from $\{M_1, M_2, M_3, M_4\}$. And then, using the network to learn the weight parameters, these features from layers $\{P_1, P_2, P_3, P_4\}$ are summed, and the structure is used as the final feature of this ROI.

In summary, AugFPN makes a series of improvements based on the FPN. First, AugFPN introduces consistency monitoring to reduce the semantic gap between features at different scales before feature fusion. Secondly, ratio-invariant contextual information is extracted by ASF in feature fusion to reduce information loss in feature mapping at the highest pyramid level. Finally, a soft RoI selection method is used to perform maximum pooling for non-uniform size inputs to obtain a fixed-size feature map for better extraction and fusion of semantic features in images. The network, after adding the AugFPN, can effectively identify the smaller clusters of dead trees with scattered distribution, which can solve the problem of difficult detection due to the high-scale differences and the too fragmented distribution of some dead tree clusters.

3. Results

3.1. Experimental Environment and Preparation

To ensure that the experiments in this paper are valid and fair, all the experiments are conducted in the same environment and use the exact same hyperparameters in the model.

The hardware environment includes a CPU with R9-5950X | 3.4GHz | 16 cores and 32 threads; the GPU is RTX2060 with 6GB of video memory. The software environment includes CUDA Toolkit 10.0; CUDNN V7.5.0; Python 3.6; torch 1.8.1; torchvision 0.9.1. The uniform input of image size is 256 × 256, and translations, rotations, scaling, and adding noise were performed during the input to extend the dataset. A total of 19,517 images were obtained. In this paper, a 10-fold cross-validation method is employed for training. Before training the network, considering the performance of hardware devices and the training effect, a stochastic gradient descent (SGD) is utilized. The batch size is set to 16, the momentum parameter is set to 0.9, and the number of epochs is set to 15. The Adam optimizer is employed in the model. Since changing the learning rate affects the convergence speed and stability of the model, a callback function is included. The learning rate is set to 0.0001 for the first 10 epochs, and the decay rate of the weights is set to 0.0005 for the last 5 epochs to improve the fitting speed.

3.2. Evaluation Metrics

To present our work clearly and robustly to the reader, in this subsection, we briefly describe the model evaluation metrics involved in this paper and how they are computed.

True Positive (*TP*): positive in prediction, and positive in real.

False Positive (*FP*): positive in prediction, but negative in real.

False Negative (*FN*): negative in prediction, but positive in real.

True Negative (*TN*): negative in prediction, and negative in real.

Accuracy: The percentage of the number of correctly detected samples to the number of all samples.

$$Accuracy = \frac{TN + TP}{FP + TN + TP + FN} \tag{6}$$

Precision: The percentage of correct positive samples detected to the number of all positive predicted samples.

$$Precisoin = \frac{TP}{TP + FP} \tag{7}$$

Recall: The percentage of correct positive samples detected to the actual number of positive samples.

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

In addition to detection accuracy, researchers often need to explore the speed of object detection algorithms, which is extremely important for some real-time detection scenarios. A common metric for evaluating speed is Frame Per Second (FPS), the number of images that can be processed per second.

3.3. Ablation Experiments

To objectively and independently evaluate the performance of each method proposed in this paper, we implemented seven sets of ablation experiments for LLAM-MDCNET and removed MDCN, LLAM, and AugFPN sequentially. The deleted implementation of MDCN should adopt DenseNet as a replacement, and similarly, the deleted implementation of AugFPN should adopt FPN as a replacement. The results are shown in Table 1.

Table 1. Experimental statistics of ablation for remote sensing image target detection of dead and old trees using different sub-modules in LLAM-MDCNet.

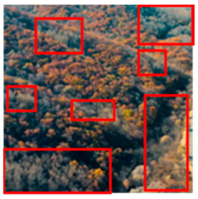
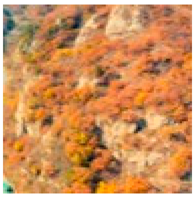
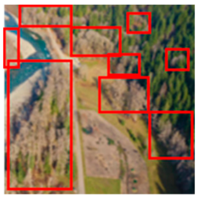
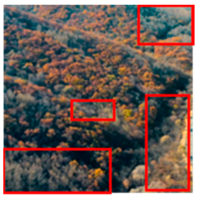
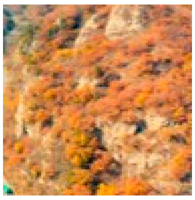
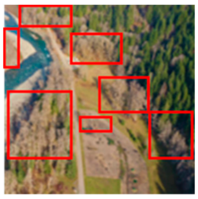
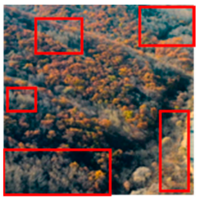
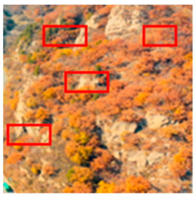
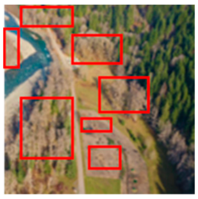
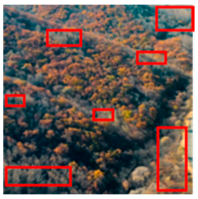
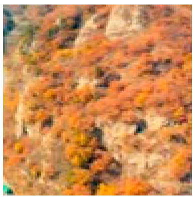
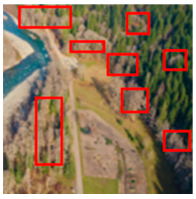
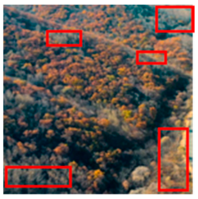
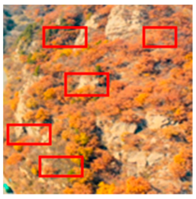
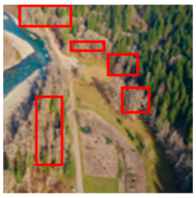
Number	Method	AP (%)	AR (%)	FPS
1	**MDCN + LLAM + AugFPN (LLAM-MDCNet)**	**87.25**	**51.91**	**66**
2	MDCN + LLAM	86.01	49.32	66
3	MDCN + AugFPN	84.46	50.76	67
4	LLAM + AugFPN	85.16	50.71	79
5	MDCN	82.27	46.29	68
6	LLAM	83.24	45.66	77
7	AugFPN	80.97	48.27	77
8	DenseNet	78.46	44.59	78

- By comparing groups 1 and 2, we find that the enhancement of AR using AugFPN is more obvious than AP. This is because using AugFPN enables the low-level texture features to significantly contrast the color features of healthy trees with dead trees, and the high-level semantic features to express the overall trend of irregular dead tree clusters. Specifically, fewer small target positive samples are ignored, and the use of AugFPN can improve the ability of the model to learn positive samples.
- Comparing groups 1 and 3, we find that the loss of accuracy by losing the LLAM is significant. This may be because the combination of different multi-directional attention mechanisms in two different paths not only makes the network stronger in detecting small clusters of interspersed dead trees with healthy trees but also makes the network much better at capturing global features of the images.
- By comparing groups 1 and 4, we find that removing the multipath dense composite network leads to a certain decrease in both AP and AR and a slight increase in FPS. This is not only because the parallel structure of the network substantially increases the underlying features and semantic features that can be extracted but also because the single-path network cannot fully exploit the LLAM.

- The comparison of group 1 and groups 5, 6, 7, and 8 shows that the removal of MDCN, LLAM, or AugFPN leads to a huge decline in AP and AR, and the superiority of LLAM-MDCNet over DenseNet is obvious.

To give the reader a concrete picture of the usefulness of the proposed method in remote sensing image detection of old and dead tree clusters, we analyze the performance of several sets of ablation experiments on individual images, as shown in Table 2.

Table 2. Visual comparison of the test results.

Experimental Method	Detection Result		
MDCN + LLAM + AugFPN (LLAM-MDCNet)			
MDCN + LLAM			
MDCN + AugFPN			
LLAM + AugFPN			
DenseNet			
	a	b	c

As shown in Table 2, the distribution of dead trees and healthy trees in Figure a is mixed, and the boundary is unclear. If DenseNet is used instead of MDCN, it is difficult to fit the dead tree clusters, resulting in their omission. Thus, it shows that MDCN can greatly improve the feature extraction ability. There are no dead and old trees in Figure b. Only maple trees suspected to be dead and bare rock faces are present, and no false detection occurs when LLAM is present. This indicates that LLAM is a powerful tool for reducing the interference of redundant information. Scattered individual dead trees appear in the forest in the upper right corner of Figure c. Small areas of dead trees are missed when AugFPN is not present. AugFPN is also significant for forest fire prevention because these small areas of dead trees can also be a fire source.

The results of eight sets of ablation experiments fully demonstrate the contribution of MDCN, the latitude-longitude attention mechanism, and AugFPN. LLAM-MDCNet is more suitable than DenseNet for the object detection task of remote sensing images of dead and old tree clusters.

3.4. Compared with the State-of-the-Art Methods

To further analyze the performance of LLAM-MDCNet, the experimental results based on LLAM-MDCNet are compared with some state-of-the-art models. The currently popular object detection methods for deep learning are mainly divided into two broad categories, namely object detection methods with CNN or Transformer as the baseline. The models involved in the comparison experiments are mainly from the post-2018 literature, and the superior methods in the Kaggle competition and the models used are carefully tuned. DANet adaptively integrates local features and global dependencies by attaching two types of attention modules on top of the traditional expanded FCN, modeling semantic interdependencies in the spatial and channel dimensions. The features at each location are selectively aggregated by a weighted sum of features at all locations [25]. Deeplabv3+ introduces the spatial pyramidal pooling (SPP) module or encoder-decoder structure to deep neural networks to refine the target edges [26]. Ding, X et al. [27] propose the ACNet with the asymmetric convolution block (ACB, Asymmetric Convolution Block) as a construction block for CNN, which uses a one-dimensional asymmetric convolution kernel to augment the square convolution kernel to improve the accuracy. DNL divides the original non-local expression into pairwise and unary terms by mathematical transformation and decouples the two so that they do not affect each other in terms of learning content and gradient propagation [28]. OCRNet computes the relationship between each pixel and target region and augments each pixel representation with object context representation [29]. UPerNet introduces unified perceptual resolution to integrate the variability between different datasets and learn different visual concepts from heterogeneous image annotations [30]. yolov4 uses Mosaic data augmentation means, SAT self-adversarial training, etc., to implement an efficient and powerful model that allows anyone to use a 1080Ti or 2080Ti GPU to train a fast and accurate object detection network [31]. YOLOv5 uses Focus and C3Net for the backbone network, and two different C3Net and Detect are designed to achieve fast detection [32]. Sparse R-CNN discards dense concepts such as anchor boxes or reference points and starts directly from a sparse set of learnable proposals without the handle of NMS [33]. RepPointsV2 introduces corner point detection and foreground heat map into the pure regression target detection algorithm, prompting joint learning between tasks to improve the network's feature representation and obtain better joint predictions [34]. ViT-FRCNN attempts to detect and localize objects in an image by adding a ViT with a detection-specific task head, demonstrating several known properties associated with transformers, including pre-training capabilities and fast fine-tuning performance [35]. SETR uses pure Transformers instead of encoders based on stacked convolutional layers to gradually reduce the spatial resolution. At the same time, it treats the input image as a sequence of image patches and transforms the sequence using global self-attentive modeling for discriminative feature representation learning [36]. YOLOS chooses a random initial DET as a proxy for the target representation to avoid inductive bias in the presence of prior knowledge in 2D structure and label assignment. At each forward propagation, it constructs an optimal even match between the DET and the real target [37]. DETR employs the idea of sequence prediction similar to machine translation to suppress repetitive predictions using self-attention, outperforming the traditional approach of target detection [38]. Table 3 gives their comparative experimental results on the dataset of forest dead tree clusters.

Table 3. Comparison experiment between LLAM-MDCNet and SOTA models.

Baseline	Method	mAP	mAP^{50}	mAP^{75}	AR	FPS
CNN	DANet [25]	65.35	68.61	61.26	38.48	52
	Deeplabv3+ [26]	64.15	67.99	60.08	37.94	47
	ACNet [27]	66.46	69.58	62.85	38.45	44
	DNL [28]	66.14	70.48	62.57	38.94	49
	OCRNet [29]	65.87	68.50	61.81	38.41	54
	UperNet [30]	79.54	81.81	75.67	39.76	57
	YOLOv4 [31]	78.52	80.48	74.19	43.48	34
	YOLOv5 [32]	70.65	72.84	68.81	42.15	122
	Sparse R-CNN [33]	76.75	73.96	72.17	46.64	62
	RepPointsV2 [34]	81.15	82.64	77.81	46.34	78
	LLAM-MDCNet	**87.25**	**89.01**	**84.34**	**50.30**	**66**
Transformer	ViT-B-FRCNN [35]	85.85	86.10	81.66	49.20	21
	SETR [36]	88.61	89.97	85.45	51.67	16
	YOLOS [37]	86.91	88.95	84.71	50.68	27
	DETR [38]	86.12	89.62	84.78	50.96	24

The results show that most of the Transformer-based models have significantly higher accuracy than the CNN-based models, probably because the transformer structure gives them a larger real-world receptive field. CNN-based models such as YOLOv5 cannot meet the high accuracy requirements of the remote sensing image object detection of dead and old tree clusters. If there is a missed detection, the ignored dead and old trees may become a potential fire source. The Transformer-based model, on the other hand, typically has very low FPS, and the real-time performance for dead tree detection is relatively unsatisfactory. While LLAM-MDCNet shows average performance in terms of FPS, the accuracy is comparable to each of the popular Transformer-based models, so the LLAM-MDCNet proposed in this paper achieves the best speed-accuracy tradeoff among the models. We identify possible reasons for the superior performance of our proposed LLAM-MDCNet model:

1. MDCN can substantially increase the extraction of underlying and semantic features to enhance its accurate extraction capability for more complex features and information-rich regions. Eventually, the object detection accuracy of the LLAM-MDCNet is improved for remote sensing images.
2. LLAM in row-column and diagonal directions can not only suppress the irrelevant and redundant information but also improve the representation of high-level semantic feature information, which is beneficial for improving the detection accuracy of irregular clusters of dead trees and the ability to detect small clusters of interspersed rows.
3. AugFPN can produce a more comprehensive representation of image features by combining low-level texture features and high-level semantic information. Consequently, the network's detection effect for dead tree cluster targets with high-scale differences is improved.

3.5. Interpretability of the Model

The experimental results in Sections 3.4 and 3.5 demonstrate that each of the submodules in LLAM-MDCNet contributes to the outcomes. However, the above experiments do not explain why LLAM-MDCNet works better than DenseNet. To visually analyze the focus of our model, we used Grad-CAM to visualize the output of the last convolutional layer of LLAM-MDCNet and DenseNet. The results of the Grad-CAM are shown in Figure 8, and the colors on the plot (from blue to red) represent the degree of contribution to the outcomes. The larger the contribution, the closer the color is to red.

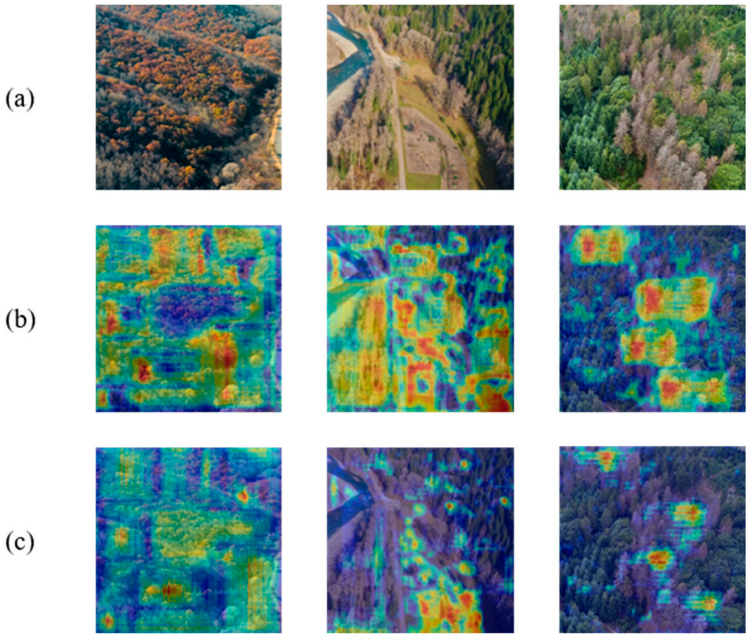


Figure 8. Class activation maps(CAM): (**a**) Original picture (**b**) LLAM-MDCNet (**c**) DenseNet.

Clearly, the proposed LLAM-MDCNet (a basic network based on DenseNet) in this paper pays close attention to the semantic regions relevant to the task. DenseNet focuses more on shallow, significant features, such as grayish close to the dead trees, and does not focus on small clusters of dead trees, as depicted in Figure 8. LLAM-MDCNet, on the other hand, not only focuses on dead trees of all sizes but also activates healthy tree clusters surrounding dead tree clusters. The outcomes of the LLAM-MDCNet demonstrate that the method can fully exploit contextual information and avoid feature confusion between valid semantic and redundant information.

4. Discussion

This paper makes an aerial dataset of dead tree clusters publicly available, which includes various healthy trees, dead trees, and objects suspected to be dead trees in the forest. Through several sets of experimental comparisons and analyses, it is verified that the proposed LLAM-MDCNet is effective for the task of remote sensing image object detection of dead and old trees, and addresses the three primary issues of “mixed distribution of adjacent different classes”, “interference of redundant information”, and “high differences in object scales”.

Hyperspectral imaging is a fine-grained technique capable of capturing and analyzing point-by-point spectra over a spatial area. Because unique spectral “features” can be detected at different spatial locations of individual objects, it can detect visually indistinguishable substances. Classification and detecting ground objects using hyperspectral or multispectral images is a typical application of computer vision technology in remote sensing. However, we believe that in this paper hyperspectral images do not apply to our method for the following reasons: (1) Hyperspectral images typically have a large number of channels (much more than the three channels of RGB images), where only partial information is useful. As a consequence, the information of these channels must be filtered before being input into the network. (2) Since hyperspectral images contain large and dense semantic information, deep learning methods are employed with a shallow structure. This results in a smaller receptive field, which is inconsistent with our original idea of using global contextual information. In conclusion, using hyperspectral images for dead tree cluster detection is a good direction because hyperspectral images have richer details than RGB images. Until then, a proven network is still required in machine learning to harness it.

It has been mentioned above that the presence of only healthy trees in images in the dataset may lead to an inter-class imbalance effect. The inter-class imbalance problem refers to the fact that when the dominance of a larger number of classes in backpropagation, the detector tends to predict a larger number of classes due to the training error minimization principle. In the practical application of actual forest fire prevention and control, the image quality is uneven. Therefore, we need to fit the real situation as much as possible to improve the robustness of the model. Obviously, we have to make a tradeoff between inter-class imbalance and the risk of the dataset not fitting the real situation. We selected all images with only healthy trees in the dataset. We added 0.5% of the total dataset with only healthy trees in each experiment to observe the possible classification surface bias, shown in Figure 9.

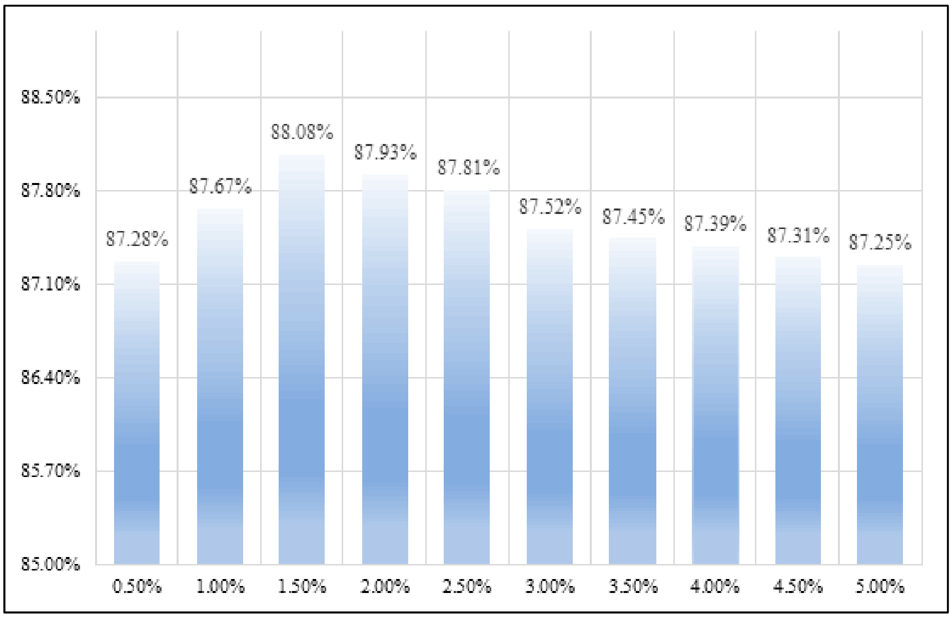


Figure 9. Inter-class imbalance test.

Figure 9 shows that it is acceptable for the results obtained according to the classification plane to have a small deviation. LLAM-MDCNet is less sensitive to inter-class imbalance, and even improves the overall accuracy by adding a small number of healthy tree images. The reason is that while MDCN substantially increases the semantic features that can be extracted by the network, the two parallel feature extraction networks can complement and correct mutually during the backpropagation.

Analyzing samples of false detections can provide considerably beneficial insights into the direction of network improvement. Therefore, we analyzed a sample of forest images that showed severe false detection during the experiment. As shown in Figure 10a, large areas of old and dead trees are ignored, and only pixels near healthy trees are detected correctly. This is because the altitude of the UAV is too high when shooting, the forest area contained in the image is too large, and there is a lack of reference between objects and noise blurring between images. CNN-based models tend to lose output resolution when predicting images with large sizes. This is because some degree of down-sampling of the feature map along the network is required to increase the amount of context, leading to blurring around the object's edges. In Figure 10b, there are little red clusters in the dataset. The UAV view is parallel to the ground when capturing this image, resulting in a lack of contextual information and shape features. The interspersed rows of dead trees and maple-red clusters suspected of being dead trees also raise various challenges for detection. In the future, image denoising techniques and Transformer-based models can be considered to improve the processing of such images. Further, increasing the diversity of data and performing data cleaning are also essential components of our future work.

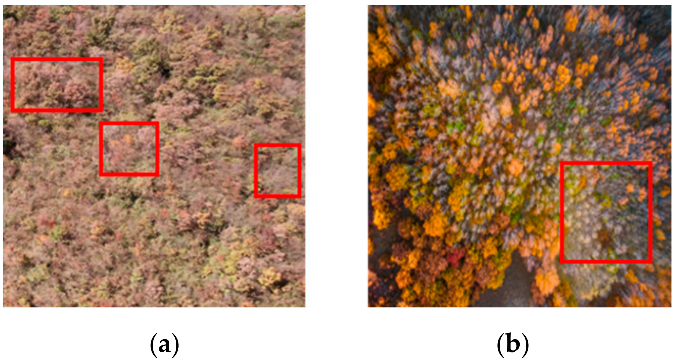


Figure 10. (**a**) The image size is too large, and there is a lot of noise. (**b**) The viewing angle is completely parallel to the ground.

5. Conclusions

In recent years, the frequent occurrence of forest fires has drawn the attention of governments from all over the world due to global warming. Preventing forest fires has become an essential means of protecting the ecosystem and maintaining people's property and safety. From finding possible clusters of dead trees as a starting point, this paper combines deep learning with remote sensing images captured by UAVs. It proposes an LLAM-MDCNet for dead and old tree clusters remote sensing image object detection to improve the accuracy of their detection. Our methods include: proposing MDCN to enhance the capability to detect information-rich regions and to address the "mixed distribution of adjacent different classes" issue; LLAM is presented and incorporated into the network, which is capable of suppressing irrelevant and redundant information and improving the representation of high-level semantic feature information, addressing the issue of "Interference of redundant information"; AugFPN can combine low-level texture features and high-level semantic information to address the "High differences in object scales" difficulty. Experiments on the dataset of forest dead tree clusters show that LLAM-MDCNet has 87.25% mAP and an FPS of 66. LLAM-MDCNet has close performance with the Transformer-related model in accuracy and is far superior in detection speed. Therefore, it can be demonstrated that our proposed method is superior to the methods compared in this paper.

In the future, we will make the collected dataset publicly available for other researchers who study forest fire prevention, which is also a major contribution of this paper. We will also focus on addressing the limitations of our work in this paper. An effective image preprocessing algorithm is needed to improve the performance of detecting blurred and noisy images. Simultaneously, more datasets are required to improve the algorithm's accuracy and performance to play a more important role in forest fire prevention and control and ecosystem protection. Lastly, we will try to apply our algorithm to fire departments to make more useful suggestions for forest fire prevention and the construction of forest firebreaks.

Author Contributions: Z.L.: methodology, writing—original R.Y.: software, data acquisition, formal analysis. W.C.: model guidance, resources. Y.X.: validation, project administration, funding acquisition, supervision. Y.H.: draft preparation, conceptualization. L.L.: visualization, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Changsha Municipal Natural Science Foundation (Grant No. kq2014160), in part by the National Natural Science Foundation in China (Grant No. 61703441), in part by the key projects of the Department of Education Hunan Province (Grant No. 19A511) and in part by Hunan Key Laboratory of Intelligent Logistics Technology (2019TP1015).

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: We are grateful to all members of the Forestry Information Research Centre for their advice and assistance in the course of this research.

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

References

1. Boyd, D.S.; Danson, F.M. Satellite remote sensing of forest resources: Three decades of research development. *Prog. Phys. Geogr.* **2005**, *29*, 1–26. [CrossRef]
2. Wu, G.; Xiao, H.; Zhao, J.; Shao, G.; Li, J. Forest ecosystem services of Changbai Mountain in China. *Sci. China Ser. C Life Sci.* **2002**, *45*, 21–32. [CrossRef] [PubMed]
3. Nijhuis, M. Forest fires: Burn out. *Nature* **2012**, *489*, 352–354. [CrossRef]
4. Tedim, F.; Xanthopoulos, G.; Leone, V. Chapter 5—Forest Fires in Europe: Facts and challenges A2. In *Wildfire Hazards, Risks and Disasters*; Elsevier: Oxford, UK, 2015; pp. 77–99.
5. Das, P.; Thomas, H.; Moeller, M.; Walther, A. Large-scale, thick, self-assembled, nacre-mimetic brick-walls as fire barrier coatings on textiles. *Sci. Rep.* **2017**, *7*, 39910. [CrossRef] [PubMed]
6. Bondi, E.; Jain, R.; Aggrawal, P.; Anand, S.; Hannaford, R.; Kapoor, A.; Piavis, J.; Shah, S.; Joppa, L. BIRDSAI: A dataset for detection and tracking in aerial thermal infrared videos. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass, CO, USA, 1–5 March 2020; pp. 1747–1756.
7. Akhloufi, M.A.; Couturier, A.; Castro, N.A. Unmanned aerial vehicles for wildland fires: Sensing, perception, cooperation and assistance. *Drones* **2021**, *5*, 15. [CrossRef]
8. Yan, S.; Jing, L.; Wang, H. A new individual tree species recognition method based on a convolutional neural network and high-spatial resolution remote sensing imagery. *Remote Sens.* **2021**, *13*, 479. [CrossRef]
9. Xu, D.; Wu, Y. Improved YOLO-V3 with DenseNet for multi-scale remote sensing target detection. *Sensors* **2020**, *20*, 4276. [CrossRef] [PubMed]
10. Guo, Y.; Du, L.; Lyu, G. SAR target detection based on domain adaptive faster R-CNN with small training data size. *Remote Sens.* **2021**, *13*, 4202. [CrossRef]
11. Welinder, P.; Branson, S.; Mita, T.; Wah, C.; Schroff, F.; Belongie, S.; Perona, P. Caltech-UCSD Birds 200. In *Technical Report CNS-TR-2010-001*; California Institute of Technology: Pasadena, CA, USA, 2010.
12. Huang, S.; Xu, Z.; Tao, D.; Zhang, Y. Part-stacked cnn for fine-grained visual categorization. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1173–1182.
13. Lin, D.; Shen, X.; Lu, C.; Jia, J. Deep lac: Deep localization, alignment and classification for fine-grained recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1666–1674.
14. Zhang, H.; Xu, T.; Elhoseiny, M.; Huang, X.; Zhang, S.; Elgammal, A.; Metaxas, D. SPDA-CNN: Unifying semantic part detection and abstraction for fine-grained recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1143–1152.
15. Guan, X.; Wang, G.; Xu, X.; Bin, Y. Learning Hierarchal Channel Attention for Fine-grained Visual Classification. In Proceedings of the 29th ACM International Conference on Multimedia, New York, NY, USA, 10–16 October 2021; pp. 5011–5019.
16. Deng, H.; Sun, X.; Liu, M.; Ye, C.; Zhou, X. Infrared small-target detection using multiscale gray difference weighted image entropy. *IEEE Trans. Aerosp. Electron. Syst.* **2016**, *52*, 60–72. [CrossRef]
17. Han, J.; Liang, K.; Zhou, B.; Zhu, X.; Zhao, J.; Zhao, L. Infrared small target detection utilizing the multiscale relative local contrast measure. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 612–616. [CrossRef]
18. Huang, H.; Sun, D.; Wang, R.; Zhu, C.; Liu, B. Ship target detection based on improved YOLO network. *Math. Probl. Eng.* **2020**, *2020*, 6402149. [CrossRef]
19. Wang, G.; Zhang, T.; Wei, L.; Sang, N. Efficient method for multiscale small target detection from a natural scene. *Opt. Eng.* **1996**, *35*, 761–768. [CrossRef]
20. Wu, W.; Liu, H.; Li, L.; Long, Y.; Wang, X.; Wang, Z.; Li, J.; Chang, Y. Application of local fully Convolutional Neural Network combined with YOLO v5 algorithm in small target detection of remote sensing image. *PLoS ONE* **2021**, *16*, e0259283. [CrossRef] [PubMed]
21. Luong, M.-T.; Pham, H.; Manning, C.D. Effective approaches to attention-based neural machine translation. In Proceedings of the Conference on Empirical Methods in Natural Languag-e Processing, Lisbon, Portugal, 17–21 September 2015.
22. Cohn, T.; Hoang, C.D.V.; Vymolova, E.; Yao, K.; Dyer, C.; Haffari, G. Incorporating structural alignment biases into an attentional neural translation model. In Proceedings of the NAACL-HLT 2016: The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016.
23. Tu, Z.; Lu, Z.; Liu, Y.; Liu, X.; Li, H. Modeling coverage for neural machine translation. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016.
24. Guo, C.; Fan, B.; Zhang, Q.; Xiang, S.; Pan, C. Augfpn: Improving multi-scale feature learning for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 12595–12604.
25. Xue, H.; Liu, C.; Wan, F.; Jiao, J.; Ji, X.; Ye, Q. Danet: Divergent activation for weakly supervised object localization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 6589–6598. [CrossRef]
26. Baheti, B.; Innani, S.; Gajre, S.; Talbar, S. Semantic scene segmentation in unstructured environment with modified DeepLabV3+. *Pattern Recognit. Lett.* **2020**, *138*, 223–229. [CrossRef]

27. Ding, X.; Guo, Y.; Ding, G.; Han, J. Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 1911–1920. [CrossRef]
28. Yin, M.; Yao, Z.; Cao, Y.; Li, X.; Zhang, Z.; Lin, S.; Hu, H. Disentangled Non-local Neural Networks. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 191–207.
29. Silva, S.M.; Jung, C.R. License Plate Detection and Recognition in Unconstrained Scenarios. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 80–596. Available online: https://linkspringer.53yu.com/chapter/10.1007/978-3-030-01258-8_36 (accessed on 3 July 2022).
30. Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; Sun, J. Unified perceptual parsing for scene understanding. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 418–434. Available online: https://arxiv.53yu.com/abs/1807.10221 (accessed on 3 July 2022).
31. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**. [CrossRef]
32. Ultralytics. Available online: https://github.com/ultralytics/yolov5 (accessed on 3 July 2022).
33. Sun, P.; Zhang, R.; Jiang, Y.; Kong, T.; Xu, C.; Zhan, W.; Tomizuka, M.; Li, L.; Yuan, Z.; Wang, C.; et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 14454–14463. Available online: https://arxiv.org/abs/2011.12450 (accessed on 3 July 2022).
34. Chen, Y.; Zhang, Z.; Cao, Y.; Wang, L.; Lin, S.; Hu, H. Reppoints v2: Verification meets regression for object detection. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 5621–5631. Available online: https://arxiv.org/abs/2007.08508 (accessed on 3 July 2022).
35. Beal, J.; Kim, E.; Tzeng, E.; Park, D.H.; Zhai, A.; Kislyuk, D. Toward transformer-based object detection. *arXiv* **2020**. [CrossRef]
36. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.S. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6881–6890. Available online: https://arxiv.org/abs/2012.15840 (accessed on 3 July 2022).
37. Fang, Y.; Liao, B.; Wang, X.; Fang, J.; Qi, J.; Wu, R.; Niu, J.; Liu, W. You Only Look at One Sequence: Rethinking Transformer in Vision through Object Detection. *arXiv* **2021**, arXiv:2106.00666.
38. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 213–229. Available online: https://arxiv.org/abs/2005.12872 (accessed on 3 July 2022).