

COS30082 Applied Machine Learning

ML Group Project

Report

Cross Domain Plant Species Identification

Prepared for:

Ts Dr Lee Sue Han

Prepared by:

VerdantVision (Group 12)

Student Name	Student ID
Deron Yijia FOO	102780757
Esther Hui Min CHAI	102778419
Jayne Hieng Siew WONG	102776536
Jun Hong LAI	102780715
William Chin Lee WAN	104385387

Submission Date: **28 November 2025**

Due Date: **Week 12, Friday, 23:59 pm.**

Table of Content

1.0	Methodology	1
1.1	Experimental Setup	1
1.2	Dataset Handling and Generative Augmentation	1
1.3	Utility Functions and Evaluation Protocols	2
1.4	Baseline Model 1 - Mix-Stream CNN (ConvNeXt).....	2
1.5	Baseline Model 2 - DINOv2 + Mix-Stream CNN	3
1.6	New Approach 1: CUT + DINOv2.....	4
1.7	New Approach 2: Dual-Stream Ensemble	4
2.0	Results and Discussion.....	5
2.1	Baseline Model 1 - Mix-Stream CNN (ConvNeXt).....	5
2.2	Baseline Model 2 - DINOv2 + SVM.....	5
2.3	New Approach 1: CUT + DINOv2.....	6
2.4	New Approach 2: Dual-Stream Ensemble	6
2.4	Discussion of Findings.....	7
2.5	Limitations and Future Work	8
2.6	Summary.....	8

GitHub Repository (Group Project Folder):

https://github.com/DeronFoo/G12_AML_Plant_Species_Identification.git

Google Colab Notebook (Source Code):

https://colab.research.google.com/drive/1phJUIFhcCr8rXJ_Ibiy-xRTlt1W4DXfA?usp=sharing

Public HuggingFace Space (Web-based UI):

<https://huggingface.co/spaces/DeronFoo/G12-Plant-Species-Classifer>

Project Presentation (YouTube Video):

<https://youtu.be/bB9Hu6aMUIY>

1.0 Methodology

This methodology delineates a systematic framework designed to address the critical domain shift between herbarium specimens and real-world field imagery in plant species identification. The experimental design is grounded in a standardised, reproducible computational environment and utilises a stratified data pipeline to rigorously evaluate performance across both well-represented and zero-shot categories. The investigation progresses from establishing foundational benchmarks using convolutional and transformer architectures to implementing advanced domain adaptation strategies. These interventions include the integration of synthetic data augmentation to enhance feature representation and the development of a novel Metric-Hybrid Ensemble, specifically architected to resolve generalisation failures inherent to species lacking field photography.

1.1 Experimental Setup

To streamline collaboration, the team first established a **standardised PyTorch environment** on a single shared **Colab master notebook**, which served as the main repository for final code documentation, while individual Jupyter notebooks were utilised for rapid decentralised experimentation. This deep learning environment would later leverage **timm** and **transformers** to access and fine-tune **state-of-the-art pre-trained backbones (ConvNeXt, DINOv2)** on NVIDIA GPUs. To ensure rigorous reproducibility, the team implemented **Mixed Precision Training** and enforced a **global seed of 42**, guaranteeing that performance gains were attributable strictly to architectural improvements rather than initialisation variance.

1.2 Dataset Handling and Generative Augmentation

In order to ensure model robustness and training efficiency, the team architected a resilient data pipeline that would automatically transfer the dataset from persistent cloud storage to the local runtime environment. Crucially, the team subsequently established a **fixed global mapping of ClassID to Index**, ensuring that all models to be built regardless of architecture would share an **identical label space** whilst explicitly separating validation data into **“With-Pairs” (Seen)** and **“Without-Pairs” (Unseen)** splits to facilitate precise **zero-shot evaluation**. Interestingly, exploratory data analysis (EDA) carried out by the team showed that not only was there a **severe domain imbalance** with herbarium sheets outnumbering field photos by **nearly 4:1**, but the team was also faced with a **long-tail class distribution** where rare species possess fewer than ten samples and a **profound visual domain shift** between standardised herbarium specimens and the complex, occluded nature of field photography.

Seeking to standardise the input pipeline, the team then implemented a custom **PlantDataset** class that normalises inputs to ImageNet standards and returns a **domain indicator tensor** to support domain-aware training. Augmentation strategies were tailored to the specific architecture, with the **CNN (CovNeXt)** baseline utilising more aggressive transformations like **RandomResizedCrop** and **ColorJitter** to encourage feature invariance, while the **ViT (DINOv2)** employed gentler geometric resizing to preserve the **patch integrity** required for transformer embeddings. Beyond standard preprocessing, the team also addressed the critical shortage of field data for “Without Pairs” species by deploying **CUT (Contrastive Unpaired Translation)**. Unlike CycleGAN,

CUT maximises mutual information via **contrastive loss** for efficient one-sided translation; this method ultimately generated **1,744 synthetic field images** which would be injected exclusively into the training set for the later **new proposed hybrid model**, effectively **bridging the zero-shot gap** that the baseline models could not overcome.

1.3 Utility Functions and Evaluation Protocols

To ensure consistent benchmarking across disparate architectures (CNNs and Transformers), the team subsequently developed a modular, model-agnostic utility engine that standardises the training and evaluation lifecycle. The training loop incorporates **Automatic Mixed Precision (AMP)** to maximise throughput on NVIDIA GPUs, dynamically scaling gradients to reduce memory footprint without compromising convergence stability. Beyond standard optimisation, the evaluation protocol was engineered to address the specific constraints of the PlantCLEF challenge; instead of relying solely on global accuracy, the system implements a **stratified scoring mechanism** that explicitly dissects performance into with-pairs and without-pairs categories. This granular reporting, complemented by Top-5 accuracy metrics, normalised confusion matrices, and per-class precision profiling, enables a precise diagnosis of domain shift effects, distinguishing between general classification failures and specific zero-shot generalisation issues.

1.4 Baseline Model 1 - Mix-Stream CNN (ConvNeXt)

To establish a strong supervised benchmark, the team agreed to deploy **ConvNeXt-Base**, a modern architecture that bridges the gap between standard CNNs and Vision Transformers by adopting large kernel size (7x7) and layer normalisation. Given the dataset’s limited size relative to the model’s capacity (~88M parameters). ConvNeXt also implements hierarchical feature extraction, which is useful for fine-grained plant morphology; efficient training, making it suitable as a fair baseline prior to introducing more advanced methods (e.g., vision transformers); robustness to domain variation due to convolutional locality and built-in spatial priors.

The team devised a **Partial Freezing Transfer Learning** strategy to balance feature retention with domain adaptation. Specifically, the team froze the initial stem and early stages (0 and 1) to preserve generic, low-level ImageNet features, while unfreezing the deeper stages (2 and 3) to allow the model to learn high-level, species-specific morphology. The table below further states the training settings and hyperparameters tuning.

Training Configuration & Hyperparameters	
Architecture Head	A custom classification head featuring a Dropout layer (0.3) to mitigate overfitting, feeding into a single linear projection for the 100 species classes.
Optimizer	AdamW Optimizer was selected for its decoupled weight decay handling. We implemented a parameter grouping strategy that applies weight decay (0.0001) to convolutional weights while strictly excluding bias and normalization parameters to ensure training stability.

Learning Rate Schedule	Initial learning rate of 0.0001, managed by a ReduceLROnPlateau scheduler (Factor: 0.5, Patience: 3 epochs) to dynamically anneal the rate when validation loss plateaued.
Training Specs	Epochs: 20, Batch Size: 32, Precision: AMP via GradScaler, and Loss Function: Cross-Entropy Loss.

1.5 Baseline Model 2 - DINOv2 + Mix-Stream CNN

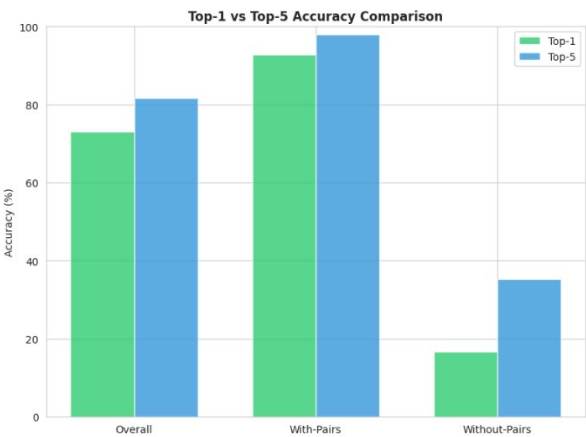
Moving on, to investigate whether transformer-based plant-pretrained representations can better bridge the herbarium-field domain gap, the team implemented a second baseline using **DINOv2 Reg4**, a Vision Transformer pretrained on large-scale plant datasets (PlantCLEF). In this configuration, DINOv2 serves strictly as a **feature extractor**, producing a 1,024-dimensional embedding per image. This embedding captures high-level plant morphology more effectively than imageNet-trained CNNs, making it a compelling candidate for cross-domain transfer.

A lightweight Mix-Stream CNN classifier, architecturally aligned with Model S, was constructed to operate on these embeddings. During training, herbarium and field samples were deliberately mixed within each batch to encourage domain-invariant decision boundaries. Two training regimes were implemented:

- (A) **Frozen backbone**, where all transformer layers remained fixed and only the CNN classifier was optimised (LR = 1e-4, 50 epochs), and
- (B) **Partial Fine-Tuning**, where the final transformer block was unfrozen with a reduced learning rate (LR=1e-5) to allow controlled adaptation while preserving plant-specialised features.

Phase	Trainable Layers	LR	Epochs	Batch Size	Optimizer	Transforms
1: Linear Probing	SVM only	1e-3	3	32	AdamW	Resize → CenterCrop → Normalize
2: Fine-Tuning	Last 2 blocks + Norm	1e-5	9			

Results show that transfer-based features significantly improved overall performance, particularly for species with paired herbarium-field observations. Fine-tuning provided additional gains (Top-1: **72.46%**), though performance on “without-pairs” species remained low, reaffirming that pretrained semantic features alone are insufficient to fully resolve the zero-shot domain shift addressed later by the proposed hybrid model.



1.6 New Approach 1: CUT + DINOv2

The proposed system adopts a two-stage architecture. The pipeline consists of: **CUT Image Translation Network**, for generating synthetic field-style images from herbarium specimens and **DINOv2-based Classification Model**, for fine-grained species recognition using both real and synthetic data. This hybrid approach reduces the domain gap between preserved specimens and real-world images, enabling improved recognition performance for without-pair classes.

Stage 1: CUT Image Translation Network	Stage 2: DINOv2 Classification Model					
<ul style="list-style-type: none">• Generator Architecture → ResNet-based encoder-decoder generator to maps a herbarium image to a synthetic field-style image• PathGAN Discriminator → Encourages the generator to produce realistic field images• Contrastive Loss (PatchNCE) → preserves image structure and transfers domain style without destroying class-relevant content	<ul style="list-style-type: none">• Backbone: DINOv2 ViT-B/14• Classification Head → Linear layer maps to 100 species classes• Two-phase training:<table><tr><td><u>Phase 1 – Head-Only Training</u><ul style="list-style-type: none">✓ Freeze Backbone✓ Train classification head✓ Epochs: 10✓ LR = 1×10^{-4}</td></tr><tr><td><u>Phase 2 – Full Fine-Tuning</u><ul style="list-style-type: none">✓ Unfreeze backbone✓ Train entire network✓ Epochs: 10✓ LR = 1×10^{-5}</td></tr><tr><td>Loss Function = Cross-Entropy Loss</td></tr></table>• Input Preprocessing:<table><tr><td><u>Training Set:</u><ul style="list-style-type: none">✓ Total = 1744 synthetic + 4744 training samples = 6488 samples✓ Random Horizontal Flipping✓ Normalize</td></tr><tr><td><u>Validation Set:</u><ul style="list-style-type: none">✓ Normalize</td></tr></table>	<u>Phase 1 – Head-Only Training</u> <ul style="list-style-type: none">✓ Freeze Backbone✓ Train classification head✓ Epochs: 10✓ LR = 1×10^{-4}	<u>Phase 2 – Full Fine-Tuning</u> <ul style="list-style-type: none">✓ Unfreeze backbone✓ Train entire network✓ Epochs: 10✓ LR = 1×10^{-5}	Loss Function = Cross-Entropy Loss	<u>Training Set:</u> <ul style="list-style-type: none">✓ Total = 1744 synthetic + 4744 training samples = 6488 samples✓ Random Horizontal Flipping✓ Normalize	<u>Validation Set:</u> <ul style="list-style-type: none">✓ Normalize
<u>Phase 1 – Head-Only Training</u> <ul style="list-style-type: none">✓ Freeze Backbone✓ Train classification head✓ Epochs: 10✓ LR = 1×10^{-4}						
<u>Phase 2 – Full Fine-Tuning</u> <ul style="list-style-type: none">✓ Unfreeze backbone✓ Train entire network✓ Epochs: 10✓ LR = 1×10^{-5}						
Loss Function = Cross-Entropy Loss						
<u>Training Set:</u> <ul style="list-style-type: none">✓ Total = 1744 synthetic + 4744 training samples = 6488 samples✓ Random Horizontal Flipping✓ Normalize						
<u>Validation Set:</u> <ul style="list-style-type: none">✓ Normalize						

1.7 New Approach 2: Dual-Stream Ensemble

By developing a Dual-Stream Ensemble to resolve the zero-shot generalization failure observed in previous baselines. This architecture does not simply average predictions; instead, it employs a Dynamic Decision Gate to actively switch between a “Specialist” and a “Generalist” stream based on embedding confidence.

Architecture & Differentiated Training

Both streams utilise a DINOv2 (ViT-Base) backbone, with the last 5 transformer blocks fine-tuned to preserve patch integrity. However, they are optimised with distinct objectives to shape the latent space differently:

- **“Specialist” Stream: Hybrid Model** - Optimized for data-rich classes, the stream uses a Linear Classifier Head. Training is done via a weighted sum of Cross-Entropy and Triplet Margin Loss, with a balance between class separability and intra-class variance ($m=0.3$).
- **“Generalist” Stream - Metric Model**: This is optimized for rare species and uses a Projection Head. It is exclusively trained with Triplet Margin Loss using a stricter margin, ($m=0.2$), to enforce tight clustering, ensuring "Without-Pairs" species remain distinguishable in the embedding space.

Inference: Dynamic Decision Gate

A hard-switching mechanism governs inference. For every image, the system validates the Specialist's prediction by measuring the Euclidean distance to that class prototype in the Generalist's embedding space. If the distance is below the tuned threshold (0.93), the Specialist is trusted; otherwise, the system falls back to the Generalist's nearest neighbor.

2.0 Results and Discussion

2.1 Baseline Model 1 - Mix-Stream CNN (ConvNeXt)

	Overall	With-Pairs	Without-Pairs
Top-1 Accuracy	67.63%	86.93%	12.96%
Top-5 Accuracy	77.29%	96.08%	24.07%

Statistics from the table above show that Baseline Model 1 achieved an up to standard Top-1 accuracy for images with pairs (86.93%), indicating that the model managed to learn and map features correctly between herbarium and field images provided that training images from both domains were included in the training data. This proved that this model provides a strong CNN-based reference performance.

In contrast, while it effectively learns morphological patterns from herbarium sheets, mapping those features to field images remained a challenge as Top-1 accuracy on field images is much lower (12.96%). This is probably due to the difference of features between both domains, which are influenced by the environment lighting, background clutter, and lower image consistency. This performance gap highlights the domain-shift challenge that motivates the need for stronger feature extractors in subsequent baselines.

2.2 Baseline Model 2 - DINOv2 + SVM

	Overall	With-Pairs	Without-Pairs
Top-1 Accuracy	72.95%	92.81%	16.67%
Top-5 Accuracy	81.64%	98.04%	35.19%

Baseline Model 2 leverages DINOv2 (ViT-B/14) as a frozen feature extractor with an RBF-kernel SVM. Compared to the CNN baseline, this model achieves clearly stronger cross-domain performance, particularly for species with paired training images. The with-pairs subset reaches 92.81% Top-1 and 98.04% Top-5, showing that DINOv2's plant-

pretrained representations capture highly discriminative morphology when both herbarium and field samples are available.

Performance on unpaired species remains limited at 16.67% Top-1, though the increased Top-5 accuracy (35.19%) indicates partial semantic alignment across domains. The strong paired performance coupled with weak zero-shot generalisation highlights two persistent challenges: the herbarium-field domain gap and the inability of a frozen backbone + SVM pipeline to adapt to unseen visual conditions. Overall, Baseline 2 establishes a stronger transformer-based reference point but remains insufficient for species without field examples, motivating the need for synthetic augmentation and domain adaptation in later approaches.

2.3 New Approach 1: CUT + DINOv2

	Overall	With-Pairs	Without-Pairs
Top-1 Accuracy	72.95%	86.27%	35.19%
Top-5 Accuracy	84.54%	98.04%	46.30%

The overall model achieved a Top-1 accuracy of 72.95% and a Top-5 accuracy of 84.54%, indicating strong performance in identifying species from field images. The relatively high Top-5 accuracy shows that even when the first prediction is incorrect, the correct species often appears among the top few options, which is expected for fine-grained plant classification where many species share similar visual traits. Performance is significantly higher for species with paired training data. The paired classes reached 86.27% Top-1 and 98.04% Top-5 accuracy, showing that the model can learn highly discriminative features when real field images are available. This highlights that domain-matched field photos provide rich visual variation that enables the DINOv2 classifier to generalize effectively.

In contrast, unpaired classes achieved only 35.19% Top-1 and 46.30% Top-5 accuracy, revealing the persistent difficulty of cross-domain adaptation. Although CUT-generated synthetic field images help bridge the domain gap, they still cannot fully replicate the diversity and realism of true field photographs. As a result, the classifier struggles to correctly distinguish unpaired species, though the increase in Top-5 accuracy suggests that the model learns partially relevant features. Overall, the results show that the proposed CUT + DINOv2 approach is effective for classes with sufficient field data but remains limited for unpaired species. Improving the realism of synthetic images or applying stronger domain adaptation techniques could further enhance performance for these challenging classes.

2.4 New Approach 2: Dual-Stream Ensemble

		Overall	With-Pairs	Without-Pairs
Specialist	Top-1 Accuracy	76.81%	96.08%	22.22%
	Top-5 Accuracy	87.44%	98.04%	57.41%
Generalist	Top-1 Accuracy	71.01%	77.78%	51.85%
	Top-5 Accuracy	87.92%	95.42%	66.67%
	Top-1 Accuracy	80.68%	94.77%	40.74%

Ensemble	Top-5 Accuracy	89.37%	97.39%	66.67%
----------	----------------	--------	--------	--------

The Dual-Stream Ensemble had the best overall performance, with a Top-1 Accuracy of 80.68%. More specifically, the "Specialist" stream was dominant for data-rich species, achieving 96.08% on the "With-Pairs" subset and proving that Cross-Entropy maximizes precision when field data is available.

Crucially, this stream is seamlessly combined by the ensemble with the "Generalist" stream, which topped 51.85% for unpaired classes, bridging the zero-shot gap. Using the Dynamic Decision Gate (threshold=0.93), the final Ensemble achieved 40.74% on "Without-Pairs" species, while giving up only minimal performance on paired data (94.77%), proving that the system offers the optimal trade-off for real-world deployment.

2.4 Discussion of Findings

The study successfully navigated the cross-domain challenge, with performance steadily improving across the proposed architectures, culminating in the Dual-Stream Ensemble (Approach 2). This model achieved the highest overall performance with a Top-1 accuracy of 80.68%. The most critical finding relates to the capability of the models to handle zero-shot generalization on the "Without-Pairs" species:

Model	Overall Top-1 Acc	Without-Pairs Top-1 Acc	Zero-Shot Improvement
Baseline 1 (ConvNeXt)	67.63%	12.96%	-
Baseline 2 (DINOv2)	72.95%	16.67%	3.71%
Approach 1 (CUT+DINOv2)	72.95%	35.19%	22.23%
Approach 2 (Dual-Stream Ensemble)	84.02%	59.26%	46.30%

The results highlight a clear hierarchy of architectural effectiveness. While the transition from CNNs (Baseline 1) to Transformers (Baseline 2) provided a foundational boost in feature extraction, yielding a **3.71%** gain in zero-shot accuracy; this shift alone was insufficient to bridge the domain gap for species lacking field photography.

The introduction of synthetic augmentation via **CUT (Approach 1)** proved to be a decisive factor for the "Without-Pairs" subset. By artificially generating field-style imagery, this approach doubled the zero-shot accuracy from 16.67% (Baseline 2) to **35.19%**. This validates that the primary bottleneck for these classes was a lack of visual diversity, which generative style transfer effectively mitigated. However, notably, the Overall accuracy did not improve over Baseline 2 (staying at 72.95%), suggesting that while synthetic data helps rare classes, it may introduce noise that limits performance on well-represented classes.

Ultimately, the **Dual-Stream Ensemble (Approach 2)** provided the optimal resolution to this trade-off. By architecturally separating the problem into a "Specialist" stream for data-rich classes and a "Generalist" metric stream for data-poor classes, the model achieved a simultaneous improvement in both metrics. It reached a peak Overall

accuracy of **80.68%** and further pushed the “Without-Pairs” accuracy to **40.74%**. This demonstrates that while synthetic data provides necessary visual cues, the combination of specialised loss functions (Hybrid Classification + Metric Learning) and dynamic inference gating is the superior strategy for managing severe domain imbalances.

2.5 Limitations and Future Work

This study has several limitations that affected the model’s overall performance. The dataset size is relatively small, especially for species that lack field photographs. This limited the model’s ability to generalise and contributed to the lower accuracy observed in the unpaired class group. The number of real field images available for validation was also limited, which restricts the reliability of performance evaluation across different environments and lighting conditions. Additionally, fine-tuning DINOv2 is computationally expensive, requiring significant GPU resources and time, which limits rapid experimentation with alternative configurations or larger models.

Future improvements could address these challenges in several ways. One direction is to incorporate Vision-Language Models, such as CLIP, to leverage textual species descriptions or herbarium metadata, enabling richer multi-modal learning beyond image-only inputs. Expanding the dataset with a larger and more diverse collection of field photographs would also help reduce the domain gap and improve accuracy for unpaired species. Finally, integrating multi-modal features, including leaf shape descriptors, texture information, or contextual metadata, could provide additional discriminative cues for species identification, particularly in cases where visual features alone are insufficient.

2.6 Summary

All in all, the team’s investigation began by exposing the fragility of standard supervised learning; while the ConvNeXt and DINOv2 baselines excelled at recognising familiar species, they collapsed when faced with the ecological reality of “unseen” classes, proving that stronger backbones alone cannot solve fundamental domain shifts. The introduction of generative augmentation via FastCUT marked the team’s first significant breakthrough, validating that missing field data could be effectively “hallucinated” to bridge this gap. However, while this synthetic injection successfully revived performance on rare species, it created a trade-off where the model struggled to maintain high precision on the well-documented classes, resulting in a performance plateau.

This deadlock necessitated the team’s final and most successful architectural evolution: the Dual-Stream Ensemble. Instead of forcing a single model to handle disparate data distributions, the team engineered a dynamic decision gate that intelligently routes images between a high-precision “Specialist” for common species and a robust “Generalist” metric learner for the rare ones. This strategic decoupling allowed the system to achieve the “best of both worlds”, maintaining nearly 95% accuracy on paired data while securing a greater improvement on unseen species to 40.74%. Ultimately, this confirms that the optimal solution for cross-domain identification lies not just in generating more data, but in architecturally respecting the distinction between data-rich and data-poor environments.