forests

MDPI

Article

AMDNet: A Modern UAV RGB Remote-Sensing Tree Species Image Segmentation Model Based on Dual-Attention Residual and Structure Re-Parameterization

Haozhe Huang [1,†], Feiyi Li [1,†], Pengcheng Fan [2], Mingwei Chen [1], Xiao Yang [1], Ming Lu [1], Xiling Sheng [1,3], Haibo Pu [1,3,*] and Peng Zhu [2,*]

[1] College of Information Engineering, Sichuan Agricultural University, Ya'an 625000, China
[2] College of Forestry, Sichuan Agricultural University, Chengdu 611134, China
[3] Sichuan Key Laboratory of Agricultural Information Engineering, Ya'an 625000, China
* Correspondence: puhb@sicau.edu.cn (H.P.); zhup@sicau.edu.cn (P.Z.)
† These authors contributed equally to this work.

Abstract: Tree species classification is an important and challenging task in image recognition and the management of forest resources. Moreover, the task of tree species classification based on remote sensing images can significantly improve the efficiency of the tree species survey and save costs. In recent years, many large models have achieved high accuracy in the task of tree species classification in an airborne remote-sensing manner, but due to their fixed geometric structure, traditional convolutional neural networks are inherently limited to the local receptive field and can only provide segmental context information. The limitation of insufficient context information greatly affects the segmentation accuracy. In this paper, a dual-attention residual network (AMDNet) and a re-parameterized model approach are proposed to capture the global context information, fuse the weight, reduce the model volume, and maintain the computational efficiency. Firstly, we propose MobileNetV2 as the backbone network for feature extraction to further improve the feature identification by modeling semantic dependencies in the spatial dimension and channel dimension and adding the output of the two attention modules. Then, the attention perception features are generated by stacking the attention modules, and the in-depth residual attention network is trained using attention residual learning, through which more accurate segmentation results can be obtained. Secondly, we adopt the approach of structure re-parameterization, use a multi-branch topology for training, carry out weighted averaging on multiple trained models, and fuse multiple branch modules into a completely equivalent module in inference. The proposed approach results in a reduction in the number of parameters and an accelerated inference speed while also achieving improved classification accuracy. In addition, the model training strategy is optimized based on Transformer to enhance the accuracy of segmentation. The model was used to conduct classification experiments on aerial orthophotos of Hongya Forest Farm in Sichuan, China, and the MIOU of tree species recognition using the test equipment reached 93.8%. Compared with current models such as UNet, our model exhibits a better performance in terms of both speed and accuracy, in addition to its enhanced deployment capacity, and its speed advantage is more conducive to real-time segmentation, thereby representing a novel approach for the classification of tree species in remote sensing imagery with significant potential for practical applications.

Keywords: tree species classification; airborne remote sensing; dual-attention residual network; re-parameterized model approach; training strategy

check for updates

Citation: Huang, H.; Li, F.; Fan, P.; Chen, M.; Yang, X.; Lu, M.; Sheng, X.; Pu, H.; Zhu, P. AMDNet: A Modern UAV RGB Remote-Sensing Tree Species Image Segmentation Model Based on Dual-Attention Residual and Structure Re-Parameterization. *Forests* **2023**, *14*, 549. https://doi.org/10.3390/f14030549

Academic Editor: Giorgos Mallinis

Received: 18 January 2023
Revised: 24 February 2023
Accepted: 27 February 2023
Published: 10 March 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

It is of great significance for sustainable forest management and ecological environment protection to obtain information on tree species and their spatial distribution [1]. Concurrently, accurate tree species identification technology can play a guiding role in

sustainable forestry management, biodiversity research and ecological environment protection [2,3]. Today, remote-sensing technology is widely used in various fields, and forestry remote-sensing technology, as an indispensable part of it, can not only facilitate the data acquisition of forestry resource management but also further reveal the ecological image of forestry management [4–6]. With the advancement of new high-resolution remote-sensing equipment and technologies such as hyper-spectral radar, thermal infrared radar, LIDAR, aperture radar, etc., remote sensing has entered a stage of rapid development in the quantitative and qualitative mapping of forestry biological ecological attributes [7,8]. Moreover, the emergence of active remote sensing at the level of a fine spectral resolution, coupled with the refinement of tree species identification technology, has facilitated the identification of forest attributes on a finer scale. In contrast to traditional remote-sensing monitoring technology, the utilization of unmanned aerial vehicles (UAVs) for remote sensing purposes offers a host of advantages, such as increased convenience, reduced weight, enhanced portability, a high resolution of the remote sensing images, and cost-effectiveness, as well as greater practicality. Many scholars have also combined this technology with multispectral and hyperspectral remote-sensing techniques. Cheng et al. [9] used UAV multispectral imagery and radiative transfer computing and image simulation of large-scale apple orchard 3D scenes, estimating the canopy-scale chlorophyll content in apple orchards. Guo and Chen et al. [10] used machine learning algorithms to estimate SPAD values of maize by collecting multispectral images of maize. Everaerts [11] provided an overview of the development of UAV remote-sensing technology in various projects and predicted that UAVs will rapidly become the preferred platform for the development of remote-sensing instruments and applications. In the context of forestry remote sensing, this technology holds the potential to capture images of a higher resolution with richer waveband information [12,13]. Due to the small and convenient attributes of UAVs, UAV remote-sensing technology has become a tool that can be used to obtain key information in the field of forestry.

In recent years, with the development of computer technology, deep learning has become increasingly pervasive in various fields, such as medicine and biology [14–17]. Due to the fact that machine learning methods such as support vector machine [18–20] and random forest [21,22] rely excessively on parameter selection and kernel function in forest species classification, the generalization capability of mainstream models is generally poor. Therefore, scholars have increasingly attempted to use deep learning technology to study tree species classification. Ferreira et al. [23] used DeepLabV3+ to monitor the number of individuals and classify the tree species of palm trees in the Amazon basin, which achieved good results and could replace manual tree surveys. Cao and Zhang [24] combined the unique residual unit of ResNet [25] with the semantic segmentation network, U-Net, to solve the problem of gradient degradation of neural networks and improve the accuracy of tree species classification while extracting multi-scale spatial features. Qi et al. [26] combined the class-balanced jigsaw resampling method with the U-NET network, aiming to solve the problem of the small number of samples in deep-learning-based tree species classification. Zhang et al. [15] applied the semantic segmentation model U2-NET to extract olive crowns, which broadened the range of applications of deep learning technology in forestry engineering, such as tree species classification. However, the basic backbone networks commonly used to extract global semantic details cannot efficiently integrate the local features with the global context information based on multiple data sources. Among the various techniques that have been proposed to overcome this problem, attention mechanics are one of the most effective ways to solve this problem. Therefore, this paper applies the attention mechanism to forestry information processing. Previously, Fu et al. [27] designed a dual-attention module, which effectively integrated the location features with the channel features and produced a marked enhancement in performance. Moreover, Wang et al. [28] first introduced residual branch into the attention mechanism and applied it to image classification, achieving great success in solving this problem. Meanwhile, with the introduction of Vision Transformers (VIT) [29] in 2020, this technology quickly

replaced ConvNets as the most advanced visual model. In the past two years, most articles produced in the CV field have been based on Transformer, and a major focus of VIT is scaling operations. Transformer can significantly outperform standard ResNet on larger models and datasets, but this application is not suitable for tree classification with small sample sizes and an uneven data distribution within classes. Although Liu et al. [30] broke trend of dominance of the Transformer by modernizing ResNet in accordance with Transformer, many designs in the experiment were still detrimental to the speed of inference, which may indicate unsuitability for forestry engineering applications. Therefore, the challenge of modernizing convolutional neural networks in the context of forestry engineering involves the need to address the issue of model size reduction while effectively integrating the local features and global context information of the dataset.

The inspiration for research often stems from real life. In real life, due to the obstacle of information processing, people tend to selectively pay attention to part of the information while ignoring other visible information. This mechanism is often referred to as the attention mechanism, which can also be found in deep learning. The attention mechanism can enhance different manifestations of a certain part of the object to augment the importance of that part of the information. In order to accomplish the task of tree species recognition efficiently, it is essential to differentiate between certain categories that may present confusion. Firstly, it is necessary to screen and extract the vast information contained in the hyperspectral data, and the manner in which sufficient and effective features are extracted directly affects the quality of the classification results. In fact, a large proportion of the information concerning various tree species growing in the same geographical location is highly coincident; thus, it is highly necessary to enhance the recognition ability of feature identification. However, to our best knowledge, there have been limited applications of the attention mechanism to forestry engineering and even fewer to deep learning semantic segmentation. To address these problems, we propose a new model: the dual-attention residual network (AMDNet). Specifically, after comprehensively considering the model volume and subsequent deployment problems, we took the lightweight model MobilNetV2 [31] as the backbone and DeepLabV3+ [32] as the segmentor. Secondly, two modules were attached to the traditional extended network to simulate the semantic dependencies of spatial and channel dimensions, respectively. The spatial attention module selectively aggregates the features of each location by summing the features of all the locations, while the channel attention module selectively emphasizes the existence of interdependent channel mappings by integrating the related features among all the channel mappings. Finally, the outputs of the two modules are integrated to further enhance the feature representation ability of the model. Inspired by the deep neural network ResNet, we constructed a dual attention residual network by stacking multiple attention modules that map semantic dependencies. Each attention module can be divided into two parts: the mask branch and trunk branch. The trunk branch is used for feature processing, and the mask branch is used to learn the attention mechanism. Through this mechanism, the two information sources are captured and queried in context, allowing the two branches to assist each other. Since the convolutional neural network inevitably loses inference speed when modernized, we propose the concept of "structural re-parameterization" [33] to optimize the model. This involves using a multi-branch model during the training, which is transformed into a VGG-like, straight-through, one-way model during the inference. In this case, this method not only ensures that enough feature information is learned during the training but also fuses the weights during the model's deployment, reducing the model volume and improving the inference speed.

Finally, to solve the classification problem of remote sensing tree species information, we tested the effect of AMDNet and conducted a comparative analysis of its speed and accuracy with other mainstream deep learning models to verify the comprehensive performance of the model. In summary, our main contributions are as follows:

- A new solution is provided to obtain tree species distribution information on the forest and to improve the tree species identification technology in forestry resource surveys.

- A dual-attention residual module is proposed, which adaptively integrates local features and global dependencies by simulating the semantic interdependence of spatial and channel dimensions, respectively. At the same time, we integrate residual learning, the stacked network structure, and the attention mechanism to enable the efficient fusion of local features and global context information.
- The structural re-parameterization is applied to solve the problem of inference speed loss of the convolutional neural network structure when it is modernized in accordance with Transformer, which reduces the model size without affecting the learning ability of the model, which can still learn sufficient information based on remote sensing tree images.

2. Materials and Methods

2.1. Study Area

Situated in the Hongya Forestry Farm, Hongya County, Meishan City, Sichuan Province of Southwest China, the study area lies on the southwest edge of Sichuan Basin, specifically the Qionglai Mountain branch. It is located on the northeast slope of Daxiangling in the E'mei Mountain System and is bordered by the south bank of the Qingyi River. The geographical location of the area is defined by the coordinates 29°24′ N–30°00′ N and 102°49′ E–103°32′ E, with an altitude ranging from 900 to 3522 m. The annual temperature of the region is measured to be 8 °C, while the annual precipitation is recorded to be 2400 mm. The Hongya Forestry Farm has a total area of 659 km^2, of which 526 km^2 is natural forests, constituting 80.9% of the forest land, and 123 km^2 is man-made forests, comprising the remaining 19.1% of the forest land. The region is characterized by deep and fertile soil with a loose structure, while the forest vegetation shows an evident vertical zonal distribution. The low mountainous area is dominated by pure man-made coniferous forest, with other main species such as *Cryptomeria japonica* (L.f.) D.Don and *Cunninghamia lanceolate* (Lamb.) Hook. Most of the central mountainous areas are natural coniferous and broad-leaved mixed forests, and the main tree species are *Abies fabri* (Mast.) Craib, *Castanopsis platyacantha* Rehd.et Wils., *Lithocarpus hancei* (Bentham) Rehd., etc. The alpine areas are dominated by scrublands, mainly *Rhododendron simsii* Planch., *Fargesia spathacea* Franch., and so on. Therefore, the species targeted for classification in this study are categorized as follows: *Cunninghamia lanceolata* (Lamb.) Hook., *Cryptomeria japonica* (L.f.) D.Don, *Metasequoia glyptostroboides* Hu & W.C.Cheng, *Abies fabri* (Mast.) Craib, and a small number of *Phoebe Nees*. Areas occupied by buildings and roads are consistently classified as non-forest land. The specific classification categories are shown in Figure 1.

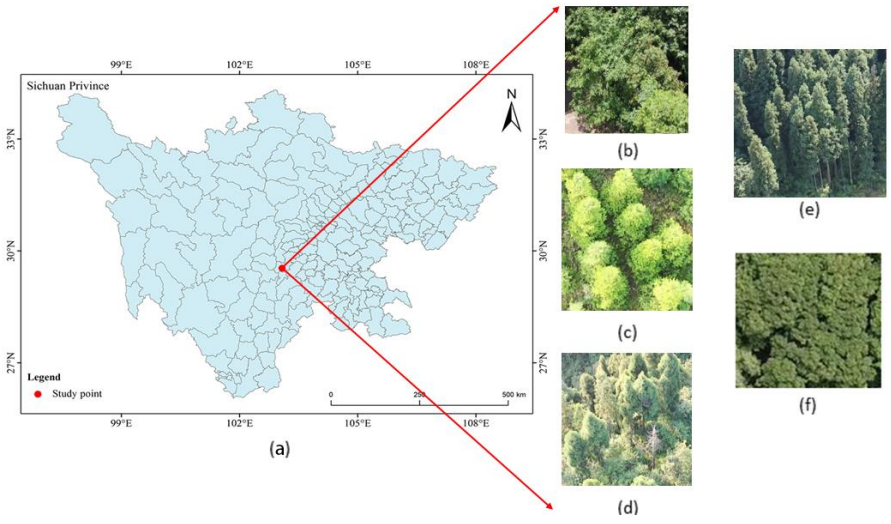


Figure 1. Geographical location of the study area: (**a**) Administrative Region Map of Sichuan Province, China; (**b**–**f**) main canopy images taken by UAV from 200 m in the study area, (**b**): *Phoebe Nees*, (**c**): *Cunninghamia lanceolata*, (**d**): *Metasequoia glyptostroboides*, (**e**): *Cryptomeria japonica*, (**f**): *Abies fabri*.

2.2. Data and Pre-Processing

UAV remote sensing images were obtained in the field on 4 July 2022 at local time. Drones were mainly used for the data collection by hovering. The equipment used was DJI Phantom 4, which is equipped with 6 vision sensors. During the data collection process, the relative distance between the UAV and the ground was always maintained at 200 m, and the resolution of all the cameras was 2 million pixels, with clear and stable imaging. The spatial resolution of the images was 1 m, including three red, green, and blue bands. In the data collection phase, we first conducted field surveys in the collection area to analyze the distribution and features of the tree species and to classify the different tree species into areas. We then performed preliminary data acquisition under clear and cloudless weather conditions, capturing a total of 1247 initial images of all categories included in the classification system. After data collection, we performed radiometric calibration on the remote sensing images using the pseudo-standard feature radiometric correction method [34] via Envi (version 5.31), an image processing software. This method achieves radiometric calibration by establishing a linear relationship between the measured ground reflectance and the actual ground reflectance coefficient. Ten aerial experimental standard reflectance reference plates were positioned uniformly around the data acquisition point. The DN values corresponding to the standard reference plates were extracted from the UAV images. The DN values of the reference plates for each band and the known reflectance values of the calibration reference plates were used to establish an equation so as to convert the DN values of the UAV images into the reflectance after radiometric calibration. The specific radiation correction obeys the following equation:

$$R_T = \frac{DN_T}{DN_R} R'$$

Here, R_T is the reflectance of the target feature, DN_T is the DN average of the target feature, DN_R is the DN average of the standard reflectance reference plate for aerial photography experiments, and R' is the reflectance value of the reference plate.

In order to enhance the model's robustness through an increase in the number of training samples, we performed data augmentation manipulations such as rotations, shifts, and flips, forming a total of 2500 images that were sent to the model for training. Meanwhile, to facilitate the subsequent AMDNet extraction of image spectral features, we used Lableme (version 4.60) to construct the corresponding tree species sample set by manual visual interpretation under the guidance of professionals. Finally, we randomly selected 4/5 samples for the model training, using the remaining samples for validation. The specific numbers of samples corresponding to tree species are shown in Table 1.

Table 1. Classification categories of the study area and the number of training and verification samples for each category.

Class	Species Name	Training Samples	Validation Sample
1	*Cunninghamia lanceolata*	462	92
2	*Abies fabri*	321	64
3	*Metasequoia glyptostroboides*	468	93
4	*Phoebe Nees*	222	44
5	*Cryptomeria japonica*	425	84
6	Other conifers	168	34
7	Other forest land	266	54
8	Construction land	62	13
9	Roads	106	22
Total	Total	2500	500

2.3. Methods

In this part of the paper, a new solution is presented to address the segmentation problem in remote sensing tree species images. Our proposed model, AMDNet, comprises multiple sensitive and targeted modules that are optimized to extract tree species information with greater accuracy. Considering the problems of the small sample size, high information coincidence, image background aggregation, and complex scenes in the classification of remote sensing tree species, we developed a series of customized modules and model strategies tailored to the distinct characteristics of tree species information. They are: (1) the attention residual module; (2) structural re-parameterization; and (3) the improvement of the training strategy.

2.3.1. Dual-Residual Attention Module

The final process of the model's implementation aims to accurately predict the tree species, which entails the resolution of a pixel-level prediction problem. Addressing pixel-level prediction problems requires the integration of both local features and global context information. The local features and global dependence of tree species information inherently show a strong correlation, which allows the attention module to play a significant role in enhancing the model.

To enable the adaptive integration of both local features and global dependence, a dual-attention module is incorporated at the head of the model; that is, two modules are attached to the traditional expansion network to simulate the semantic dependencies of spatial and channel dimensions, respectively.

As shown in the position attention module in Figure 2, the position attention module sends the feature map A ($C \times H \times W$) to the three convolutional layers, respectively, to obtain three feature maps and then reshapes the three feature maps to obtain $C \times N$ ($N = H \times W$). Next, the transpositions ($N \times C$) of the first reshaped feature ($C \times H \times W$) and the second reshaped feature ($C \times N$) are multiplied to yield a spatial attention map ($H \times W \times H \times W$), which is computed using Softmax. Following this step, the reshaped third feature ($C \times H \times W$) is multiplied with the transposition of the map ($H \times W \times H \times W$) and then further multiplied by the scaling coefficient α (initialized to 0 and gradually learned to obtain larger weights). After reshaping it to its original shape, it is then added to the original map to produce the final output E ($C \times H \times W$).

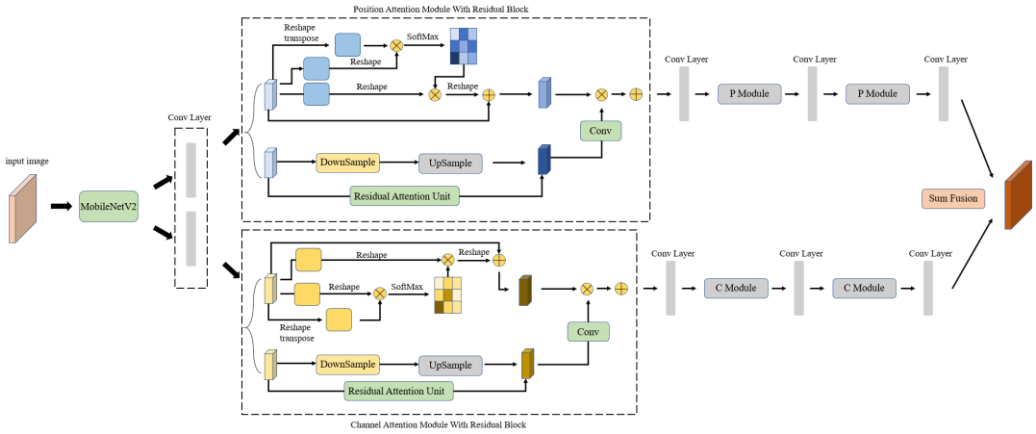


Figure 2. A structure diagram of the dual-attention residual module.

Similarly, as illustrated in Figure 2, the channel attention module reshapes ($C \times N$) and reshapes and transposes ($N \times C$) on the original image ($C \times H \times W$), respectively, and then multiplies the two obtained feature maps using Softmax to obtain the channel attention map ($C \times C$). Next, the transposition of the map ($C \times C$) is multiplied with the original map ($C \times N$) and is then further multiplied with the scaling coefficient β (initialized to 0 and

gradually learned to obtain larger weights). The resulting feature map is then reshaped back into the original shape and added to the initial image to generate the output E (C × H × W).

The dual-attention residual network is essentially a stack of multiple dual attention modules. Each attention module is divided into two parts: the residual branch and the trunk branch. The trunk branch can be any current convolutional neural network model. In our case, MobileNetV2 is selected for feature processing. The residual branch uses a combination of bottom-up and top-down attention to learn an attention feature map that is consistent with the dimension of the output of the trunk feature processing. Then, the feature maps of the two branches are combined using the dot product operation to obtain the final output feature map.

If the output feature map of the trunk branch is Qi, F(x), and the output feature map of the mask branch is Pi, F(x), the final output feature of the attention module is:

$$\mathrm{Hi, F(x)} = \mathrm{Qi, F(x)} \times \mathrm{Pi, F(x)}$$

In the attention module, the attention mask functions as a forward feature selector and as a gradient update filter for backpropagation.

$$\frac{\partial \mathrm{P}(x, \theta)\mathrm{Q}(x, \Phi)}{\partial \Phi} = \mathrm{P}(x, \theta)\frac{\partial \mathrm{Q}(x_2 \Phi)}{\partial \Phi}$$

In practice, since there are background occlusions and complex scenes in the training images for remote sensing tree species classification, the fusion of multiple attentions is required to process the feature information. If the method of stacking attention modules is not used, a larger number of channels will be needed to cover the combined attention of different factors. Additionally, due to the limited capacity of a single attention module to modify the feature, the model's fault tolerance rate is decreased, and increasing the number of training iterations may not improve the overall robustness of the model.

While attention modules play a significant role in target segmentation, their incorporation into a model without consideration will lead to the degradation of the model's performance. This can be attributed to two main reasons:

1. In order to generate a feature map with normalized weights, the addition of the Sigmoid activation function to the mask branch is necessary. However, when the output is normalized between 0 and 1 before conducting a dot product with the main branch, the output response of the feature map will be weakened. Multi-layer stacking of this structure will lead to a decrease in the values at each point of the final output feature map.
2. The feature map output from the masked branch may adversely affect the benefits provided by the trunk branch. For example, replacing the shortcut mechanism in the residual connection with the masked branch may cause inadequate gradient transmission in the deep network.

To solve the problems mentioned above, we use a method similar to residual learning to conduct an element-wise addition of the obtained attention feature map to the trunk feature map, so that the output is expressed as follows:

$$\mathrm{H_{i,c}}(x) = (1 + \mathrm{P_{i,c}}(x)) \times \mathrm{Y_{i,c}}(x)$$

We refer to this learning mechanism as dual-attention residual learning. By capturing the global feature dependencies in spatial and channel dimensions, this module can build a rich context dependencies model on local features so as to significantly improve the segmentation results. In addition, it has been observed that employing a decomposition structure to increase the size of the convolution kernel or introducing an effective coding layer at the top of the network can capture richer global information. The addition of more attention modules contributes to a linear improvement of the classification performance of the network. Furthermore, additional attention models can be extracted from feature maps

at different depths. With this strategy, the network can easily be extended to hundreds of layers, due to the residual structure, while still maintaining its robustness to noisy labels.

2.3.2. Structural Re-Parameterization

The backbone is the baseline network of the entire model during training and inference, which largely determines the upper limit of the model. In the case of AMDNet, MobileNetV2 was selected due to its relatively balance between speed and accuracy, effected through the inverse residual structure. However, the overall training and inference speed of the model still falls short of expectations. To address this problem, we propose modifying MobilNetV2 using structural re-parameterization.

The specific modifications are shown in Figure 3. Firstly, we introduce an identity and residual branch into the original MobileNetV2 during the training process to re-parameterize it into a one-way structure, which is equivalent to adding the advantages of ResNet. Moreover, we modify the location of the branches by directly connecting each layer instead of using cross-layer connections. It is also demonstrated that both residual branching and conv_1 × 1 can improve the network's performance compared to the native model. Finally, in the model inference stage, the residual structure used in training is transformed into a 3 × 3 convolution layer through the Op fusion strategy to facilitate subsequent model deployment and acceleration.

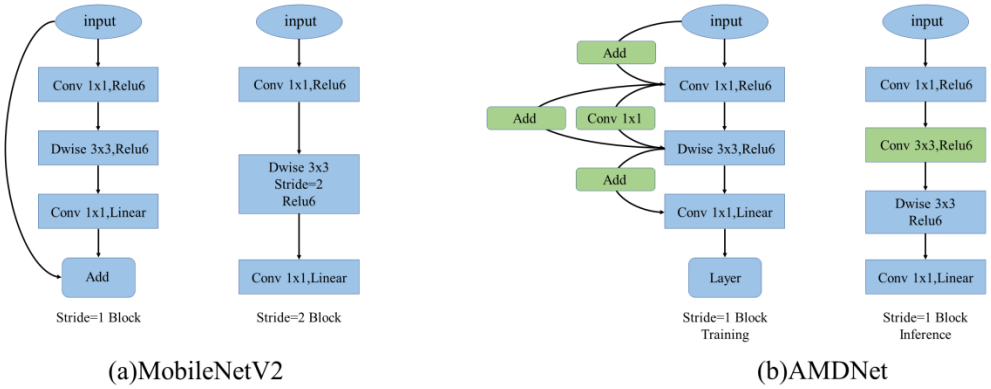


Figure 3. The diagram of MobilenetV2's improvement.

The re-parameterization process in the model inference stage is essentially a process of OP fusion and OP replacement. Firstly, the Conv3 × 3 + BN layer and Conv1 × 1 + BN layer are fused, respectively, and the formula can be expressed as follows:

$$D'_{i,:,:,:} = \frac{\gamma_i}{\sigma_i} D_{i,:,:,:}, \; c'_i = -\frac{\mu_i \gamma_i}{\sigma_i} + \beta_i$$

The D_i refers to the parameters of the convolution layer before the conversion, μ_i refers to the mean value of the BN layer, and σ_i refers to the variance of the BN layer. The scale factor and the offset factor of the BN layer are represented by γ_i and β_i, respectively, and the weight and bias of the convolution layer after fusion are represented by D′ and c′, respectively.

After this step, the fused convolutional layer is converted to Conv3 × 3. For the Conv1 × 1 branch, the conversion process aims to replace the 1 × 1 convolution kernel with the 3 × 3 convolution kernel, which involves the transfer of the value in the 1 × 1 convolution kernel to the center point of the 3 × 3 convolution kernel. For the identity branch, it is necessary to set a 3 × 3 convolution kernel and assign a weight value of 1 to each of the 9 positions so that it will keep the original value after the multiplication with the input feature map.

Finally, the weight W and bias B of all the branches are stacked together to obtain a Conv3 × 3 network layer after the fusion.

Since the introduction of the residual structure with multiple branches in network improvement has the effect of adding multiple gradient flow paths to the network, training this network is similar to training multiple networks at the same time, which are later integrated into a single network. This can be viewed as an example of model integration, which also improves the model robustness to a certain extent. Moreover, the addition of the 1×1 convolution branch and identity mapping branch can also enhance the benefits of multi-branch model training. The simpler the network structure in the model inference phase is, the more effective the model acceleration will be. Therefore, in the inference stage, the model is transformed into a single-branch structure to improve the memory utilization of the device and thus improve the inference speed of the model. In this way, the benefit of multi-branch model training (a high performance) and the advantages of single-path model inference (a fast speed and saving memory) can be leveraged simultaneously.

2.3.3. Model "Modernize"

In 2022, Liu et al. [31] proposed the idea of modernizing the CNN-style model; that is, aligning the ConvNets-style network model with the Transformer-style model. The aim of this proposal was to explore the design space and test the limits of ConvNets and break the trend of the monopoly of the Transformer. It was also suggested that the performance gap between traditional ConvNets and the Vision Transformer may be largely attributable to the training strategy level, but the authors did not elaborate much on what design can be used to optimize ConvNets. However, it is clearly stated in the Vision Transformer paper that the Transformer structure lacks some of the innate inductive bias of CNN, such as translation without deformation and the inclusion of local relations; thus, its performance on small- and medium-sized datasets is not particularly good. In the task of remote sensing tree species recognition, however, due to the limitations of the datasets and the uneven distribution of data within the class, a model structure that is suitable for large datasets, such as Transformer, may not be suitable for practical applications. Nevertheless, the training techniques corresponding to the Transformer may still have room for improvement.

Therefore, aiming to address the problems of the small sample size, uneven distribution of feature information, and large amount of repetitive information in tree species classification, we proposed a customized and optimized training strategy. We first added HSV_AUG, Cutmix, RandomScale, and other common training strategies to the original training strategy while modifying the training method of the traditional CNN-style model. Prior to Transformer, the main training method of CNN was based on SGD and learning rate decay. In the experiment, we attempted to take SGD and learning rate decay as the training strategies for Transformer, and the final results were relatively poor, which also explained the difference in training skills between CNN and Transformer to a certain extent. To modernize the CNN model, it is necessary to analyze whether Transformer's training techniques are feasible for CNN, namely, the LR Decay of Warmup and AdamW-style training (the improvement is not simply focused on the training methods but, due to the space limitations on this paper, emphasis is placed on the feasibility of applying Warmup + AdamW to remote sensing tree species datasets).

The inclusion of Warmup in the training is necessary for several reasons. Firstly, the learning rate, which determines the step size, plays a crucial role in achieving the optimal performance. Incorrect learning rate settings can lead to issues such as divergence or slow convergence. In neural networks, if the learning rate is set to be too low, it will fall into the local optimal solution. To address these concerns, a conventional approach is to begin with a higher learning rate during the initial phase of training, which is then reduced in a linear manner. Specific strategies are shown in Figure 4 below:



Figure 4. Learning rate decay and Warmup learning rate decay curve.

Warmup increases the learning rate and then decreases it linearly, as shown in the figure above. The use of Warmup is very important in the training of Vision Transformer. If Warmup is used, with AdamW as the optimizer, the training can converge normally and generate the desired results. If not, it will be difficult to converge the training process and adjust the setting of the learning rate. Similarly, here, we conducted a comparative experiment between the experimental group utilizing Warmup + AdamW and CNN trained using SGD, and the loss change during training is shown in Figure 5 below:

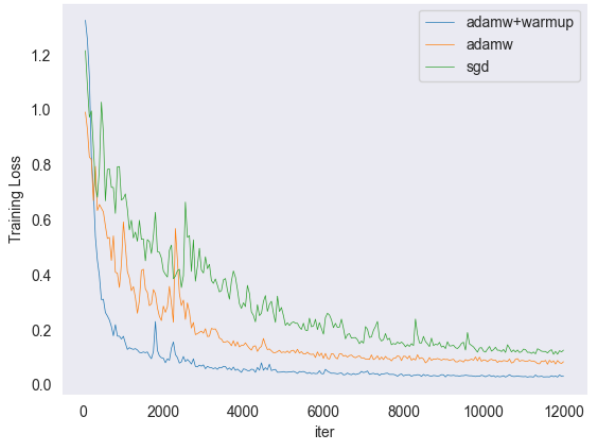


Figure 5. The loss changes during the training of SGD, AdamW, and AdamW + Warmup.

It can be seen that since AdamW contains both first-order momentum and second-order momentum, it has a clear advantage in terms of convergence speed compared with SGD. However, in the experiment, the final loss of the AdamW group was higher than that of the AdamW + Warmup group, and the final performance of the model was worse. This discrepancy can be attributed to AdamW's unstable initial learning rate. The Warmup strategy starts with a low learning rate for training epochs or iterations, followed by a linear or non-linear increase in the learning rate towards a preset value as the training process progresses. In addition, AdamW dynamically and adaptively adjusts the learning rate during the process. If the gradient is more likely to deviate from the truly optimal direction, the adaptive adjustment will mitigate the deviation, and if the deviation is in the wrong direction, a subsequent adjustment will be made to prevent the normal distribution from becoming distorted.

In the application of the model, such training configuration is not only helpful in order to mitigate the phenomenon of early over-fitting of the mini-batch in the initial stage and keep the distribution stable but also to maintain the stability of the model depth. Compared

with the SGD series, AdamW has a faster convergence speed, which can greatly reduce the training time, and has more practical significance for the collection and analysis of tree species information; thus, it has more advantages in forestry engineering applications. Furthermore, the inclusion of Warmup allows for a low learning rate within the first few epochs of the training, which facilitates the stabilization of the model. Once the model reaches a state of relative stability, the preset learning rate is selected for training, which increases the model convergence speed and improves the model effect.

Moreover, ablation experiments were conducted on the data enhancement technology and regularization scheme of Transformer, as follows:

First, the enhancement of the HSV color model is undertaken. The brightness and saturation features of each species in the proposed tree species dataset contain a large amount of feature information. In the HSV color model, colors are not represented by three channels in the RGB but according to their hue, saturation, and value. Based on this, the HSV model can improve the discriminability between different species, so that the model can more easily locate the distinguishing features in the effective parts so as to achieve the effect of improving the accuracy of tree species identification. The second is RandomSCALE, which essentially uses the specified scale factor to filter the image in order to construct the scale space so as to change the size of the image content or the degree of the blur. When applied to the task of this paper, this technique can enhance the randomness of the image and, in effect, expand the dataset size, ultimately achieving the result of improving the robustness of the model. Finally, the Cutmix is a technique of fusing mixup with cutout. The mixup method is likely to generate an image that is fuzzy and unnatural in the local area, which will confuse the model and cause the model to learn the noise information, ultimately diminishing the model's effect. In contrast, the part deleted by cutout is usually replaced with 0 or random noise, which leads to the waste of some information in the training process and results in a reduction in the amount of information obtained by the model and a decline in the model's ability. Therefore, in the experiment, we used Cutmix to enhance the tree species information. It filled the cutout part of a certain image with the cutout part of another image, which helped to retain the advantages of cutout, for example, by enabling the model to learn the characteristics of the target from the partial view and causing the model to focus on the less discriminative parts while being more efficient than cutout. The cutout part is filled with some parts from another image, so that the model can learn the features of two targets at the same time, avoiding the problem of confusing the model, as in the case of mixup, and enhancing the final effect of the model to a certain extent.

Our complete set of hyperparameters can be found in Table 2. In terms of the remote sensing tree species dataset produced by AMDNet, most of the training configurations resulted in significant and positive performance improvements for the whole model. The success of Vision Transformer is also related to its special training mode, which has potential implications for future research in forestry remote sensing.

Table 2. AMDNet training parameters.

Training Config	AMDNet
Batch size	8
Training epochs	120
Optimizer	AdamW
Optimizer momentum	0.9
Weight decay	1×10^{-4}
Base learning rate	3×10^{-4}
Warmup epochs	20
Learning rate schedule	Poly update
Warmup ratio	0.1
Warmup schedule	Linear

2.3.4. Implementation Details

Based on the dual-attention module and the structural re-parameterized feature extraction network mentioned above, we introduce AMDNet, a new model for tree species feature information processing that builds upon the DeepLabV3+ framework. The overall structure is shown in Figure 6 below. Next, the implementation process of the model is explained.

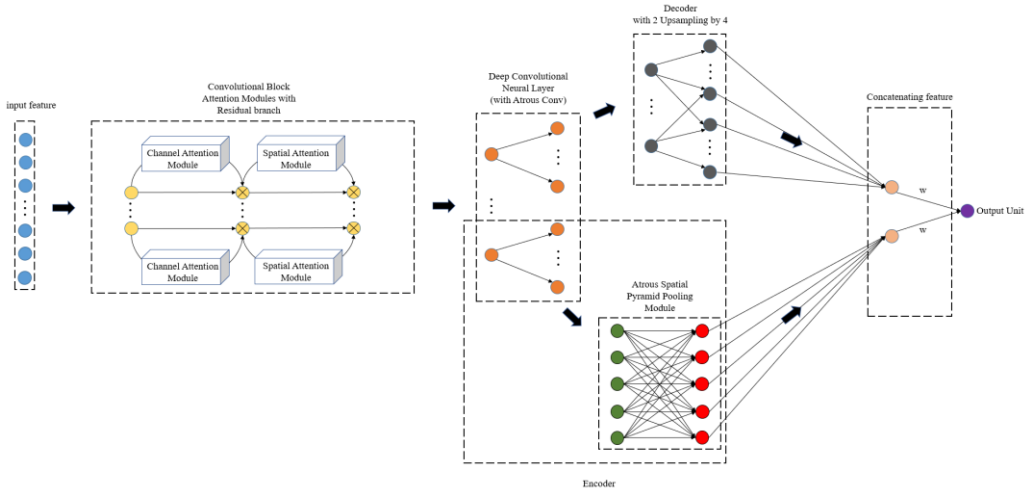


Figure 6. Model Diagram.

Firstly, the dual-residual-attention module is positioned at the head of the model to facilitate the identification and reinforcement of feature information. After the external data are fed into the model for training, the initial information is channeled, respectively, into two feature information enhancement paths, the position and channel. In each path, the feature information of each branch block is integrated with that of the preceding block to obtain the enhanced outputs of the position and channel, respectively. Upon merging the two outputs, they are fused with the initial up-sampling information. From a macro-perspective, this module uses two convolutional channels to process the initial feature information and weighs the unique information of different tree species in the image so that the segmentor will spend more "energy" on the unique information in the subsequent segmentation learning and improve the classification performance of the model.

Next, the output of the feature map from the attention module is transmitted to the segmentor section. Although it has entered the segmentation module, the feature extraction is still needed. Here, we opted for the lighter MobilenetV2 instead of the ResNet used by Deeplabv3+. Firstly, due to the original encoder–decoder structure, the attention module is added to the head. In this case, although the accuracy is significantly improved, it also induces a substantial decrease in speed. Considering the problems involved in the subsequent model deployment and practical applications, we chose to carry out lightweight processing on the model. Secondly, as the attention module of the head has strengthened the feature information, the feature extraction in the segmentor exerts a minimal influence on the final effect. Even in extreme cases (such as those involving a small number of species, i.e., less than 3 species), the straight-through network may function better as a feature extraction network. However, as the proposed dataset has 7 categories, the accuracy of the straight-through network is expected to lag behind to some degree. Consequently, taking into account the two reasons mentioned above, we finally chose MobilenetV2 as the feature extraction network of the segmentor.

After passing through the feature extraction network, the feature map is fed into an encoder–decoder structure. The spatial pyramid pooling module in the encoder structure outputs both high-level semantic features and low-level semantic features. After being up-sampled once, the high-level semantic features are then spliced with the low-level semantic features that were down-sampled to produce the final predicted output through up-sampling.

The elaboration outlined above basically covers the whole training stage of the model from the data input to the output of the prediction results. As mentioned in Section 2.3.2, the performance of our model in the ablation experiment did not meet expectations, especially at the level of the inference speed, which still lagged significantly behind that of the SegFormer and other models. Therefore, we suggest re-parameterizing the structure of the feature extraction network. In the training stage, the addition of the 1×1 convolution and the directly connected residual branches increases the flow of feature information in the model, which enhances the feature extraction ability of the model, to a certain extent, and improves the robustness of the model. Moreover, this helps to compensate for the slight drop in accuracy due to the abandonment of ResNet. In the inference stage, weight fusion is performed on the residual branches to achieve a single-branch network with which to extract feature information, which greatly reduces the time consumed by the segmentor in the inference stage. In this way, even with the inclusion of the attention mechanism in the model head, the speed of AMDNet is comparable to that of the SegFormer.

Next, we turn to the training strategy section. As mentioned in Section 2.3.3 above, while an initial attempt was made to take the SGD as the optimizer and the poly as the strategy to train the network in the beginning, this did not yield satisfactory results. The metrics, such as the MIOU, and the results cannot surpass those of the Transformer. In the subsequent ablation experiments, through the application of a series of training strategies and the modification, replacement, and debugging of the optimizer, a training strategy more suitable for the classification of remote sensing tree species, namely Adamw + Warmup, was finally obtained. Furthermore, some techniques were applied to the model (the detailed procedures and experimental data are explained in Section 3), which, compared with the original training strategy, showed a significant improvement. Ultimately, the proposed approach outperformed the Transformer when applied to the tree species dataset.

3. Results

In order to comprehensively evaluate the model for the task of tree species classification, we evaluated the effectiveness and significance of the key modules of the model and some training techniques based on the forestry dataset mentioned above. Our experiment consisted of three parts. In the first part, we explored the impacts of the improved training methods on the model accuracy through modernizing the model. Our analysis was not restricted to the VIT but, instead, aimed to identify the optimal training strategy applicable to the classification of remote sensing tree species. In the second part, we conducted an ablation experiment on the attention mechanism to assess the effectiveness of each component of the dual-attention residual network. Furthermore, here, we discuss the influence of structural re-parameterization on the model, including its influence on the model size and training speed. Meanwhile, we also compare the results with those of other commonly used segmentation models to demonstrate the superiority of the former model. Finally, we use our own dataset as a benchmark dataset to evaluate the consistent performance gained by using the dual-attention residual network for the classification of remote sensing woodland tree species. In the following sections, we first introduce our implementation details to better demonstrate our results.

3.1. Effects of Different Training Strategies on Tree Species Recognition

The success of Vision Transformer is attributed not only to its new module and network structure but also to the introduction of different training methods into the traditional convolutional neural network. Therefore, our research begins with our original model, AMDNet, and explores whether modern training techniques can result in optimization in the remote sensing identification of tree species. The specific ablation experiment results are shown in Table 3 below:

Table 3. The ablation experiment of the training strategy for MIOU.

AdamW	Warmup	HSV_AUG	Cutmix	RandomScale	MIOU
×	×	×	×	×	85.25
√	×	×	×	×	86.63
√	√	×	×	×	87.41.
√	√	√	×	×	87.66
√	√	×	√	×	87.2
√	√	×	×	0.5~1.5	87.69
√	√	√	√	×	87.74
√	√	√	√	05~1.5	87.02

The results indicate that the optimized training strategy led to an improvement in MIOU, increasing from 85.25 to 87.02 in comparison to the original training strategy. At the same time, our study also reveals that the training strategy of Transformer shows promise for application to small datasets. In the subsequent improvement, we use the improved training strategy to conduct experiments, and the accuracy is the average value obtained after three different random seed trainings.

3.2. Effect of the Improved Module on the Tree Species Identification Accuracy

3.2.1. Ablation Experiment of the Dual-Attention Residual Module

In this paper, we suggest that the dual-residual-attention network adaptively integrates local features and global dependencies to better fuse local features and global context information. In order to verify the performance of the attention module, we set up an ablation experiment, and the specific results are shown in Table 4 below.

Table 4. Study on the ablation of the tree species classification attention module. PAM is the spatial attention module, CAM is the channel attention module, and ARL is the attention residual module.

PAM	CAM	ARL	MIOU
×	×	×	87.02
√	×	×	89.25
×	√	×	88.69
√	√	×	89.91
√	×	√	91.03
×	√	√	90.2
√	√	√	92.32

As shown in Table 4, the dual-attention residual module can significantly improve the identification accuracy of neural networks, and using the channel or spatial attention module alone improves the performance, compared to the base model, by 1.67% and 2.23%, respectively. When the outputs of the two attention modules are fused, the MIOU is further improved to 89.91. In addition, when the residual learning is added to the dual-attention module, the improvement is 5.3% in comparison to the base model and 2.89% in comparison to the dual-attention module without the residual learning. This suggests that residual learning can significantly improve the ability of the attention module to capture global information. Moreover, when adding residual learning to a separate channel or spatial

attention module, the model still yields a considerable improvement, which demonstrates the universality of residual learning in the attention module.

Using the spatial attention module enhances the discriminative capacity of local features, thereby rendering certain details more apparent, such as the junctions of forest trees with roads and streams. Meanwhile, some incorrect classifications are corrected by the channel attention mechanism, which is helpful for capturing global context information and improving semantic dependency. The specific effects will be outlined later. With the introduction of residual learning, the model can improve the classification performance of the network in a linear manner through a greater number of attention modules. Moreover, feature maps based on different depths can also extract additional attention models. With this strategy, on account of the existence of the residual structure, the network could easily be extended to hundreds of layers and thus achieve classification accuracy for other large networks while significantly reducing the computational effort.

3.2.2. Influences of Structural Re-Parameterization on the Model Volume, Training Speed, and Accuracy

We first compared AMDNet with the basic model and took the remote sensing tree species dataset that we constructed as the standard dataset to compare the changes in the model volume, training speed, and accuracy after applying structural re-parameterization. We then compared AMDNet with classical and state-of-the-art models, including FCN [35], Unet [16], PSPNet [36], EncNet [37], and SegFormer, based on remote sensing tree species datasets.

As shown in Table 5, we demonstrated our AMDNet's importance through ablation experiments using structural re-parameterization. With the same model, the network using structural re-parameterization improved all the metrics, with not fewer parameters and faster inference but also a higher classification accuracy. It makes full use of the advantages of its high performance in multi-branch model training and leverages the advantages of its fast speed and capacity for memory saving in one-way inference. Meanwhile, as shown in Table 6 below, we evaluated the performance indices of AMDNet, FCN, UNet, and other common networks based on the same dataset. The performance indicators evaluated included the accuracy, model inference speed, number of model parameters, and FLOPS. Through observation, the following preliminary conclusions could be drawn: (1) Under the same test conditions, AMDNet, compared with FCN and PSPNet, exhibits a significant performance improvement across all the performance metrics evaluated. AMDNet not only has fewer parameters and a faster inference speed but also obtains a higher classification accuracy. (2) Compared with UNet and EncNet, the model performs equally well. (3) Compared with the lightweight model SegFormer, the model exhibits a higher precision rate that is increased by 1.52%, and its inference speed is basically the same, which fully demonstrates the superior performance of the model. The MIOU changes of each model are shown in Figure 7 below.

Table 5. Influences of structural re-parameterization on the MIOU, speed, params, and Theo FLOPS.

Phase	Re-Parameterization	MIOU	Speed (fps)	Params (M)	Theo FLOPs (G)
Train	×	92.32	1.28	43.58	176.65
Val	×		12.93		
Train	√	93.8	2.68	15.23	68.31
Val	√		26.04		

Table 6. Comparison of accuracy, speed, and parameters between AMDNet and other mainstream segmentation models.

Model	MIOU	Speed	Params(M)	Theo FLOPs(B)
AMDNet	93.8	26.04	18.54	89.26
FCN	81.61	8.67	49.49	197.98
PSPNet	82.16	8.72	48.96	178.73
UNet	91.75	7.41	29.06	203.43
EncNet	93.11	13.93	35.87	141.07
SegFormer	92.28	26.06	3.72	6.4

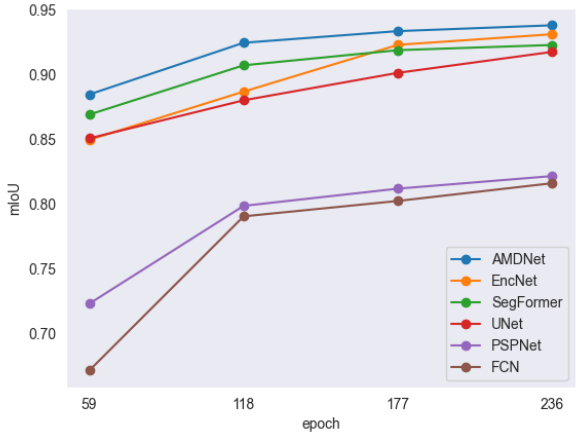


Figure 7. MIOU changes during each model training.

3.3. The Classification Results

Figure 8, below, shows the classification results of our model based on remote sensing tree species datasets. According to the comparison and analysis of the classification results and the original ground classification images, AMDNet has a good segmentation ability for each tree species. With the continuous refinement of this method, there is significant improvement in addressing the problems of the demarcation of roads and woodlands and small areas of intersecting tree species. The results of the classification of each species are shown in Table 7 below. The MIOU of our model finally reaches 93.8. Compared with the current commonly used semantic segmentation models, it achieves a good balance between accuracy, computational complexity, and model size which is more in line with the requirements for real-time segmentation and practical deployment.

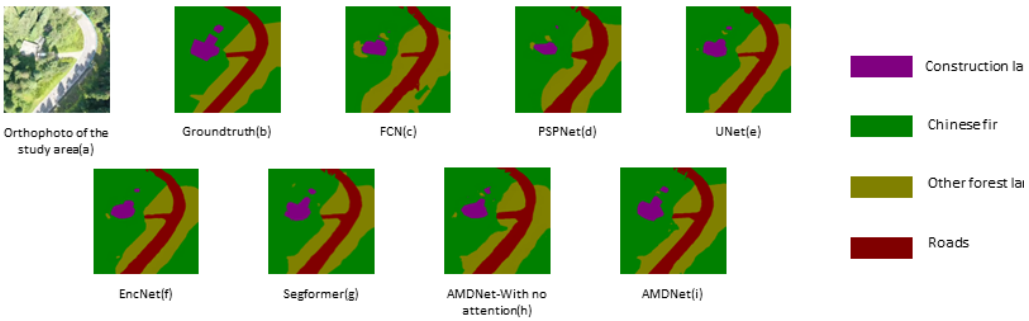


Figure 8. Classification results of tree species with different classification methods. (**a**) Orthophoto of the study area. (**b**) Groundtruth. (**c**) FCN classification results. (**d**) PSPNet classification results. (**e**) UNet classification results. (**f**) EncNet classification results. (**g**) SegFormer classification results; (**h**) Classification results for AMDNet with no attention. (**i**) AMDNet classification results.

Table 7. Effects of different tree types on different models.

Species Name	PSPNet	FCN	UNet	EncNet	SegFormer	AMDNet
Cunninghamia lanceolata	92.31	90.58	97.02	97.94	96.34	98.19
Abies fabri	80.26	76.34	87.42	88.55	90.87	92.74
Metasequoia glyptostroboides	86.28	89.82	95.58	96.66	95.64	96.73
Phoebe Nees	77.19	74.26	88.61	87.23	90.74	89.61
Cryptomeria japonica	83.13	85.52	93.26	92.95	94.98	93.57
Other conifers	67.07	63.24	85.73	89.77	90.63	91.96
Other forest land	76.39	78.21	90.45	93.91	91.32	92.98
Construction land	86.21	87.67	94.04	91.27	92.97	94.55
Roads	91.15	88.46	93.37	92.39	94.09	93.87

4. Discussion

Based on the current visual model background, we designed a semantic segmentation framework (AMDNet) which is suitable for the identification of remote sensing tree species. For the classification of tree species in remote sensing images, compared with the previous deep learning methods applied for tree species classification, our approach is more targeted. The experiment results show that our model has an excellent performance in all aspects and has a good balance between accuracy, computational speed, and model size which can better meet the requirements for tree species classification and practical deployment.

4.1. The Influence of the Attention Module on the Classification Results

Before discussing the performance improvements that the application of the attention module confers upon the model, we first separate the internal modules. The basic components of the dual-attention residual module are the channel and space dimensions. Firstly, we begin by elaborating on the channel attention. Each channel of a feature represents a special detector. Therefore, the focus of channel attention is based on what kind of information the feature semantics contain. In order to aggregate spatial features, we selectively emphasize the existence of interdependent channel maps by integrating the related features among all the channel maps. Based on this, the spatial attention is introduced to highlight the significance of features at meaningful positions. Moreover, we selectively aggregate the features in each position by weighting the sum of the features in all the positions. Similar features will correlate with each other regardless of the distance. Finally, the outputs of the two attention modules are combined to further improve the feature representation, thus completing the adaptive integration of local features and global dependencies. Given the characteristics of complex scenes and clustered image backgrounds of the remote sensing tree species dataset, we need to fuse different categories of attention to achieve the function of "attention focus". In this case, features from different layers require different types of attention. Therefore, we combine the deep neural network, stack the attention mechanisms, and adopt the attention residual branch for feature processing. This kind of application can improve the segmentation performance of the model in a linear manner, and additional feature semantic information can be extracted from feature maps based on different depths.

The results show that the classification performance of the network with dual residual attention is significantly improved in comparison with the basic network. This substantiates the promising applications of the attention module for the classification of remote sensing tree species and its great potential for improvement with the development of deep learning technology.

4.2. The Influence of Structural Re-Parameterization on the Model

In recent years, model re-parameterization has become an important topic in the context of network training for classification tasks. Similarly, this new proposed concept meets the current direction of optimizing the application of deep learning to the remote sensing of tree species using faster and stronger network structures, more effective feature integration methods, more accurate segmentation techniques, and more effective training approaches. The main proposed improvements of the model through structural re-parameterization are mainly as follows: (1) the addition of identity and residual branches to the block of the network so as to fully leverage the training advantages of the multi-branch structure; and

(2) the conversion of all the network layers into Conv3 × 3 through the Op fusion strategy in the model inference stage to facilitate the deployment and acceleration of the model. This improvement also results in a very intuitive change to the network structure; that is, different network structures are used during the training and inference stages. The training stage prioritizes accuracy, while the inference stage prioritizes speed. Why do we state that it is intuitive? There are two reasons: (1) the multi-branch structure ensures that the model can learn enough information during training, which is greatly helpful for the identification of remote sensing tree species with limited datasets; and (2) the weight is integrated into the deployment to reduce the size of the model. Additionally, different network frames significantly improve the deployment speed of the model, which is more conducive to its application in forestry engineering.

The results show that the model with structural re-parameterization is much better than the classical and advanced semantic segmentation models in terms of accuracy, the model inference speed, number of model parameters, and FLOPS, demonstrating the superior performance of the model. In addition to its application in forestry, structural re-parameterization has broad application prospects in medical imaging and agricultural applications. It is believed that in the future, the convolutional neural network will take precedence over Transformer in research on inference speed, owing to its simple structure and significant deployment advantages.

4.3. Directions for Improvement

The poor accuracy of the model's segmentation in the edge part of the segmentation and the poor segmentation of shaded roads and other woodlands can be attributed to the impacts of bright light and the direction of light during data collection. Optimizing the dataset may alleviate these issues and improve the segmentation performance of the model. At the same time, although the model shows a superior performance in the identification of remote sensing tree species, due to the characteristics of the network frame of this method, there are still some limitations which need to be improved, as follows:

1. In addition to ensuring clear and windless conditions during data collection, attention should also be paid to the light intensity and light direction in order to mitigate the adverse impacts of bright light and shadows on the segmentation effect.
2. While the dual-attention residual module and structural re-parameterization result in a significant improvement in accuracy in the classification of remote sensing tree species, they also impose an excessively high training burden due to the substantial application of the multi-branch structure. If the application involves a very large dataset, the tuning optimization may become a lengthy and time-consuming process, and the subsequent performance of the actual deployment is a time-proven problem.
3. The dual-attention residual module offers a key advantage in providing a global field of view and exhibits strong applicability to remote sensing forestry engineering and medical image segmentation. Nonetheless, for the training and testing of large images, the deformable-size serial self-attention module has more advantages than the dual-attention residual module.
4. Due to the application of re-parameterization, the model results in some additional computation, and the number of parameters in the network also increases due to the large-scale application of residual branching.

5. Conclusions

In this paper, a dual-attention residual network (AMDNet) was proposed to integrate UAV remote sensing images for tree species identification based on the current visual model background and application scenarios of remote sensing tree species identification. The proposed network optimizes the existing training strategies, adaptively integrates local features and global dependencies, and fuses local features and global context information. At the same time, it can improve the network classification performance in a linear manner. Moreover, the network also applies structural re-parameterization, which improves the

classification accuracy, reduces the model volume, maintains the computational efficiency, and significantly improves the inference speed, rendering the model more suitable for deployment and demonstrating its superior performance. The results of the experiment show that, under the same conditions, AMDNet has obvious advantages over the other models in terms of tree species classification, and its MIOU reaches 93.8, surpassing the existing mainstream models. In summary, this model is an efficient, high-speed, and easy-to-deploy solution for the identification of tree species in forests, which achieves the high-precision identification of tree species in remote sensing images and provides a new approach for remote-sensing tree species classification technology.

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

Funding: This research is funded by Natural Science Foundation Project of Sichuan Province (2022NSFSC0091).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data are available from the corresponding author upon request.

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

References

1. Ottosen, T.B.; Petch, G.; Hanson, M.; Skjoth, C.A. Tree cover mapping based on Sentinel-2 images demonstrate high thematic accuracy in Europe. *Int. J. Appl. Earth Obs. Geoinf.* **2020**, *84*, 101947. [CrossRef]
2. Liu, X.S.; Zhang, X.L. Research advances and countermeasures of remote sensing classification of forest vegetation. *For. Resour. Manag.* **2004**, *1*, 61–64.
3. Zeng, Q.W.; Wu, H.G. Development of hyperspectral remote sensing application in forest species identification. *For. Resour. Manag.* **2009**, *28*, 109.
4. Cho, M.A.; Mathieu, R.; Asner, G.P.; Naidoo, L.; van Aardt, J.; Ramoelo, A.; Debba, P.; Wessels, K.; Main, R.; Smit, I.P.J.; et al. Mapping tree species composition in South African savannas using an integrated airborne spectral and LiDAR system. *Remote Sens. Environ.* **2012**, *125*, 214–226. [CrossRef]
5. Turner, W.; Spector, S.; Gardiner, N.; Fladeland, M.; Sterling, E.; Steininger, M. Remote sensing for biodiversity science and conservation. *Trends Ecol. Evol.* **2003**, *18*, 306–314. [CrossRef]
6. Axelsson, A.; Lindberg, E.; Reese, H.; Olsson, H. Tree species classification using Sentinel-2 imagery and Bayesian inference. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *100*, 102318. [CrossRef]
7. Mei, S.; Song, C.; Ma, M.; Xu, F. Hyperspectral Image Classification Using Group-Aware Hierarchical Transformer. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
8. Dalponte, M.; Ørka, H.O.; Gobakken, T.; Gianelle, D.; Næsset, E. Tree Species Classification in Boreal Forests with Hyperspectral Data. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 2632–2645. [CrossRef]
9. Cheng, J.; Yang, H.; Qi, J.; Sun, Z.; Han, S.; Feng, H.; Jiang, J.; Xu, W.; Li, Z.; Yang, G.; et al. Estimating canopy-scale chlorophyll content in apple orchards using a 3D radiative transfer model and UAV multispectral imagery. *Comput. Electron. Agric.* **2022**, *202*, 107401. [CrossRef]
10. Guo, Y.; Chen, S.; Li, X.; Cunha, M.; Jayavelu, S.; Cammarano, D.; Fu, Y. Machine Learning-Based Approaches for Predicting SPAD Values of Maize Using Multi-Spectral Images. *Remote. Sens.* **2022**, *14*, 1337. [CrossRef]
11. Everaerts, J. The use of unmanned aerial vehicles (UAVs) for remote sensing and mapping. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2008**, *37*, 1187–1192.
12. Schiefer, F.; Kattenborn, T.; Frick, A.; Frey, J.; Schall, P.; Koch, B.; Schmidtlein, S. Mapping forest tree species in high resolution UAV-based RGB-imagery by means of convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2020**, *170*, 205–215. [CrossRef]
13. Li, Y.; Chai, G.; Wang, Y.; Lei, L.; Zhang, X. ACE R-CNN: An Attention Complementary and Edge Detection-Based Instance Segmentation Algorithm for Individual Tree Species Identification Using UAV RGB Images and LiDAR Data. *Remote Sens.* **2022**, *14*, 3035. [CrossRef]
14. Wu, J.; Yang, G.; Yang, H.; Zhu, Y.; Li, Z.; Lei, L.; Zhao, C. Extracting apple tree crown information from remote imagery using deep learning. *Comput. Electron. Agric.* **2020**, *174*, 105504. [CrossRef]

15. Ye, Z.; Wei, J.; Lin, Y.; Guo, Q.; Zhang, J.; Zhang, H.; Deng, H.; Yang, K. Extraction of Olive Crown Based on UAV Visible Images and the U2-Net Deep Learning Model. *Remote Sens.* **2022**, *14*, 1523. [CrossRef]
16. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2015; pp. 234–241.
17. Geras, K.J.; Wolfson, S.; Shen, Y.; Wu, N.; Kim, S.; Kim, E.; Heacock, L.; Parikh, U.; Moy, L.; Cho, K. High-resolution breast cancer screening with multi-view deep convolutional neural networks. *arXiv* **2017**, arXiv:1703.07047.
18. Modzelewska, A.; Fassnacht, F.E.; Stereńczak, K. Tree species identification within an extensive forest area with diverse management regimes using airborne hyperspectral data. *Int. J. Appl. Earth Obs. Geoinf.* **2020**, *84*, 101960. [CrossRef]
19. Fassnacht, F.E.; Latifi, H.; Stereńczak, K.; Modzelewska, A.; Lefsky, M.; Waser, L.T.; Straub, C.; Ghosh, A. Review of studies on tree species classification from remotely sensed data. *Remote Sens. Environ.* **2016**, *186*, 64–87. [CrossRef]
20. Dalponte, M.; Frizzera, L.; Gianelle, D. Individual tree crown delineation and tree species classification with hyperspectral and LiDAR data. *PeerJ* **2019**, *6*, e6227. [CrossRef]
21. Rana, P.; St-Onge, B.; Prieur, J.-F.; Cristina Budei, B.; Tolvanen, A.; Tokola, T. Effect of feature standardization on reducing the requirements of field samples for individual tree species classification using ALS data. *ISPRS J. Photogramm. Remote Sens.* **2022**, *184*, 189–202. [CrossRef]
22. Liu, H. Classification of urban tree species using multi-features derived from four-season RedEdge-MX data. *Comput. Electron. Agric.* **2022**, *194*, 106794. [CrossRef]
23. Ferreira, M.P.; de Almeida, D.R.; de Almeida Papa, D.; Minervino, J.B.; Veras, H.F.; Formighieri, A.; Santos, C.A.; Ferreira, M.A.; Figueiredo, E.O.; Ferreira, E.J. Individual tree detection and species classification of Amazonian palms using UAV images and deep learning. *For. Ecol. Manag.* **2020**, *475*, 118397. [CrossRef]
24. Cao, K.; Zhang, X. An improved Res-UNet model for tree species classification using airborne high-resolution images. *Remote Sens.* **2020**, *12*, 1128. [CrossRef]
25. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; IEEE: Piscataway, NJ, USA, 2016; pp. 770–778.
26. Qi, T.; Zhu, H.; Zhang, J.; Yang, Z.; Chai, L.; Xie, J. Patch-U-Net: Tree species classification method based on U-Net with class-balanced jigsaw resampling. *Int. J. Remote Sens.* **2022**, *43*, 532–548. [CrossRef]
27. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; IEEE: Piscataway, NJ, USA, 2019; pp. 3146–3154.
28. Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; Tang, X. Residual Attention Network for Image Classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; IEEE: Piscataway, NJ, USA, 2017; pp. 3156–3164.
29. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
30. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A convnet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; IEEE: Piscataway, NJ, USA, 2022; pp. 11976–11986.
31. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; IEEE: Piscataway, NJ, USA, 2018; pp. 4510–4520.
32. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV); IEEE: Piscataway, NJ, USA, 2018; pp. 801–818.
33. Ding, X.; Zhang, X.; Ma, N.; Han, J.; Ding, G.; Sun, J. Repvgg: Making vgg-style convnets great again. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; IEEE: Piscataway, NJ, USA, 2021; pp. 13733–13742.
34. Xu, S.; Xu, X.; Blacker, C.; Gaulton, R.; Zhu, Q.; Yang, M.; Yang, G.; Zhang, J.; Yang, Y.; Yang, M.; et al. Estimation of Leaf Nitrogen Content in Rice Using Vegetation Indices and Feature Variable Optimization with Information Fusion of Multiple-Sensor Images from UAV. *Remote Sens.* **2023**, *15*, 854. [CrossRef]
35. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; IEEE: Piscataway, NJ, USA, 2015; pp. 3431–3440.
36. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; IEEE: Piscataway, NJ, USA, 2017; pp. 2881–2890.
37. Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; Agrawal, A. Context encoding for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; IEEE: Piscataway, NJ, USA, 2018; pp. 7151–7160.

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