forests

MDPI

Article

Image Classification of Tree Species in Relatives Based on Dual-Branch Vision Transformer

Qi Wang [1], Yanqi Dong [1], Nuo Xu [1], Fu Xu [1,2], Chao Mou [1,2,*] and Feixiang Chen [1,2,*]

[1] School of Information Science and Technology, Beijing Forestry University, Beijing 100083, China; wangqi1822@bjfu.edu.cn (Q.W.); yanqidong@bjfu.edu.cn (Y.D.); xunuoo@bjfu.edu.cn (N.X.); xufu@bjfu.edu.cn (F.X.)

[2] Engineering Research Center for Forestry-Oriented Intelligent Information Processing, National Forestry and Grassland Administration, Beijing 100083, China

* Correspondence: chao_m@bjfu.edu.cn (C.M.); bjfxchen@bjfu.edu.cn (F.C.); Tel.: +86-10-6233-6511 (F.C.)

Abstract: Tree species in relatives refer to species belonging to the same genus with high morphological similarity and small botanical differences, making it difficult to perform classification and usually requiring manual identification by experts. To reduce labor costs and achieve accurate species identification, we conducted research on the image classification of tree species in relatives based on deep learning and proposed a dual-branch feature fusion Vision Transformer model. This model is designed with a dual-branch architecture and two effective blocks, a Residual Cross-Attention Transformer Block and a Multi-level Feature Fusion method, to enhance the influence of shallow network features on the final classification and enable the model to capture both overall image information and detailed features. Finally, we conducted ablation studies and comparative experiments to validate the effectiveness of the model, achieving an accuracy of 90% on the tree relatives dataset.

Keywords: image classification; tree species classification; Vision Transformer

check for updates

Citation: Wang, Q.; Dong, Y.; Xu, N.; Xu, F.; Mou, C.; Chen, F. Image Classification of Tree Species in Relatives Based on Dual-Branch Vision Transformer. *Forests* **2024**, *15*, 2243. https://doi.org/10.3390/f15122243

Academic Editor: Giorgos Mallinis

Received: 9 October 2024
Revised: 13 December 2024
Accepted: 18 December 2024
Published: 20 December 2024

Copyright: © 2024 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

As a key component of forests, trees play a vital role in human survival and ecosystem. Species classification is essential for understanding biodiversity and supporting ecological and environmental protection. Traditionally, this process has relied on expert knowledge, which often incurs significant labor costs [1]. In recent years, the rapid advancement of deep learning has led to remarkable progress in image recognition technology, offering innovative solutions for tree and plant species classification. Many experts and researchers have focused on using various types of images, including bark [2], crowns [3,4], leaves [5,6], flowers [7,8], seeds [9,10], and leaf diseases [11,12], to improve the efficiency of addressing issues of plant classification.

Current research is predominantly based on convolutional neural networks (CNNs) or Transformers, both of which have demonstrated significant potential and achieved promising results. According to Chen et al. [13], classical CNNs such as ResNet and VGG have achieved high accuracy in various forestry applications, including tree species and tree organ image classification. Kong et al. [14] utilized CSPNet as a backbone network, designing three parallel sub-networks to fuse the features of different granularities and the prediction results from each sub-model using a cross-level fusion strategy, achieving higher accuracy in fine-grained classification tasks. Zhao et al. [15] addressed the large inter-class and small intra-class differences in leaf disease images by incorporating the residual module into the Inception structure and adding the modified CBAM module after each RI-Block. This improved the localization of key information in the image, enhancing the model's recognition accuracy. Syeda Alleena Riaz et al. [16] presented a multi-path deep convolutional network designed for plant species identification. This approach processes different versions of plant images through separate network paths, effectively

capturing both shallow and deep features to improve classification performance. Although CNN-based methods have made significant strides in addressing challenges in the forestry domain, the growing use of Transformer models in computer vision has provided notable advantages [17]. The unique self-attention mechanism of Transformers improves the model's ability to understand global context, enhancing performance in dealing with image noise and scale variations and increasing robustness. Transformers have shown superior potential compared with CNNs in handling similar features [18] and have also been proven effective in fine-grained image classification in forestry contexts. Reedha et al. [19] used the Vision Transformer (ViT) model for weed and crop classification from UAV-captured images, outperforming other state-of-the-art CNN models. Chang et al. [20] designed an edge-feature guidance module based on ViT to improve feature fitting for leaf disease classification, showing excellent results in their experiments. Li et al. [21] proposed the Swin Routiformer Block, using the Swin Transformer as the baseline and incorporating a two-layer Routing Attention Block to reduce network complexity and improve image processing speed, achieving better performance in classifying common biological images in forest environments. Some scholars have integrated both CNN and Transformer methods to combine their respective strengths. For instance, Lee et al. [22] proposed the Plant-CNN-ViT integrated model, which combines the advantages of four pre-trained models—ViT, ResNet-50, DenseNet-201, and Xception—demonstrating superior performance on four publicly available leaf datasets.

Current research still faces certain limitations. The species included in existing public tree and plant image datasets often lack relevance to plant taxonomy, and there is a common phenomenon of using a single background in the images. Therefore, it is worthwhile to conduct further research on the classification of tree relatives' images in a forest context, which have similar traits. In plant taxonomy, following the principles and methods of biological classification, the hierarchy is organized from high to low as "kingdom, phylum, class, order, family, genus, and species". Species within the same genus are referred to as plant relatives. The botanical characteristics of plants across different families tend to be more distinct, which is why family-level classification is commonly used [23]. In contrast, plants within the same family or genus are more morphologically similar, exhibiting smaller botanical differences, which often requires expert identification [24]. As a result, feature extraction becomes more challenging and the modeling capacity is higher. In response to these research gaps and challenges, we conduct fine-grained image classification of tree species within the same genus. We have collected a dataset of tree relatives images and selected trees from the Taxus genus as representative species. The differences in leaf morphology within the Taxus genus can serve as a basis for species classification [25]. Therefore, to capture the fine-grained differences between images, we hypothesize that the Visual Transformer, which excels at processing similar features, could serve as an effective backbone. We propose a dual-branch feature fusion ViT method that integrates both global and local multi-scale feature information. Simultaneously, inspired by [26] and previous studies on the importance of shallow features on model performance [27,28], we explore the hypothesis that plant datasets are more dependent on shallow features. Based on this, we design two effective blocks to enhance the influence of shallow features on tree relatives image classification. The main contributions of our work are as follows:

- We constructed a tree relatives dataset through both shooting and network collection. This dataset includes 900 RGB images of five Taxus species. It covers images of entire plants, leaves, and fruits, providing essential data support for our study.
- To better integrate the overall morphological features and specific features of trees, we propose a dual-branch ViT architecture that combines global features and detailed features. We also introduce a Residual Cross-Attention Transformer Block and a Multi-level Feature Fusion method to realize the enhancement and fusion of dual-branch features to achieve better feature extraction results.
- We propose a Residual Cross-Attention Transformer Block. We concatenate multiple Transformer Encoders and add the attention in the first Transformer Encoder as

residual to the later Encoders with a specified ratio to focus more on shallow attention information. In addition, we incorporate a cross-attention mechanism to realize the interaction of token information between two branches, enriching the learning process of each branch.

- To more effectively integrate shallow, middle, and deep features, we design a Multi-level Feature Fusion method. It makes full use of the advantages of different levels features in the network and takes the fused features as the final features of the incoming classification header to realize the complementary advantages of each level and improve the overall performance of the model.

The rest of this paper is organized as follows. Section 2 describes the proposed method. Section 3 reports the dataset and experimental results. Section 4 discusses and analyzes the experimental results. Finally, Section 5 provides conclusions.

2. Methods

2.1. Dual-Branch Feature Fusion Vision Transformer

Based on research and experiments, we propose a dual-branch Feature Fusion ViT model with ViT as the backbone. The architecture of the model consists of a global branch and a detail branch formed by different patch sizes. To enhance the impact of shallow features on tree relatives image classification, we designed the Residual Cross-Attention Transformer Block and Multi-level Feature Fusion method, which are described in detail in Sections 2.2 and 2.3. Figure 1 illustrates the structure of the dual-branch Feature Fusion ViT and its constituent parts.

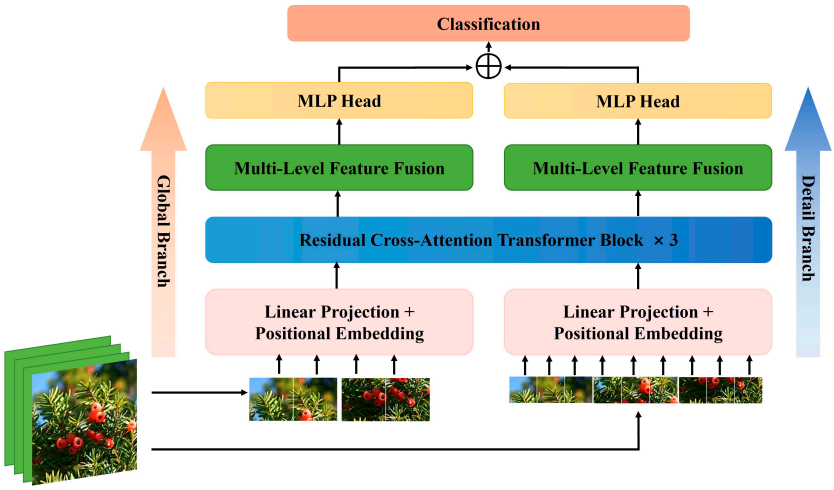


Figure 1. Overall architecture diagram of dual-branch Feature Fusion ViT. The constituent parts and process diagrams of the global branch and detail branch are described.

In ViT, the input image is divided into multiple non-overlapping patches based on the specified patch size. Each patch undergoes Flattening and Linear Projection, resulting in a one-dimensional sequence of embedding vectors. Position coding is then added to each patch to enable the model to utilize the relative position information of the patches. In this stage, the difference in patch sizes leads to variations in feature extraction results. To address this, we segment the input image at two different scales and process them through two branches: the global branch, which uses smaller patch size, and the detail branch, which uses larger patch size. We ultimately chose the parameters for the dual branch, with a global branch patch size of 12 and embedding dimension of 384 and a detail branch patch size of 16 and embedding dimension of 768. Each branch sequentially passes through Linear Projection, Position Embedding, Residual Cross-Attention Transformer Block, Multi-level Feature Fusion, and MLP Head. The outputs from both branches are classification tensors, denoted as $logits_1$ and $logits_2$. In order to integrate the information

from the two branches, we splice and average $logits_1$ and $logits_2$ to obtain the output that combines the learning information from the two branches, which is performed to receive the classification results under multiple scales.

2.2. Residual Cross-Attention Transformer Block

To better utilize the advantages of the proposed dual-branch architecture, make full use of the multi-scale features learned from different patch sizes, and realize the enhancement of global and detail branch features, we propose the Residual Cross-Attention Transformer Block (RC-Attention Transformer Block). For each branch, we concatenate a number of base Transformer Encoders as the core, on which we design the combined structure of the attention residual connection and cross-attention mechanisms. Figure 2 depicts the structure of the dual-branch RC-Attention Transformer Block.

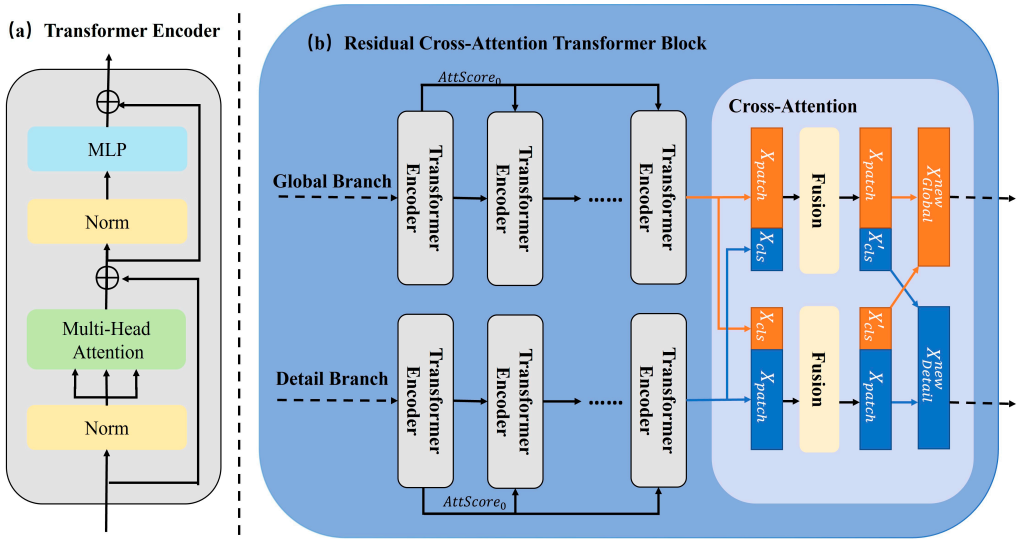


Figure 2. Residual Cross-Attention Transformer Block and related internal blocks. (**a**) The described Transformer Encoder structure. (**b**) The described Residual Cross-Attention Transformer Block structure. Each branch is concatenated with several Transformer Encoders, and the first attention score is added as the residual to the following Transformer Encoders. Pass the results into the Cross Attention step to achieve the interaction of dual-branch information.

The Transformer Encoder is mainly composed of LayerNorm, a multi-head attention mechanism, an MLP module, and a residual connection for the multi-head attention mechanism and MLP module, as shown in Figure 2a. The multi-head attention mechanism, serving as the core feature extraction module of the Transformer Encoder, progressively captures more abstract and higher-order attention information with increasing network depth. However, this progression often results in the attenuation or loss of shallow attention features. To address this problem, we are inspired by [29] and propose the shallow attention residual connection mechanism, which establishes a residual connection between Transformer Blocks connected in series for multi-head attention scores. The multi-head attention mechanism achieves the capture of complex dependencies between different locations by splitting the embedding vectors of the input sequence into multiple heads, computing the attention scores with normalized weights independently in each subspace, and merging the results of multiple heads. In our study, we choose the attention score obtained from the scaled dot product attention formula as the residual term. At first, it is determined whether the current Transformer Block is the first layer in the RC-Attention Transformer Block. If the current layer is the first layer, it is obtained by the formula $AttScore_0$; if the current layer is not the first layer, take $AttScore_0$ as the attention residual and add it with the attention score of the current layer in the proportion of 25% and 75%, as the final attention score of the current layer. The calculation formula is shown as follows:

Q and K are the query and key, respectively, $\frac{1}{\sqrt{d}}$ is the scaling factor, and n is the position of the current Transformer Block in the RC-Attention Transformer Block.

$$AttScore_n = \begin{cases} \frac{Q_n \cdot K_n^T}{\sqrt{d}}, & n = 0 \\ 0.75 \times \frac{Q_n \cdot K_n^T}{\sqrt{d}} + 0.25 \times AttScore_0, & n! = 0 \end{cases}, \quad (1)$$

After confirming the current layer attention score, the attention weight is obtained by normalization through $SoftMax(AttScore_n)$. Compared with the attention weight, the attention score more directly reflects the correlation between the query and the key and retains more of the original correlation strength, so we choose the attention score as the residual value to propagate in the network. This design balances the preservation of shallow attention features with the capacity for deeper abstraction, enhancing the overall feature representation capabilities of the network. The structure of the attention residual connection in the RC-Attention Transformer Block is shown in Figure 2b.

At the end of each RC-Attention Transformer Block, we add the cross-attention mechanism proposed in [30], which performs a dual-branch feature interaction operation for the outputs that proceed through the shallow attention residual connection mechanism. We segment the features of the global and detail branch outputs to obtain the independent class token vector (X_{cls1} and X_{cls2}) and patch token vector (X_{patch1} and X_{patch2}). We exchange the class token information of the two branches and align the dimensions of X_{cls1} and X_{patch2} using the dimension projection function to obtain X'_{cls1}, which is used as the query vector (Q) in the self-attention computation. X'_{cls1} is spliced with X_{patch2} as the key vector (K) and the value vector (V). Subsequently, the attention score is computed using the scaled dot product attention formula, and the attention weight is obtained by normalization, which in turn yields the self-attention output. Finally, the output is reduced to the same dimension as X_{cls1} using the dimensional inverse projection function as a way to gain the new class token X_{cls1}^{new} after the learning of the patch token of the other branch, which is spliced with X_{patch1} to obtain the new features of this branch in the subsequent network propagation. This cross-attention fusion process is symmetrically applied to both the global and detail branches, ensuring bidirectional feature interaction and alignment. By enabling each branch to assimilate information from the complementary scale, this mechanism fosters efficient dual-branch feature integration, enhancing the representation learning capacity of the network for multi-scale information.

2.3. Multi-Level Feature Fusion

According to the network propagation characteristics, the increase in network depth can access more advanced semantic information, but at the same time, the more specific shallow information will be lost. Therefore, we propose the Multi-level Feature Fusion method, which integrates feature information from multiple levels to form a richer feature representation. In the proposed architecture, three RC-Attention Transformer Blocks are concatenated, and as the network propagates, each block outputs the feature matrices of the global branch and the detail branch, respectively. As a result, three layers of feature matrices are obtained for each branch. According to the model location they belong to, we name them shallow features, middle features, and deep features, respectively. Figure 3 illustrates the architecture of the Multi-level Feature Fusion module, which is designed to effectively aggregate these multi-level feature representations.

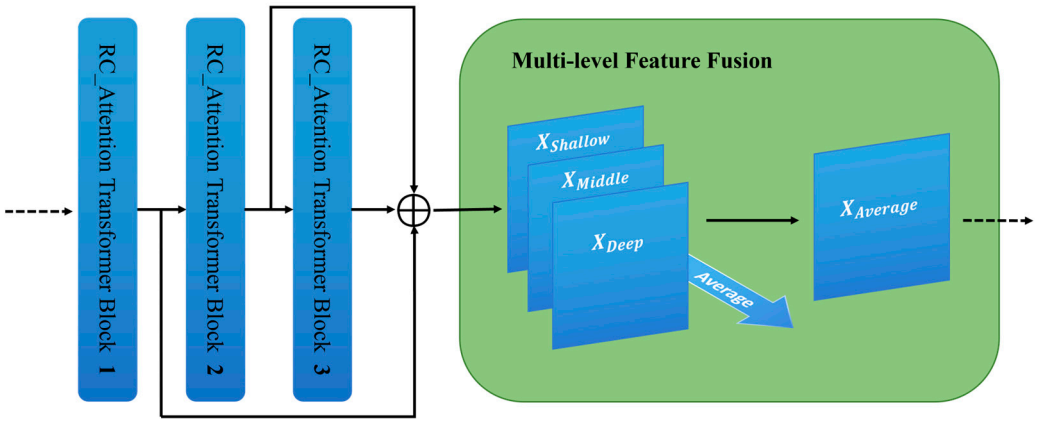


Figure 3. Description of the structure of Multi-level Feature Fusion. Concatenate and fuse the output features of three RC-Attention Transformer Blocks to obtain the averaged features.

For each branch, three levels of feature matrices are input into the Multi-level Feature Fusion block, and then the three feature matrices are stacked along a new dimension to realize the combination of multiple independent features into a unified tensor structure, which finally forms the integrated feature matrix X_{All}. The resulting integrated feature matrix encapsulates information across different levels of abstraction within the network.

$$X_{All} = \begin{bmatrix} X_{Shallow} \\ X_{Medium} \\ X_{Deep} \end{bmatrix}, \tag{2}$$

To synthesize the multi-level features into a cohesive representation, an operation is performed on X_{All} to average the data at each level by performing mean computation over the stacked dimensions as a way to achieve multiple feature fusion and output as a single, averaged feature matrix that eliminates possible redundant information and highlights common features. Mathematically, this process can be represented as:

$$X_{Aveage} = \frac{1}{H} \sum_{i=1}^{H} X_{All[i]}, \quad H = 3, \quad X_{All[i]} \in R^{B \times C}, \tag{3}$$

In the above equation, H represents the number of RC-Attention Transformer Blocks, which is set to 3 in our implementation. $X_{All[i]} \in R^{B \times C}$ represents the categorization tensor output from each RC-Attention Transformer Block, and B and C represent the batch size and the number of categories, respectively. The stacking operation will splice the multi-level output $X_{All[i]}$ along the new dimension to obtain a multi-level classification tensor with shape (H, B, C). The averaging operation averages the multi-level output tensor along the dimension of the number of levels and ultimately yields an X_{Aveage} of shape (B, C). This Multi-level Feature Fusion method effectively consolidates features from different layers, leveraging the complementary information across levels. By combining low-level, mid-level, and high-level features into a unified representation, the proposed approach enables the extraction of a more comprehensive and resilient feature set, enhancing the overall representational power of the network.

3. Experiments and Results

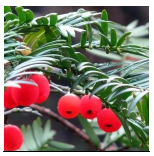
3.1. Dataset and Experiment Settings

3.1.1. Dataset

In order to better carry out the fine-grained image classification study of tree species in relatives, we independently constructed an RGB tree image dataset of the relatives of the genus Taxus with reference to the *Flora Reipublicae Popularis Sinicae* (FRPS, Chinese Edition), which contains five species, including *Taxus cuspidata*, *Taxus chinensis*, *Taxus wallichiana*, *Taxus yunnanensis*, and *Taxus wallichiana* var. *mairei*. *Taxus wallichiana* var. *mairei* is a variety

of *Taxus chinensis*. Most of the data in the dataset were obtained through our filming, covering simple and complex backgrounds, various lighting conditions such as strong and soft light, as well as local close-up and panoramic shooting methods, aiming to simulate various observation conditions that may be encountered in practical applications. In order to obtain more sufficient training data and improve the model generalization ability, we tried to balance the data of each category by means of network collection [31]. After data screening and organization, we finally obtained 200 images each of *Taxus cuspidata*, *Taxus chinensis*, *Taxus wallichiana*, and *Taxus wallichiana* var. *mairei* and 100 images of *Taxus yunnanensis*, totaling 900 images. On the basis of the collected data, we divided the training set and validation set at a ratio of 80% and 20%. Table 1 shows example images of five species and data collection information.

Table 1. Dataset information of closely related species of the Taxus genus.

Scientific Name	***Taxus cuspidata***	***Taxus chinensis***	***Taxus wallichiana***	***Taxus yunnanensis***	***Taxus wallichiana* var. *mairei***
Image Example					
Number of Images	200	200	200	100	200

In order to enhance the model robustness, we used the transforms module under the torchvision library in Python to implement image enhancement operations. For the images in the training set, we chose random scaling cropping and random horizontal flipping methods. The random scaling cropping size was set to 224 × 224 pixels, and the probability of random horizontal flipping was 50%. The combination of these two techniques can help improve the model's ability to learn from different positions, sizes, and directions and reduce the overfitting phenomenon. Finally, image normalization was performed using the mean (0.485, 0.456, 0.406) and variance (0.229, 0.224, 0.225) of the Imagenet dataset. For the images in the validation set, we chose the methods of image size reset and center cropping. We adjusted the image size to 256 × 256 pixels and then cropped out a 224 × 224 pixel area from the center. Finally, we performed image standardization processing.

To validate the effectiveness of the proposed method, we conducted three types of experiments and analyses using the Taxus dataset: ablation studies, comparative experiments, and computational cost analysis. We adopted Top1 accuracy as the evaluation metric and used only classification labels without any additional annotations for supervised training.

3.1.2. Experiment Settings

All of the experiments were conducted using PyTorch 2.0, Python 3.8, and Cuda11.8, and the whole process of training and validation was carried out on a server configured with an NVIDIA GeForce RTX 3090 (24GB) GPU. In the training phase of the proposed method, we use the Adam optimizer for training and set the initial learning rate to 0.0005. In order to stabilize the beginning phase of the training and to prevent overfitting in the later phase, we applied five rounds of learning rate warm-up and 0.05 weight decay. In each epoch, we processed a sample with a batch size of 16. The entire training process consisted of 70 rounds.

3.2. Ablation Studies

To verify the performance of the methods proposed in this paper, we conducted a series of ablation studies, using the Taxus dataset for the relevant experiments and using Top1 Accuracy as the test criterion. Because ViT was chosen as the backbone, we used the accuracy of the original ViT as the reference result in the ablation studies. In this section, we focus on the experiments and validity analysis of the three proposed methods, dual-branch

ViT stands for Dual-Scale Branch Integration ViT. RC-Attention Transformer Block is the method proposed in this paper in Section 2.2. Multi-level Feature Fusion is the method proposed in this paper in Section 2.3. The backbone model accuracy and the ablation experimental results of the three proposed methods are summarized in Table 2. We use √ to represent the methods added in this round of experiments and the optimal results are highlighted in bold.

Table 2. Ablation studies of the proposed dual-branch ViT, RC-Attention Transformer Block, Multi-level Feature Fusion.

Dual-Branch ViT	RC-Attention Transformer Block	Multi-Level Feature Fusion	Top1 Accuracy
-	-	-	85%
√	-	-	86.11%
√	√	-	87.78%
√	-	√	88.89%
√	√	√	**90.00%**

3.2.1. Dual-Branch ViT

The dual-branch ViT is increased by 1.11% compared with the backbone model. In ViT, we capture information at different scales by specifying the size of patches. Smaller size patches capture more global information, making it easier for the model to obtain the overall structure and global features in the image. Larger size patches retain more detailed information due to the fact that each patch covers a smaller area of the image and is able to characterize the local features of the image in more detail. The dual-scale branch integration ViT, which is designed with different patch sizes in the two branches, retains the advantages of large-size patches and small-size patches at the same time, which enables the model to acquire richer feature information and therefore achieves better results compared with single-scale ViT.

3.2.2. RC-Attention Transformer Block

The RC-Attention Transformer Block is added to the dual-branch ViT and achieved 87.78% accuracy on the experimental dataset, demonstrating the effectiveness of the secondary enhancement mechanism targeting attention in network propagation on this dataset. This result is attributed to the multi-head self-attention mechanism in the Transformer encoder, where attention scores determine the focus areas of the model. By emphasizing shallow attention, critical features captured in the early stages of the model are directly propagated to later stages, increasing the contribution of early-stage features in network propagation. Additionally, the cross-attention mechanism facilitates multi-scale information exchange between the dual branches, enabling richer and more comprehensive feature learning. Therefore, the addition of this method achieves better results on this experimental dataset compared with the dual-branch ViT.

We plotted heat maps of the three-stage RC-Attention Transformer Block, as shown in Figure 4. From the perspective of the dual branches, the global branch focuses more on the background and the overall features of the image, while the detail branch focuses more on the details of the tree, such as the leaf and fruit parts in this figure. From the perspective of three different stages, as the depth of the network increases, the selection of features is gradually optimized and the dual branches interact with each other more obviously, and all of them achieve better feature focus in the final stage.

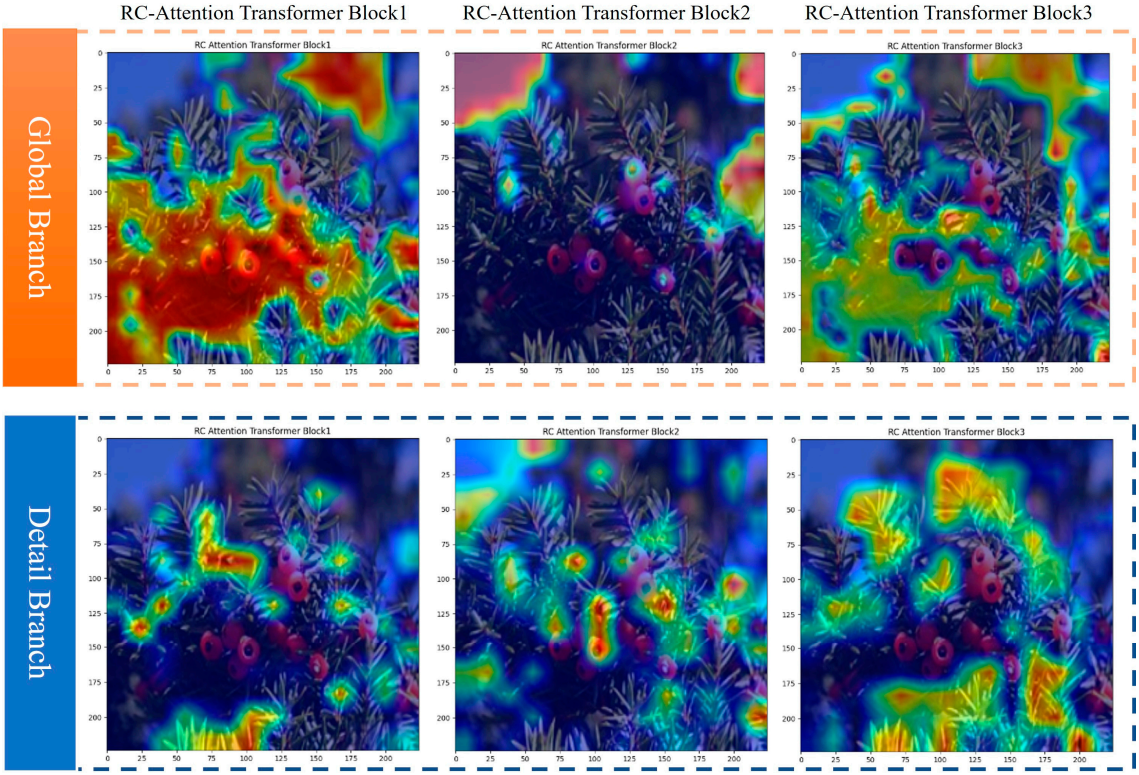


Figure 4. Heat maps of the RC-Attention Transformer Block in three stages with dual branches.

3.2.3. Multi-Level Feature Fusion

We added the Multi-level Feature Fusion method separately to the dual-branch ViT and achieved 88.89% accuracy on the experimental dataset, demonstrating that using the features output from this block inputted to the MLP Head for classification achieved better results compared with the deep features output from the regular ViT. This suggests that such an improved strategy of using fusion features that synthesize shallow, middle, and deep features is applicable to Taxus datasets. In the network, shallow features usually contain low-level detail information such as edges and textures, mid-level features mainly capture mid-level structural information, and deep features mainly capture abstract semantic information. The average combination of three levels of features replaces the use of only deep features for final classification, which is more suitable for identifying tree related species images.

Our proposed dual-branch feature fusion ViT combines the above three methods, and the experiment acquires an overall significant improvement, achieving 90% accuracy on the experimental dataset. Summarizing the above results, it can be concluded that the dual-branch ViT, compared with the original single-scale ViT, has more space for improvement. On the basis of dual-branch ViT, when adding an RC-Attention Transformer Block alone, there is no significant improvement in the accuracy, and when it is used in combination with Multi-level Feature Fusion, the RC-Attention Transformer Block can further supplement the shallow information and the information interaction between different branches. These two blocks complement each other, thus realizing further improvement in accuracy.

3.3. Comparison with Existing Methods

To validate the effectiveness of our proposed Dual-branch Feature Fusion ViT, we conducted experiments on the Taxus datasets and compared its performance with CNN-based and ViT-based networks, as well as related improved methods. Top1 accuracy was used as the evaluation criterion, with the highest accuracy being highlighted in bold.

According to Table 3, our proposed method achieves the best performance on the Taxus dataset compared with other methods. The experimental results show that ViT-Base

performs the best among the nine classical network models we chose, indicating that the Transformer-based approach has better potential for fine-grained image classification of Taxus relatives compared with the CNN-based approach. Therefore, we finally chose ViT as the backbone and improved upon it. Finally, the Top1 accuracy of our proposed method in the validation set of Taxus datasets is 90%, which is significantly improved by 5% compared with the backbone. The loss and accuracy curves on the proposed approach are shown in Figure 5.

Table 3. Results of the Taxus dataset.

Method	Top1 Accuracy
VGG16	81.67%
GoogLeNet	81.67%
ResNet50	82.22%
DenseNet121	84.44%
ConvNeXt-Base	83.33%
EfficientnetV2-B0	83.33%
ViT-Base-16	85%
SwinT-B	83.89%
CrossViT-B	87.22%
Ours	**90%**

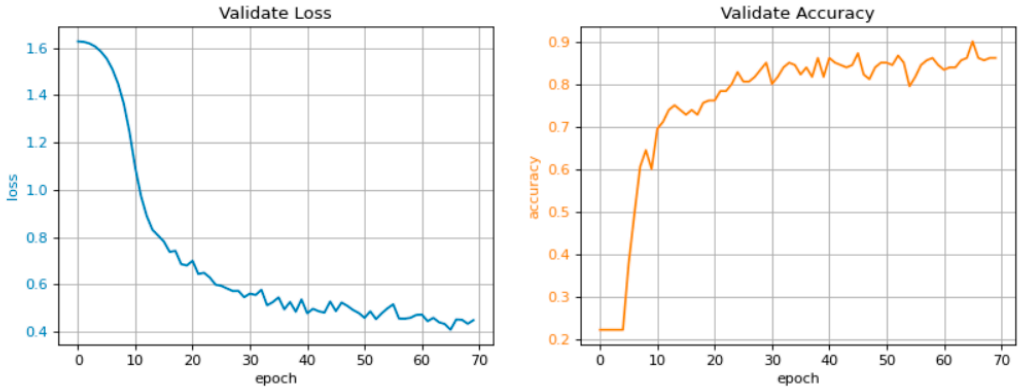


Figure 5. The loss and accuracy curves of the Taxus dataset on the proposed approach.

3.4. Computational Cost Discussion

Due to the complexity of the dual-branch model, we selected the representative CNN-based methods and Transformer-based methods mentioned in this paper. The number of model parameters (#Params) and the amount of model computations (FLOPs) for these methods are summarized in Table 4 for the image input size of 224 × 224.

Table 4. Comparison of the model capacity and the computational cost. We used #Params(M) to evaluate the model capacity and FLOPs(G) to evaluate the computational cost.

Method	#Params(M)	FLOPs(G)
VGG16	138.4	15.6
GoogLeNet	5.6	1.51
ResNet50	25.6	3.53
DenseNet121	6.96	2.9
ConvNeXt-Base	89	15.4
EfficientnetV2-B0	52.8	5.44
ViT-Base-16	86.6	17.6
SwinT-B	88	15.4
CrossViT-B	104.7	21.2
Ours	**103.57**	**20.13**

Among these representative methods, with the continuous development of CNN-based approaches, the number of parameters and FLOPs have been decreasing. In contrast, Transformer-based methods generally have higher parameter counts and FLOPs compared with CNN-based methods. To capture multi-scale information, our method adopts a dual-branch architecture and incorporates additional blocks. As a result, it is foreseeable that the computational cost of our model is higher compared with the backbone model, ViT. However, when compared with CrossViT, which is also a dual-branch model, our model has slightly fewer parameters.

4. Discussion

This study validates the effectiveness of the proposed dual-branch feature fusion ViT method in addressing the classification challenges of tree relatives, achieving a significant improvement in Top1 accuracy on the Taxus dataset. Based on the series of experiments conducted and the results obtained, we further discuss the advantages of the dual-branch ViT structure and the strengthening of shallow features in ViT.

Compared with the single-branch ViT structure, the dual-branch ViT structure offers greater flexibility and optimization potential, with several effective methods having been proposed by other researchers [32–34]. Inspired by these approaches, we extended and optimized the dual-branch framework by introducing a second branch with a different patch size, enabling the model to train on two patch sizes simultaneously. We also designed the RC-Attention Transformer Block, which incorporates a shallow attention residual mechanism to better preserve shallow feature information. Additionally, the introduced cross-attention mechanism [30] enhances the interaction between the two branches. In the final decision-making stage, we averaged the decision outputs of both branches to fully leverage the potential and advantages of the dual-branch ViT.

In traditional CNN architectures, feature fusion is typically achieved through methods such as the Feature Pyramid Network (FPN) [35], which effectively combines multi-scale features to enhance model performance. To further investigate the hypothesis that plant images are more dependent on shallow features, we designed a ViT-specific feature fusion method. This method uses an aggregated output of shallow, middle, and deep features as the basis for classification, replacing the traditional approach of relying solely on deep features. This design is particularly suited for the classification of tree relatives images. Unlike CNNs, however, the feature size in ViT remains unchanged before and after passing through the Transformer Encoder. Therefore, we adopted a stacking and averaging operation to fuse features across different layers. This approach ensures a balanced integration of multi-level features, effectively capturing both shallow and deep information to improve model performance.

Through ablation studies and comparative experiments, we demonstrated the importance and effectiveness of each proposed module. Additionally, we validated the applicability of enhancing shallow features and Multi-level Feature Fusion in the classification tasks of tree relatives' images. While this approach predictably increases computational costs, the benefits in performance justify the trade-off.

5. Conclusions

In this paper, we propose a new Dual-branch Feature Fusion ViT for the image classification of tree relatives, a task characterized by high classification difficulty due to similar features. The method employs ViT as a backbone model and designs a dual-branch architecture to realize the network's learning of multi-scale features. We added an RC-Attention Transformer Block and Multi-level Feature Fusion method to the dual-branch architecture. The former adds a cross-attention mechanism while realizing the propagation of shallow attention scores through the network, providing an interaction channel for multiscale branch output features; the latter realizes the fusion of shallow, middle, and deep features to provide richer feature representations for model classification. The combined use of the two methods achieves the fusion and enhancement of multi-scale features. We conducted

ablation and comparative experiments on the proposed Taxus dataset to demonstrate the effectiveness of the model's modules, achieving a Top1 accuracy of 90%. This result indicates that the proposed method is suitable for the classification of tree relatives image classification tasks.

Author Contributions: Conceptualization, Q.W. and F.X.; data curation, Q.W. and Y.D.; formal analysis, N.X. and C.M.; funding acquisition, F.C.; investigation, Q.W. and C.M.; methodology, Q.W. and F.C.; project administration, F.C.; resources, F.X.; software, Q.W. and N.X.; supervision, F.C.; validation, Q.W. and F.X.; visualization, Q.W. and Y.D.; writing—original draft, Q.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Outstanding Youth Team Project of Central Universities: QNTD202308; the National Key R&D Program of China: 2022YFF1302700; and the Emergency Open Competition Project of the National Forestry and Grassland Administration: 202303.

Data Availability Statement: Data will be made available upon request.

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

References

1. Ahmad, M.; Abdullah, M.; Moon, H.; Han, D. Plant Disease Detection in Imbalanced Datasets Using Efficient Convolutional Neural Networks With Stepwise Transfer Learning. *IEEE Access* **2021**, *9*, 140565–140580. [CrossRef]
2. Carpentier, M.; Giguere, P.; Gaudreault, J. Tree Species Identification from Bark Images Using Convolutional Neural Networks. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; pp. 1075–1081.
3. Egli, S.; Höpke, M. CNN-Based Tree Species Classification Using High Resolution RGB Image Data from Automated UAV Observations. *Remote Sens.* **2020**, *12*, 3892. [CrossRef]
4. Ferreira, M.P.; Almeida, D.R.A.D.; Papa, D.D.A.; Minervino, J.B.S.; Veras, H.F.P.; Formighieri, A.; Santos, C.A.N.; Ferreira, M.A.D.; Figueiredo, E.O.; Ferreira, E.J.L. Individual Tree Detection and Species Classification of Amazonian Palms Using UAV Images and Deep Learning. *For. Ecol. Manag.* **2020**, *475*, 118397. [CrossRef]
5. Koklu, M.; Unlersen, M.F.; Ozkan, I.A.; Aslan, M.F.; Sabanci, K. A CNN-SVM Study Based on Selected Deep Features for Grapevine Leaves Classification. *Measurement* **2022**, *188*, 110425. [CrossRef]
6. Tavakoli, H.; Alirezazadeh, P.; Hedayatipour, A.; Banijamali Nasib, A.H.; Landwehr, N. Leaf Image-Based Classification of Some Common Bean Cultivars Using Discriminative Convolutional Neural Networks. *Comput. Electron. Agric.* **2021**, *181*, 105935. [CrossRef]
7. Tang, Y.; Yang, J.; Xu, Q.; Wang, H.; Ding, H. Flower Classification Model Based on Improved Attention. In Proceedings of the 2023 IEEE 3rd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA), Chongqing, China, 26 May 2023; pp. 495–499.
8. Zhang, M.; Su, H.; Wen, J. Classification of Flower Image Based on Attention Mechanism and Multi-Loss Attention Network. *Comput. Commun.* **2021**, *179*, 307–317. [CrossRef]
9. Li, X.; Feng, X.; Fang, H.; Yang, N.; Yang, G.; Yu, Z.; Shen, J.; Geng, W.; He, Y. Classification of Multi-Year and Multi-Variety Pumpkin Seeds Using Hyperspectral Imaging Technology and Three-Dimensional Convolutional Neural Network. *Plant Methods* **2023**, *19*, 82. [CrossRef]
10. Luo, T.; Zhao, J.; Gu, Y.; Zhang, S.; Qiao, X.; Tian, W.; Han, Y. Classification of Weed Seeds Based on Visual Images and Deep Learning. *Inf. Process. Agric.* **2023**, *10*, 40–51. [CrossRef]
11. Sarkar, C.; Gupta, D.; Gupta, U.; Hazarika, B.B. Leaf Disease Detection Using Machine Learning and Deep Learning: Review and Challenges. *Appl. Soft Comput.* **2023**, *145*, 110534. [CrossRef]
12. Sharma, V.; Tripathi, A.K.; Mittal, H. DLMC-Net: Deeper Lightweight Multi-Class Classification Model for Plant Leaf Disease Detection. *Ecol. Inform.* **2023**, *75*, 102025. [CrossRef]
13. Chen, Y.; Huang, Y.; Zhang, Z.; Wang, Z.; Liu, B.; Liu, C.; Huang, C.; Dong, S.; Pu, X.; Wan, F.; et al. Plant Image Recognition with Deep Learning: A Review. *Comput. Electron. Agric.* **2023**, *212*, 108072. [CrossRef]
14. Kong, J.; Wang, H.; Wang, X.; Jin, X.; Fang, X.; Lin, S. Multi-Stream Hybrid Architecture Based on Cross-Level Fusion Strategy for Fine-Grained Crop Species Recognition in Precision Agriculture. *Comput. Electron. Agric.* **2021**, *185*, 106134. [CrossRef]
15. Zhao, Y.; Sun, C.; Xu, X.; Chen, J. RIC-Net: A Plant Disease Classification Model Based on the Fusion of Inception and Residual Structure and Embedded Attention Mechanism. *Comput. Electron. Agric.* **2022**, *193*, 106644. [CrossRef]
16. Riaz, S.A.; Naz, S.; Razzak, I. Multipath Deep Shallow Convolutional Networks for Large Scale Plant Species Identification in Wild Image. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–7.
17. Huy, L.V.; Tuong, N.H.; Han, L.H.N.; Hieu, N.V. Masked Autoencoder-Based Self-Supervised Learning for Forest Plant Classification. *Cybern. Phys.* **2024**, *13*, 32–41. [CrossRef]

18. Maurício, J.; Domingues, I.; Bernardino, J. Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review. *Appl. Sci.* **2023**, *13*, 5521. [CrossRef]
19. Reedha, R.; Dericquebourg, E.; Canals, R.; Hafiane, A. Transformer Neural Network for Weed and Crop Classification of High Resolution UAV Images. *Remote Sens.* **2022**, *14*, 592. [CrossRef]
20. Chang, B.; Wang, Y.; Zhao, X.; Li, G.; Yuan, P. A General-Purpose Edge-Feature Guidance Module to Enhance Vision Transformers for Plant Disease Identification. *Expert Syst. Appl.* **2024**, *237*, 121638. [CrossRef]
21. Li, P.; Wang, H.; Wang, Z.; Wang, K.; Wang, C. Swin Routiformer: Moss Classification Algorithm Based on Swin Transformer With Bi-Level Routing Attention. *IEEE Access* **2024**, *12*, 53396–53407. [CrossRef]
22. Lee, C.P.; Lim, K.M.; Song, Y.X.; Alqahtani, A. Plant-CNN-ViT: Plant Classification with Ensemble of Convolutional Neural Networks and Vision Transformer. *Plants* **2023**, *12*, 2642. [CrossRef]
23. Haider, N. A Brief Review on Plant Taxonomy and Its Components. *J. Plant Sci. Res.* **2018**, *34*, 277–292. [CrossRef]
24. Fan, B.; Wang, Q.; Wu, W.; Zhou, Q.; Li, D.; Xu, Z.; Fu, L.; Zhu, J.; Karimi-Maleh, H.; Lin, C.-T. Electrochemical Fingerprint Biosensor for Natural Indigo Dye Yielding Plants Analysis. *Biosensors* **2021**, *11*, 155. [CrossRef]
25. Wu, X.; Wang, M.; Li, X.; Chen, Y.; Liao, Z.; Zhang, D.; Wen, Y.; Wang, S. Identification and Characterization of a New Species of Taxus—Taxus Qinlingensis by Multiple Taxonomic Methods. *BMC Plant Biol.* **2024**, *24*, 658. [CrossRef]
26. Xu, Q.; Wang, J.; Jiang, B.; Luo, B. Fine-Grained Visual Classification via Internal Ensemble Learning Transformer. *IEEE Trans. Multimed.* **2023**, *25*, 9015–9028. [CrossRef]
27. Wang, S.; Han, Y.; Song, S.; Zhu, H.; Zhang, L.; Dong, A.; Yu, J. SFRSwin: A Shallow Significant Feature Retention Swin Transformer for Fine-Grained Image Classification of Wildlife Species. In Proceedings of the Pattern Recognition and Computer Vision, Urumqi, China, 18–20 October 2024; Liu, Q., Wang, H., Ma, Z., Zheng, W., Zha, H., Chen, X., Wang, L., Ji, R., Eds.; Springer Nature: Singapore, 2024; pp. 232–243.
28. Chen, Y.; Liu, P.; Zhao, J.; Huang, K.; Yan, Q. Shallow-Guided Transformer for Semantic Segmentation of Hyperspectral Remote Sensing Imagery. *Remote Sens.* **2023**, *15*, 3366. [CrossRef]
29. Diko, A.; Avola, D.; Cascio, M.; Cinque, L. ReViT: Enhancing Vision Transformers with Attention Residual Connections for Visual Recognition. *arXiv* **2024**, arXiv:2402.11301.
30. Chen, C.-F.; Fan, Q.; Panda, R. CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. *arXiv* **2021**, arXiv:2103.14899.
31. Matheson, C.A. iNaturalist. *Ref. Rev.* **2014**, *28*, 36–38. [CrossRef]
32. Zeng, X.; Zeng, P.; Tang, C.; Wang, P.; Yan, B.; Wang, Y. DBTrans: A Dual-Branch Vision Transformer for Multi-Modal Brain Tumor Segmentation. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2023*; Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R., Eds.; Lecture Notes in Computer Science; Springer Nature: Cham, Switzerland, 2023; Volume 14223, pp. 502–512. ISBN 978-3-031-43900-1.
33. Liu, K.; Du, X.; Liu, S.; Zheng, Y.; Wu, X.; Jin, C. DDT: Dual-Branch Deformable Transformer for Image Denoising. *arXiv* **2023**, arXiv:2304.06346.
34. Ding, M.; Xiao, B.; Codella, N.; Luo, P.; Wang, J.; Yuan, L. DaViT: Dual Attention Vision Transformers. In *Computer Vision—ECCV 2022*; Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T., Eds.; Lecture Notes in Computer Science; Springer Nature: Cham, Switzerland, 2022; Volume 13684, pp. 74–92. ISBN 978-3-031-20052-6.
35. Lin, T.-Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.

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.