GeoIdTree: Domain-adaptive Multi-temporal Plant Identification from Geospatial Orthomosaics and RGB Images*

Doan-Duc Phan[1][0009−0005−6672−5869], Quang-Duy Pham[1], Minh-Anh Cu[1], Thi-Ngoc-Anh Nguyen[1], Van-Nam Hoang[1], Mai-Viet-Hoang Do[1], Van-Sam Hoang[2], Thanh-Hai Tran[1], Hai Vu[1], Michiel Vlaminck[3], Hiep Luong[3], and Thi-Lan Le[1][0000−0001−9541−3905]

[1] SigM Laboratory, School of Electrical and Electronic Engineering, Hanoi University of Science and Technology, Hanoi 100000, Vietnam
lan.lethi1@hust.edu.vn

[2] Vietnam National Forestry University, Hanoi 100000, Vietnam

[3] Image Processing and Interpretation (IPI), UAV Research Centre (URC), IMEC, Ghent University, Ghent 9000, Belgium

Abstract. Forest biodiversity conservation requires reliable plant identification across time, yet multi-period recognition remains challenging due to variations in illumination and plant phenology that significantly alter plant appearance. In this paper, we propose GeoIdTree, a unified framework for multi-period plant recognition from UAV imagery, integrating multi-temporal orthomosaics and RGB images. The contributions are two-fold. First, we introduce a geospatial-alignment mechanism that propagates tree-level annotations from a reference orthomosaic to new temporal acquisitions, enabling dataset expansion without repeated manual labeling; using this mechanism we built VNUF, a multi-temporal UAV dataset of four subsets (VNUF-1, VNUF-2, VNUF-3A, VNUF-3B) covering a tropical forest in Vietnam. Second, to mitigate temporal domain shifts, we combine Center Loss, Triplet Loss, and CORAL Loss for cross-period feature consistency. Experiments with five backbones (DenseNet201, EfficientNet-B4, ViT-B/16-224, ViT-B/16-DINOv3, Swin-T) reach up to 85.03% Top-1 in in-dataset evaluation on VNUF-3A but drop to 25.51–36.73% in in-dataset evaluation on VNUF-2, confirming a strong temporal gap. The domain adaptation technique raises VNUF-2 F1-score from 21–35% to 40–45% and improves Top-1 to 86.42% on VNUF-3A and 56.24% on VNUF-1, demonstrating effective domain adaptation without target-domain labels.

Keywords: Plant identification · Biodiversity · UAV · Geospatial · Photogrammetry · Orthomosaic · Domain adaptation · Forest

* Supported by VLIR SI project "DORA: open Drone data platfOrm for biodiveRsity mApping to support the sustainable management of forests", grant number: VN2023SIN403A103.

1 Introduction

According to the Ministry of Agriculture and Environment - 2025, Vietnam is one of the world's biodiversity hotspots, ranking 16th globally with over 12,000 documented plant species [10]. As key components of forest ecosystems, plant species play a crucial role in shaping forest composition, stability, and resilience. Forests are vital for ecological balance and the livelihoods of millions, especially in rural and mountainous areas. However, despite increased forest cover from afforestation, biodiversity continues to decline due to deforestation, habitat degradation, land-use change, and climate change. These pressures highlight the urgent need for effective, science-based conservation strategies.

Monitoring biodiversity has become an urgent and essential task in the current context. Among related activities, plant identification represents a key component in the assessment of forest biodiversity, detecting the presence of invasive or endangered species, and supporting the evaluation of impacts caused by environmental changes and human activities [13]. However, traditional manual-based approaches are time-consuming, costly, labor-intensive, and expose practitioners to significant risks due to harsh forest terrain and the presence of dangerous wild. Nowadays, advances in artificial intelligence, particularly deep learning models combined with image processing techniques, have partially addressed challenges in plant identification [21]. Unmanned Aerial Vehicles (UAVs) are emerging as an effective solution for image data acquisition, simplifying plant identification tasks due to the increasing availability and affordability of consumer-grade UAV platforms for forest research. Canopy-based plant identification using UAV-acquired data has been widely proposed as a promising approach [16] as UAVs are compact and capable of capturing hundreds of images and videos from multiple altitudes while operating from a fixed location. However, processing large-scale forest areas in a single workflow, as well as accounting for temporal variations in forest plant throughout the day or across different seasons rather than relying solely on individual images, remains a major challenge for researchers. Addressing these challenges requires interdisciplinary collaboration among experts in computer science, botany, and geodesy, supported by deep learning models and existing image processing and geospatial analysis tools.

In this study, we propose an integrated framework – GeoIdTree, which provides a comprehensive pipeline ranging from forest map reconstruction based on photogrammetric techniques to plant identification across different periods of the year. The main contributions of this paper are as follows:

– We propose a mechanism for constructing and expanding a plant identification dataset across multiple time periods by leveraging orthomosaic mapping and geospatial alignment. By georeferencing orthomosaics acquired at different dates into a unified coordinate system, tree-level annotations from a reference map can be automatically propagated to subsequent temporal datasets, significantly reducing the need for repeated manual labeling. Using this approach, we collect four multi-temporal datasets from the same forest area at four distinct time points.

- We introduce domain adaptation techniques to address domain shifts arising from variations in lighting, seasonal phenology, camera sensors, and acquisition times in multi-temporal UAV datasets. Specifically, we combine CORrelation ALignment (CORAL), Center Loss and Triplet Loss into main loss to enhance feature alignment and discriminability across domains. This enables models trained on one temporal dataset to generalize effectively to imagery captured at different time periods.
- We conduct a comprehensive evaluation of multiple deep learning architectures, including CNNs, Vision Transformers, and their variants, on the aforementioned four multi-temporal UAV datasets. The results demonstrate the effectiveness of the proposed framework for robust cross-temporal plant species identification in tropical forest environments.

2 Related works

UAVs have become an essential platform for high-resolution environmental monitoring due to their flexibility, low operational cost, and ability to capture imagery at user-defined temporal intervals. UAV-based remote sensing has been widely adopted for forest inventory, biodiversity monitoring, wildfire assessment, and precision agriculture, providing detailed spatial information that is difficult to obtain using satellite imagery alone [1], [6]. Compared to satellite platforms, UAV imagery offers centimeter-level spatial resolution, enabling fine-scale observation of vegetation structure, canopy conditions, and species-level characteristics, which are crucial for ecological research and conservation planning [1].

Structure-from-Motion (SfM) photogrammetry has emerged as a standard approach for reconstructing 3D surface models from overlapping UAV images. The technique estimates camera poses and scene geometry simultaneously, producing dense point clouds, Digital Surface mMdels (DSMs), and orthorectified imagery products [20]. Among these outputs, orthomosaics are particularly important because they provide geometrically corrected, seamless 2D maps generated by stitching multiple orthophotos together after removing distortions caused by camera perspective, terrain relief, and sensor tilt. Such orthomosaic maps preserve spatial consistency and are therefore widely used for Geographic Information System (GIS) analysis, forest mapping, and annotation-based machine learning tasks [20]. The integration of ground control points (GCPs) is commonly employed to improve the geospatial accuracy of the reconstructed products, ensuring reliable alignment with real-world coordinate systems [17].

Recent advances in deep learning have significantly improved automatic plant identification, particularly through Convolutional Neural Networks (CNNs) and transformer-based models trained on large-scale image datasets. [16] has presented a method using plant images extracted from orthophotos created by specialized photometric instruments with certain results. However, single-time imagery often fails to capture phenological variations such as seasonal leaf color changes, flowering stages, or canopy density differences. The study by [11], conducted in a botanical garden in Pakistan, applies Machine Learning for seasonal plant recognition. Nevertheless, dataset expansion is challenging in large, dense forest areas, and variations in lighting conditions and aerial sensors can affect identification accuracy and performance.

3 Methodology

To ensure that the entire workflow from data acquisition to plant identification and can be reused across multiple experiments, we propose a comprehensive framework: GeoIdTree that includes data acquisition from the study area, orthomosaic-based forest map generation, orthophoto mapping, tree labeling and preprocessing, and plant identification. The overall workflow of the proposed framework is illustrated in Fig. 1.

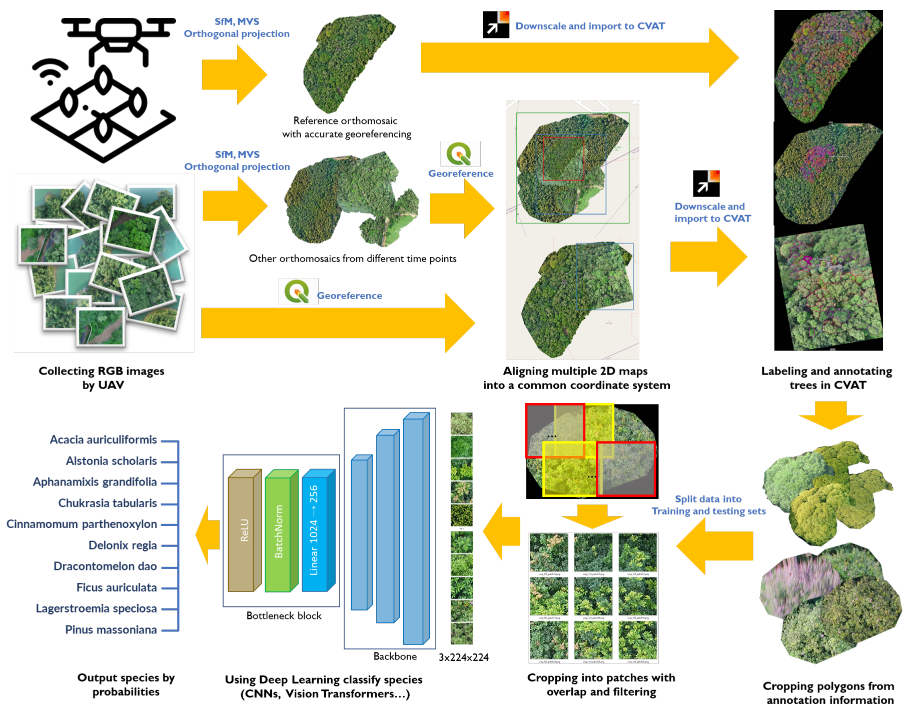

Fig. 1. Key components of GeoIdTree framework

The data acquisition stage, described in Section 3.1, provides detailed information about the implementation of UAV flight missions, including flight schedules and the characteristics of the collected datasets. Section 3.2 presents the process of generating forest maps using photogrammetric techniques to reconstruct high-resolution orthomosaic images from UAV imagery. To improve spatial consistency among reconstructed maps, a morphological transformation approach is introduced in Section 3.3. Based on these aligned maps, tree canopy–based labeling and data preprocessing procedures are conducted as described in Section 3.4 to prepare the dataset for machine learning tasks. Section 3.5 presents plant identification using deep learning techniques. In addition, this section also introduces the integration of domain adaptation techniques to improve training performance and model generalization for multi-temporal datasets.

It is worth noting that the proposed framework GeoIdTree is generalizable to forest environments beyond the study site. In this work, we demonstrate its application in a tropical forest area in Vietnam.

3.1 Study area & Data acquisition

In this study, the research area for conducting UAV flight campaigns is a forest located at Vietnam National University of Forestry (VNUF), Xuan Mai Commune, Hanoi, Vietnam as illustrated in Fig. 2. With the area of interest approximately 20 hectares, it is an experimental forest with a high level of biodiversity in both flora and fauna. Multiple flight missions were carried out to collect data using different coverage areas, UAV platforms, and times of day. However, a common characteristic of all data collection campaigns is that they focused on the northwestern part of the forest, where forestry experts can more easily identify tree species to support subsequent processing stages.

Fig. 2. Map of Vietnam with study area: VNUF Forest - an area with high tree density and rich biodiversity

A total of four UAV image datasets were used in this study. Some images in these datasets are illustrated in Fig. 3. The datasets are named according to the figure captions using the following format: VNUF (forest name) – X (temporal order of data acquisition). Among them, two datasets (VNUF-3A and VNUF-3B) were acquired on the same day, August 24, 2024, using a DJI Air 3 UAV. However, the dataset VNUF-3B was captured in the late afternoon using the 1× camera, whereas VNUF-3A was collected around noon with the 3×

camera, providing higher spatial resolution and clearer leaf texture details. The VNUF-2 dataset was collected two weeks before the aforementioned datasets, also using a 1× camera, but the flight altitude was set at 211m (approximately 60m higher than VNUF-3B). For the dataset VNUF-1, the acquisition platform was a DJI Phantom 4 RTK. This UAV is equipped with a 20 MP camera; therefore, the resulting leaf texture quality is more limited compared to the other three datasets. Detailed information on the four datasets primarily used in this study is summarized in Table 1.

With high-quality leaf detail and favorable acquisition conditions, VNUF-3A is selected as the reference dataset for the others. The remaining datasets (VNUF-1, VNUF-2 and VNUF-3B) are used for cross-evaluation to provide insights into the influence of time and seasonal variations.

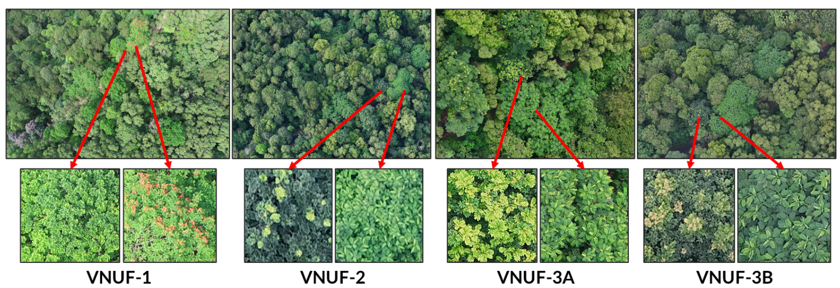

Fig. 3. RGB images captured by UAV by 4 points of time. Each image includes detailed close-up illustrations of leaf structures for two representative species at the same location: *aphanamixis grandifolia* (left) and *delonix regia* (right).

Table 1. Information about RGB images collected at VNUF Forest by multiple points of time

Dataset	Resolution (px)	Time	Device	Flight altitude
VNUF-1	5472 × 3648	14:40–14:50 May 22nd, 2024	DJI Phantom 4 RTK	80m
VNUF-2	8064 × 6048	16:50–17:15 Aug 10th, 2024	DJI Air 3	211m
VNUF-3A	8064 × 6048	12:30–13:30 Aug 24th, 2024	DJI Air 3 (Camera 3×)	100m
VNUF-3B	8064 × 6048	14:10–16:00 Aug 24th, 2024	DJI Air 3	150m

3.2 Orthomosaic generation

Images collected from the flight campaign are used to generate orthomosaic images. A protocol similar to [7] was adopted in which OpenDroneMap [18] is used to reconstruct 3D models and generate orthomosaics from overlapping UAV images. First, feature matching and SfM are applied to estimate camera poses and produce a sparse point cloud, followed by Multi-View Stereo (MVS) for dense reconstruction and DSM. Finally, orthorectification and image stitching are performed to obtain a geometrically consistent orthomosaic.

To enhance the positional accuracy of objects represented in the map, the use of Ground Control Points (GCPs) is essential. However, for large-scale forest areas, deploying GCPs that adequately cover the study region is often impractical. Therefore, in addition to utilizing GCPs visible from UAV-acquired imagery, this study incorporates the locations of individual tree canopies for which botanical experts have precisely identified both the central coordinates and species information as auxiliary "pseudo-GCPs". These supplementary points are leveraged to further improve the geospatial accuracy of the resulting orthomosaic.

3.3 Orthophotos mapping

In our study, we focus on robust multi-period plant recognition using a UAV-captured images. However, the annotation cost for building a multi-period plant recognition dataset may be high. Therefore, we propose a mechanism for constructing and expanding a plant identification dataset across multiple time periods by leveraging orthomosaic mapping and geospatial alignment. By using the proposed mechanism, tree-level annotations from a reference map can be automatically propagated to subsequent temporal datasets, significantly reducing the need for repeated manual labeling. However, during the construction of forest maps based on orthomosaics, positional inaccuracies commonly arise due to errors in sensor parameters, GPS measurements, and GCPs, resulting in geospatial misalignments among the generated orthomosaic products. To transform all orthomosaics into a common coordinate reference system, QGIS [5], a powerful geospatial processing software, is employed through its Georeferencer tool. This process is based on an affine transformation, which preserves parallelism and relative spatial relationships between objects, and is combined with Nearest Neighbors resampling method to geometrically align the orthomosaics within a unified spatial framework. This approach enables the consistent transfer and mapping of annotations from the reference orthomosaic generated from the VNUF-3A dataset to other orthomosaic maps, thereby significantly reducing the time and effort required, compared to re-annotating and re-labeling by experts for each individual map.

The georeferencing process starts with a set of GCPs, defined as correspondences between image coordinates (x_i, y_i) and map coordinates (X_i, Y_i). The parameters of an affine transformation are estimated by minimizing the reprojection error using a least-squares approach:

$$\min \sum_i \left[(X_i - \hat{X}_i)^2 + (Y_i - \hat{Y}_i)^2 \right] \tag{1}$$

The resulting transformation is then applied to map all image pixels into the target coordinate system:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & t_x \\ a_{21} & a_{22} & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{2}$$

where the 2×2 submatrix models linear transformations (scaling, rotation, and shear), and (t_x, t_y) represents translation. Finally, resampling (e.g., nearest neighbor) is performed to obtain the output image.

Although the orthomosaic generation methods for the VNUF-2, VNUF-3A, and VNUF-3B datasets generally produce good results with minimal errors, the VNUF-1 dataset, due to its lower level of detail and certain limitations during data acquisition, especially wind conditions, still exhibits some issues. To address this, a proposed approach is to utilize nadir-view images to replace regions where trees are incorrectly reconstructed in the orthomosaic. Similarly, for the orthomosaics, both real GCPs and pseudo-GCPs are employed. As a result, regions affected by reconstruction errors or ghost pixels in the orthomosaic can be replaced with the corresponding areas from the RGB images.

3.4 Tree labeling and preprocessing

Once the orthomosaic image is generated, two forestry experts annotate the VNUF-3A dataset using CVAT [3]. The annotation information is then automatically propagated to the VNUF-3B, VNUF-2, and VNUF-1 datasets using the aforementioned mechanism. From these annotations, we select the 10 most representative species in the study area, as illustrated in Fig. 4.

Fig. 4. Example images of 10 species in VNUF-3A dataset

Based on the annotated polygons, we extract 224×224 image patches with a 30% overlap. Patches containing less than 15% non-tree pixels, typically those near boundaries or heavily occluded are discarded. The data splitting strategy is based on the number of tree canopy polygons, following a $0.6 : 0.2 : 0.2$ ratio for training, validation, and testing, respectively. The final number of image patches used for training and testing is presented in Table 2. The distribution of image patches across species is illustrated in Fig. 5, where the total number of samples per species is indicated above each bar.

3.5 Plant identification and Domain Adaptation

Plant identification is an important task in monitoring the quality of biodiversity conservation. Despite its importance, a key challenge in multi-temporal plant

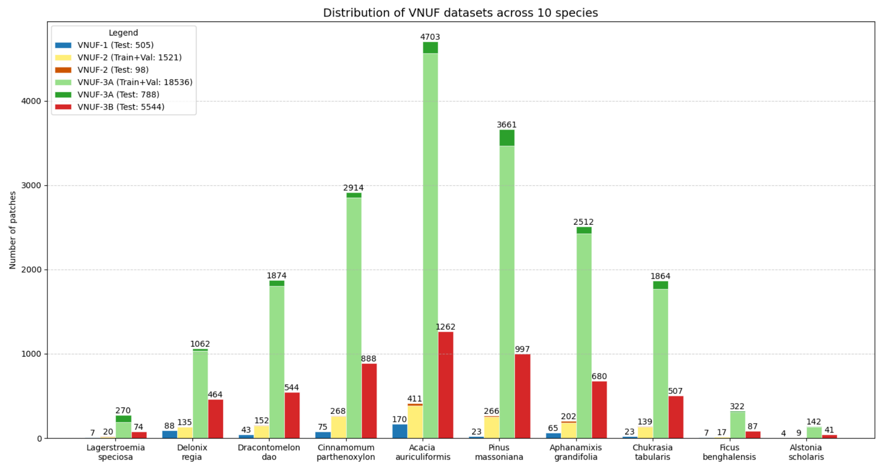

Fig. 5. Distribution of patch-level data from the VNUF forest datasets

Table 2. Detailed distribution of patches for training and testing across 4 datasets

Species	VNUF-1 (test)	VNUF-2 (train+val)	VNUF-2 (test)	VNUF-3A (train+val)	VNUF-3A (test)	VNUF-3B (test)
Lagerstroemia speciosa	7	12	8	188	82	74
Delonix regia	88	133	2	1033	29	464
Dracontomelon dao	43	144	8	1806	68	544
Cinnamomum parthenoxylon	75	256	12	2848	66	888
Acacia auriculiformis	170	388	23	4564	139	1262
Pinus massoniana	23	249	17	3464	197	997
Aphanamixis grandifolia	65	185	17	2425	87	680
Chukrasia tabularis	23	132	7	1769	95	507
Ficus benghalensis	7	14	3	305	17	87
Alstonia scholaris	4	8	1	134	8	41

identification is the domain shift between datasets acquired at different time points. Variations in lighting conditions (e.g., midday versus late afternoon), seasonal phenological changes (leaf color, canopy density, etc.), camera sensors (DJI Phantom 4 RTK, DJI Air 3, etc.), and flight altitudes introduce significant distributional differences between training and test data. A model trained on one temporal dataset often suffers substantial performance degradation when applied to another. To mitigate this domain gap, we incorporate domain adaptation techniques into the training pipeline. Regarding the backbone, five off the shelf models that are Dense-Net201 [8], Efficient-Net-B4 [15], ViT-B/16-DINO-V3 [2], ViT-B/16-224 [4], Swin-T [9] are used.

Focal Loss The class imbalance is illustrated in Fig. 5. Therefore, Focal Loss is used as the primary loss function to address this issue by focusing on hard examples, where p_t is the predicted probability for the true class, α_t is a balancing factor, and γ is the focusing parameter.

$$\mathcal{L}_{\mathrm{FL}} = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \tag{3}$$

In addition, metric and domain alignment losses, including CORAL, Center and Triplet Loss, are incorporated to enhance feature discrimination and improve robustness across different data distributions.

CORAL Loss CORAL [14] is an unsupervised domain adaptation method that minimizes the discrepancy between the source and target domain distributions by aligning their second-order statistics (covariance matrices) in the feature space. Given the source domain feature representations $\mathbf{C}_S$ and target domain feature representations $\mathbf{C}_T$ extracted from the penultimate layer of the deep network, the CORAL loss is defined as:

$$\mathcal{L}_{\mathrm{CORAL}} = \frac{1}{4d^2} \|\mathbf{C}_S - \mathbf{C}_T\|_F^2 \tag{4}$$

where d is the feature dimension and $\| \cdot \|_F$ denotes the Frobenius norm. In our framework, the source domain corresponds to the training dataset (VNUF-3A), which provides the highest image quality with clear leaf textures captured at noon using the $3\times$ camera. The target domain corresponds to each of the remaining three datasets captured under different conditions. By minimizing the CORAL loss during training, the model learns feature representations that are more invariant to the domain-specific variations across temporal acquisitions.

Center Loss To enhance intra-class compactness in the learned feature space, we incorporate center loss [19] alongside standard cross-entropy loss. Center loss penalizes the distance between deep features and their corresponding class centers, encouraging features of the same plant species to cluster tightly regardless of the temporal acquisition conditions. The center loss is formulated as:

$$\mathcal{L}_{\mathrm{center}} = \frac{1}{2} \sum_{i=1}^{m} \|\mathbf{x}_i - \mathbf{c}_{y_i}\|_2^2 \tag{5}$$

where $\mathbf{x}_i$ is the feature representation of sample i, $\mathbf{c}_{y_i}$ is the learnable center of class y_i, and m is the mini-batch size. The class centers are updated jointly with the network parameters during training.

Triplet Loss As an alternative to center loss, we also investigate triplet loss [12] to enforce discriminative feature learning. Triplet loss operates on triplets of samples consisting of an anchor, a positive sample (same species), and a negative sample (different species), encouraging the model to learn an embedding space where intra-class distances are smaller than inter-class distances by a margin α:

$$\mathcal{L}_{\mathrm{triplet}} = \sum_{i=1}^{N} \max\left(0, \|\mathbf{x}_i^a - \mathbf{x}_i^p\|_2^2 - \|\mathbf{x}_i^a - \mathbf{x}_i^n\|_2^2 + \alpha\right) \tag{6}$$

where $\mathbf{x}_i^a$, $\mathbf{x}_i^p$, and $\mathbf{x}_i^n$ denote the anchor, positive, and negative feature embeddings, respectively.

Combined Training Objective The overall training objective combines the classification loss with the domain adaptation and metric learning losses. The configurations is used for evaluating:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{FL}} + \lambda_1 \mathcal{L}_{\text{CORAL}} + \lambda_2 \mathcal{L}_{\text{center}} + \lambda_3 \mathcal{L}_{\text{triplet}} \tag{7}$$

where $\lambda_1 = 0.1$, $\lambda_2 = 0.1$, $\lambda_3 = 0.1$ are weighting hyperparameters that balance the contribution of each loss component.

4 Experimental results

4.1 Experiments setting

The model was trained for 50 epochs with 224×224 inputs. All experiments were conducted on a Google Colab A100 GPU with five backbones DenseNet201, EfficientNet-B4, ViT-B/16-224, ViT-B/16-DINOv3 and Swin-T. We conduct two experiments to assess the plant identification performance with and without domain adaptation. Regarding domain adaptation, we train the model with loss function defined in Equation 7 with images in training set of VNUF-3A is used as source and that of VNUF-2 used as target domain.

4.2 Plant identification results without domain adaptation

Table 3 reports plant identification performance for five backbones that are DenseNet201, EfficientNet-B4, ViT-B/16-DINOv3, ViT-B/16-224 and Swin-T trained on either the VNUF-3A or the VNUF-2 set and evaluated across the four VNUF datasets.

In-dataset performance: On VNUF-3A, all models reach competitive scores (Top-1 between 80.84% and 85.03%), with DenseNet201 and ViT-B/16-DINOv3 leading at 85.03% and 84.94% respectively, and Swin-T close behind at 82.99%. In-dataset performance on VNUF-2 is markedly lower (Top-1 between 51.02% and 73.47%), confirming that VNUF-2 is the more challenging acquisition split despite the identical class set. ViT-B/16-DINOv3 clearly dominates with 73.47% Top-1 and 71.05% F1-score, followed by DenseNet201 at 70.41%.

Cross-dataset performance. Two consistent patterns emerge. First, transfer between closely related subsets is largely preserved: VNUF-3A → VNUF-3B retains 75.81–84.76% Top-1, and VNUF-2 → VNUF-3B even climbs to 65.75–77.78%, indicating that the visual variability captured by VNUF-2 generalises well to images of the same species drawn from a closely related collection. Second, transfer to the more distant VNUF-2 and VNUF-1 splits drops sharply for every model, revealing a clear domain gap induced purely by acquisition variation (e.g., VNUF-3A → VNUF-2 falls to 25.51–36.73% Top-1, and VNUF-3A → VNUF-1 to 28.71–55.25%). Across models, the transformer-based backbones particularly ViT-B/16-DINOv3 and Swin-T exhibit the most robust generalisation: ViT-B/16-DINOv3 obtains the best cross-dataset scores on VNUF-1, while ViT-B/16-224 is the most fragile under domain shift. Overall, the results suggest that DenseNet201 remains a strong in-dataset baseline, but self-supervised and hierarchical transformer backbones are preferable when robustness to acquisition variability across the same species set is required.

Table 3. Plant identification results on four datasets when training the models on either VNUF-3A or VNUF-2 training sets

Metric (%)	**Testing dataset**	**Models trained on VNUF-3A training set**					**Models trained on VNUF-2 training set**				
		DenseNet201 [8]	**EfficientNet-B4** [15]	**ViT-B/16-DINO-V3** [2]	**ViT-B/16-224** [4]	**Swin-T** [9]	**DenseNet201** [8]	**EfficientNet-B4** [15]	**ViT-B/16-DINO-V3** [2]	**ViT-B/16-224** [4]	**Swin-T** [9]
Precision	**VNUF-3A**	**85.94**	82.40	85.30	81.39	83.73	61.20	58.88	69.86	**77.77**	76.35
	VNUF-3B	81.87	84.26	**86.65**	82.39	86.64	77.68	70.84	76.57	**80.92**	73.44
	VNUF-2	45.11	32.19	45.27	22.59	**47.35**	68.15	58.55	**71.03**	70.93	67.85
	VNUF-1	53.63	48.28	59.44	**65.41**	56.27	43.89	32.88	**58.14**	40.58	52.32
F1-score	**VNUF-3A**	84.68	81.45	**85.28**	80.09	83.63	46.44	46.36	47.12	57.61	**58.33**
	VNUF-3B	79.49	80.20	81.40	76.72	**82.44**	69.97	64.15	70.25	**76.25**	69.00
	VNUF-2	33.57	28.35	34.37	21.06	**34.56**	68.43	50.67	**71.05**	62.71	61.61
	VNUF-1	41.98	29.61	53.30	25.65	**55.25**	30.30	22.53	**45.60**	30.52	36.16
Top-1 acc.	**VNUF-3A**	**85.03**	81.73	84.94	80.84	82.99	45.94	42.39	42.13	52.28	**53.81**
	VNUF-3B	80.93	81.82	84.05	75.81	**84.76**	72.67	65.75	73.38	**77.78**	72.31
	VNUF-2	31.63	31.63	35.71	25.51	**36.73**	70.41	51.02	**73.47**	59.18	65.31
	VNUF-1	45.54	28.71	54.06	29.90	**55.25**	40.20	20.59	**51.09**	27.92	44.16
Top-3 acc.	**VNUF-3A**	96.32	94.54	96.83	96.07	**97.34**	78.43	71.57	**79.57**	78.93	77.28
	VNUF-3B	95.62	96.55	**98.11**	96.70	97.96	92.86	91.31	90.95	88.60	**93.45**
	VNUF-2	69.39	66.33	**70.41**	58.16	67.35	**90.82**	82.65	89.80	86.73	89.80
	VNUF-1	69.90	64.95	**87.13**	73.86	85.94	63.56	41.98	**69.70**	65.74	64.75

4.3 Plant identification results with domain adaptation

Table 4 shows the results when applying domain adaptation. When incorporating CORAL, Triplet Loss, and Center Loss, the performance on VNUF-2 improves notably compared to results without domain adaptation as shown in Table 3. For instance, F1-score increase from approximately 21-35% (Table 3) to around 40–45% (Table 4), depending on the model, indicating that domain adaptation successfully reduces feature distribution discrepancies between the two domains. Although these results still do not reach the in-dataset performance, the improvement is consistent across all architectures and achieved without requiring labeled data from the target domain.

Moreover, domain adaptation also contributes to improved performance on datasets with similar characteristics. DenseNet201 model achieves notably high precision on the VNUF-3A, with Precision, F1-score, and Top-1 accuracy reaching 87.12%, 85.97%, and 86.42%, respectively. On the other hand with the VNUF-3B, all models benefit from domain adaptation, showing consistent improvements. In particular, the ViT-B/16-224 model achieves 86.94%, 85.74%, and 86.87% for these metrics. Finally, it is worth noting that performance on VNUF-1 shows some improvement in Top-1 accuracy across all settings, although the gains remain modest, likely due to differences in the UAV platform used for

Table 4. Plant identification results on four datasets when applying domain adaptation with VNUF-3A is source and VNUF-2 is target

Metric (%)	Testing dataset	DenseNet -201 [8]	EfficientNet -B4 [15]	ViT-B/16-DINOv3 [2]	ViT-B/16-224 [4]	Swin-T [9]
Precision	**VNUF-3A**	**87.12**	83.72	86.23	86.22	84.57
	VNUF-3B	84.74	83.14	86.16	**86.94**	84.98
	VNUF-2	51.21	47.74	44.07	**51.71**	50.80
	VNUF-1	51.07	39.40	57.26	51.67	56.67
F1-score	**VNUF-3A**	**85.97**	82.92	85.36	85.26	84.86
	VNUF-3B	83.53	82.27	82.53	**85.74**	80.60
	VNUF-2	40.80	44.16	39.20	**44.39**	40.51
	VNUF-1	41.47	27.05	**52.93**	51.07	48.81
Top-1 acc.	**VNUF-3A**	**86.42**	83.38	85.91	85.53	85.41
	VNUF-3B	84.92	82.23	84.81	**86.87**	84.13
	VNUF-2	39.80	43.88	40.82	**44.9**	41.84
	VNUF-1	40.59	26.14	**56.24**	53.07	51.29
Top-3 acc.	**VNUF-3A**	95.43	94.42	96.32	95.94	96.07
	VNUF-3B	97.60	97.17	97.71	97.01	96.88
	VNUF-2	70.41	**72.45**	67.35	69.39	58.16
	VNUF-1	80.99	63.56	77.82	75.84	70.50

data acquisition. This suggests that sensor characteristics and imaging conditions introduce additional domain discrepancies, further emphasizing the necessity of robust domain adaptation techniques.

Fig.6 illustrated the confusion matrix of ViT-B/16-DINOv3 model for VNUF-3A dataset. Although there is a slight increase in confusion for some samples within certain species compared to using only the original dataset, the *ficus benghalensis* class is correctly recognized in several cases, and other species (*chukrasia tabularis*, *pinus massoniana*) also show improved performance. This demonstrates the effectiveness of domain adaptation for multi-temporal plant identification.

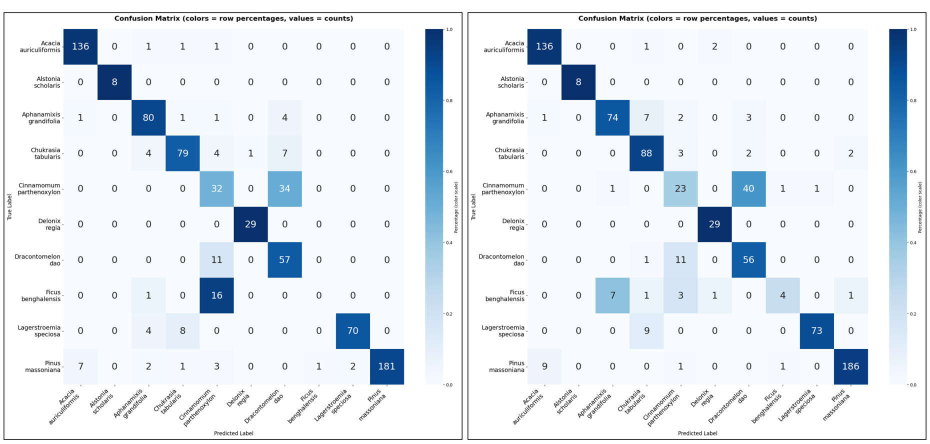

Fig. 6. Confusion matrix of ViT-B/16-DINOv3 model before (left) and after (right) using domain adaptation

5 Conclusions and future works

In this paper, we have presented GeoIdTree, a unified framework for plant identification from multi-temporal UAV imagery. The framework integrates three key components: orthomosaic construction via photogrammetry, geospatial alignment for cross-temporal annotation transfer, and deep learning-based plant identification with domain adaptation. A central contribution is the proposed mechanism for building and expanding plant identification datasets across time by leveraging georeferenced orthomosaic maps, enabling automatic label propagation from a reference map to new temporal acquisitions without repeated manual annotation. To address the domain shift arising from variations in lighting, season, camera sensor, and flight parameters across different acquisition periods, we incorporated CORAL-based domain adaptation combined with center loss and triplet loss, demonstrating improved generalization on temporally diverse datasets. Extensive experiments on four UAV datasets collected at the VNUF forest across different dates and conditions validate the effectiveness of the proposed approach. Future work will focus on enlarging the dataset with additional seasonal acquisitions, exploring more advanced domain adaptation strategies, and integrating spectral information for enhanced species discrimination.

Acknowledgements This work is funded by VLIR SI project "DORA: open Drone data platfOrm for biodiveRsity mApping to support the sustainable management of forests", grant number: VN2023SIN403A103

References

1. Anderson, K., Gaston, K.J.: Lightweight unmanned aerial vehicles will revolutionize spatial ecology. Frontiers in Ecology and the Environment **11**(3), 138–146 (2013). https://doi.org/https://doi.org/10.1890/120150, https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/120150
2. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the International Conference on Computer Vision (ICCV) (2021)
3. CVAT.ai Corporation: Computer vision annotation tool (cvat). https://doi.org/10.5281/zenodo.19711134 (2026), zenodo, Apr. 23, 2026
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2021), https://arxiv.org/abs/2010.11929
5. Graser, A., Sutton, T., Bernasocchi, M.: The qgis project: Spatial without compromise. Patterns **6**(7), 101265 (2025). https://doi.org/https://doi.org/10.1016/j.patter.2025.101265, https://www.sciencedirect.com/science/article/pii/S2666389925001138
6. Guimarães, N., Pádua, L., Marques, P., Silva, N., Peres, E., Sousa, J.J.: Forestry remote sensing from unmanned aerial vehicles: A review focusing on the data, processing and potentialities. Remote Sensing **12**(6) (2020). https://doi.org/10.3390/rs12061046, https://www.mdpi.com/2072-4292/12/6/1046

7. Hoang Do, M.V., Phan, D.D., Minh Le, C.T., Hoang, V.N., Nguyen, H.Q., Tran, T.H., Vu, H., Vu, X.D., Hoang, V.S., Vlaminck, M., Maes, W., Luong, H., Le, T.L.: A unified framework for plant identification from uav images. In: 2024 Tenth International Conference on Communications and Electronics (ICCE). pp. 421–426 (2024). https://doi.org/10.1109/ICCE62051.2024.10634639
8. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks (2018), https://arxiv.org/abs/1608.06993
9. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows (2021), https://arxiv.org/abs/2103.14030
10. Ministry of Agriculture and Environment: Pride and responsibility – the driving force for biodiversity conservation (2025), https://en.mae.gov.vn/pride-and-responsibility–the-driving-force-for-biodiversity-conservation-8873.htm, accessed: 2026-04-26
11. Qasim, H., Ding, X., Usman, M., Abbas, S., Shahzad, N., Keshk, H.M., Bilal, M., Ahmad, U.: Advancing tree species classification with multi-temporal uav imagery, geobia, and machine learning. Geomatics **5**(3) (2025). https://doi.org/10.3390/geomatics5030042, https://www.mdpi.com/2673-7418/5/3/42
12. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). p. 815–823. IEEE (2015). https://doi.org/10.1109/cvpr.2015.7298682
13. Sesay, R.E.V., Sesay, F., Azizi, M.I., Rahmani, B.: Invasive species and biodiversity: Mechanisms, impacts, and strategic management for ecological preservation. Asian Journal of Environment amp; Ecology **23**(9), 82–95 (Sep 2024)
14. Sun, B., Feng, J., Saenko, K.: Correlation alignment for unsupervised domain adaptation (2016), https://arxiv.org/abs/1612.01939
15. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks (05 2019). https://doi.org/10.48550/arXiv.1905.11946
16. Tariku, G., Ghiglieno, I., Gilioli, G., Gentilin, F., Armiraglio, S., Serina, I.: Automated identification and classification of plant species in heterogeneous plant areas using unmanned aerial vehicle-collected rgb images and transfer learning. Drones **7**(10) (2023). https://doi.org/10.3390/drones7100599, https://www.mdpi.com/2504-446X/7/10/599
17. Turner, D., Lucieer, A., Wallace, L.: Direct georeferencing of ultrahigh-resolution uav imagery. IEEE Transactions on Geoscience and Remote Sensing **52**, 2738–2745 (2014), https://api.semanticscholar.org/CorpusID:206691801
18. Vacca, G.: Overview of open source software for close range photogrammetry. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences **XLII-4/W14**, 239–245 (2019)
19. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016. pp. 499–515. Springer International Publishing, Cham (2016)
20. Westoby, M., Brasington, J., Glasser, N., Hambrey, M., Reynolds, J.: 'structure-from-motion' photogrammetry: A low-cost, effective tool for geoscience applications. Geomorphology **179**, 300–314 (2012)
21. Wäldchen, J., Mäder, P.: Plant species identification using computer vision techniques: A systematic literature review. Archives of Computational Methods in Engineering **25**, 507–543 (04 2018). https://doi.org/10.1007/s11831-016-9206-z