FFGF-Net: A Forestry Fine-Grained Fusion Network for Tree Species Classification from UAV Imagery

Xuanzi Zhou[a], Jingyi Liu[a], Zhulin Chen[b,c], Hongxin Yang[d,e,*], Sheng Xu[a,*]

[a]*College of Information Science and Technology & Artificial Intelligence, Nanjing Forestry University, Nanjing, 210037, Jiangsu, China*
[b]*Institute of Forest Resource Information Techniques, Chinese Academy of Forestry, Beijing, 100091, China*
[c]*State Forestry and Grassland Administration, Key Laboratory of Forest Management and Growth Modelling, Beijing, 100091, China*
[d]*Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai, 200241, China*
[e]*Key Laboratory of Spatial-temporal Big Data Analysis and Application of Natural Resources in Megacities, Ministry of Natural Resources, East China Normal University, Shanghai, 200241, China*

Abstract

Accurate classification of tree species is crucial for biodiversity conservation, forest resource management, and intelligent breeding. Traditional manual survey methods are inefficient and costly, while the AI-driven paradigm offers a new approach for intelligent forestry through computer vision technologies, enabling large-scale, high-precision tree species identification. However, fine-grained classification of tree species remains challenging due to high intra-class variation, low inter-class distinction, and complex environmental noise in real-world forest scenarios. To bridge the intra- and inter-class discrepancy, this study proposes a Forestry Fine-Grained Fusion Network (FFGF-Net), which integrates three key innovations: a CanopyPatchExtractor for local feature extraction, a CanopyTextureAnalyzer for multi-scale texture modeling, and a Cascaded Canopy Attention mechanism for noise suppression. For comprehensive evaluation, we introduce a newly collected UAV dataset, NJFUTreeData, and utilize established public datasets, including SZUTreeData and ETH dataset. Experimental results demonstrate that FFGF-Net achieves state-of-the-art performance, with Overall Accuracy (OA) of 75.00% on NJFUTreeData and 94.75% on the ETH dataset. Notably, our method consistently outperforms a wide range of existing model families, including representative CNN backbones (ConvNeXt-Tiny), lightweight architectures (e.g., EfficientNet-B0, MobileNet_V3-Small), and advanced Vision Transformers (e.g., Swin-Tiny). The proposed network provides a robust technical foundation for automated forest resource surveys, ecological monitoring, and smart forestry applications.

Keywords: Computer vision, Forest resource monitoring, Fine-Grained visual classification, Tree species, UAV remote sensing

1. Introduction

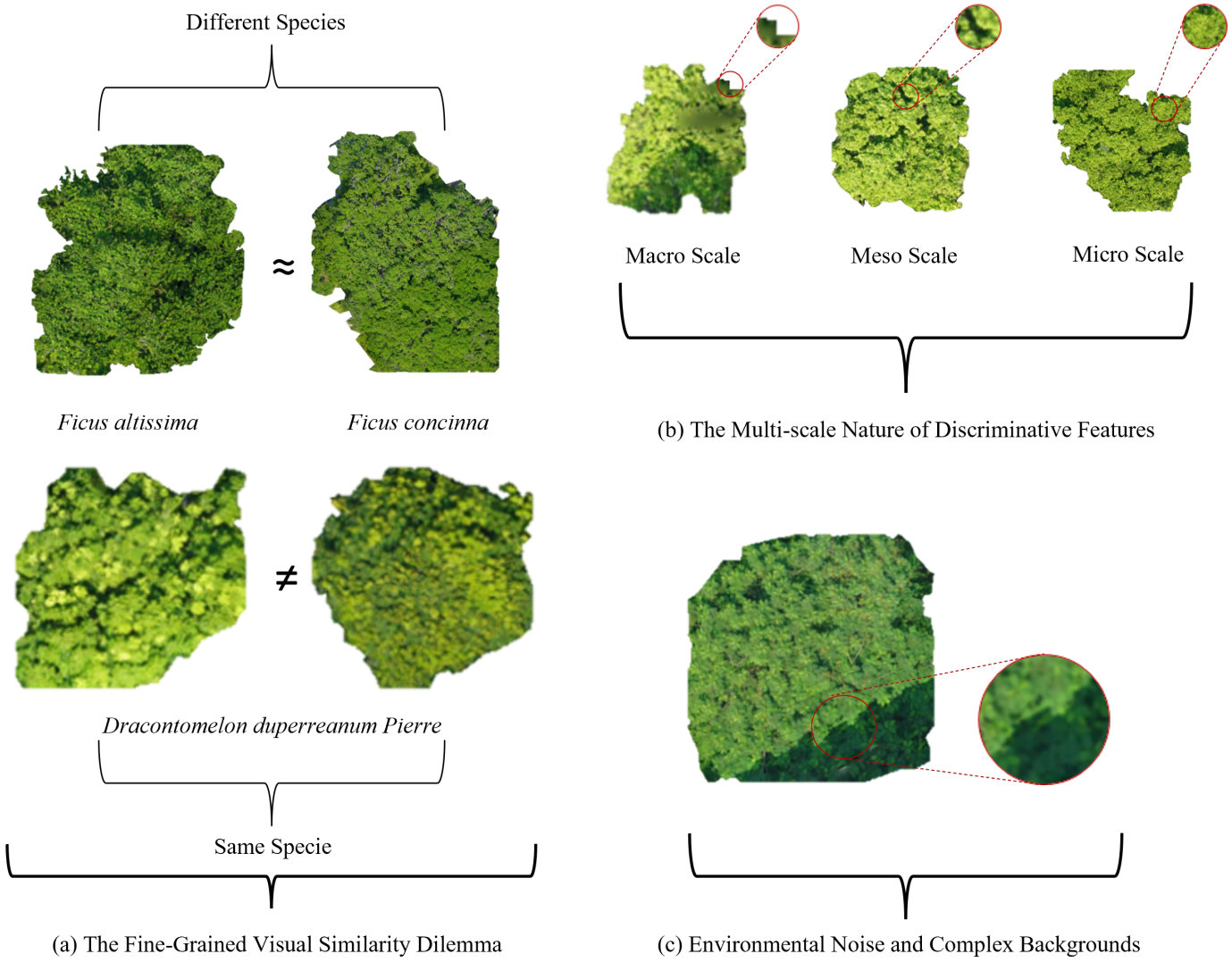
Accurate tree species classification is crucial for biodiversity conservation (Sun et al., 2019; Poorter et al., 2014), forest inventory (Fang et al., 2020), and forest breeding (Mace et al., 2012). Traditional manual survey methods suffer from low efficiency and high costs (Zou et al., 2020). The AI-for-Science paradigm offers a new path for forestry intelligence by enabling large-scale, high-precision tree species identification through computer vision technology, thereby fostering interdisciplinary innovation between forestry and artificial intelligence (Zhang et al., 2023).

Fine-grained tree species classification in forestry faces dual challenges: environmental complexity and biodiversity (Yu et al., 2023). In real-world forest scenarios, species identification must overcome dynamic factors such as varying illumination and occlusion interference, while simultaneously distinguishing between morphologically highly similar, closely related species (e.g., *Pinus spp.*, *Quercus spp.*). This places higher demands on the discriminative ability of classification models (Chen et al., 2023).

Current remote sensing-based tree species classification research primarily focuses on vegetation parameter inversion at the stand level or individual tree crown (contour) extraction. There is comparatively less involvement in fine-grained classification using high-resolution imagery at the crown scale. Current fine-grained classification in forestry primarily faces the following challenges: (1) Significant intra-class variance and small inter-class differences. (2) Discriminative features exist across multiple scales. (3) Complexity and noise in field environments (see Fig. 1).

To address these challenges, various deep learning approaches have been explored. In general visual classification, standard architectures like ResNet (He et al., 2016) and Vision Transformers (Dosovitskiy et al., 2020) extract global contextual features but often fail to adequately capture subtle local feature differences between sub-categories. Consequently, fine-grained visual classification (FGVC) has evolved to distinguish visually similar sub-categories (Zhao et al., 2017). In forestry remote sensing, initial works transferred pre-trained CNNs to tree crown imagery (Fricker et al., 2019) and evaluated

*Corresponding authors:
E-mail addresses:
hxyang@geoai.ecnu.edu.cn (H. Yang), xusheng@njfu.edu.cn (S. Xu)



(a) The Fine-Grained Visual Similarity Dilemma

(b) The Multi-scale Nature of Discriminative Features

(c) Environmental Noise and Complex Backgrounds

Figure 1: Challenges in fine-grained tree species classification

data augmentation strategies to improve adaptability (Neupane et al., 2021). As understanding deepened, specialized architectures emerged, such as the Bilinear Squeeze-and-Excitation Network (BiSENet) (He Z. & He D., 2020) and Transformer-based Multi-scale Feature Fusion (Dong et al., 2025). Furthermore, recent trends have explored multi-source data fusion, combining LiDAR with optical imagery for urban tree inventories (Ferreira et al., 2024; Liu et al., 2025).

Despite this significant progress, existing deep learning-based forestry identification research still lacks specific improvements for crown-scale imagery captured under complex field conditions. Most current methods are either limited to coarse-grained categorization or exhibit limited discriminative capability when differentiating visually similar, closely-related species. They predominantly rely on generic feature extraction mechanisms that struggle with severe background occlusion and are constrained by their inability to comprehensively capture and model subtle yet discriminative features across multiple scales.

In addition to overcoming the limitations of fine-grained visual similarity and texture, it is also necessary to consider the role of 3D structural clues. Structural attributes, such as crown geometry, canopy height variation, and vertical forest structure, are often crucial for accurate tree species discrimination. Recent large-scale datasets integrating structural forest attributes further emphasize the importance of 3D information for improving species discrimination (Ali et al., 2026; Calders et al., 2022; Hollaus & Chen, 2022; Weiser et al., 2022). Furthermore, recent studies have demonstrated that incorporating structural features derived from LiDAR significantly improves vegetation characterization and biomass estimation, highlighting the limitations of purely image-based approaches (Ali et al., 2024; Wild et al., 2026). Recognizing these structural dimensions provides a more comprehensive perspective for forest inventory challenges. However, despite the undeniable value of 3D data, its acquisition is often costly and complex to operate. Therefore,

maximizing the representativeness of cost-effective 2D optical images remains a highly practical and urgent research focus.

To bridge this specific gap and overcome the aforementioned challenges, this work introduces the Forestry Fine-Grained Fusion Network (FFGF-Net). The methodological novelty of our framework lies in its targeted design to address the shortcomings of existing models: (1) A CanopyPatchExtractor module that regularly segments canopy images into multiple local regions, mitigating large intra-class differences caused by growth environments while amplifying subtle local differences between species; (2) A CanopyTextureAnalyzer module, which utilizes parallel convolutional layers with different dilation rates to capture multi-scale features simultaneously, extracting texture information across different scales; (3) A Cascaded Canopy Attention mechanism that suppresses noise features representing background, shadows, and occlusions. On the NJFUTreeData dataset, the FFGF-Net achieved an accuracy of 75.00%, representing a noticeable improvement over existing state-of-the-art models and providing reliable technical support for forestry intelligence.

2. Study Area and Data

This section overviews the three forestry remote sensing datasets from distinct geographical regions: SZUTreeData (SZU) (http://szu-hsilab.com/szu-tree-dataset/), NJFUTreeData (NJFU), and ETH dataset (ETH) (https://form.ethz.ch/research/tree-ai-global-database/treeai-competition.html), the locations of which are mapped in Fig. 2. These datasets encompass a variety of tree species and forest conditions, providing a comprehensive basis for validation.

SZUTreeData represents the world's first multimodal remote sensing dataset designed for individual tree segmentation and species identification, featuring hyperspectral imagery, LiDAR data, and high-resolution optical imagery. Comprising tens of millions of annotated pixels across 25 common tree species, this dataset was collected and publicly shared by the Hyperspectral Remote Sensing Team at Shenzhen University using integrated UAV and ground surveys. It covers a total area of approximately 1.5 square kilometers and includes detailed information on various land cover types such as roads, buildings, forest stands, grassland, and water bodies. SZUTreeData serves as a benchmark for precise species identification and multi-modal classification research. Fig. 3 (a1-a2) displays the true-color images of SZUTreeRGB, the individual tree annotation map, and the tree species category map derived from the SZUTreeData (Jia et al., 2024; Li et al., 2024; Long et al., 2024; Zhang et al., 2024).

The ETH dataset, released by ETH Zurich, is a public dataset widely adopted in forestry remote sensing and computer vision research (Beloiu Schwenke et al., 2025). It integrates multi-source remote sensing data, including high-resolution optical imagery, LiDAR point clouds, and hyperspectral information, covering multiple tree species and stand types. Noted for its data diversity and annotation accuracy, the ETH dataset provides rich tree-level annotations such as crown contours, species labels, and height models, making it suitable for deep learning-based tasks including tree detection, segmentation, and classification. Its rigorous data acquisition and annotation protocols have established it as a benchmark dataset widely recognized by the international academic and industrial communities. Sample images from this dataset and their corresponding annotation schematics are exhibited in Fig. 3 (b1-b2).

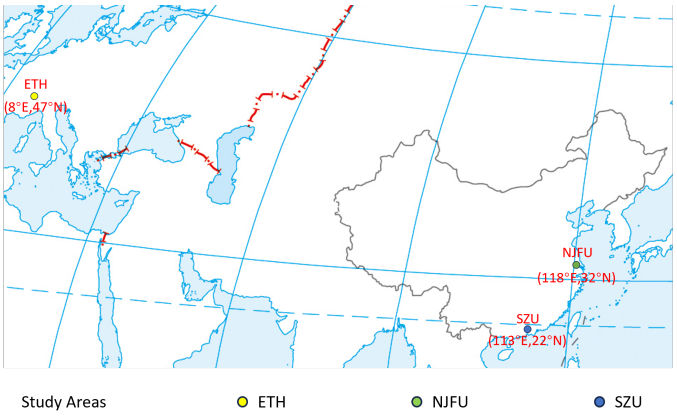


Figure 2: The geographical locations of study areas. The base map is sourced from the Standard Map Service website of the Ministry of Natural Resources of the People's Republic of China (Approval Number: GS(2016)2945).

Newly constructed specifically for this study, NJFUTreeData is a high-resolution UAV remote sensing dataset designed for forest resource investigation, focusing on the tree distribution area in Xuanwu District, Nanjing, China. Data acquisition was conducted using a DJI Matrice 3E series UAV equipped with an RTK module, flying at an altitude of 120 meters with a criss-cross flight plan. This rigorous acquisition strategy yielded an ultra-high spatial resolution of approximately 2 to 3 centimeters per pixel, ensuring the precise capture of subtle structural and textural characteristics of tree canopies. Furthermore, the dataset exhibits remarkable taxonomic diversity, encompassing 400 distinct plant samples distributed across 94 families and 222 genera. The dataset includes various data types, such as high-resolution orthoimagery and Digital Surface Models, providing detailed information on tree morphological structure, distribution density, and growth status. It is suitable for forestry remote sensing applications including individual tree segmentation, species identification, and biomass estimation, offering high-quality benchmark data for the precise monitoring and management of forest resources. In this paper, Fig. 3 (c1-c3) presents the overall orthomosaic of the dataset, sample cropped patches along with their corresponding annotation maps, and the final individual tree segmentation and species classification results.

3. Methodology

3.1. Problem Formulation

The task of automated tree species classification is fundamentally a Fine-Grained Visual Categorization (FGVC) problem. In contrast to generic image classification, which distinguishes broadly divergent categories (e.g., cats vs. dogs),

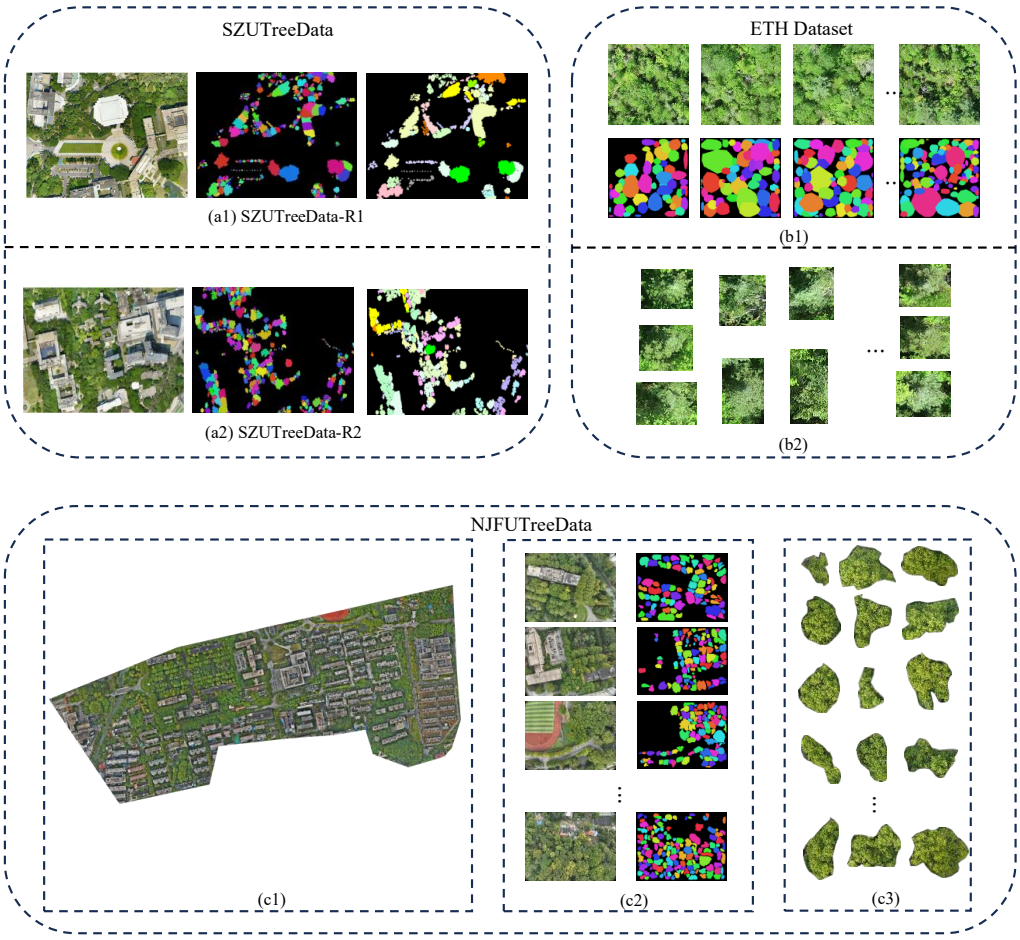


Figure 3: Three forestry remote sensing datasets used in this study. (a1) True-color images of SZUTreeRGB-R1 data, along with individual tree annotation map and tree species category map for SZUTreeDataR1 data; (a2) true-color images of SZUTreeRGB-R2 data, along with individual tree annotation map and tree species category map for SZUTreeDataR2 data. (b1) Individual tree annotation map for ETH dataset; (b2) the final map of individual tree segmentation. (c1) The overall orthomosaic of the NJFUTreeData dataset; (c2) individual tree annotation map for NJFUTreeData dataset; (c3) individual tree segmentation and crown contours.

FGVC requires discriminating between visually similar sub-categories under a common super-category. In the context of forestry, this entails differentiating among tree species (all belonging to the macro category "trees") based on subtle and localized discriminative features, such as leaf texture, crown structure, or branching pattern. These distinguishing characteristics are often confined to specific regions of the canopy, while significant intra-class variations may arise due to factors including growth environment, health status, and imaging conditions. Given a forestry image dataset $D = \{(I_n, y_n)\}_{n=1}^{N}$, where $I_n \in \mathbb{R}^{H\times W\times 3}$ represents the input image and $y_n \in \{1, 2, \ldots, C\}$ is the corresponding tree species label, Here, N denotes the total number of samples in the dataset, H and W represent the spatial height and width of the images, respectively, and C indicates the total number of tree species categories. The goal of this study is to construct a mapping function $F : \mathbb{R}^{H\times W\times 3} \rightarrow \{1, 2, \ldots, C\}$, such that the input image I is accurately mapped to the predicted class $\hat{y}$.

As illustrated in Fig. 4, the FFGF-Net executes the mapping function F through a multi-stage pipeline. A Backbone CNN (F_b) performs initial feature learning, succeeded by Gated Feature Reduction (F_d) for channel pruning. Multi-scale pattern analysis and feature recalibration are then concurrently managed by the CanopyTextureAnalyzer (F_t) and Cascaded Canopy Attention (F_a), respectively. The CanopyPatchExtractor (F_p) subsequently divides the enhanced features into localized tokens, with their mutual dependencies captured by Position-aware Patch Attention (F_s). The classification (F_c) component finalizes the process by integrating these tokens into a potent representation for species categorization.

In summary, the tree species classification problem in forestry demands that the model possesses the capability to discover and utilize subtle, localized discriminative features from highly similar categories, which is the core focus of fine-grained visual recognition research.

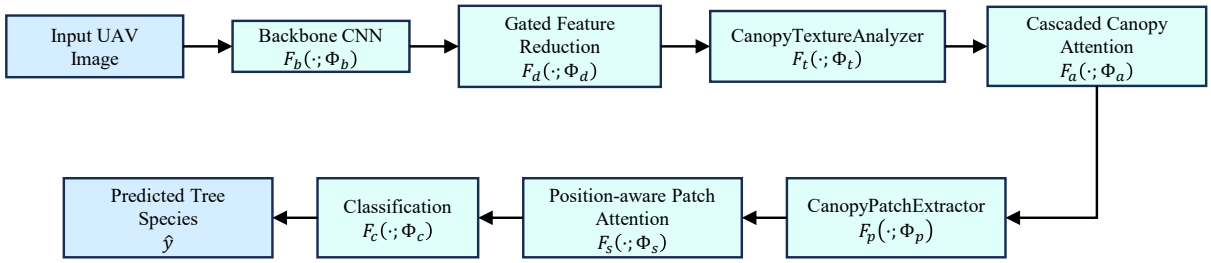


Figure 4: FFGF-Net architecture overview

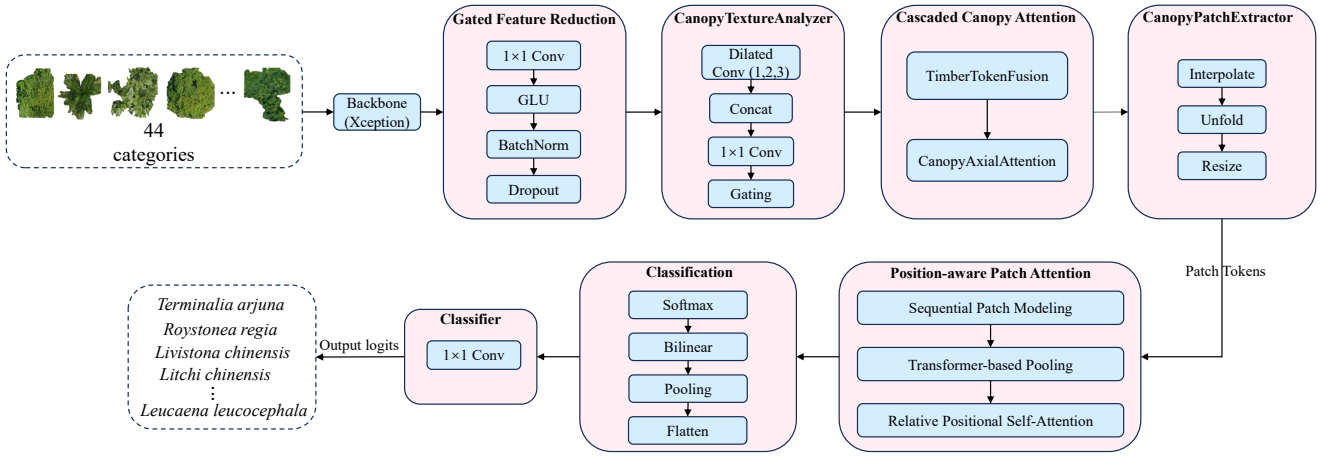


Figure 5: FFGF-Net: a local-global feature fusion network for fine-grained tree species classification from UAV imagery. The framework comprises a backbone CNN for feature extraction, a Gated Feature Reduction module for channel compression, a CanopyTextureAnalyzer for multi-scale texture modeling, a Cascaded Canopy Attention mechanism for discriminative region emphasis and noise suppression, a CanopyPatchExtractor for local token generation, a Position-aware Patch Attention block for spatial-semantic relation reasoning, and a Bilinear Attention Pooling module for feature aggregation and classification.

3.2. Gated Feature Reduction

High-dimensional features pose challenges in computational complexity and noise interference. Gated Feature Reduction (GFR) is inspired by GLU (Dauphin et al., 2017), strengthening the representation of foreground targets by compressing the feature dimensionality (as shown in Fig. 6). For an input feature $X \in R^{H\times W\times 3}$, GFR first maps it to a transitional channel space using a 1×1 convolution:

$$U = W_p * X + b_p, U \in R^{h\times w\times c'} \quad (1)$$

Here, $*$ denotes the convolution operation, $W_p \in R^{1\times 1\times c\times 2c'}$ represents the convolutional kernel parameters, b_p is the bias term, and c' is the designated target number of channels. Subsequently, the tensor U is split evenly along the channel dimension into two parts $U = [U_1, U_2]$, where $U_1, U_2 \in R^{h\times w\times c'}$.

The core gating operation is:

$$\hat{U} = \sigma(U_1) \odot U_2 \quad (2)$$

Where $\sigma(\cdot)$ is the Sigmoid function, which compresses the values of U_1 between 0 and 1, forming a soft selection mask. $\odot$ denotes element-wise multiplication. The features in U_2 are reweighted by this mask, suppressing background responses irrelevant to classification and highlighting feature patterns relevant to tree species discrimination.

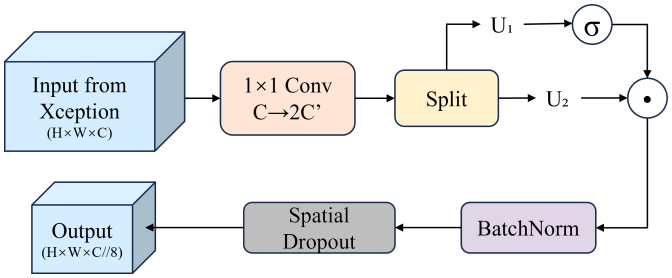


Figure 6: Gated Feature Reduction

Finally, standardization and regularization are applied via Batch Normalization (BatchNorm) and Spatial Dropout to obtain the compressed feature output.

This module significantly reduces computational resource consumption while endowing the model with powerful feature selection capabilities.

3.3. CanopyTextureAnalyzer

Leaf texture is an important basis for distinguishing tree species. To explicitly extract multi-scale texture patterns, we introduce the CanopyTextureAnalyzer (CTA) module. This process can be seen in Fig. 7. The module takes the input feature $\hat{X} \in R^{h\times w\times c}$, and extracts multi-scale contextual features through three parallel dilated convolution paths:

$$F_1 = \mathrm{ReLU}\left(W_{d1} * \hat{X}\right), \mathrm{Dilation} = 1 \tag{3}$$

$$F_2 = \mathrm{ReLU}\left(W_{d2} * \hat{X}\right), \mathrm{Dilation} = 2 \tag{4}$$

$$F_3 = \mathrm{ReLU}\left(W_{d3} * \hat{X}\right), \mathrm{Dilation} = 3 \tag{5}$$

Here, W_{d1}, W_{d2}, W_{d3} are the convolution weights. Different dilation rates allow the kernels to have varying receptive fields without increasing the number of parameters, thereby capturing texture structures from fine to coarse. Subsequently, the multi-scale features are concatenated and fused via convolution:

$$F_{\mathrm{concat}} = [F_1, F_2, F_3], F_{\mathrm{fused}} = W_f * F_{\mathrm{concat}} \tag{6}$$

Here, $[\cdot]$ denotes the concatenation operation along the channel dimension, and W_f is a 1×1 convolution kernel used to restore the number of channels to c.

Simultaneously, an independent gating pathway utilizes Global Average Pooling (GAP) to generate channel-wise statistical summaries and produces modulation weights through a two-layer fully connected network:

$$g = \sigma\left(W_2\left(\delta\left(W_1\left(\mathrm{GAP}\left(\hat{X}\right)\right)\right)\right)\right) \tag{7}$$

where δ is the ReLU activation function, W_1 and W_2 are the weights of the fully connected layers, and σ is the Sigmoid function. The final output of the module is a weighted residual sum of the input feature and the gated weighted texture features:

$$X_{\mathrm{textured}} = \hat{X} + \alpha \cdot (g \odot F_{\mathrm{fused}}) \tag{8}$$

α is a trainable scaling coefficient that controls the contribution strength of the texture features. This design injects enhanced multi-scale texture representations while preserving the integrity of the original information.

3.4. Cascaded Canopy Attention

Similar to Xu & Wan (2024) and Zhang et al. (2024), we construct a Cascaded Canopy Attention (CCA) mechanism composed of two cascaded attention mechanisms to enable the model to autonomously focus on discriminative regions.

1) TimberTokenFusion. This component achieves spatial and channel-aware feature modulation through convolution and gating. The input features are first projected into Query (Q), Key (K), and Value (V) triplets. Subsequently, Q and K are modulated via spatial and channel operations respectively (see Fig. 8) and fused via Depthwise Convolution (DWConv):

$$\mathrm{TTF}(X) = \gamma \cdot \mathrm{DWConv}\left(\mathcal{M}_{\mathrm{spa}}(Q) + \mathcal{M}_{\mathrm{cha}}(K)\right) \odot V + X \tag{9}$$

$\mathcal{M}_{\mathrm{spa}}$ and $\mathcal{M}_{\mathrm{cha}}$ represent Spatial and Channel modulation. γ is a learnable scalar parameter, and $\odot$ denotes element-wise multiplication. This operation allows the model to efficiently integrate spatial context and recalibrate channel responses.

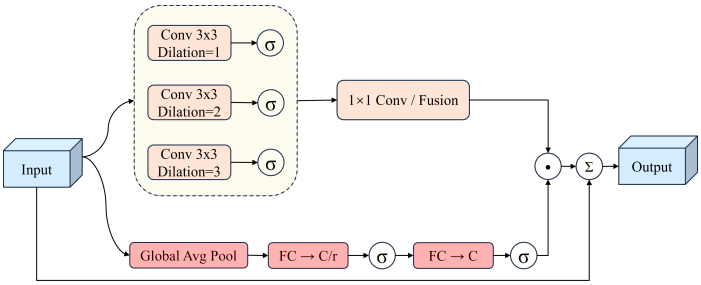


Figure 7: CanopyTextureAnalyzer

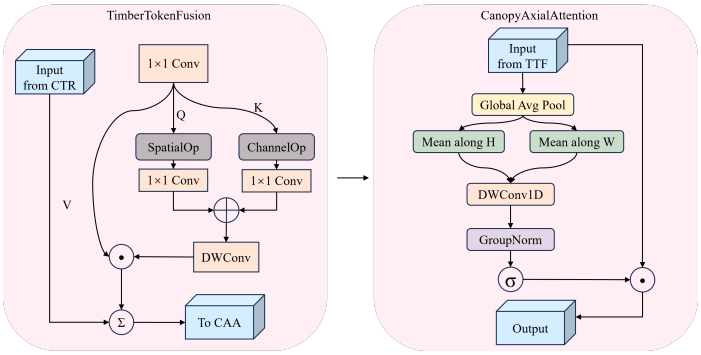


Figure 8: Cascaded Canopy Attention

2) CanopyAxialAttention. This component computes attention weights along the image height and width directions separately to establish long-range dependencies. First, Global Average Pooling is applied to obtain global context:

$$g = \mathrm{GAP}(X) \in \mathbb{R}^{c\times 1\times 1} \tag{10}$$

Then, directional features are extracted with background suppression:

$$x_h = \mathrm{Mean}_w(X \odot g) \in \mathbb{R}^{c\times h}, \quad x_w = \mathrm{Mean}_h(X \odot g) \in \mathbb{R}^{c\times w} \tag{11}$$

Shared-weight 1D depthwise convolution and Group Normalization are applied:

$$a_h(x) = \sigma\left(\mathrm{GroupNorm}\left(f_{\mathrm{dw1d}}(x_h)\right)\right) \tag{12}$$

$$a_w(x) = \sigma\left(\mathrm{GroupNorm}\left(f_{\mathrm{dw1d}}(x_w)\right)\right) \tag{13}$$

The final output is the product of the attention maps from the two directions:

$$\mathrm{CAA}(X) = a_h(x) \cdot a_w(x) \cdot X \tag{14}$$

This cascaded design enables the model to sequentially filter and reinforce features from different dimensions, thereby more accurately locating key regions.

3.5. CanopyPatchExtractor

Inspired by the image patching strategy in Vision Transformers (Dosovitskiy et al., 2020; Touvron et al., 2021), we design

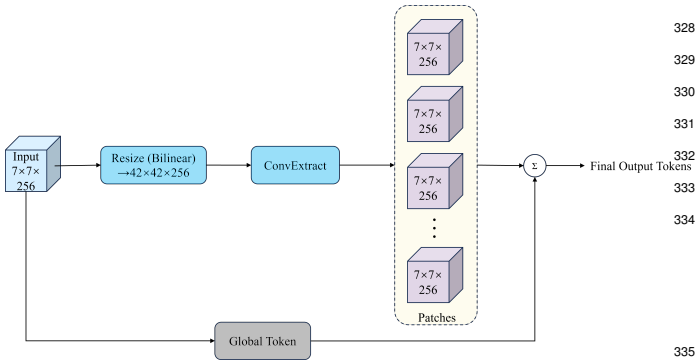


Figure 9: CanopyPatchExtractor

the CanopyPatchExtractor (CPE) module to structurally extract a series of overlapping local region tokens from the optimized feature map (see Fig. 9).

First, the feature map is upscaled to a preset resolution using an interpolation method to retain richer spatial details. A convolutional sliding-window operation with kernel size k then extracts multiple overlapping patches from the upscaled map, where the number of output channels corresponds to the desired patch count N. These patches are reorganized into a tensor and resized to a fixed spatial dimension $p \times p$ via bilinear interpolation, yielding $P \in \mathbb{R}^{B\times N\times d\times p\times p}$.

To preserve global context, we augment the patch sequence with an additional token derived from global average pooling of the original feature map, resulting in $N+1$ tokens for subsequent processing. This design enables the model to capture both local discriminative details and holistic scene statistics, effectively mitigating environment-induced intra-class variance while enhancing fine-grained inter-species cues.

3.6. Feature Encoding and Classification

In the final stage of our methodology, region tokens are transformed into discriminative representations for fine-grained classification. The Position-aware Patch Attention module employs sequential modeling to capture global semantic dependencies, utilizes Transformer-based pooling for feature compression, and incorporates relative positional encoding to establish spatial relationships, forming a hierarchical processing mechanism that maintains structural awareness while enabling comprehensive contextual reasoning (Behera et al., 2021). The classification module adopts bilinear interaction modeling to aggregate features, fusing second-order statistics between semantic distributions and token characteristics to generate compact global representations, which are subsequently projected to produce final classification outputs.

3.7. Evaluation Metrics

To comprehensively and objectively evaluate the performance of the proposed model on the fine-grained tree species classification task, this study employs a set of widely accepted evaluation metrics. Overall Accuracy (OA) is adopted as the primary and core metric for assessing model performance. Furthermore, to enable an in-depth analysis of the classification effectiveness and stability across different categories, supplementary metrics including Precision, Recall, and F1-Score are also utilized. OA directly measures the overall proportion of correctly classified samples out of the total predictions. The calculation formula is as follows:

$$\text{OA} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{15}$$

This metric reflects the overall classification correctness of the model and is the most intuitive and core performance measure for classification tasks. Precision is used to evaluate the accuracy of the model's positive predictions, i.e., the proportion of predicted positive samples that are truly positive. The calculation formula is:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{16}$$

where TP (True Positive) denotes the number of correctly predicted positive samples, and FP (False Positive) denotes the number of negative samples incorrectly predicted as positive. A high Precision indicates that the model is highly reliable when making positive predictions. Recall is used to assess the model's coverage ability for positive samples, i.e., the proportion of actual positive samples that are correctly predicted by the model. The calculation formula is:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{17}$$

where FN (False Negative) denotes the number of true positive samples incorrectly predicted as negative. A high Recall indicates that the model can effectively identify the majority of the target samples. F1-Score is the harmonic mean of Precision and Recall, used to comprehensively reflect the overall performance of the model. It is particularly important in fine-grained classification tasks where the sample distribution across classes may be imbalanced. The calculation formula is:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{18}$$

The overall performance values for the above metrics are obtained by calculating the Macro-average across all classes. This ensures an equal contribution from each category to the final metric, prevents dominant classes from skewing the evaluation results, and better aligns with the characteristics of fine-grained classification tasks.

3.8. Experiments Settings

We utilized a pre-trained Xception network as the backbone feature extractor. The input image resolution was uniformly adjusted to 42×42 pixels. The number of feature channels was compressed to 256, and the number of regional patches was set to 32. Regarding training hyperparameters: the batch size was set to 8; the SGD optimizer (with a momentum of 0.99) and cross-entropy loss function were employed; the initial learning

rate was set to 1e-4 and dynamically decayed using a Lambda scheduler; the model was trained for a total of 200 epochs; and the random seed was fixed to ensure reproducibility. During training, the latest model was saved every 10 epochs, and the best model was automatically saved based on the highest validation accuracy.

4. Results

4.1. Performance Comparison

To comprehensively evaluate the overall performance of FFGF-Net, we conducted extensive experiments on three challenging public datasets for fine-grained tree species classification: SZUTreeData, NJFUTreeData, and ETH dataset. These datasets vary in scale, image provenance, and habitat complexity, constituting a comprehensive testing platform. We selected representative advanced methods as baseline comparisons, including generic deep network backbones (e.g., EfficientNet-B0, ConvNeXt-Tiny) and recently published state-of-the-art (SOTA) methods. As shown in Table 1, compared to current state-of-the-art methods, the FFGF-Net demonstrates consistent and significant performance improvements across all evaluation metrics. This validates that our proposed architectural design, which combines the CanopyTextureAnalyzer and Cascaded Canopy Attention mechanisms, is more effective than existing techniques when handling the highly similar features between tree species.

To provide a visual comparison of the performance trends across forestry datasets, we plot the overall accuracy of each method on NJFUTreeData, SZUTreeData, and ETH dataset in Fig. 10. The line graph clearly illustrates that FFGF-Net achieves the highest accuracy on SZUTreeData and ETH dataset, with a notable lead over other methods, especially on ETH dataset where it reaches 94.75%. This graphical representation reinforces the superiority of FFGF-Net in forestry fine-grained classification.

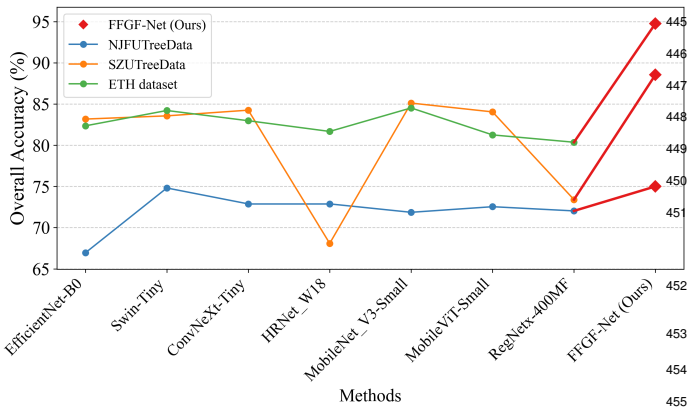


Figure 10: Performance comparison across forestry datasets

To evaluate the broad applicability of our method, FFGF-Net was evaluated against recent specialized models on standard fine-grained benchmarks including CUB-200-2011 (CUB), FGVC-Aircraft (Aircraft), Stanford Cars (Cars), and Stanford Dogs (Dogs). These models include Context-Aware Attentional Pooling (CAP) (Behera et al., 2021), Counterfactual Attention Learning (CAL) (Rao et al., 2021), Feature Fusion Vision Transformer (FFVT) (Wang et al., 2021), Spatial Relation-Aware Graph Neural Network (SR-GNN) (Bera et al., 2022)), MetaFormer (Diao et al., 2022), and Inter and Intra Region High-Order Feature Interaction (I2-HOFI) (Sikdar et al., 2025). As shown in Table 2, FFGF-Net delivers favorable performance in all evaluated domains, attaining state-of-the-art results on CUB and maintaining robust accuracy across the other datasets. It should be acknowledged, however, that despite its consistent outcomes on general fine-grained benchmarks, the method does not achieve top performance on certain generic tasks, suggesting potential for further optimization in broader applications. These findings nonetheless validate that the integration of multi-scale texture modeling with a cascaded attention architecture yields a generalizable and efficient strategy, which is primarily designed for forestry contexts while retaining adaptability to various fine-grained visual recognition challenges.

In addition to accuracy metrics, we further evaluated the computational efficiency of the compared models. Although generic methods provide powerful classification performance, they often come with higher computational costs. In contrast, our proposed method achieves highly competitive accuracy on all datasets, while significantly reducing the required parameter count and computational cost. The significant reduction in computational complexity means smaller memory usage and shorter theoretical inference time. Consequently, our method maintains a high balance between classification performance and model efficiency, demonstrating its strong feasibility in practical deployment.

Fig. 15 presents four line graphs comparing FFGF-Net with other state-of-the-art methods on general fine-grained benchmarks (CUB, Aircraft, Cars, Dogs). As illustrated, FFGF-Net achieves leading performance on CUB while maintaining consistent results across diverse tasks, further validating the rationality and practicality of our method's design in forestry image recognition. In this comparison, we adopted a strategy of directly citing literature results to ensure that all comparison models were in their optimal state after being optimized by their original authors. Although this results in missing metrics not mentioned in some literature, it provides a broader macro performance reference for our method.

4.2. Visualization Analysis

4.2.1. Confusion matrix analysis

Through an examination of the confusion matrix, the classification behavior of the FFGF-Net is elucidated for eight characteristic subtropical tree species in the NJFUTreeData collection from East China (refer to Fig. 11, which displays a representative subset of the species for visual clarity). The network effectively identifies species with distinctive morphological features, as demonstrated by the strong diagonal distribution observed in *Ginkgo biloba* and *Cinnamomum camphora*, indicating suc-

Table 1: FFGF-Net vs. SOTA in forestry fine-grained classification. The proposed Forestry Fine-Grained Fusion Network (FFGF-Net) is compared against various baseline models. Note: OA = Overall Accuracy. GFLOPs stands for Giga Floating-point Operations Per Second, Param. indicates the number of model parameters in millions, and IPS indicates Images Per Second. To provide a rigorously fair, head-to-head evaluation, all models in this table were locally re-implemented and evaluated by us under strictly identical training and testing conditions.

Methods	**OA (%)**			**Precision (%)**			**Recall (%)**			**F1-score (%)**			**GFLOPs**	**Param. (M)**	**IPS**
	NJFU	**SZU**	**ETH**	**NJFU**	**SZU**	**ETH**	**NJFU**	**SZU**	**ETH**	**NJFU**	**SZU**	**ETH**			
EfficientNet-B0 (2019)	66.95	83.17	82.35	59.61	73.67	76.68	57.31	68.59	73.30	57.86	71.73	74.08	0.4	5.3	742
Swin-Tiny (2021)	74.81	83.56	84.21	58.13	76.23	81.09	56.38	70.84	75.11	56.78	73.14	77.67	2.9	19.6	374
ConvNeXt-Tiny (2022)	72.88	84.25	82.97	54.35	77.29	75.64	50.82	72.80	74.58	52.26	74.83	74.90	4.5	28.6	451
HRNet-W18 (2019)	72.88	68.09	81.67	53.73	50.37	78.06	54.47	51.25	72.31	53.83	51.09	74.69	4.4	21.3	134
MobileNetV3-Small (2019)	71.86	85.11	84.52	56.52	63.32	82.26	54.40	60.68	77.42	54.59	61.34	79.63	0.1	1.5	1116
MobileViT-Small (2021)	72.54	84.04	81.25	53.64	70.81	77.83	52.68	68.26	75.19	52.65	69.53	76.87	1.4	5.6	570
RegNetx-400MF (2020)	72.03	73.40	80.36	40.26	58.14	71.47	49.68	54.99	70.64	44.37	56.53	70.52	4.0	22.0	338
FFGF-Net (Ours)	**75.00**	**88.54**	**94.75**	**67.05**	**81.25**	**89.24**	**62.25**	**74.08**	**83.61**	**64.30**	**77.26**	**85.86**	4.7	22.7	260

Table 2: FFGF-Net vs. SOTA in fine-grained visual classification. Note: To accurately reflect the optimal capabilities of these specialized architectures, baseline results are directly cited from their respective original publications. As underlying experimental conditions and hardware environments may vary across these independent studies, this table serves as a broader literature reference rather than a strictly controlled evaluation. Consequently, missing data points indicate specific metrics that were unavailable in the original papers.

Methods	**CUB OA (%)**	**Aircraft OA (%)**	**Cars OA (%)**	**Dogs OA (%)**	**GFLOPs**	**Param. (M)**
CAP (Behera et al. 2021)	91.8	95.7	94.1	96.1	10.2	34.2
generic CAL (Rao et al. 2021)	90.6	94.2	95.5	-	40.0	55.7
FFVT (Wang, Yu, and Gao 2021)	91.6	-	-	91.5	16.3	87.5
SR-GNN (Bera et al. 2022)	91.9	95.4	96.1	**97.3**	9.8	30.9
MetaFormer (Diao et al. 2022)	91.8	-	95.1	-	16.9	81.0
I2-HOFI (Sikdar et al. 2024)	91.6	**96.0**	**96.4**	-	14.2	**22.6**
FFGF-Net (Ours)	**92.4**	95.1	95.9	96.7	**4.7**	22.7

cessful learning of their discriminative leaf patterns and canopy structures.

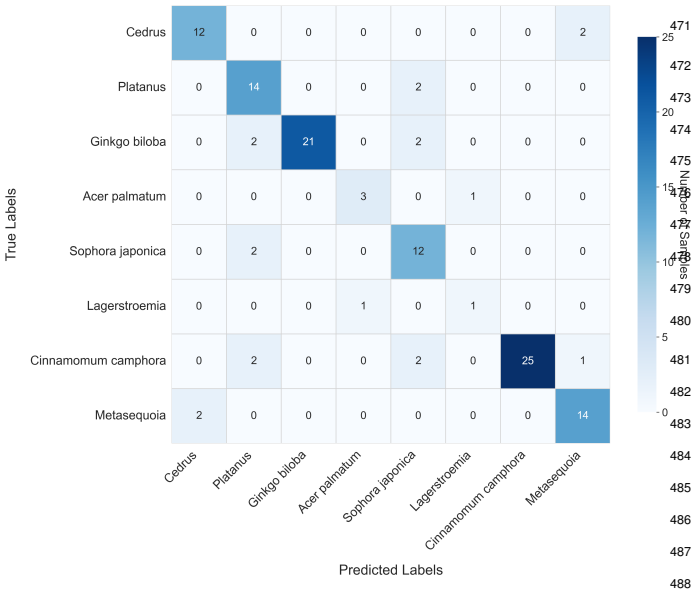


Figure 11: Confusion matrix of FFGF-Net tree species classification on the NJFUTreeData

The analysis identifies mutual misclassification between coniferous species *Cedrus* and *Metasequoia glyptostroboides*, highlighting the classification challenges posed by similar structural characteristics in needle-leaved trees within UAV imagery. Broadleaf species *Platanus* and *Sophora japonica* also exhibit reciprocal misclassification, revealing limitations in distinguishing deciduous species with comparable canopy textures.

Notably, the confusion between ornamental species *Acer palmatum* and *Lagerstroemia indica* is exacerbated by limited sample availability. *Cinnamomum camphora*, despite having the largest sample size, shows misclassification across multiple categories, suggesting either significant intra-species morphological variation or feature overlap with other broadleaf species. These systematic error patterns provide clear directions for model optimization, particularly in enhancing the recognition of subtle features among morphologically similar tree species.

As shown in Fig. 12, the confusion matrix analysis was conducted on a selected, representative subset of tree species from the SZUTreeData. The results demonstrate that the model exhibits strong and consistent discriminative capabilities for most of the selected species. For instance, all test samples of *Firmiana simplex*, *Litchi chinensis*, *Araucaria cunninghamii*, *Livistona chinensis*, and *Cinnamomum camphora* tree were classified correctly without any misidentification, underscoring the highly discriminative representations learned by the model for these categories.

However, the matrix also reveals subtle confusion between certain visually similar species pairs. Specifically, mutual misclassification occurs between *Flcus spp.* and *Flcus microcarpa*, and one sample of *Flcus microcarpa* is incorrectly predicted as *Livistona chinensis*. Furthermore, one misclassification case is observed between *Melaleuca cajuputi* and *Cinnamomum cam-*

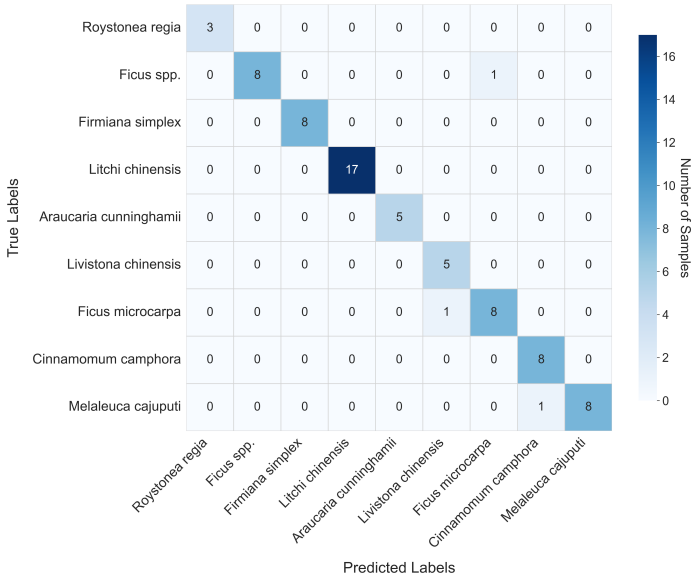


Figure 12: Confusion matrix of FFGF-Net tree species classification on the SZUTreeData

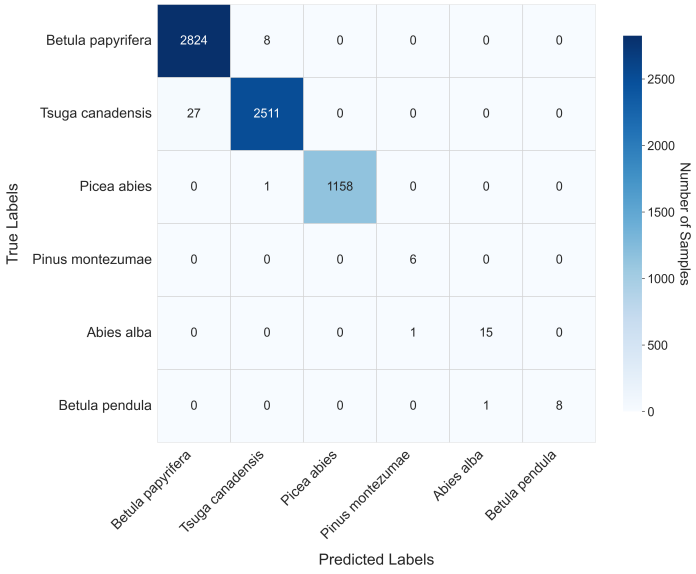


Figure 13: Confusion matrix of FFGF-Net tree species classification on the ETH dataset

phora. These errors likely stem from the intrinsic similarities in fine-grained features such as canopy texture and structure among these species, compounded by environmental noise common in UAV imagery, such as occlusion and varying illumination conditions, which exacerbates the challenge of fine-grained distinction.

Although the analysis is confined to a representative subset, the results robustly indicate that the FFGF-Net possesses commendable classification consistency and robustness within complex forestry environments.

The confusion matrix analysis on the ETH dataset reveals the exceptional performance and specific challenges of the FFGF-Net in fine-grained classification of temperate tree species (see Fig. 13). Due to the large number of categories, only a representative subset of the tree species is visualized in the matrix to ensure readability. The network achieved a leading OA of 94.75% on the test set, demonstrating its powerful overall discriminative capability. The analysis indicates that the model exhibits near-perfect distinguishability for most species. Nearly all samples of *Betula papyrifera*, *Tsuga canadensis*, and *Picea abies* were classified correctly, with only minimal misidentifications. This confirms that the canopy features extracted by the FFGF-Net are highly robust for differentiating these common, yet morphologically diverse, species in temperate forests.

Nevertheless, the matrix captures rare misclassification cases, which are highly representative and concentrated on rare species with extremely limited samples. For instance, all 6 samples of *Pinus montezumae* were incorrectly predicted as *Abies alba*; conversely, one *Abies alba* sample was also misclassified as *Pinus montezumae*. This indicates the model's difficulty in learning discriminative features for very uncommon species. This confusion likely stems from a "long-tail distribution" problem, where insufficient training samples for rare categories prevent the model from adequately learning their unique fine-grained characteristics.

In conclusion, the FFGF-Net achieves state-of-the-art classification performance on the ETH dataset, with particularly accurate and stable identification of dominant tree species. Simultaneously, the results clearly highlight the current model's limitation in handling extremely imbalanced, rare species. This provides a clear direction for future work, suggesting the integration of targeted re-sampling strategies or few-shot learning techniques to enhance the recognition capability for rare species.

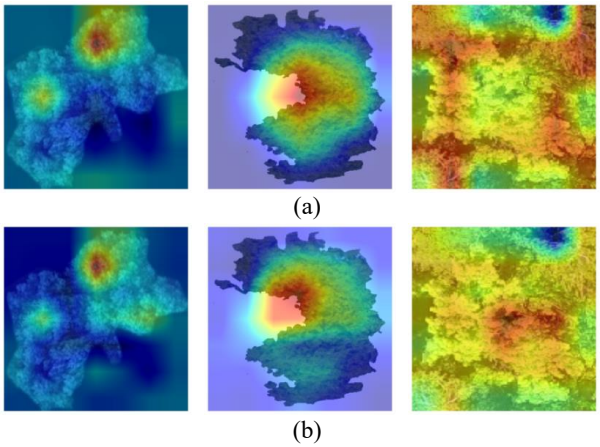


Figure 14: Cascaded Canopy Attention (CCA) Module Attention Visualization. (a) TTF Attention Responses. (b) CAA Attention Responses.

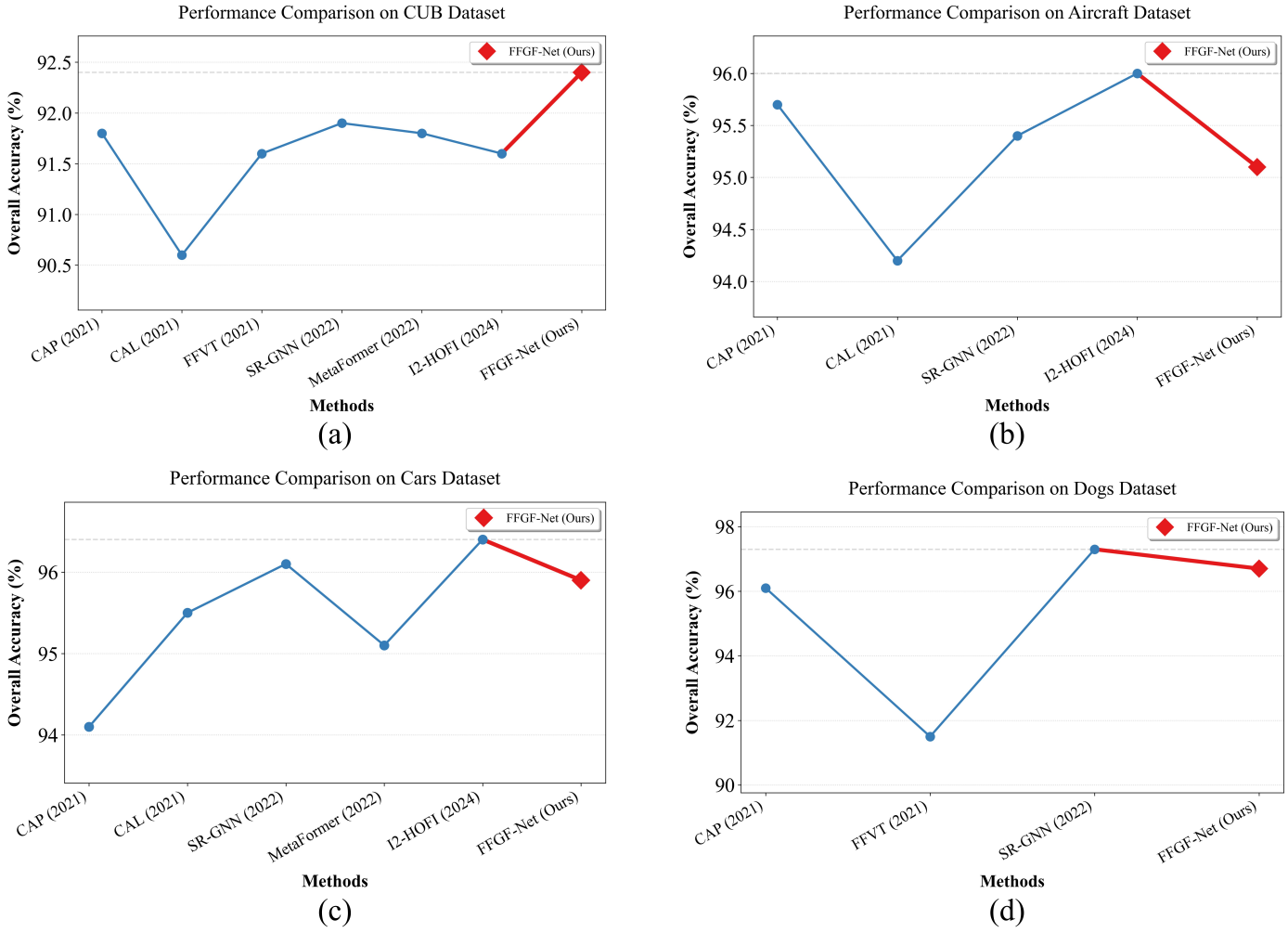


Figure 15: Performance comparison across general datasets. Note: Missing data points for certain baseline algorithms on specific datasets indicate that those performance metrics were not reported in their original publications. To ensure a fair and standardized evaluation, all baseline results presented here are directly cited from their respective original papers.

4.2.2. *Module response visualization*

To empirically validate the capability of the Cascaded Canopy Attention (CCA) module in enhancing discriminative regions and suppressing noise, we performed a systematic visualization of the attention heatmaps generated by its two core sub-modules: TimberTokenFusion (TTF) and CanopyAxialAttention (CAA). The results are presented in Fig. 14.

An analysis of the spatial distribution characteristics within the attention heatmaps reveals that TTF and CAA exhibit complementary and cascading modulation effects. The TTF module, employing spatial-channel joint modulation, primarily amplifies local textural regions inside the canopy that demonstrate high discriminative power. Subsequently, the CAA module refines the global structural coherence of the attention distribution through axial attention computation along the height (H) and width (W) dimensions. By integrating TTF's local discriminative enhancement with CAA's global structural regularization, the CCA module establishes a progressive attention refinement mechanism. This architecture provides essential support for fine-grained classification within complex forestry environments.

4.2.3. *Comparison of t-SNE visualization*

To quantitatively analyze the contribution of each module in the FFGF-Net to the enhancement of feature discriminability, this study employed the t-SNE (t-distributed Stochastic Neighbor Embedding) algorithm to project high-dimensional features into a two-dimensional space. A comparison was conducted between the feature distributions of the "Xception Backbone Only" and the "Full FFGF-Net", as illustrated in Fig. 16.

The visualization results clearly demonstrate that the features learned by the full FFGF-Net exhibit more distinct intra-class clustering and inter-class separation compared to those extracted using only the backbone network. Specifically, when solely relying on the backbone, feature points of different tree species show substantial overlap with blurred boundaries, indicating limited discriminative power. In contrast, after process-

ing with FFGF-Net, samples belonging to the same class form tighter clusters, while the margins between different categories are markedly enlarged. These findings empirically confirm that the introduced mechanisms, including local texture modeling, attention modules, and feature fusion strategies, collectively contribute to enhancing the model's ability to capture subtle inter-class distinctions. The FFGF-Net effectively improves the discriminative quality of the learned features, thereby providing a more reliable representation for high-accuracy fine-grained classification.

4.2.4. Comparative performance visualization

As illustrated in Fig. 17, this study conducts comprehensive visual comparisons with three baseline methods on the SZUTreeData dataset, which comprises 25 tree species characterized by complex canopy structures and diverse morphological features. The visualization scheme employs a standardized color coding: ground truth labels are displayed in blue, correct predictions in green, and incorrect predictions in red. Representative samples were randomly selected for analysis, with each sample including both the original canopy image and an enlarged region of interest (ROI) that highlights texture details crucial for species discrimination. The experimental results demonstrate that FFGF-Net exhibits superior performance in recognizing species with distinctive morphological signatures and maintains high confidence scores for correct predictions, indicating robust feature representation learning. Furthermore, texture ROI analysis confirms that FFGF-Net effectively focuses on discriminative canopy regions by leveraging both global structural patterns and local texture details.

However, the analysis reveals persistent challenges in handling taxonomically complex scenarios. A notable limitation emerges in distinguishing congeneric species with high visual similarity, such as *Ficus virens* and *Ficus microcarpa* within the Moraceae family. These closely related species share numerous morphological traits inherited from common ancestors, causing models to preferentially learn these dominant shared characteristics while overlooking discriminative local features (e.g., leaf size, apex morphology) when training samples are insufficient for fine-grained differentiation. This phenomenon leads to "intra-genus confusion" in classification.

A more complex challenge arises from convergent evolution, where phylogenetically distant species develop similar morphological structures due to adaptation to analogous environments. For instance, *Terminalia arjuna* (Combretaceae) and *Cinnamomum camphora* (Lauraceae) both exhibit "full and regular" broad-ovate crown structures as a convergent adaptation to ecological niches such as light competition. During training, models tend to prioritize capturing these macroscopic, dominant "ecotype" features while struggling to focus on taxonomically significant subtle distinctions like bark texture and phyllotaxy, resulting in "inter-family confusion".

The underlying mechanism for these misclassifications can be attributed to the FFGF-Net's feature learning priorities. Many critical discriminative features belong to fine-grained or local characteristics. When the backbone network's receptive field is either too large or too small, it may fail to effectively capture these local high-resolution patterns, instead overrelying on macroscopic features such as overall crown shape. This fundamental limitation explains the systematic misclassification of different species that share similar macroscopic morphological structures.

These challenging cases highlight promising directions for future research in developing more sophisticated feature representation techniques that can better handle fine-grained botanical classification, particularly through enhanced local feature extraction and multi-scale representation learning.

4.3. Ablation Study

This section presents an ablation study (see Table 3), evaluating the contributions of the CanopyTextureAnalyzer (CTA), Cascaded Canopy Attention (CCA), CanopyPatchExtractor (CPE), and Bilinear Attention Pooling (BAP) modules. The baseline model (utilizing only BAP) achieved the lowest accuracy. Incorporating CTA alone, which extracts multi-scale textures, yielded a marginal performance gain. The combination of CTA and CPE (without BAP) led to improved performance on the SZU dataset, highlighting CPE's capability in discerning local discrepancies. The CCA+BAP combination further enhanced accuracy, demonstrating CCA's efficacy in suppressing noise inherent in field environments. The full configuration of our model achieved superior performance on the SZUTreeData.

To rigorously verify the effectiveness of the proposed modules, we conducted robustness analysis on the ablation results. Specifically, all models were evaluated in three independent runs, each using a different random seed. The results demonstrate that the performance improvements yielded by our proposed components significantly exceed the standard deviations. This variability measurement confirms that the observed improvements in accuracy are robust and empirically significant, rather than just a result of random variance.

Table 3: FFGF-Net ablation study. To ensure statistical rigor, all models were evaluated over three independent runs with different random seeds. Results are reported as Mean ± Std.

CTA	CCA	CPE	BAP	NJFU	SZU	ETH
			✓	70.42±0.14	81.25±0.11	85.36±0.12
✓			✓	71.83±0.10	82.29±0.15	86.91±0.13
✓		✓		72.54±0.14	84.38±0.11	88.25±0.16
		✓	✓	73.16±0.13	84.38±0.13	89.47±0.12
	✓		✓	73.89±0.15	85.42±0.15	91.32±0.13
✓	✓	✓	✓	**75.00±0.14**	**88.54±0.17**	**94.75±0.14**

5. Discussion

5.1. Discussion

The proposed FFGF-Net was developed to overcome the inherent challenges of fine-grained tree species classification using UAV imagery, contributing a novel methodological approach to the broader field of Earth Observation (EO). Our framework shifts the paradigm towards domain-specific structural modeling. Experimental evaluations confirm that domain-specific feature extraction modules, namely the CanopyPatchExtractor and CanopyTextureAnalyzer, achieve substantial performance improvements over standard CNN and Vision

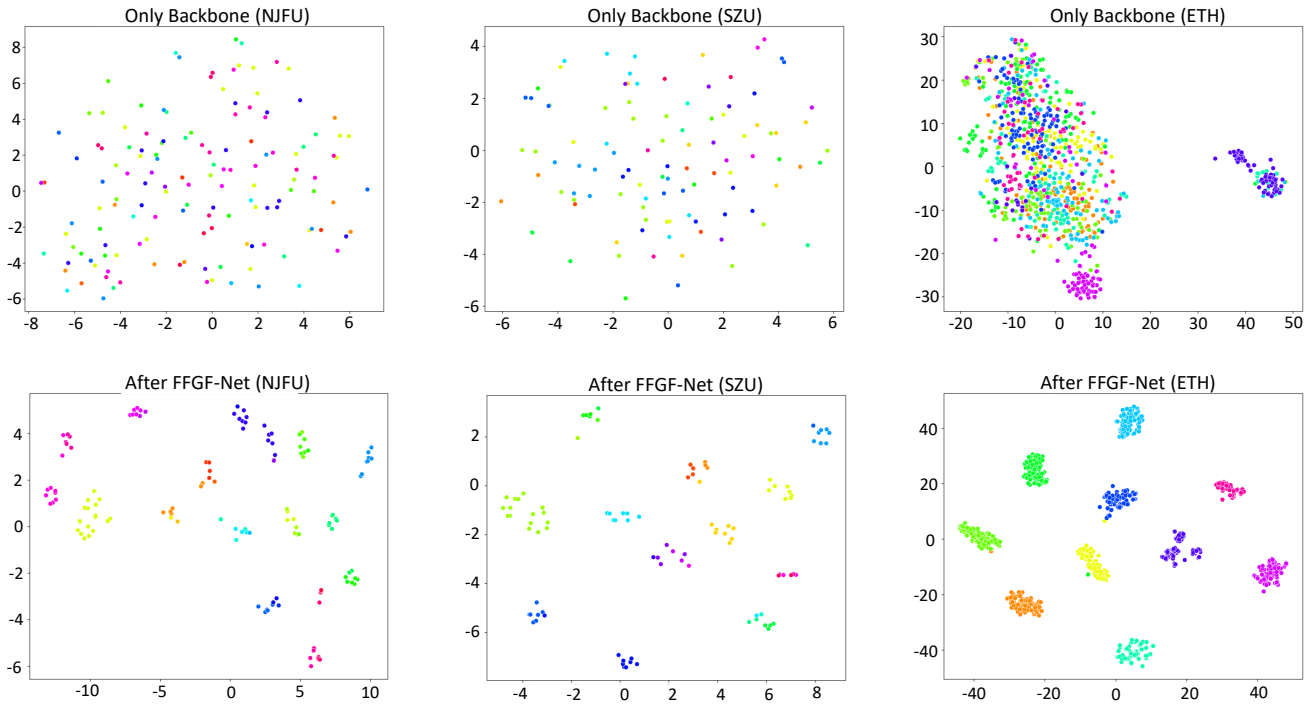


Figure 16: t-SNE visualization of the feature embedding space

Transformer baselines. By isolating subtle discriminative characteristics from complex canopy backgrounds and explicitly modeling multi-scale textures, this framework effectively differentiates morphologically similar and closely related species.

Regarding operational deployment and existing technologies, recent state-of-the-art forestry classification research increasingly relies on multi-modal fusion strategies (e.g., combining LiDAR point clouds with optical imagery) to capture highly discriminative 3D canopy structures (Ferreira et al., 2024). As highlighted by Ali et al. (2024) and Wild et al. (2026), the integration of such 3D structural features offers unparalleled advantages in vegetation characterization and biomass estimation over purely image-based methods. Although these fusion methods achieve exceptional accuracy in multi-layered canopies, they inherently suffer from prohibitive data acquisition costs, the need for bulky specialized UAV payloads, and highly complex, computationally intensive preprocessing pipelines. As a purely image-based framework, our proposed FFGF-Net provides a highly accessible and cost-effective alternative for large-scale surveys. By directly extracting crucial spectral and textural features from high-resolution UAV imagery, our method maximizes the potential of two-dimensional visual characteristics to document species-specific phenotypic traits. This image-centric solution maintains practical feasibility while offering an efficient technical pathway for routine forest resource monitoring (Sothe et al., 2019; Wallace et al., 2020).

Beyond classification accuracy, the practical application value of FFGF-Net lies in its computational efficiency. Forestry inventory applications require rapid processing of large orthomosaic datasets on resource-limited devices. FFGF-Net maintains a highly favorable balance between performance and overhead, requiring only 4.7 GFLOPs and 22.7 million parameters. This lightweight architecture enables deployment on edge platforms like UAV onboard processors or mobile field tablets. An inference speed of 260 images per second delivers near real-time processing. This high-throughput efficiency overcomes the bottleneck of manual visual interpretation, positioning FFGF-Net as a deployable tool for automated large-area forest management.

Crucially, the extensibility of the FFGF-Net framework warrants emphasis. While currently optimized and validated on high-resolution UAV RGB datasets, the modular architecture holds significant potential for integrating emerging remote sensing data sources. For instance, coupling this texture-aware framework with hyperspectral imagery or sub-meter commercial satellite data could effectively address existing spectral data gaps, providing the critical physiological dimensions currently absent in standard RGB sensors. In addition, the core mechanism of localizing and analyzing fine-grained structural textures demonstrates high potential for broader transferability. Beyond forestry inventories, this EO-centric framework could be adapted to support wider applications, such as agricultural

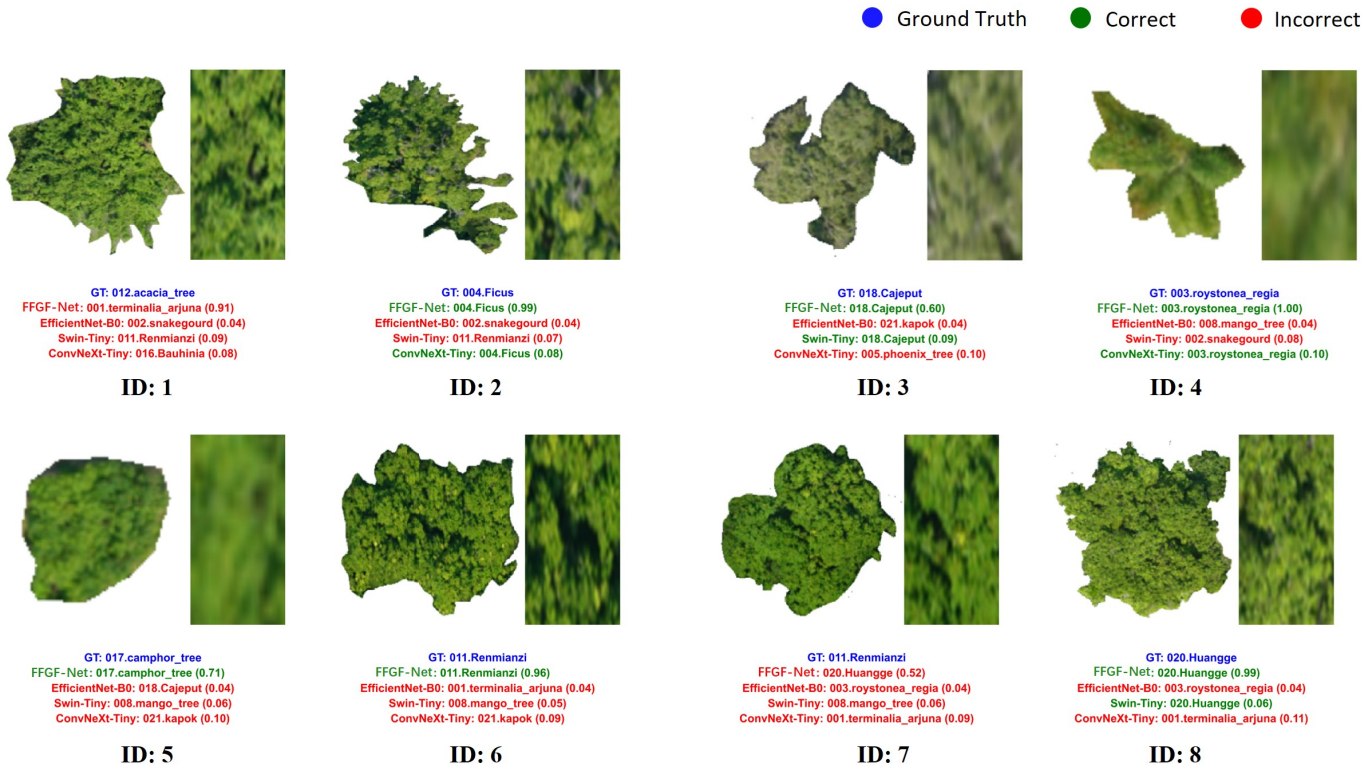


Figure 17: Comparative visualization of classification performance across different methods. In the visualization, blue text indicates the ground truth labels, while green and red texts represent correct and incorrect predictions, respectively.

yield estimation, urban green space phenotyping, and invasive plant species tracking across diverse geographical datasets.

5.2. Limitations

Despite the robust performance achieved by the FFGF-Net, several limitations remain, which also delineate directions for future research. Currently, as a purely image-based model, our framework's primary drawback compared to multi-modal approaches is the absence of explicit three-dimensional spatial metrics. Consequently, in extremely dense forests with severe canopy overlap, the performance ceiling of FFGF-Net may still be constrained. Future research should explore lightweight multi-modal fusion strategies that combine our efficient 2D architecture with LiDAR point-cloud data to further enhance classification accuracy without entirely sacrificing deployment convenience.

Beyond structural constraints, a significant challenge is the prevalent "long-tail distribution" problem in forest environments, where numerous rare tree species are represented by extremely limited samples. Future research must focus on few-shot or even zero-shot learning paradigms to enhance the model's capability for recognizing rare species.

Furthermore, the current model primarily relies on static imagery for identification and does not incorporate dynamic information related to tree canopy phenological changes. Developing temporal-aware models capable of integrating multi-temporal remote sensing data will be crucial for achieving more stable classification and enabling early warning of pest and disease outbreaks.

Ultimately, the scope of the present study is restricted to taxonomic identification. Advancing toward comprehensive ecological monitoring necessitates extending the current architecture to assess physiological status. Developing inversion models to estimate biophysical parameters, directly from high-resolution UAV imagery represents a crucial trajectory for subsequent research.

6. Conclusion

This study focused on addressing the critical challenges of fine-grained tree species classification in forestry using computer vision techniques. Specifically, we aimed to overcome the inherent difficulties of crown-scale remote sensing imagery, including low feature discriminativity, high inter-class similarity, and complex noise in field environments.

To achieve this objective, we proposed a novel Forestry Fine-Grained Fusion Network (FFGF-Net) and constructed a new UAV-based dataset, NJFUTreeData. Our primary scientific contributions are twofold. First, the introduction of the NJFUTreeData, combined with public datasets such as SZUTreeData and ETH dataset, establishes a comprehensive multi-source evaluation framework, offering invaluable data assets for the research community. Second, the FFGF-Net effectively addresses the aforementioned challenges through its tailored ar-

chitecture. By integrating the CanopyPatchExtractor for local feature isolation, the CanopyTextureAnalyzer for multi-scale perception, and a Cascaded Canopy Attention mechanism for noise suppression, the framework significantly enhances the capacity to capture subtle discriminative features. Consequently, the model demonstrates robust and highly accurate performance in distinguishing visually similar tree species under challenging in-situ conditions.

The implications of this research are substantial from both theoretical and practical perspectives. Theoretically, the FFGF-Net offers a novel technical strategy for tackling common bottlenecks in fine-grained visual classification, shifting the paradigm towards domain-specific structural modeling. Practically, this purely image-based framework provides a reliable algorithmic foundation for high-precision, automated forest resource inventories and ecological monitoring. Ultimately, this research fosters the deep integration of computer vision technology within forestry applications, accelerating the digital transformation of intelligent forest management practices. Furthermore, a detailed analysis of real-world application scenarios, including the Hegyi competition index calculations for specific thinning strategies, is provided in the Supplementary Material.

7. Acknowledgments

This work was supported by the Key Research Projects of Yibin, Research and Integrated Demonstration of Key Technologies for Smart Bamboo Industry (YBZD2024-1), the Practice Innovation Training Program Projects for Jiangsu College Students under Grant 202510298056Z, the Biological Breeding-National Science and Technology Major Project (2023ZD0405605), the Shanghai Science and Technology Program (Grant No.25ZR1402133) and the China Postdoctoral Science Foundation (Grant No.GZC20250239).

8. Data Availability

To support the transparency of this research, a representative subset of the NJFUTreeData dataset and the core network implementation of the FFGF-Net are currently available at https://github.com/zxz-xy/anonymous_code. The complete full-scale dataset, comprehensive training scripts, and pretrained models will be made fully publicly available upon the formal publication of this article.

References

Sun, Y., Huang, J., Ao, Z., et al., 2019. Deep learning approaches for the mapping of tree species diversity in a tropical wetland using airborne LiDAR and high-spatial-resolution remote sensing images. Forests 10(11), 1047. https://doi.org/10.3390/f10111047

Poorter, L., McNeil, A., Hurtado, V. H., et al., 2014. Bark traits and life-history strategies of tropical dry- and moist forest trees. Funct. Ecol. 28(1), 232-242. https://doi.org/10.1111/1365-2435.12158

Fang, F., McNeil, B.E., Warner, T.A., et al., 2020. Discriminating tree species at different taxonomic levels using multi-temporal WorldView-3 imagery in Washington DC, USA. Remote Sens. Environ. 246, 111811. https://doi.org/10.1016/j.rse.2020.111811

Mace, G.M., Norris, K., Fitter, A.H., 2012. Biodiversity and ecosystem services: A multilayered relationship. Trends Ecol. Evol. 27(1), 19-26. https://doi.org/10.1016/j.tree.2011.08.006

Zou, F., Li, A., Wang, Q., 2020. Comparative analysis of new forest resources survey methods. J. Phys. Conf. Ser. 1646(1), 012007. https://doi.org/10.1088/1742-6596/1646/1/012007

Zhang, X., Zhang, T., Wang, G., et al., 2023. Remote sensing object detection meets deep learning: A metareview of challenges and advances. IEEE Geosci. Remote Sens. Mag. 11(4), 8-44. https://doi.org/10.1109/MGRS.2023.3312347

Yu, F., Zhang, Q., Xiao, J., et al., 2023. Progress in the application of CNN-based image classification and recognition in whole crop growth cycles. Remote Sens. 15(12), 2988. https://doi.org/10.3390/rs15122988

Chen, M., Lin, M., Li, K., et al., 2023. CF-ViT: A general coarse-to-fine method for vision transformer. Proc. AAAI Conf. Artif. Intell. 37(6), 7042-7052. https://doi.org/10.1609/aaai.v37i6.25860

He, K., Zhang, X., Ren, S., et al., 2016. Deep residual learning for image recognition. Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 770-778. https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html

Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. https://doi.org/10.48550/arXiv.2010.11929

Zhao, B., Feng, J., Wu, X., et al., 2017. A survey on deep learning-based fine-grained object classification and semantic segmentation. Int. J. Autom. Comput. 14(2), 119-135. https://doi.org/10.1007/s11633-017-1053-3

Fricker, G.A., Ventura, J.D., Wolf, J.A., et al., 2019. A convolutional neural network classifier identifies tree species in mixed-conifer forest from hyperspectral imagery. Remote Sens. 11(19), 2326. https://doi.org/10.3390/rs11192326

Neupane, B., Horanont, T., Aryal, J., 2021. Deep learning-based semantic segmentation of urban features in satellite images: A review and meta-analysis. Remote Sens. 13(4), 808. https://doi.org/10.3390/rs13040808

He, Z., He, D., 2020. Bilinear squeeze-and-excitation network for fine-grained classification of tree species. IEEE Geosci. Remote Sens. Lett. 18(7), 1139-1143. https://doi.org/10.1109/LGRS.2020.2994952

Dong, Y., Ma, Z., Zi, J., et al., 2025. Multiscale feature fusion and enhancement in a transformer for the fine-grained visual classification of tree species. Ecol. Inform. 86, 103029. https://doi.org/10.1016/j.ecoinf.2025.103029

Liu, Y., Chen, Y., Liu, Z., et al., 2025. A multi-feature fusion network for tree species classification based on ground-based LiDAR data. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. https://doi.org/10.1109/JSTARS.2025.3527808

Ferreira, M.P., dos Santos, D.R., Ferrari, F., et al., 2024. Improving urban tree species classification by deep-learning based fusion of digital aerial images and LiDAR. Urban For. Urban Green. 94, 128240. https://doi.org/10.1016/j.ufug.2024.128240

Weiser, H., Schäfer, J., Winiwarter, L., Krašovec, N., Seitz, C., Schimka, M., et al., 2022. Terrestrial, UAV-borne, and airborne laser scanning point clouds of central European forest plots, Germany, with extracted individual trees and manual forest inventory measurements [dataset]. PANGAEA. https://doi.org/10.1594/PANGAEA.942856

Hollaus, M., Chen, Y.-C., 2022. SilviLaser 2021 Benchmark Dataset - Terrestrial Challenge [dataset]. TU Wien. https://doi.org/10.48436/afdjq-ce434

Ali, M., Biswas, A., Iglseder, A., Kumar, V., Kumar, S., Gupta, S., et al., 2026. Terrestrial and airborne laser scanning dataset of trees in the Shivalik Range, India with field measurements and leaf-wood classifications. Sci. Data 13, 420. https://doi.org/10.1038/s41597-026-06674-w

Calders, K., Verbeeck, H., Burt, A., Origo, N., Nightingale, J., Malhi, Y., et al., 2022. Terrestrial laser scanning data Wytham Woods: individual trees and quantitative structure models (QSMs) [dataset]. Zenodo. https://doi.org/10.5281/zenodo.7307956

Ali, M., Lohani, B., Hollaus, M., Pfeifer, N., 2024. Benchmarking geometry-based leaf-filtering algorithms for tree volume estimation using terrestrial LiDAR scanners. Remote Sens. 16(6), 1021. https://doi.org/10.3390/rs16061021

Wild, B., Özkan, T., Ali, M., Pöppl, F., Milenković, M., Hofhansl, F., et al., 2026. Evaluating RayCloudTools to estimate single-tree volume. Forestry, cpaf087. https://doi.org/10.1093/forestry/cpaf087

Jia, S., Jiang, S., Zhang, S., et al., 2024. Graph-in-graph convolutional network for hyperspectral image classification. IEEE Trans. Neural Netw. Learn. Syst. 35(1), 1157-1171. https://doi.org/10.1109/TNNLS.2022.3182715

Li, N., Jiang, S., Xue, J., et al., 2024. Texture-aware self-attention model for hyperspectral tree species classification. IEEE Trans. Geosci. Remote Sens. 62, 1-15. Art no. 5502215. https://doi.org/10.1109/TGRS.2023.3344787

Long, Y., Ye, S., Wang, L., et al., 2024. Scale pyramid graph network for hyperspectral individual tree segmentation. IEEE Trans. Geosci. Remote Sens. 62, 1-14. Art no. 5526014. https://doi.org/10.1109/TGRS.2024.3439094

Zhang, S., Xu, M., Zhou, J., et al., 2022. Unsupervised spatial-spectral CNN-based feature learning for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 60, 1-17. Art no. 5524617. https://doi.org/10.1109/TGRS.2022.3153673

Beloiu Schwenke, M., Xia, Z., Novoselova, I., et al., 2025. TreeAI Global Initiative - Advancing tree species identification from aerial images with deep learning [dataset]. Zenodo, Version TreeAI.V1.2. https://doi.org/10.5281/zenodo.15351054

Dauphin, Y.N., Fan, A., Auli, M., et al., 2017. Language modeling with gated convolutional networks. Int. Conf. Mach. Learn., pp. 933-941. https://proceedings.mlr.press/v70/dauphin17a.html

Xu, W., Wan, Y., 2024. ELA: Efficient local attention for deep convolutional neural networks. arXiv preprint arXiv:2403.01123. https://arxiv.org/abs/2403.01123

Zhang, T., Li, L., Zhou, Y., et al., 2024. CAS-ViT: Convolutional additive self-attention vision transformers for efficient mobile applications. arXiv preprint arXiv:2408.03703. https://arxiv.org/abs/2408.03703

Touvron, H., Cord, M., Douze, M., et al., 2021. Training data-efficient image transformers & distillation through attention. Int. Conf. Mach. Learn., pp. 10347-10357. https://doi.org/10.48550/arXiv.2012.12877

Behera, A., Wharton, Z., Hewage, P.R.P.G., et al., 2021. Context-aware attentional pooling (CAP) for fine-grained visual classification. Proc. AAAI Conf. Artif. Intell. 35(2), 929-937. https://doi.org/10.1609/aaai.v35i2.16176

Rao, Y., Chen, G., Lu, J., et al., 2021. Counterfactual attention learning for fine-grained visual categorization and re-identification. Proc. IEEE/CVF Int. Conf. Comput. Vis., pp. 1025-1034. https://doi.org/10.1109/ICCV48922.2021.00109

Wang, J., Yu, X., Gao, Y., 2021. Feature fusion vision transformer for fine-grained visual categorization. arXiv preprint arXiv:2107.02341. https://arxiv.org/abs/2107.02341

Bera, A., Wharton, Z., Liu, Y., et al., 2022. SR-GNN: Spatial relation-aware graph neural network for fine-grained image categorization. IEEE Trans. Image Process. 31, 6017-6031. https://doi.org/10.1109/TIP.2022.3205215

Diao, Q., Jiang, Y., Wen, B., et al., 2022. MetaFormer: A unified meta framework for fine-grained recognition. arXiv preprint arXiv:2203.02751. https://arxiv.org/abs/2203.02751

Sikdar, A., Liu, Y., Kedarisetty, S., et al., 2025. Interweaving insights: High-order feature interaction for fine-grained visual recognition. Int. J. Comput. Vis. 133(4), 1755-1779. https://doi.org/10.1007/s11263-024-02260-y

Sothe, C., Dalponte, M., Almeida, C. M., et al., 2019. Tree species classification in a highly diverse subtropical forest integrating UAV-based photogrammetric point cloud and hyperspectral data. Remote Sens. 11(11), 1338. https://doi.org/10.3390/rs11111338

Wallace, L., Hally, B., Hillman, S., et al., 2020. Terrestrial image-based point clouds for mapping near-ground vegetation structure: Potential and limitations. Fire 3(4), 59. https://doi.org/10.3390/fire3040059