Synergizing Foundation Model Transfer and Phenological Information for Fine-Grained Forest Segmentation

Youcef Ben Ghorbel, Guneet Mutreja, Mareike Weishaupt, Jiaojiao Tian

German Aerospace Center (DLR), Earth Observation Center (EOC),
Münchener Str. 20, 82234 Oberpfaffenhofen, Germany
(youcef.benghorbel@dlr.de, guneet.mutreja@dlr.de, mareike.weishaupt@dlr.de, jiaojiao.tian@dlr.de)

Keywords: tree species classification, UAV imagery, phenology, semantic segmentation, deep learning, foundation models

Abstract

Accurate mapping of tree species is fundamental for sustainable forest management, biodiversity monitoring, and ecological research. Recent advances in Uncrewed Aerial Vehicle (UAV) photogrammetry provide rich spatial and spectral information at unprecedented detail (0.02 m Ground Sampling Distance). Meanwhile, the development of large-scale foundational models, pre-trained on expansive, multi-modal remote sensing datasets, offers highly transferable representations critical for advancing geoscience applications. This paper investigates the underexplored integration of foundation model pre-training with multi-temporal high-resolution UAV imagery. We introduce a novel two-phase framework: first, leveraging a Vision Transformer (ViT)-based foundation model (FoMo-Net) pre-trained on the multi-scale FoMo-Bench benchmark to initialize a DeepLabv3+ architecture; and second, fusing multi-temporal UAV data (May and September RGB) through change composites and pseudo-labeling to exploit species-specific phenology. The proposed approach, tested on the multi-class Québec Trees Dataset, yields an Overall Accuracy (OA) of 78.21%, demonstrating that foundation model initialization significantly boosts feature generalization, while multi-temporal cues are essential for disambiguating closely related species.

1. Introduction

Understanding the composition of tree species is essential for studying forests and efficient management. It supports biodiversity conservation, carbon estimation, and ecosystem modeling (Gamfeldt et al., 2013; Wessely et al., 2024; Vorster et al., 2020). However, collecting species-level information across large and diverse forest areas remains a major challenge. Traditional field inventories are reliable but require a lot of time and effort, and they cover only small areas (Tomppo et al., 2010), which limits their use for large-scale monitoring.

Remote sensing has become a key tool for forest research because it allows large-scale and repeated observations of forest structure and composition. In recent years, Uncrewed Aerial Vehicles (UAVs) have become especially useful. They can capture very high-resolution RGB images (0.02 m GSD), which makes it possible to identify individual tree crowns and their characteristics (Kattenborn et al., 2019; Schiefer et al., 2020; Franklin et al., 2022). When data from multiple seasons of the year are available, these images can also show seasonal changes in color and leaf density, helping to distinguish species that look similar at a single time point.

Deep learning models have shown strong performance in tree species classification, but most approaches, especially Convolutional Neural Networks (CNNs), need large annotated datasets that are expensive to create (Brodrick et al., 2019; Kattenborn et al., 2021). Their performance can also drop in complex temperate forests, where crowns overlap and lighting conditions vary (Zhang et al., 2022; Beloiu et al., 2023; Gan et al., 2023). The introduction of large foundation models (FMs) trained on global remote sensing data offer a new way forward. These models can learn general, transferable features and reduce the need for large local training sets. While some studies have investigated how phenology and multi-temporal data influence classification (Cloutier et al., 2024; Liang et al., 2025), the combination of FM pre-training and multi-temporal UAV imagery for detailed species mapping has not yet been fully explored.

In this work, we propose a two-phase framework for fine-grained tree species segmentation that explicitly decouples structural feature learning from temporal semantic refinement. In the first phase, a DeepLabv3+ model is initialized using a Vision Transformer-based foundation model (FoMo-Net), enabling the transfer of generalized forest representations across scales. In the second phase, we introduce an explicit temporal fusion strategy based on difference composites and pseudo-label refinement to capture species-specific phenological variations.

Unlike previous approaches that directly combine multi-temporal inputs, our framework separates spatial representation learning from temporal discrimination, allowing each component to be learned more effectively. We evaluate this design on the Québec Trees Dataset (Cloutier et al., 2024) and demonstrate consistent improvements over both single-date and multi-temporal baselines.

2. Related Work

2.1 Phenology and Multi-temporal UAV Data

Seasonal changes in leaf color and canopy structure carry valuable information that can help distinguish species. Several studies using satellite time-series data from Landsat and Sentinel-2 have shown that including this kind of phenological information can significantly improve forest classification (Zhang et al., 2018).

For example, Cloutier et al. (2024) analyzed UAV images acquired at seven times during the growing season in a temperate

mixed forest in Québec. They found that early autumn images performed best because of stronger color differences between species while late autumn images were less effective due to uneven leaf fall. Similar results have been found using satellite data, where the best month for classification often depends on the forest type and local climate (Zhang et al., 2018).

Using several acquisitions together can also make models more stable. Liang et al. (2025) tested four deep learning architectures on UAV images from six different time points and showed that object-based approaches performed better than pixel-based ones, especially when spring and autumn data were combined.

Overall, these studies agree that including phenological information, either by choosing the best acquisition time or by combining data from multiple dates, helps improve species classification.

2.2 Foundation Models for Remote Sensing and Forest Monitoring

Traditional CNNs have been successful in tree species classification, but depend on limited datasets and struggle to adapt to new forest types or sensors. Recently, Remote Sensing Foundation Models (RSFMs) (Sun et al., 2022; Cong et al., 2022) have introduced a major shift by using large-scale multimodal pretraining to learn general and transferable representations across diverse remote sensing data.

From CNNs to Foundation Models RSFMs learn transferable spatial and spectral features that improve performance in specialized applications by boosting stability and speeding up convergence. Most RSFMs are trained using masked autoencoding (MAE) (Reed et al., 2023), which requires the network to reconstruct masked input patches, thereby forcing the model to learn holistic context and fundamental data structure, typically utilizing a Vision Transformer (ViT) (Neil and Dirk, 2020) encoder. Other approaches use contrastive learning to align features across different sensors, such as optical and SAR, or across multiple time periods.

Examples of RS Foundation Models Recent RSFMs differ in architecture and application focus:

- **FOMO-Net:** specifically developed for forest monitoring and pre-trained on the Forest Monitoring Benchmark (FoMo-Bench), which includes 15 datasets from various sensors and spatial resolutions for both classification and segmentation tasks (Bountos et al., 2025). Its single, flexible backbone can handle up to 36 different remote sensing modalities, from a few centimeters to 60 meters GSD.

- **SkySense:** A multimodal model (Guo et al., 2024) that performs well across geospatial tasks. However, its use of separate encoders for spectral groups can make it less adaptable than FoMo-Net when working with new data types.

- **OmniSat:** A self-supervised model (Astruc et al., 2024) that fuses VHR, Sentinel-1, and Sentinel-2 data. Its success in tree species classification supports the importance of combining temporal and multimodal information for fine-grained forest mapping.

Why FoMo-Net Was Chosen We selected FoMo-Net for this study because its design and training data align closely with our application. It includes UAV-scale imagery (below 5 cm GSD) and focuses directly on forest monitoring, making it a strong starting point for high-resolution tree crown segmentation. It yields a more effective initialization than models trained solely on coarser-resolution satellite imagery or general vision tasks.

3. Dataset

3.1 Québec Trees Dataset

This study uses the Québec Trees Dataset introduced by Cloutier et al. (2024). The dataset covers a temperate mixed forest at the Montmorency Research Station in Saint-Hippolyte, Québec, Canada. It includes three separate areas (zones 1–3), each with 0.02 m resolution RGB orthomosaics tree crown annotations.

In total, 22,139 crowns were manually annotated and categorized into 14 classes, including 11 species, 2 genera, and one class representing dead trees. A summary of these classes and their abbreviations is provided in Table 1.

Table 1. Tree species, classes, and pixel percentages in the Québec Trees Dataset.

Abbreviation	Species / Class	Pixel %
BEPA	*Betula papyrifera*	31.60
ACSA	*Acer saccharum*	7.45
ACRU	*Acer rubrum*	23.20
ABBA	*Abies balsamea*	4.82
THOC	*Thuja occidentalis*	2.74
Picea	*Picea* sp.	2.02
Mort	Dead trees	1.06
BEAL	*Betula alleghaniensis*	2.74
TSCA	*Tsuga canadensis*	0.36
ACPE	*Acer pensylvanicum*	1.79
FAGR	*Fagus grandifolia*	1.36
PIST	*Pinus strobus*	6.27
LALA	*Larix laricina*	0.54
Populus	*Populus* sp.	14.10

3.2 Data preprocessing

For this study, we selected the UAV acquisitions from 28-05-2021 and 02-09-2021, as combining spring and autumn imagery has been shown to improve model stability and performance(Liang et al., 2025).

Before training, the orthomosaics and their corresponding polygon annotations were divided into smaller image tiles and raster masks. Each tile measures 256×256 pixels, corresponding to an area of roughly 5×5 m on the ground. To avoid including empty or ambiguous regions, only tiles with at least 25% annotated area were retained.

The dataset was split into training, validation, and test subsets according to geographic location. Zones 1 and 2 were used for training and validation, while zone 3 was used as an independent test set area to evaluate the model's ability to generalize to unseen areas.

Rasterized masks were generated directly from the GPKG vector polygons, and perfectly aligned with their corresponding image tiles. Each pixel in a mask was assigned to one of the 14 species classes or labeled as background.

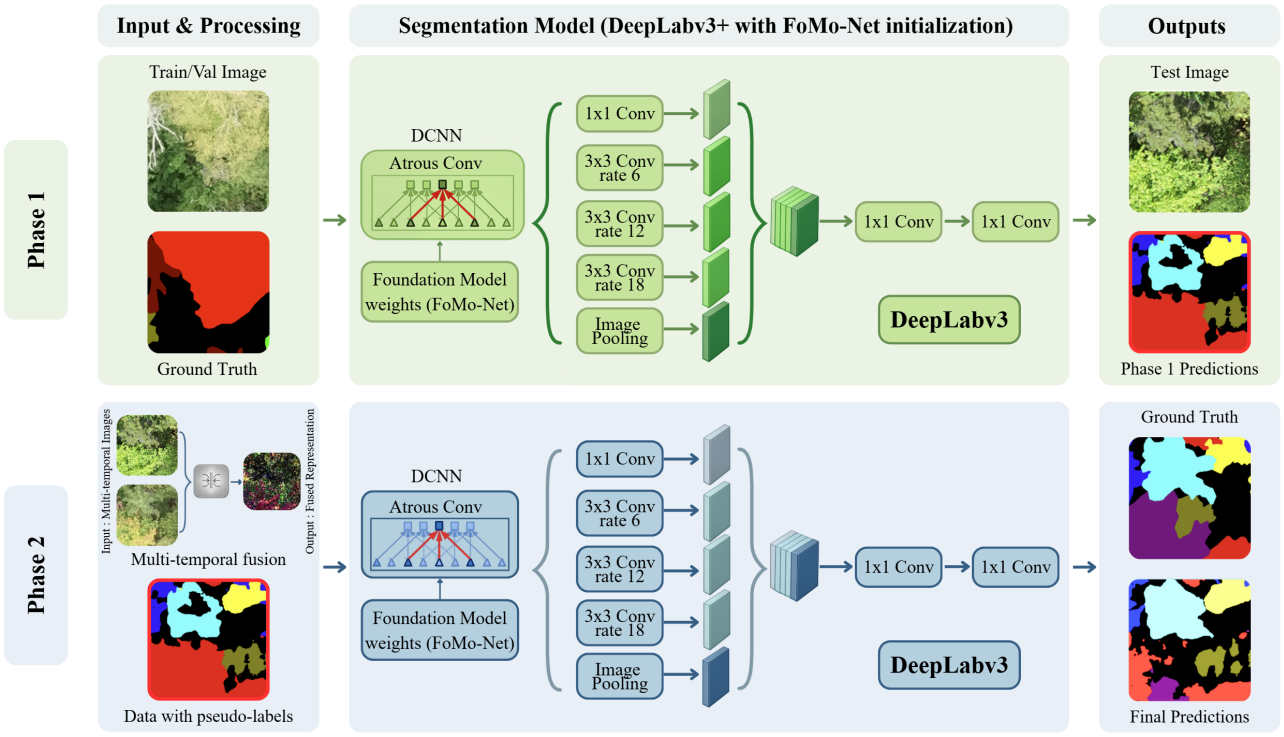


Figure 1. Overview of the proposed two-phase framework. Phase 1 performs foundation model initialization and fine-tuning on single-date imagery to learn structural features. Phase 2 introduces multi-temporal difference composites (D_c) and pseudo-label refinement to incorporate phenological information and improve species discrimination.

4. Methodology

4.1 General Overview: Integration of Foundation Model Pre-training

The overall workflow, illustrated conceptually in Figure 1, is designed to enhance species-level crown segmentation performance by combining generalized feature representation learning through foundation model initialization with domain-specific knowledge integration via a self-training, multi-temporal fusion procedure.

4.2 System Architecture and Initialization

The core network architecture employed is DeepLabv3+. DeepLabv3+ is a state-of-the-art convolutional neural network for dense semantic segmentation, combining atrous convolutions and the Atrous Spatial Pyramid Pooling (ASPP) module to effectively capture multi-scale context, which is essential for delineating complex, overlapping crown boundaries in VHR imagery.

The DeepLabv3+ encoder backbone, typically a ResNet, is initialized using weights derived from FoMo-Net, which utilized a ViT encoder trained on the multi-scale FoMo-Bench. This transfer process maps the robust, high-level structural and spectral features learned by the ViT-based foundation model across diverse scales (0.02 m to 60 m GSD) to the convolutional parameters of the DeepLabv3+ encoder. This cross-architectural knowledge transfer provides a significantly stronger starting point than standard ImageNet pre-training or training from scratch, effectively mitigating the constraints imposed by the limited size of high-resolution annotated datasets available for the Québec region. The necessity for this initialization is underscored by the poor performance of the model trained entirely from scratch (52.79% OA).

4.3 The Proposed Two-Phase Training Strategy

The training is structured into two sequential phases to strategically leverage the strengths of the foundation model and the multi-temporal data. This modular design decouples the learning of generalized spatial features (Phase 1) from the resolution of highly specific phenological ambiguities (Phase 2).

Phase 1: Foundational Fine-tuning (Spatial Feature Learning) In the first phase, the FoMo-Net initialized DeepLabv3+ model is fine-tuned exclusively on images from a single acquisition date (e.g., September), along with the corresponding ground truth segmentation masks. This step rapidly leverages the pre-trained foundation model features, optimized for general forest structure, texture, and geometry, to establish a robust baseline for crown delineation and segmentation. The resulting substantial increase in performance (OA 71.21% compared to 52.79% baseline) confirms the success of the cross-scale feature transfer, demonstrating that FMs generalize structural understanding downward from low/medium resolution satellite data to solve VHR problems.

Phase 2: Multi-temporal Fusion and Prediction Refinement (Semantic Feature Resolution) The second phase introduces temporal information, exploiting phenological differences between May (t_1) and September (t_2) acquisitions to refine species discrimination.

- **Input Augmentation using Temporal Composites:** The standard RGB input is expanded to include synthesized multi-temporal features. For each RGB channel, an absolute difference composite (D_c) is computed as:

$$D_c = |I_{t_1,c} - I_{t_2,c}|. \quad (1)$$

These difference images mathematically highlight subtle, pixel-wise phenological changes, such as the magnitude and timing of leaf senescence, which are crucial for distinguishing between species that appear visually similar at any single date. The ability of the network to process these derived, non-standard inputs is inherently supported by the FoMo-Net's training philosophy, which favors a single, sensor-agnostic projection layer ($FoMo-Net_1$ outperformed $FoMo-Net_m$).

Figure 2 shows an example of the May and September images and their derived difference composites. These representations make seasonal changes in leaf coloration and canopy density clearly visible, which are key indicators for differentiating tree species.

2021-05-28 2021-09-02 Diff

Figure 2. Illustration of the multi-temporal fusion process. From left to right: May RGB image (I_{t_1}), September RGB image (I_{t_2}), and absolute difference composite (D). The derived product emphasizes phenological and spectral variations that help distinguish tree species.

- **Pseudo-Label Integration (Self-Training):** In Phase 2, the model leverages pseudo-labels generated from Phase 1 predictions on the *held-out test area* (Zone 3). These pseudo-labels are used together with the corresponding *difference images* (D_c), rather than single-date RGB inputs. This self-training process allows the network to refine its feature representations by exploiting temporal variations in canopy structure and color, while still preserving the robust spatial-semantic priors learned during Phase 1.

4.4 Implementation details

The DeepLabv3+ architecture(Chen et al., 2018) was initialized with FoMo-Net foundation model weights, pre-trained on FoMo-Bench, which integrates data from seven satellite-based datasets (e.g., BigEarthNet-MM, TalloS) and eight aerial-based datasets (e.g., NeonTree, FLAIR #1, Woody). This comprehensive pre-training allows the model to develop general representations of tree crown geometry, canopy texture, and species-specific radiometric traits across varied forest environments and imaging conditions.

Through this second phase, the model integrates temporal variation and prior knowledge, enabling it to recognize phenological cues that are not evident when training with a single image date.

4.5 Training Setup

Data augmentation was applied during training to improve model robustness. Each image had a 50% chance of being randomly rotated, and horizontal and vertical flips were applied with a probability of 0.75 each. Training was performed using the Adam optimizer with a learning rate of 1×10^{-3} for 40 epochs. We used an NVIDIA TITAN RTX GPU (24 GB) with a batch size of 16. Early stopping based on validation loss was employed to prevent overfitting.

4.6 Evaluation Metrics

Model performance was evaluated using **overall accuracy (OA)** and **per-class F1-score**. Overall accuracy is given by:

$$\mathrm{OA} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{2}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

The **F1-score** for each class is computed as:

$$\mathrm{F1} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}, \tag{3}$$

with $\mathrm{Precision} = \frac{TP}{TP+FP}$ and $\mathrm{Recall} = \frac{TP}{TP+FN}$. In addition, normalized confusion matrices were produced to analyze patterns of misclassification, particularly among visually similar taxa such as the *Acer* species and closely related conifers.

5. Results

5.1 Quantitative Results

5.1.1 Per-Class and Overall Accuracy The evaluation demonstrated significant performance gains achieved through the foundation model initialization and the subsequent multi-temporal fusion, as summarized in the table 2.

Table 2. Tree species recognition performance using DeepLabv3+ with FoMo-Net foundation weights, compared with the two-timepoint baseline. OA is reported for each configuration.

Source	**Liang et al. (2025)**		**Ours**	
Tree Species	**B (%)**	**2tpB (%)**	**P1 (%)**	**P2 (%)**
Betula papyrifera (BEPA)	43.10	**83.42**	71.85	72.33
Acer saccharum (ACSA)	38.56	53.82	82.19	**90.10**
Acer rubrum (ACRU)	68.47	**79.09**	53.21	74.82
Abies balsamea (ABBA)	25.16	53.45	**73.48**	72.73
Thuja occidentalis (THOC)	35.38	69.83	71.56	**72.72**
Picea spp. (Picea)	23.04	70.22	75.06	**75.58**
Mort (dead trees)	43.59	48.77	46.69	**73.17**
Betula alleghaniensis (BEAL)	25.16	55.90	52.05	**63.12**
Tsuga canadensis (TSCA)	2.24	28.09	63.45	**65.46**
Acer pensylvanicum (ACPE)	38.56	10.94	40.52	**78.84**
Fagus grandifolia (FAGR)	10.20	**70.52**	31.85	54.29
Pinus strobus (PIST)	49.11	**75.41**	66.82	62.61
Larix laricina (LALA)	41.03	82.70	81.95	**92.96**
Populus spp. (Populus)	54.25	83.85	16.19	**89.90**
Overall Accuracy (OA)	52.79	63.84	71.21	**78.21**
Gain vs B	–	+11.05	+18.42	**+25.42**

B: baseline (single-date model). **2tpB:** two-timepoint baseline (trained on combined 02-09-2021 and 28-05-2021 acquisitions). **P1:** Phase 1 (FoMo-Net fine-tuning). **P2:** Phase 2 (multi-temporal fusion refinement).

The baseline model trained from scratch achieved a moderate OA of 52.79%, exhibiting high variability and poor prediction

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-19-2026 | © Author(s) 2026. CC BY 4.0 License.

for structurally complex or rare classes such as *Tsuga canadensis* (2.24%) and *Fagus grandifolia* (10.20%). The reference two-timepoint baseline improved OA to 63.84%, showing that incorporating temporal information can already provide meaningful benefits even without foundation model pre-training.

Phase 1 fine-tuning, utilizing the FoMo-Net foundation weights, resulted in a dramatic improvement, with OA rising to 71.21%. This represents an absolute gain of 18.42% over the one-timepoint baseline and 7.37% over the two-timepoint baseline, confirming that pre-training captures transferable forest-related features, particularly crown texture and shape, which generalize well to the VHR UAV domain.

Phase 2 , which incorporated multi-temporal imagery and Phase 1 predictions, further enhanced performance, achieving a final OA of 78.21%. This surpasses the two-timepoint baseline by +14.37%, demonstrating that the combination of foundation model initialization and targeted temporal fusion enables more effective exploitation of phenological cues. The most significant improvements were achieved for species frequently misclassified in both single-date and traditional two-date approaches, validating the necessity of incorporating foundation-model-based feature priors.

5.1.2 Species-Specific Performance Analysis The application of the two-phase approach yielded highly targeted improvements, especially for deciduous species characterized by strong seasonal variation. For instance, *Acer saccharum* (ACSA) increased from 53.82% (two-timepoint baseline) to 90.10% (Phase 2), while *Larix laricina* (LALA) improved from 82.70% to 92.96%. Even rare classes benefited significantly; *Fagus grandifolia* (1.36% pixel count) increased from 10.20% (baseline) to 54.29% (Phase 2). These results emphasize that integrating foundation model knowledge with temporal difference features produces superior generalization compared to conventional temporal baselines.

5.1.3 Confusion Matrix Analysis To better understand the relationships between species and the main sources of confusion, normalized confusion matrices for Phase 1 and Phase 2 are shown in Figure 3. Phase 1 reveals notable overlap between visually or structurally similar species, such as *Acer rubrum* and *Acer saccharum*. In contrast, Phase 2 substantially reduces these confusions, confirming that integrating temporal cues helps the model separate species with similar appearance at a single date.

5.2 Qualitative Visualization

Figure 4 illustrates four representative examples from the test set, comparing the predictions from Phase 1 and Phase 2.

Phase 2 predictions generally display sharper crown boundaries and fewer misclassifications, especially in mixed stands where several species co-occur. The improvement is particularly visible for deciduous species that undergo strong seasonal color variations. The legend associates class indices with species names and colors for clarity.

6. Discussion: Synergies in Foundational Transfer and Phenological Cues for Fine-Grained Forest Segmentation

6.1 Foundational Feature Transferability and Initialization Efficacy

The magnitude of the performance improvement,an 18.42% absolute increase in OA attributed solely to the switch from training from scratch to foundation model initialization (Baseline vs. Phase 1), is a powerful validation of the efficacy of foundational feature transfer in remote sensing. This result confirms the core hypothesis that RSFMs, pre-trained on massive, diverse datasets like FoMo-Bench, capture representations of forest structure and spectral variability that are highly transferable across sensor types and scales, even when transitioning from low-resolution satellite imagery to VHR UAV photogrammetry.

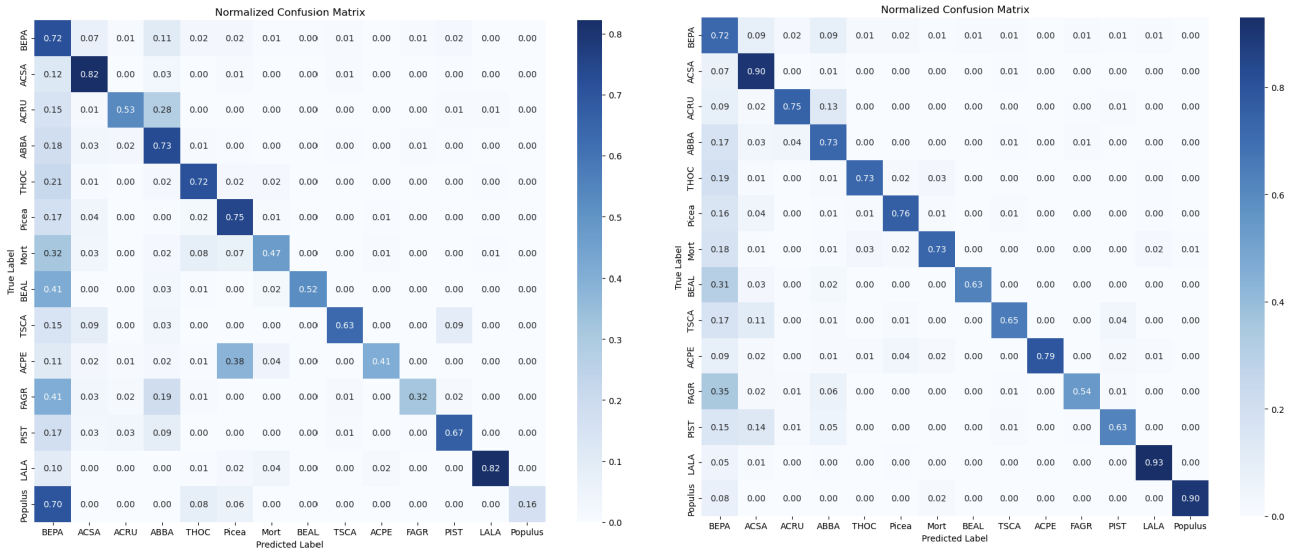
The transferred knowledge from FoMo-Net, which included aerial and LiDAR data during pre-training, provided the necessary scale-invariant understanding of canopy geometry and texture. This structural knowledge allowed the network to efficiently delineate individual crown boundaries, thereby overcoming the significant challenge posed by the structural complexity of temperate forests, which often feature overlapping and interlocking canopies. For high-resolution classification, where collecting large volumes of local, labeled data is infeasible, utilizing the FM as a powerful regularization and generalization mechanism successfully mitigates the data scarcity problem.

6.2 The Critical Role of Multi-Temporal Fusion in Resolving Semantic Ambiguity

While foundation model pre-training provides robust structural features (Phase 1 OA 71.21%), the subsequent 7.00% absolute gain in Phase 2 (OA 78.21%) demonstrates that specialization through phenological cues is mandatory for achieving state-of-the-art results in the fine-grained, species-specific domain of temperate forests.

The most notable gains occurred in deciduous taxa like Populus spp. and Acer pensylvanicum, which rely heavily on temporal spectral shifts for reliable differentiation. For example, the massive increase in Populus accuracy (from 16.19% in Phase 1 to 89.90% in Phase 2) warrants specific attention. This initial poor performance after FM initialization suggests that the generalized structural features learned globally by FoMo-Net may have introduced a bias or misclassification pattern resistant to single-date fine-tuning. However, this bias was effectively overcome when the model was presented with the explicit temporal contrast (Difference, D_c composite) in Phase 2 This indicates that for species with strong seasonal variation, temporal signals become the dominant discriminative features, overriding the structural representations learned during foundation model initialization.

The choice to utilize explicit, engineered change composites (D_c) rather than merely stacking raw multi-temporal RGB bands proved pivotal. These composites translate the complex concept of seasonal change into a direct, mathematically quantifiable spectral signature. For example, a high difference value in the red channel between May (green leaves) and September (autumn colors) definitively signals deciduous senescence, providing the classification network with a highly discriminative feature for species separation where simple spectral inputs alone would fail. The design of the two-phase



(a) Phase 1 (single-date). (b) Phase 2 (multi-temporal fusion).

Figure 3. Normalized confusion matrices for (a) Phase 1 and (b) Phase 2. Multi-temporal fusion reduces confusion between visually similar species.

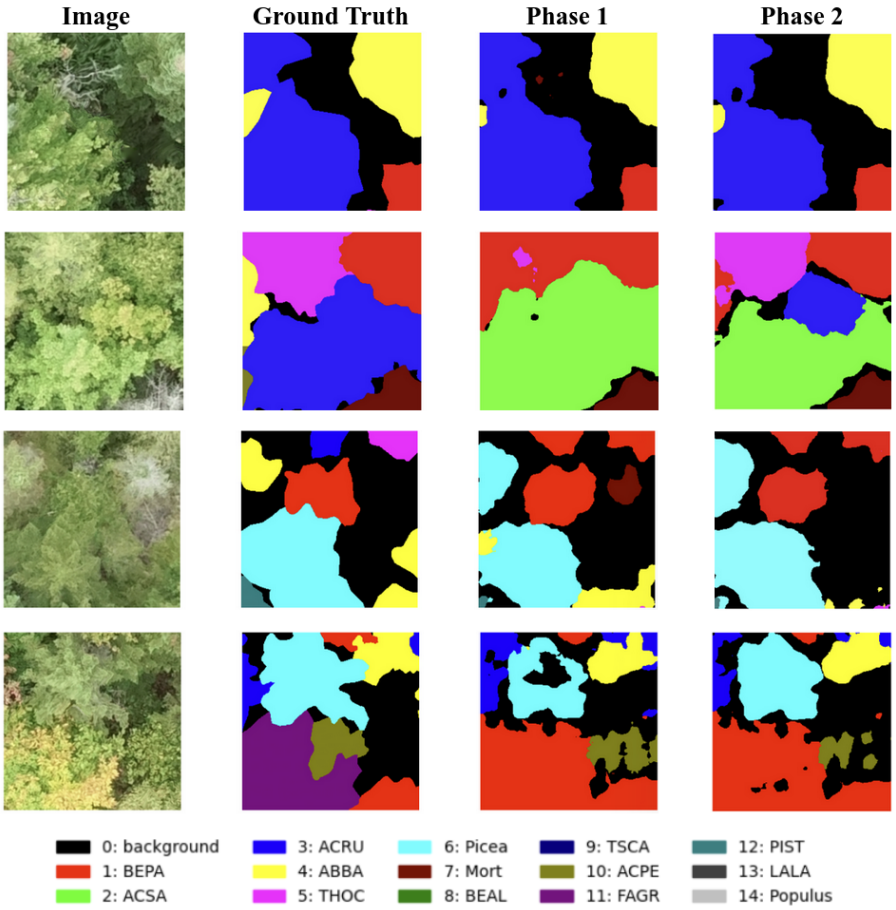


Figure 4. Comparison of semantic segmentation results for four representative input images. The figure contrasts predictions from single-timepoint training (Phase 1) with those from multi-temporal fusion finetuning (Phase 2). The legend maps species names and colors.

strategy, which decouples structural learning (Phase 1) from semantic/phenological refinement (Phase 2), ensures stability and maximizes the utility of both the pre-trained features and the domain-specific temporal data.

In this paper, we present a two-phase framework for fine-

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-19-2026

grained tree species segmentation that explicitly separates temporal semantic refinement from structural feature learning. A Vision Transformer-based foundation model (FoMo-Net) is used to initialize a DeepLabv3+ model in the first phase, allowing generalized forest representations to be transferred across scales. To capture species-specific phenological variations, we employ an explicit temporal fusion strategy based on difference composites and pseudo-label refinement in the second phase. Our framework separates temporal discrimination from spatial representation learning, enabling each component to be learned more efficiently, in contrast to earlier methods that directly combine multi-temporal inputs. We show consistent improvements over single-date and multi-temporal baselines by evaluating this design on the Québec Trees Dataset.

6.3 Architectural Context and Limitations

This study successfully demonstrated the viability of cross-architecture knowledge transfer: mapping ViT-FM learned representations to a CNN segmentation head (DeepLabv3+ encoder). This hybrid approach effectively leverages the global context understanding typical of transformers (learned efficiently during pre-training) with the computational efficiency and fine spatial refinement capabilities of CNNs (specifically, DeepLabv3+'s ASSP module) during fine-tuning.

A key challenge noted is the disparity between the FoMo-Net pre-training size (64×64 pixels for efficiency) and the fine-tuning requirement (256×256 pixels at 0.02 m GSD). Although the structural knowledge successfully transferred, performance for rare classes remains a limitation. Despite substantial gains for Fagus grandifolia (10.20% to 54.29%), extremely rare classes such as Tsuga canadensis (0.36% pixel count) still exhibited lower performance, demonstrating that data imbalance remains a significant hurdle in highly fine-grained multi-class segmentation, even with foundation model initialization. Furthermore, while segmentation accuracy was high, the difficulty in perfectly delineating crown boundaries due to heavy occlusion, internal shadows, and variable lighting in dense temperate forests persists as an inherent limitation of VHR optical imagery.

7. Conclusion

In order to enhance tree species segmentation and classification from UAV photos, we proposed a two-step strategy in this work. First, we initialized DeepLabv3+ with FoMo-Net, which boosted accuracy from 52.79% to 71.21%. Next, we combined May and September images, using spectral differences and pseudo labels to refine predictions, reaching 78.21%. Seasonal changes, like leaf color and canopy density, helped separate similar species, especially *Acer pensylvanicum*. Trees with strong seasonal patterns, such as *Populus* and *Larix laricina*, improved the most. Even rare species, like *Tsuga canadensis*, achieved better results. Overall, combining pre-trained models with multi-temporal data improves both accuracy and generalization for tree species segmentation.

References

Astruc, G., Gonthier, N., Mallet, C., Landrieu, L., 2024. Omnisat: Self-supervised modality fusion for earth observation. *European Conference on Computer Vision*, Springer, 409–427.

Beloiu, M., Heinzmann, L., Rehush, N., Gessler, A., Griess, V. C., 2023. Individual tree-crown detection and species identification in heterogeneous forests using aerial RGB imagery and deep learning. *Remote Sensing*, 15(5), 1463.

Bountos, N. I., Ouaknine, A., Papoutsis, I., Rolnick, D., 2025. Fomo: Multi-modal, multi-scale and multi-task remote sensing foundation models for forest monitoring. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39number 27, 27858–27868.

Brodrick, P. G., Davies, A. B., Asner, G. P., 2019. Uncovering ecological patterns with convolutional neural networks. *Trends in ecology & evolution*, 34(8), 734–745.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. *Proceedings of the European conference on computer vision (ECCV)*, 801–818.

Cloutier, M., Germain, M., Laliberté, E., 2024. Influence of temperate forest autumn leaf phenology on segmentation of tree species from UAV imagery using deep learning. *Remote Sensing of Environment*, 311, 114283.

Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D., Ermon, S., 2022. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35, 197–211.

Franklin, H., Veras, P., Pinheiro, M., Paula, A., Corte, D., Roberto, C., 2022. Fusing Multi-Season UAS Images with Convolutional Neural Networks to Map Tree Species in Amazonian Forests. Ecol. *Inform*, 71, 101815.

Gamfeldt, L., Snäll, T., Bagchi, R., Jonsson, M., Gustafsson, L., Kjellander, P., Ruiz-Jaen, M. C., Fröberg, M., Stendahl, J., Philipson, C. D., Mikusiński, G., 2013. Higher levels of multiple ecosystem services are found in forests with more tree species. *Nature Communications*, 4, 1340.

Gan, Y., Wang, Q., Iio, A., 2023. Tree crown detection and delineation in a temperate deciduous forest from UAV RGB imagery using deep learning approaches: Effects of spatial resolution and species characteristics. *Remote Sensing*, 15(3), 778.

Guo, X., Lao, J., Dang, B., Zhang, Y., Yu, L., Ru, L., Zhong, L., Huang, Z., Wu, K., Hu, D. et al., 2024. Skysense: A multimodal remote sensing foundation model towards universal interpretation for earth observation imagery. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27672–27683.

Kattenborn, T., Eichel, J., Fassnacht, F. E., 2019. Convolutional Neural Networks enable efficient, accurate and fine-grained segmentation of plant species and communities from high-resolution UAV imagery. *Scientific reports*, 9(1), 17656.

Kattenborn, T., Leitloff, J., Schiefer, F., Hinz, S., 2021. Review on Convolutional Neural Networks (CNN) in vegetation remote sensing. *ISPRS journal of photogrammetry and remote sensing*, 173, 24–49.

Liang, X., Chen, J., Gong, W., Puttonen, E., Wang, Y., 2025. Influence of data and methods on high-resolution imagery-based tree species recognition considering phenology: The case of temperate forests. *Remote Sensing of Environment*, 323, 114654.

Neil, H., Dirk, W., 2020. Transformers for image recognition at scale. *Online: https://ai. googleblog. com/2020/12/transformers-for-image-recognitionat. html*.

Reed, C. J., Gupta, R., Li, S., Brockman, S., Funk, C., Clipp, B., Keutzer, K., Candido, S., Uyttendaele, M., Darrell, T., 2023. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4088–4099.

Schiefer, F., Kattenborn, T., Frick, A., Frey, J., Schall, P., Koch, B., Schmidtlein, S., 2020. Mapping forest tree species in high resolution UAV-based RGB-imagery by means of convolutional neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 170, 205–215.

Sun, X., Wang, P., Lu, W., Zhu, Z., Lu, X., He, Q., Li, J., Rong, X., Yang, Z., Chang, H. et al., 2022. RingMo: A remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1–22.

Tomppo, E., Gschwantner, T., Lawrence, M., McRoberts, R. E., Gabler, K., Schadauer, K., Vidal, C., Lanz, A., Ståhl, G., Cienciala, E. et al., 2010. National forest inventories. *Pathways for Common Reporting. European Science Foundation*, 1, 541–553.

Vorster, A. G., Evangelista, P. H., Stovall, A. E. L., Ex, S., . . . , 2020. Variability and uncertainty in forest biomass estimates from the tree to landscape scale: the role of allometric equations. *Carbon Balance and Management*, 15, 8.

Wessely, J., Essl, F., Fiedler, K., Gattringer, A., Hülber, B., Ignateva, O., Moser, D., Rammer, W., Dullinger, S., Seidl, R., 2024. A climate-induced tree species bottleneck for forest management in Europe. *Nature Ecology & Evolution*, 8(6), 1109–1117.

Zhang, C., Zhou, J., Wang, H., Tan, T., Cui, M., Huang, Z., Wang, P., Zhang, L., 2022. Multi-species individual tree segmentation and identification based on improved mask R-CNN and UAV imagery in mixed forests. *Remote Sensing*, 14(4), 874.

Zhang, M., Lin, H., Wang, G., Sun, H., Fu, J., 2018. Mapping paddy rice using a convolutional neural network (CNN) with Landsat 8 datasets in the Dongting Lake Area, China. *Remote Sensing*, 10(11), 1840.