Environmental Data Science (2025), 4: e37, 1–23
doi:10.1017/eds.2025.10013

CAMBRIDGE UNIVERSITY PRESS

APPLICATION PAPER

Tree semantic segmentation from aerial image time series

Venkatesh Ramesh[1,2], Arthur Ouaknine[1,3] and David Rolnick[1,3]

[1]Mila, Quebec AI Institute, Montréal, QC, Canada
[2]Département d'informatique et de recherche opérationnelle, Université de Montréal, Montréal, QC, Canada
[3]School of Computer Science, McGill University, Montréal, QC, Canada
Corresponding author: Venkatesh Ramesh; Email: venka97@gmail.com

Received: 17 July 2024; **Revised:** 14 May 2025; **Accepted:** 13 June 2025

Keywords: deep learning; forest monitoring; phenology; remote sensing; time series

Abstract

Earth's forests play an important role in the fight against climate change and are in turn negatively affected by it. Effective monitoring of different tree species is essential to understanding and improving the health and biodiversity of forests. In this work, we address the challenge of tree species identification by performing tree crown semantic segmentation using an aerial image dataset spanning over a year. We compare models trained on single images versus those trained on time series to assess the impact of tree phenology on segmentation performance. We also introduce a simple convolutional block for extracting spatio-temporal features from image time series, enabling the use of popular pretrained backbones and methods. We leverage the hierarchical structure of tree species taxonomy by incorporating a custom loss function that refines predictions at three levels: species, genus, and higher-level taxa. Our best model achieves a mean Intersection over Union (mIoU) of 55.97%, outperforming single-image approaches particularly for deciduous trees where phenological changes are most noticeable. Our findings highlight the benefit of exploiting the time series modality via our Processor module. Furthermore, leveraging taxonomic information through our hierarchical loss function often, and in key cases significantly, improves semantic segmentation performance.

Impact Statement

This work advances forest monitoring using deep learning on aerial imagery time series. By leveraging phenological information and taxonomic hierarchies, our proposed methods improve tree species segmentation performance. The introduction of a compact spatio-temporal feature extraction module enables the use of pretrained models for this task. Our findings highlight the importance of incorporating temporal data and hierarchical knowledge in forest monitoring, and we hope our work will offer valuable insights for biodiversity conservation and climate change mitigation efforts.

1. Introduction

Climate change and biodiversity loss in forests are closely intertwined, with each potentially exacerbating the other. As the climate changes, the suitable habitat for many tree species shifts geographically, with ranges expanding in some regions while contracting or disappearing in others, leading to changes in forest composition and potential biodiversity loss (Lenoir et al., 2008; Allen et al., 2010; Mahecha et al., 2024).

This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

Conversely, biodiversity loss in forests can reduce their ability to absorb and store carbon, further contributing to climate change. Different tree species have varying tolerances to changes in temperature, precipitation, and other environmental factors. As a result, climate change can cause variable phenological changes (Visser and Gienapp, 2019), shifts in species distribution (Babst et al., 2019) and differential growth responses due to increased atmospheric CO_2 (Bonan, 2008; Anderegg et al., 2012). Phenology in trees refers to the timing of seasonal events such as leaf emergence, color change, and leaf fall. These cyclical changes are influenced by environmental factors like temperature and day length, and often vary between tree species. Understanding phenological patterns can potentially enhance our ability to distinguish between tree species and monitor their responses to environmental changes.

Increasingly, deep learning–based methods, alongside remote sensing applications (*e.g.*, land-use and land-cover mapping (Hamdi et al., 2019; Helber et al., 2019; Vali et al., 2020; Hamedianfar et al., 2022), change detection (Khelifi and Mignotte, 2020), have helped with advancing the field of forest monitoring in tree species classification (Fricker et al., 2019), biomass estimation (Zhang et al., 2019), and tree crown semantic segmentation (Schiefer et al., 2020; Weinstein et al., 2020).

The use of temporal data as inputs to these methods has also shown successes in other tasks such as crop mapping (Sainte Fare Garnot et al., 2020; Cai et al., 2023; Tarasiou et al., 2023) and forest health mapping (Hamdi et al., 2019). Semantic segmentation of tree crowns is a crucial task in forest monitoring as it provides valuable information about forest composition and health. It could be further explored by leveraging time-series inputs to learn phenological changes that occur between seasons according to each tree species throughout the years.

In this work, we evaluate multiple models on the task of tree crown semantic segmentation using a rich dataset recorded in the Laurentides region of Québec, Canada (Cloutier et al., 2024). Among the numerous datasets available for tree crown semantic segmentation (Ouaknine et al., 2025), we chose this one for its unique characteristics: high-resolution time-series data and a number of closely related classes. This allows us to investigate the impact of phenological (seasonal) changes on tree species identification and assess the ability of the model to distinguish between closely related species.

To this end, we employ state-of-the-art models in semantic segmentation for single-image and time-series segmentation. Additionally, we introduce a lightweight module to extract spatio-temporal features from a time-series input, allowing it to be used with backbones that typically operate on single images. The dataset we use lacks fine-grained species-level labels for all trees, as it is challenging to accurately identify tree species at a granular level. As a result, it is often easier to identify them on a coarser (genus or family) level. To address this, we propose a custom hierarchical loss function that incorporates labels from all three levels (species, genus, and family) and penalizes incorrect predictions at each level. Overall, our work can be summarized as follows:

- We introduce a simple yet effective module for extracting spatio-temporal features, enabling the use of pretrained models for segmenting tree crowns with time series.
- We find that time-series data improves species identification performance, particularly for deciduous trees.
- We demonstrate that models achieve better accuracy when leveraging taxonomic hierarchies through our proposed loss function.

2. Related Work

2.1. Semantic segmentation

Deep learning applications for computer vision have been widely explored over the years, including various methods based on convolutional neural networks (CNNs) such as Fully Convolutional Networks (FCNs) (Long et al., 2015), U-Net (Ronneberger et al., 2015), and DeepLab (Chen et al., 2018a).

The "dilated" (also named "atrous") convolution (Yu and Koltun, 2016; Chen et al., 2018a), has been introduced to increase the receptive field of CNNs, while attention mechanisms (Oktay et al., 2018; Fu et al., 2019) have been incorporated to focus on relevant regions. Multi-scale and pyramid pooling

approaches, such as PSPNet (Zhao et al., 2017) and DeepLabV3+ (Chen et al., 2018b), have been employed to capture context at different scales. Specific methods have also been designed to exploit temporal information for semantic segmentation, *e.g.*, with 3D U-Net (Çiçek et al., 2016) and V-Net (Milletari et al., 2016).

Recently, transformer-based models have gained popularity in semantic segmentation, showing impressive results, *e.g.*, Mask2Former (Cheng et al., 2022), combining the strengths of CNN-based and transformer-based architectures. It employs a hybrid approach with a CNN backbone for feature extraction and a transformer decoder for capturing global context and generating high-resolution segmentation masks. Other transformer-based models, such as SETR (Zheng et al., 2021), TransUNet (Chen et al., 2024), and SegFormer (Xie et al., 2021), have also been proposed, leveraging the self-attention mechanism to capture long-range dependencies and global context effectively. These latter methods have demonstrated competitive or improved performance on various semantic segmentation benchmarks compared to traditional CNN-based models.

2.2. Satellite image time series (SITS)

Leveraging the temporal information with satellite and aerial imagery provides information on land dynamics and phenology. Researchers have used convolutional neural networks (CNNs) in temporal convolutions for land cover mapping (Lucas et al., 2021) and crop classification (Rußwurm and Körner, 2018). Attention-based methods have been used for encoding time series, which have proven to be well-suited for satellite imagery (Garnot and Landrieu, 2021; Sainte Fare Garnot et al., 2020; Rußwurm et al., 2023). More recently, transformer-based methods have proven their merit using satellite image time series (SITS) with self-supervised learning, exploiting unlabeled data to improve performance on downstream tasks (Cong et al., 2022; Tarasiou et al., 2023; Tseng et al., 2023; Reed et al., 2023).

A recent method has also proposed a new encoding scheme for SITS in order to fit popular pretrained backbones rather than creating task-specific architectures (Cai et al., 2023).

2.3. Forest monitoring

Deep learning methods have helped advance the field of vegetation monitoring using remote sensing, including both satellite and aerial imagery Kattenborn et al. (2021), enabling progress in forest monitoring for accurate and efficient analysis at scale (Bae et al., 2019; Reichstein et al., 2019; Beloiu et al., 2023; Nguyen et al., 2024). Such models have achieved state-of-the-art performance in classifying tree species from high-resolution remote sensing imagery (Fricker et al., 2019; Onishi and Ise, 2021).

Mapping deforestation at a large scale using satellite imagery has also been explored (Adarme et al., 2020; Maretto et al., 2021). Computer vision and remote sensing have also been leveraged in applications to plant phenology (Katal et al., 2022). Global vegetation phenology has been modeled with satellite imagery alongside meteorological variables as inputs of a 1D CNN (Zhou et al., 2021). Automated monitoring of forests has also been investigated to accurately identify key phenological events (Cao et al., 2021; Song et al., 2022; Wang et al., 2023).

Deep learning–based segmentation methods have been applied to automatically delineate individual tree crowns from high-resolution remote sensing imagery (Brandt et al., 2020; Schiefer et al., 2020; Weinstein et al., 2020; Li et al., 2023). In a similar vein, a U-Net architecture has been used for fine-grained segmentation of plant species using aerial imagery (Kattenborn et al., 2019). A foundation model trained on datasets from multiple sources is also able to perform decently on a variety of downstream tasks for forest monitoring, including classification, detection, and semantic segmentation (Bountos et al., 2025).

2.4. Hierarchical losses

Hierarchical loss functions have been extensively explored in various tasks to leverage the inherently hierarchical structure of object classes. By incorporating information from different levels of granularity, such loss functions aim to improve the ability of the model to make fine-grained distinctions and enhance

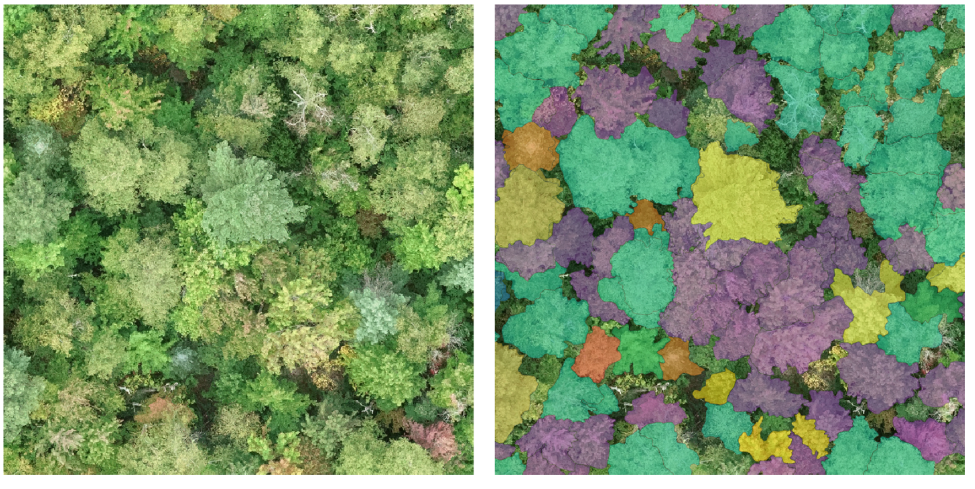
overall performance. For classification tasks, a curriculum-based hierarchical loss, gradually increasing the specificity of the target class, was explored by Goyal and Ghosh (2021). Similarly, a loss function evaluated at multiple operating points within the class hierarchy has helped to capture information at various levels of this hierarchy (Valmadre, 2022). In contrast, one may encourage the model to make better mistakes by assigning different weights to the misclassified samples based on their position in the hierarchy, promoting more semantically meaningful errors (Bertinetto et al., 2020).

Hierarchical loss functions have also been applied to object detection (Katole et al., 2015; Zwemer et al., 2022) and semantic segmentation (Sharma et al., 2015; Muller and Smith, 2020; Li et al., 2022) demonstrating the effectiveness of incorporating a more structured and informative signal during the learning process.

3. Dataset

The dataset used in our work (Cloutier et al., 2024) consists of high-resolution RGB imagery from unmanned aerial vehicles (UAVs) at seven different acquisition dates over a temperate-mixed forest in the Laurentides region of Québec, Canada, during the year 2021. The acquisitions were conducted monthly from May to August, with three additional acquisitions in September and October to capture color changes during autumn. The dataset contains a total of 23,000 individual tree crowns that were segmented and annotated, mostly at the species level, with 1,956 trees annotated only at the genus level due to the difficulty in accurately identifying species-level labels. This dataset offers a unique combination of time-series data and a large number of fine-grained tree species. This allows us to leverage the temporal information to investigate the impact of phenological changes on tree species identification. An example of this dataset is shown in Figure 1.

We perform three-fold cross-validation using spatially separated splits for training, validation, and test, while ensuring balanced distribution of tree species classes across splits. The spatial separation between splits, with a consistent test region across all folds, allows us to evaluate how well our models generalize to new geographic areas, a critical requirement for real-world applications. An example of one cross-validation fold is illustrated in Figure 2.



(a) Sample from 2nd September 2021. (b) Labels overlaying the sample image.

Figure 1. *Example of an annotated sample from the studied dataset. The image 1 shows a scene captured on September 2nd, while the image 1a overlays the tree species labels on the same scene. Each tree species is represented by a distinct color, as seen in Table 1.*

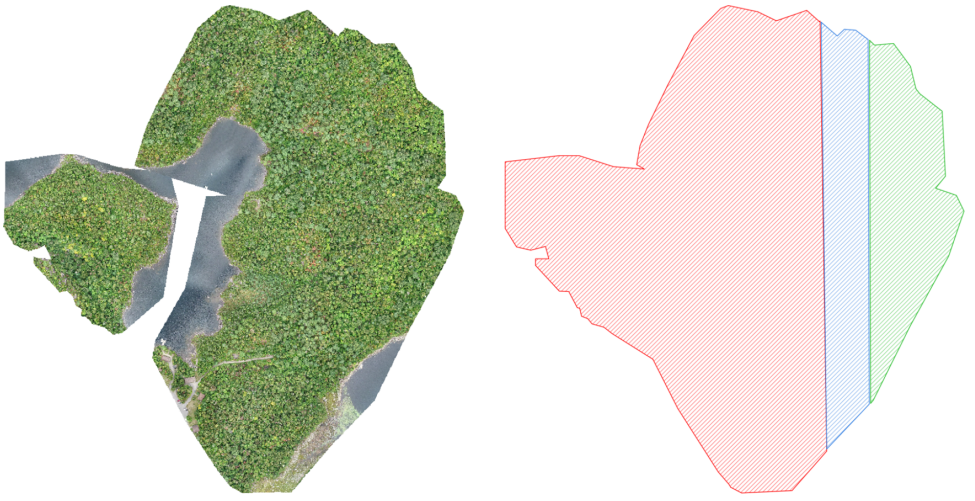


Figure 2. *Spatial splits of the dataset. The image on the left depicts the entire region where the aerial imagery was captured, while the image on the right shows the different subregions used to train, evaluate, and test models from one fold of cross-validation. The training region is represented by* ■*, the validation region by* ■*, and the test region by* ■*. To prevent data leakage between the subsets, a buffer tile is omitted between the adjacent regions. This spatial partitioning ensures that the model's performance is assessed on geographically distinct areas, simulating real-world scenarios where the model would be applied to unseen locations.*

For our three-fold cross-validation, we maintain a consistent test region across all folds to ensure reliable performance comparisons. In the remaining area, we create train and validation splits by systematically shifting their positions from left to right. Both training and validation regions maintain approximately equal sizes while their positions shift in each fold. Buffer tiles separate all regions (test, train, and validation) to prevent spatial autocorrelation, which is crucial for aerial imagery where neighboring pixels typically share similar characteristics. This method ensures no data leakage between splits while preserving the distribution of tree species across the heterogeneous forest ecosystem.

We ensure that each split maintains approximately the same proportion of tree species as the overall dataset, addressing potential sampling biases while preserving the natural spatial patterns of the forest.

For our experiments, we use an image size of $768 \times 768 \times 3$, providing sufficient spatial context to include multiple tree crowns and to learn relationships between different regions in the image. The labels are annotated using recordings from September 2 as reference (representing a date before most leaves change colour), which is also used as the input for our single-image models. For the models that take time series as input, we select one image from June, two from September, and one from October to reduce redundant information, as most phenological changes occur between September and October.

As a design choice, from the initial 28 classes, we merged those with less than 50 occurrences (mostly species with fewer than 10 samples) into the background class, leaving us with a total of 15 classes, excluding the background class. This ensures the selected classes have sufficient samples in each split in order to effectively train and evaluate each model. The tree species distribution is illustrated in Figure 3.

The dataset is split into train, validation, and test sets with 64%, 16%, and 20% of the samples, respectively. We opted for a larger test set (20%) compared to conventional splits to ensure robust evaluation across all tree species classes, particularly given the class imbalance in our dataset. This split ratio maintains adequate representation of less frequent species in the test set while preserving sufficient training data. The validation set (16%) remains large enough for effective model selection and hyper-parameter tuning. This is kept approximately consistent across all three folds of cross-validation. Given that this dataset has a mix of coarse (genus) and fine-grained (species) labels, we leverage this information

Figure 3. *Distribution of the selected classes in the dataset. We observe that there is a substantial difference in the frequency of occurrence of each tree species. The common and scientific names used for the abbreviations are detailed in Table 1.*

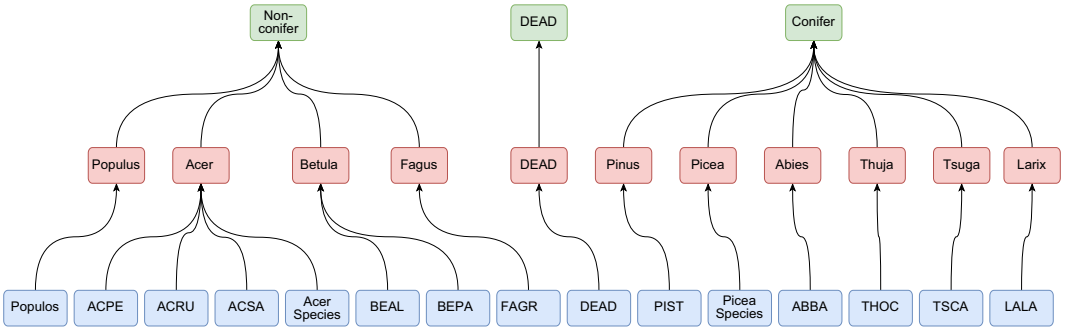


Figure 4. *Taxonomic hierarchy of tree species. The hierarchical structure is visually represented using a tree diagram. Blue nodes represent the species level, the most fine-grained classification in the hierarchy. Red nodes denote the genus level, which groups together closely related species. Finally, green nodes group the higher-level taxon, the broadest classification level, which encompasses multiple genera and families. This structure of labels allows the models to learn more comprehensive relationships between different tree species at multiple levels of granularity. The full names of each abbreviation are detailed in Table 1.*

to create a complete taxonomy of the classes used, as seen in Figure 4. This taxonomic hierarchy is incorporated in our proposed loss function as detailed in Section 4.3.

4. Methods

In this section, we provide more details on the methods used to perform semantic segmentation, either with single-image or time-series inputs. We will also describe the proposed hierarchical loss used to exploit the tree label taxonomy.

Table 1. *Tree species names and their abbreviations*

Common name (Scientific name)	Abbreviation
Balsam fir (Abies balsamea)	ABBA
Striped maple (Acer pensylvanicum)	ACPE
Red maple (Acer rubrum)	ACRU
Sugar maple (Acer saccharum)	ACSA
Maple (Acer sp.)	Acer
Swamp birch (Betula alleghaniensis)	BEAL
Paper birch (Betula papyrifera)	BEPA
American beech (Fagus grandifolia)	FAGR
Tamarack (Larix laricina)	LALA
Dead tree	DEAD
Spruce (Picea sp.)	Picea
Eastern white pine (Pinus strobus)	PIST
Aspen (Populus sp.)	Populus
Northern white-cedar (Thuja occidentalis)	THOC
Eastern hemlock (Tsuga canadensis)	TSCA

Note. The color we use to depict each species is highlighted in the second column and is consistent for all the plots and figures.

4.1. Single image semantic segmentation

The single-image semantic segmentation experiments are conducted with diverse methods detailed in the following sections.

4.1.1. U-Net

U-Net (Ronneberger et al., 2015) is a widely adopted convolutional neural network (CNN) architecture (Dong et al., 2017; Falk et al., 2018; Li et al., 2018) designed for efficient image segmentation tasks. The architecture consists of an encoder path and a decoder path, which together form a U-shaped structure. The encoder path follows the typical structure of a CNN, consisting of successive CNN layers, rectified linear units (ReLU), and max-pooling operations, which gradually reduce the spatial dimensions while increasing the number of feature maps. The decoder path utilizes transposed convolutions to upsample the feature channels, enabling the network to construct segmentation maps at the original input resolution. The U-Net architecture uses skip connections (He et al., 2016) to concatenate feature maps from the encoder path with the corresponding upsampled feature maps in the decoder path.

4.1.2. DeepLabv3+

The DeepLabv3+ architecture (Chen et al., 2018b) is an image segmentation method built upon the strengths of pyramid pooling with an encoder–decoder structure (Chen et al., 2018a). The encoder module of the DeepLabv3+ utilizes "dilated" (also named "atrous") convolutions to extract dense feature maps at multiple scales with larger receptive fields while keeping the computation costs lower. The encoder incorporates atrous spatial pyramid pooling (ASPP), which applies atrous convolutions with different dilation rates in parallel to further capture multi-scale context (Chen et al., 2018a).

The decoder module of the DeepLabv3+ combines the output of the encoder with low-level features from the encoder. This information is refined with 3×3 convolutions to produce the final output segmentation maps.

4.1.3. Mask2Former

The Mask2Former architecture (Cheng et al., 2022) is a versatile method that applies binary masks to focus attention only on the areas with foreground features. The architecture consists of three parts: a

backbone network, a pixel decoder, and a transformer decoder. Universal backbones (ResNet (He et al., 2016) or Swin Transformer (Liu et al., 2021)) are used to extract features from the input image. The low-resolution features are then used in a pixel decoder and upsampled to higher resolution. The masked attention is finally applied to the pixel embeddings in the transformer decoder.

To reduce the computational burden of using high-resolution masks, the transformer decoder processes the multi-scale features per resolution one at a time. The Mask2Former architecture performs well across a variety of tasks, like semantic, instance, and panoptic segmentation, which makes it a popular choice.

4.2. Time-series semantic segmentation

We compare various methods for semantic segmentation with time-series data, including 3D-UNet (Çiçek et al., 2016), specialized for medical images, and U-Net with temporal attention encoder (U-TAE) (Garnot and Landrieu, 2021) specialized for SITS. Additionally, we propose a simple, yet effective, module composed of 3D convolutional layers, referred to as "Processor", to preliminarily process the time series and use its representation as input for mainstream single-image segmentation methods.

4.2.1. 3D-UNet

The 3D-UNet method (Çiçek et al., 2016) is composed of successive 3D convolutions with a $3 \times 3 \times 3$ kernel, followed by batch normalization and a leaky ReLU activation. The 3D-UNet downsampling part is composed of five blocks, separated by spatial downsampling after the second and fourth blocks. The upsampling part consists of five blocks with transposed convolutions, while features from the downsampling part are concatenated similarly to U-Net (Ronneberger et al., 2015).

4.2.2. U-TAE

The U-TAE architecture (Garnot and Landrieu, 2021) has been introduced for panoptic segmentation of SITS. It consists of three main parts: a multi-scale spatial encoder, a temporal encoder, and a convolutional decoder that produces a single feature map with the same spatial resolution as the input. The sequence of images is processed in parallel by the spatial encoder, and the temporal attention encoder (TAE) is applied at the lowest resolution features to generate attention masks. These masks are interpolated and applied to each feature map, allowing the extraction of spatial and temporal information at multiple scales. The decoder uses a series of transposed convolutions, ReLU, and batch normalization layers to produce the final feature map.

4.2.3. Processor module

Our proposed Processor module is composed of 3D convolutions and is designed to extract spatio-temporal features from time-series data, enabling the use of pretrained models for semantic segmentation. The motivation behind the Processor architecture is to capture spatio-temporal patterns while maintaining the spatial resolution to fit established models pretrained on single-image datasets. This approach differs from task-specific models relying on specialized architectures for processing time-series data in particular contexts, such as land use and land cover mapping (Garnot and Landrieu, 2021; Tarasiou et al., 2023).

The module is composed of two 3D convolutional layers. The first layer has a kernel size of $3 \times 3 \times 3$, followed by a second layer with a kernel size of $2 \times 3 \times 3$. The padding in these layers is set to $(0,1,1)$, and the number of output channels is set to 32 and 64, respectively. This configuration will collapse the temporal dimension of the input while simultaneously increasing the number of channels.

Since the kernel sizes are designed for a specific time-series length, they must be adjusted for a different application, yet our lightweight module is easily trainable from scratch.

Formally, $\mathbf{x} \in \mathbb{R}^{T \times C \times H \times W}$ let be an input time series, where T is the length of the time series, C the number of channels of each image, H and W their respective height and width dimensions. Our Processor module $p_\Theta(.)$, parameterized by Θ, can be used prior to any semantic segmentation model f_θ parameterized by θ, via $f_\theta(p_\Theta(\mathbf{x}))$. To evaluate the effectiveness of our approach, we used the Processor alongside U-Net and DeepLabv3+. The results of our experiments are detailed in Section 6.

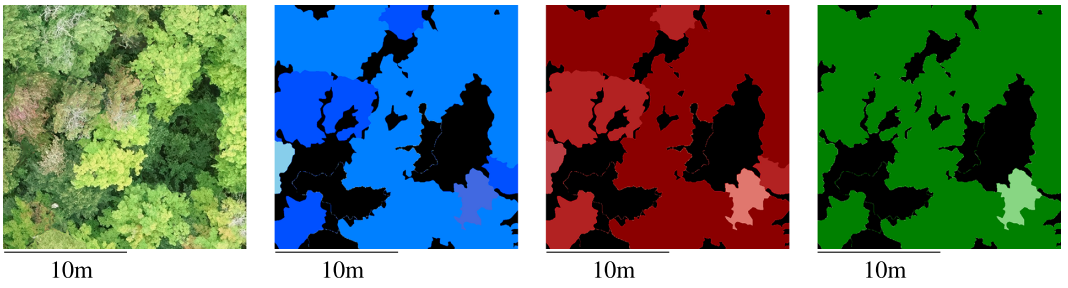


Figure 5. *Example of the proposed three-level hierarchical label structure. The labels are concatenated to form semantic segmentation masks where each channel corresponds to a specific taxonomic level: species 5b, genus 5c, and higher-level taxon 5d. Each image represents an area of approximately 20.1m × 20.1m. In this example, there are three classes at the species and genus levels. However, the higher-level taxon only has two classes due to the aggregation of different trees under one class. Note that the colors used in this image do not conform to the color code shown in Table 1.*

4.3. Hierarchical loss

This section details the proposed hierarchical loss that leverages information about taxonomic hierarchies of tree species, genus, and families.

The dataset detailed in Section 3 groups a mix of finer (species) and coarser (genus) level labels.

The taxonomic structure of these labels offers an opportunity to train a model while benefiting from such a hierarchical structure.

To exploit this hierarchy, we extend each label to multiple levels: species, genus, and higher-level taxon. The taxonomic hierarchy is illustrated in Figure 4, and a visual example of these labels is illustrated in Figure 5.

During training, the model predicts only the species-level labels for each pixel. These softmax probabilities at the species level are then aggregated according to our knowledge of the label taxonomies (see Figure 4) to generate first the genus level predictions (see Equation 4.3) and second the higher-level predictions (see Equation 4.5).

Note that our implementation of the hierarchical loss differs from certain related work presented in Section 2, where classes at all levels are predicted separately to compute the loss (Turkoglu et al., 2021).

Formally, let $\mathbf{x} \in \mathbb{R}^{C \times H \times W}$ be a training example, $\mathbf{y}_S \in \{0,1\}^{S \times H \times W}$ its one-hot ground truth where S is the number of classes at the species level, and $f_\theta(\mathbf{x}) = \mathbf{p}_S$ the associated predictions. The cross-entropy loss function at the species level is defined as normal via

$$\boldsymbol{L}_{\text{species}} := -\frac{1}{S}\sum_{s=1}^{S}\sum_{(h,w)\in\Omega}\mathbf{y}_S[h,w,s]\log\mathbf{p}_S[h,w,s], \tag{4.1}$$

where $\Omega = [\![1,H]\!] \times [\![1,W]\!]$. The cross-entropy loss function at the genus level is then computed using the ground truth and predictions at the species level, as

$$\boldsymbol{L}_{\text{genus}} := -\frac{1}{G}\sum_{g=1}^{G}\sum_{(h,w)\in\Omega}\mathbf{y}_G[h,w,g]\log\mathbf{p}_G[h,w,g] \tag{4.2}$$

$$= -\frac{1}{G}\sum_{g=1}^{G}\sum_{(h,w)\in\Omega}\left[\sum_{s=1}^{S_g}\mathbf{y}_S[h,w,s]\right]\log\left[\sum_{s=1}^{S_g}\mathbf{p}_S[h,w,s]\right], \tag{4.3}$$

where G is the number of classes at the genus level and S_g is the number of classes at the species level corresponding to a given genus class g. In the same vein, the cross-entropy loss function at the higher-level taxon is also obtained via the ground truth and predictions at the species level, as:

$$\boldsymbol{L}_{\text{taxon}} := -\frac{1}{T}\sum_{t=1}^{T}\sum_{(h,w)\in\Omega}\mathbf{y}_T[h,w,t]\log\mathbf{p}_T[h,w,t] \tag{4.4}$$

$$= -\frac{1}{T}\sum_{t=1}^{T}\sum_{(h,w)\in\Omega}\left[\sum_{g=1}^{G_t}\sum_{s=1}^{S_g}\mathbf{y}_S[h,w,s]\right]\log\left[\sum_{g=1}^{G_t}\sum_{s=1}^{S_g}\mathbf{p}_S[h,w,s]\right], \tag{4.5}$$

where T is the number of classes at the higher-level taxon and G_t the number of classes at the genus level corresponding to a given higher-level class t.

The hierarchical loss function is formulated as

$$\boldsymbol{L}_{\text{HLoss}} = \lambda_1 \cdot \boldsymbol{L}_{\text{species}} + \lambda_2 \cdot \boldsymbol{L}_{\text{genus}} + \lambda_3 \cdot \boldsymbol{L}_{\text{taxon}}, \tag{4.6}$$

where λ_1, λ_2, and λ_3 are the weights for the species, genus, and higher-level taxon losses, respectively, and $\boldsymbol{L}_{\text{species}}$, $\boldsymbol{L}_{\text{genus}}$, and $\boldsymbol{L}_{\text{taxon}}$ are the corresponding cross-entropy losses.

We set empirically $\lambda_1 = 1$, $\lambda_2 = 0.3$, and $\lambda_3 = 0.1$ since we observed that giving more weight to the species-level loss helps the model to prioritize the fine-grained predictions while still benefiting from the hierarchical information. However, we have not attempted to fully optimize these values.

5. Experiments

5.1. Experimental setup

All methods detailed in Section 4 have been trained with normalized input data, either with the means and standard deviations of our dataset to train models from scratch, or with statistics of the datasets used for pretraining for models based on MS-COCO and ImageNet weights. All these experiments are performed on three-fold cross-validation sets to get a better understanding of model performance.

We employ the Adam optimizer (Kingma and Ba, 2015) for all models except Mask2Former, which is trained with the AdamW optimizer (Loshchilov and Hutter, 2019) to maintain consistency with the original training methodology. We trained all models with a learning rate of $1e-4$ with exponential learning rate decay for 300 epochs.

We included rotation (in multiples of 90°) with horizontal flips as data augmentation to enhance the diversity of the training data. The batch sizes used for each model are detailed in Table A2. These were set to the largest size that could fit within an NVIDIA RTX 8000 GPU. We train our models either using our proposed hierarchical loss, noted HLoss, and described in Section 4.3, or using a combination of dice and cross-entropy losses, noted Dice + CE (Figure 6).

The latter is a popular choice for segmentation tasks since the dice loss measures the overlap between the predicted and ground truth masks, while the cross-entropy loss quantifies the dissimilarity between the predicted and true class probabilities. We trained the Mask2Former model with the loss function proposed by its authors (Cheng et al., 2022). This loss function improves the training efficiency by randomly sampling a fixed number of points in the labels and predictions.

Model	Batch Size
U-TAE	4
Unet-3D	6
Processor+U-Net	16
Processor+DeepLabv3+	16
U-Net	16
DeepLabv3+	16
Mask2former	16

Figure 6. *Batch sizes used for training.*

The loss weighing scheme and other implementation details are kept consistent with the original implementation to ensure a fair comparison. Note that we did not run Mask2Former with HLoss and Dice + CE loss as the training would be much more computationally expensive, resulting in a smaller batch size.

The performance of our models are evaluated with the Intersection over Union (IoU) metric, also known as the Jaccard index, which measures the overlap between the predicted and ground truth masks. Letting A and B be two sets, the IoU score is defined as

$$\mathrm{IoU}(A,B) := \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}. \tag{5.1}$$

The mean IoU (mIoU) is computed by averaging the IoU scores across all classes. This metric provides a comprehensive assessment of the segmentation performance of a model, taking into account both the precision and recall.

5.2. Experiment configuration

We conduct a comprehensive set of experiments to thoroughly evaluate the performance of the considered methods:

- We compared models using either single-image or time-series inputs to evaluate the contribution of the phenological information on the tree species segmentation task. The time series is composed of images at four different periods of the year (see Section 3). Note that both methods predict segmentation masks corresponding to a single image.
- We compared models with two different loss functions to demonstrate the value of leveraging taxonomic information through the HLoss against a standard combination of loss functions (Dice + CE).
- We conduct ablation studies to investigate the impact of different pretrained backbones on the segmentation performance. For the CNN-based models, we experiment with ResNet-34, ResNet-50, and ResNet-101 backbones, whereas for the Mask2Former model, we use Swin-T and Swin-S backbones (Liu et al., 2021).

The results of these experiments are discussed in Section 6 where we compare results both quantitatively and qualitatively.

6. Results

6.1. Single-image input for semantic segmentation

For the single-image segmentation model, we compare the performance of DeepLabv3+ and U-Net architectures with ResNet backbones of varying depths (ResNet-34, ResNet-50, ResNet-101) and the Mask2Former architecture with the Swin-T and Swin-S backbones.

As seen in Table 2, both DeepLabv3+ and U-Net architectures show a consistent increase in performance with increasing backbone size, where the ResNet-101 model achieves the highest mIoU score. Comparing the loss functions, the proposed HLoss generally leads to higher mean IoU scores than the Dice+CE loss across most backbones and architectures. For instance, with the U-Net ResNet101 backbone, HLoss achieves a statistically significant improvement over Dice+CE (55.15 ± 0.29 vs 54.6 ± 0.21). However, for some configurations, such as DeepLabv3+ with ResNet101, the performance difference between HLoss and Dice+CE is smaller and not statistically significant, given the overlapping error margins. This suggests that while leveraging taxonomic information via HLoss is often beneficial, its impact can vary.

We also observe in Table 2 that training models from scratch results in significantly lower mIoU scores compared to using pretrained ImageNet weights, highlighting the importance of transfer learning. When trained from scratch, both U-Net and DeepLabv3+ with ResNet50 backbones achieved comparable results using either HLoss or Dice+CE, with all differences falling within the error margins.

Table 2. *Comparison of single image methods with different losses and backbones*

Model	Backbone	Dice+CE	HLoss	Mask2Former Loss
DeepLabv3+	ResNet34	52.30 ± 0.30	**52.62 ± 0.38**	–
	ResNet50	52.46 ± 0.36	**53.08 ± 0.32**	–
	ResNet101	54.14 ± 0.51	**54.81 ± 0.44**	–
	ResNet50[a]	42.87 ± 0.83	**42.91 ± 0.90**	–
U-Net	ResNet34	52.19 ± 0.65	**52.55 ± 0.64**	–
	ResNet50	52.79 ± 0.58	**53.42 ± 0.44**	–
	ResNet101	54.61 ± 0.21	**55.15 ± 0.29**	–
	ResNet50[a]	42.12 ± 0.75	**42.17 ± 0.69**	–
Mask2Former	Swin-t[b]	–	–	45.78 ± 1.57
	Swin-s[b]	–	–	45.74 ± 1.58

Note. Performances are compared with IoU averaged over all the classes of the dataset (mIoU) for single-image models.
[a]Indicates models trained from scratch without using ImageNet weights (Deng et al., 2009).
[b]Indicates Swin-based models using weights from the MS-COCO dataset (Lin et al., 2014). All results are averaged across three-fold cross-validation, and the best result for each backbone is shown in bold text. The best model overall is highlighted in red.

The Mask2Former models, trained with the loss of the original implementation and with pretrained weights from the MS-COCO dataset, perform better than the models trained from scratch; however, their performance is not comparable to the CNN-based architectures.

While mIoU provides insights into spatial segmentation accuracy, we also evaluated the models using classification metrics (F1-score, precision, and recall). These results follow similar trends to the mIoU scores and are detailed in Appendix A.2.

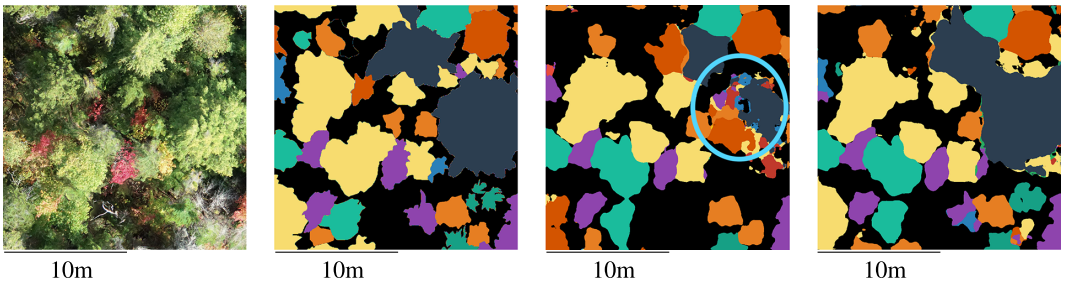
6.2. Time-series input for semantic segmentation

For time-series inputs, we make use of the Processor module, detailed in Section 4.3, to extract spatio-temporal features and evaluate its performance with DeepLabv3+ and U-Net architectures. Among the time-series models incorporating the Processor module, the use of HLoss often results in mean IoU scores similar to those from Dice + CE loss (Table 3). For instance, with the U-Net + Processor architecture

Table 3. *Comparison of time-series methods with different losses and backbones*

Model	Backbone	Dice+CE	HLoss
DeepLabv3+ + Processor	ResNet34	**52.71 ± 0.69**	52.69 ± 0.56
	ResNet50	52.54 ± 0.64	**53.41 ± 0.75**
	ResNet101	54.66 ± 0.43	**54.93 ± 0.31**
	ResNet50[a]	48.34 ± 0.76	**49.08 ± 0.88**
U-Net + Processor	ResNet34	52.89 ± 0.58	**52.91 ± 0.58**
	ResNet50	53.36 ± 0.83	**53.85 ± 0.85**
	ResNet101	55.04 ± 0.47	**55.97 ± 0.48**
	ResNet50[a]	48.96 ± 0.37	**49.07 ± 0.46**
UNet 3D[a]	—	37.94 ± 0.58	**38.46 ± 0.26**
U-TAE[a]	—	36.84 ± 0.90	**38.28 ± 0.33**

Note. Performances are compared with IoU averaged over all the classes of the dataset (mIoU) for single-image models.
[a]Indicates models trained from scratch. All results are averaged across three-fold cross-validation, and the best result for each backbone is shown in bold text. The best model overall is highlighted in red.



(a) A sample image. (b) Annotation. (c) Results w/ Dice+CE. (d) Results w/ HLoss.

Figure 7. *Qualitative results of the Dice + CE loss versus HLoss. This example compares the best-performing Processor + UNet (ResNet101) models trained with the Dice + CE loss and the proposed hierarchical loss (HLoss). Each image represents an area of approximately 20.1m × 20.1m. First, 7a shows a sample image from the sequence, while 7b displays the corresponding ground truth annotation. Then, 7c depicts the segmentation output obtained by the model trained with the Dice + CE loss, and finally, 7d illustrates the output from the model trained with HLoss. The colors of the labels and predicted segments correspond to specific tree species, as indicated by the legend in Table 1. Upon closer inspection of the regions highlighted by the cyan circle (◯), the model trained with the Dice + CE loss exhibits some confusion among classes, whereas the model trained with HLoss demonstrates improved discrimination between classes.*

(ResNet101 backbone), HLoss (55.97 ± 0.48) provides only a marginal and likely nonsignificant improvement compared to Dice+CE (55.04 ± 0.47). Similarly, for the DeepLabv3+ + Processor architecture with the identical backbone, the mean IoU achieved with HLoss is not statistically distinguishable from that of Dice+CE when accounting for their respective standard deviations.

Qualitative results comparing HLoss with Dice + CE loss are illustrated in Figure 7, where HLoss demonstrates the ability to better discriminate between classes. Models trained using the Dice + CE loss exhibit some confusion among classes. Using HLoss would reduce confusion among classes that do not belong in the same genus or higher-level taxon as the model is penalized for incorrect predictions at all levels. The U-Net + Processor with ResNet-101 backbone trained with HLoss achieves the best mIoU score among all models. Furthermore, the time-series models slightly outperform their single-image counterparts, indicating the importance of leveraging phenological patterns by incorporating temporal information for tree species segmentation. The classification metrics further support the advantages of temporal information, with the Processor+U-Net models showing balanced performance across F1-score, precision, and recall metrics. A detailed analysis of these classification metrics is provided in Appendix A.2.

To gain a deeper understanding of how leveraging time-series data affects the performance of our models for individual species, we conduct a detailed analysis of the class-wise results for our best-performing single-image and time-series models. For the single-image model, we select the U-Net architecture with a ResNet-101 backbone, while for the time-series model, we choose the Processor +U-Net architecture, also with a ResNet-101 backbone. This allows for a fair comparison between the two approaches, as the main difference lies in the incorporation of temporal information through the Processor module. Table 4 presents the class-wise Intersection over Union (IoU) scores for both models, with the classes grouped into non-coniferous and coniferous categories. Note that we omit a class from this analysis: "Acer sp.," a class composed of trees belonging to Striped Maple (ACPE), Red Maple (ACRU), or Sugar Maple (ACSA) that have not been assigned a fine-grained label by the annotators due to low confidence.

While the overall mIoU shows a statistically significant advantage for the time-series approach, the class-wise results reveal a more complex picture with considerable variability. This class-level analysis reveals where the overall statistically significant mIoU improvement for the time-series model originates.

Table 4. *The table shows the IoU for individual classes for our best-performing Processor + U-Net and U-Net models, both with ResNet-101 as backbone*

Class	Processor + U-Net (R101)	U-Net (R-101)
Non-coniferous trees		
Populus	**76.09 ± 1.02**	74.01 ± 2.26
ACPE	28.34 ± 0.96	**28.38 ±0.98**
ACRU	**56.82 ± 0.60**	55.17 ± 0.49
ACSA	**45.87 ±1.22**	43.03 ± 2.47
BEAL	**59.95 ± 2.52**	57.63 ± 2.19
BEPA	**71.33 ± 0.87**	69.80 ± 0.98
FAGR	53.30 ± 2.09	**54.19 ± 0.44**
Coniferous trees		
PIST	74.01 ± 1.20	**74.16 ± 3.39**
Picea	**76.09 ± 1.02**	74.01 ± 2.26
ABBA	**62.06** ± 0.60	**63.82 ± 0.24**
THOC	**58.87 ± 1.09**	57.82 ± 1.85
TSCA	**61.84 ± 0.96**	59.29 ± 1.28
LALA	**74.53 ± 1.25**	73.25 ± 1.71
Others		
DEAD	40.40 ± 0.89	**42.77 ± 0.62**
Overall results	**55.97 ± 0.48**	55.15 ± 0.29

Note. All results are averaged across three-fold cross-validation. The classes are grouped into non-coniferous and coniferous categories, with the color shown for each class corresponding to the color code in Table 1. The last row presents the metrics from Table 2 and Table 3 as a reference. These metrics represent the average performance across all classes over three seeds, not the average of the values shown in this table. We observe that incorporating time-series data improves the segmentation performance for most of the individual tree species. This performance gain is more pronounced for non-coniferous trees.

While the performance advantage was statistically significant for specific classes like Red Maple (ACRU) and Eastern Hemlock (TSCA), the time-series model achieved comparable performance to the single-image model for the majority of other species (e.g., Populus, ACPE (Striped Maple), ACSA (Sugar Maple), BEAL (Yellow Birch), BEPA (Paper Birch), FAGR (American Beech), PIST (Eastern White Pine), Picea, THOC (Eastern White Cedar), LALA (Tamarack)), with differences not being statistically significant based on our analysis.

This indicates that for many classes, the temporal information allowed the model to maintain a high level of accuracy similar to the strong single-image baseline. Although the single-image model did perform better for Balsam Fir (ABBA) and the DEAD tree class, the overall significant mIoU improvement for the time-series approach stems from the combination of specific significant gains and comparable performance across most other classes, which supports the value of incorporating temporal data for this task.

An example of the results comparing single-image and time-series models is illustrated in Figure 8, where using temporal information helps the model differentiate between tree species that undergo senescence at slightly different times. Red maple trees are among the earliest trees to show color changes in the fall, and the single-image model misclassifies a Swamp Birch as a Red Maple. This misclassification can be attributed to the lack of temporal context, which is necessary to understand the correlation between tree species and the timing of their senescence.

We also test the generalization capability of our best-performing time-series model on a dataset from a different region of Quebec, which has similar ecological characteristics as the training area. An in-depth explanation has been provided in Appendix A, and the results can be seen in Figure A1.

(a) A sample image. (b) Annotation. (c) Results with SI (d) Results with TS

Figure 8. *Qualitative results of the single-image versus time-series inputs. This example compares the best-performing models with single-image (SI) and time-series (TS) inputs for tree species segmentation. Each image represents an area of approximately 20.1m × 20.1m. First, 8a shows a sample image from the sequence, while 8b displays the corresponding ground truth annotation. Then 8c depicts the segmentation output obtained by the single-image model, and finally 8d illustrates the output from the time-series model. The colors of the labels and predicted segments correspond to specific tree species, as indicated by the legend in Table 1. Upon comparing the results, we observe that here the time-series model outperforms the single-image model in correctly predicting the classes. In the instance highlighted by the cyan circle (◯), the time-series model accurately identifies the Swamp Birch, while the single-image model misclassifies it as Red Maple.*

7. Discussion

This work advances the field of forest monitoring through several contributions that build upon and extend previous research in tree semantic segmentation using a time series of images. The slight yet statistically significant gain in performance observed for U-Net models with HLoss when comparing our best time-series model to its single-image counterpart validates the importance of incorporating phenological information. (Zhou et al., 2021; Wang et al., 2023). Previous studies have achieved success in tree species classification using single high-resolution images (Fricker et al., 2019; Zhang et al., 2020). Our results, however, demonstrate that incorporating temporal data can improve discrimination, with this enhancement being particularly significant for specific species such as Red Maple (ACRU) and Eastern Hemlock (TSCA). The ability to capture distinct seasonal changes makes this approach especially valuable for deciduous trees like Red Maple.

The lightweight Processor module offers a practical solution to a key challenge in remote sensing: the need for specialized architectures that can handle temporal data while leveraging pretrained models (Cao et al., 2021; Kattenborn et al., 2021). The significant performance gap between models trained from scratch versus those using pretrained weights reinforces the value of transfer learning in forest monitoring applications (Bountos et al., 2025).

The effectiveness of our hierarchical loss function, which often led to performance gains compared to a standard Dice+CE loss, builds upon the previous work in hierarchical classification (Bertinetto et al., 2020; Muller and Smith, 2020; Valmadre, 2022), addressing the specific challenges in forest monitoring where species-level identification may not always be possible or necessary. This observed tendency for improvement suggests that the HLoss approach could be particularly valuable for large-scale forest monitoring applications.

Our work could enable more accurate forest inventories and better monitoring of species distribution changes in response to climate change. However, some manual intervention is still required, particularly in the initial data acquisition phase, where high-quality aerial imagery must be collected at specific temporal intervals to capture phenological changes. The collection of ground truth data for model training also remains a labor-intensive process, requiring expert knowledge for accurate annotation for species identification.

While our results demonstrate promising capabilities for automated forest monitoring, several practical challenges remain. Crown delineation accuracy can vary significantly with canopy density and image quality (Weinstein et al., 2020). We used a uniform learning rate across architectures for experimental consistency; however, an exhaustive ablation study exploring architecture-specific learning rates could potentially yield improved results, particularly for transformer-based models. The computational costs of such a study, coupled with the need to establish reliable standard deviations, led us to adopt our current approach, though we acknowledge this may result in conservative performance estimates. The Processor module's fixed time-step requirement, while effective for our dataset, may limit applicability to regions with different temporal sampling frequencies or irregular acquisition patterns (Rußwurm et al., 2023). Future work could explore integrating attention mechanisms to better handle longer time series (Garnot and Landrieu, 2021, Sainte Fare Garnot et al., 2020), extending the hierarchical loss approach to incorporate additional ecological relationships beyond taxonomic structure, and developing more flexible temporal processing architectures (Cai et al., 2023; Tarasiou et al., 2023). Such improvements would enhance the model's ability to handle variable-length time series and irregular sampling patterns, making it more adaptable to different forest monitoring scenarios and geographic regions.

Despite these limitations, the Processor module offers a simple yet effective approach to leveraging temporal information in tree species segmentation. Moreover, the compact design of the Processor module allows for efficient computation and reduces the overall complexity of the model, making it suitable for resource-constrained scenarios.

8. Conclusion

In this work, we developed a comprehensive approach for tree species segmentation using aerial image time series, demonstrating the advantages of incorporating temporal information and taxonomic knowledge. By combining a lightweight temporal processing module with a hierarchical loss, our approach often improved species discrimination, achieving statistically significant gains in key comparisons while maintaining the benefits of existing pretrained models. The framework's ability to effectively leverage phenological changes and taxonomic relationships provides a robust foundation for large-scale forest monitoring applications.

Climate change affects different tree species in varying ways, from altered phenological patterns to shifts in habitat suitability, making it essential to track changes at both species and broader taxonomic levels to understand ecosystem-wide responses. Our framework's ability to work with both species-level and higher taxonomic classifications enables monitoring at multiple scales, supporting both detailed species-specific studies and broader assessments of forest composition change. The proposed methods have significant implications for forest monitoring and biodiversity conservation, enabling accurate mapping of tree species composition, crucial for understanding forest ecosystems, monitoring changes over time, and informing conservation strategies. Future research could explore the incorporation of additional data modalities, addressing the limitations of the Processor module mentioned in Section 7, and the extension of the methods to other applications in forest ecology and management.

This work opens new possibilities for integrating remote sensing with ecological research. The combination of temporal analysis and hierarchical classification could serve as a foundation for studying species distribution shifts, phenological changes, and ecosystem responses to environmental stressors. These capabilities will be essential for developing evidence-based conservation strategies and understanding the ongoing impacts of climate change on forest ecosystems.

Open peer review. To view the open peer review materials for this article, please visit http://doi.org/10.1017/eds.2025.10013.

Acknowledgments. The authors are grateful for support from M. Cloutier for her guidance in understanding the intricacies of the dataset. In addition, we acknowledge material support from the Mila Quebec AI Institute and from NVIDIA Corporation in the form of computational resources.

Author contribution. Conceptualization: V.R.; A.O.; D.R. Methodology: V.R.; A.O.; D.R. Data curation: V.R. Data visualization: V.R. Writing original draft: V.R. Writing—Review and Editing: V.R.; A.O.; D.R. All authors approved the final submitted draft.

Competing interests. The authors declare none.

Data availability statement. The dataset used in our work is published by the original authors here: https://doi.org/10.5281/zenodo.8148479. Our code can be found at https://github.com/RolnickLab/Forest-Monitoring.

Ethics statement. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

Funding statement. This work was funded through the IVADO program on "AI, Biodiversity and Climate Change" and the Canada CIFAR AI Chairs program.

References

Adarme MO, **Feitosa RQ**, **Happ PN**, **De Almeida CA and Gomes AR** (2020) Evaluation of deep learning techniques for deforestation detection in the brazilian amazon and cerrado biomes from remote sensing imagery. *Remote Sensing 12*(6), 910.

Allen CD, **Macalady AK**, **Chenchouni H**, **Bachelet D**, **McDowell N**, **Vennetier M**, **Kitzberger T**, **Rigling A**, **Breshears DD**, **Hogg EH(Ted)**, **Gonzalez P**, **Fensham R**, **Zhang Z**, **Castro J**, **Demidova N**, **Lim J-H**, **Allard G**, **Running SW**, **Semerci A and Cobb N** (2010) A global overview of drought and heat-induced tree mortality reveals emerging climate change risks for forests. *Forest Ecology and Management 259*(4), 660–684.

Anderegg WRL, **Kane JM and Leander DLA** (2012) Consequences of widespread tree mortality triggered by drought and temperature stress. *Nature Climate Change 3*(1), 30–36.

Babst F, **Bouriaud O**, **Poulter B**, **Trouet V**, **Girardin MP and Frank DC** (2019) Twentieth century redistribution in climatic drivers of global tree growth. *Science Advances 5*(1), eaat4313.

Bae S, **Levick SR**, **Heidrich L**, **Magdon P**, **Leutner BF**, **Wöllauer S**, **Serebryanyk A**, **Nauss T**, **Krzystek P**, **Gossner MM**, **Schall P**, **Heibl C**, **Bässler C**, **Doerfler I**, **Schulze E-D**, **Krah F-S**, **Culmsee H**, **Jung K**, **Heurich M**, **Fischer M**, **Seibold S**, **Thorn S**, **Gerlach T**, **Hothorn T**, **Weisser WW and Müller J** (2019) Radar vision in the mapping of forest biodiversity from space. *Nature Communications 10*(1), 4757.

Beloiu M, **Heinzmann L**, **Rehush N**, **Gessler A and Griess VC** (2023) Individual tree-crown detection and species identification in heterogeneous forests using aerial rgb imagery and deep learning. *Remote Sensing 15*(5), 1463. https://doi.org/10.3390/rs15051463.

Bertinetto L, **Mueller R**, **Tertikas K**, **Samangooei S and Lord NA** (2020) Making better mistakes: Leveraging class hierarchies with deep networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June.

Bonan GB (2008) Forests and climate change: Forcings, feedbacks, and the climate benefits of forests. *Science 320*(5882), 1444–1449.

Bountos NI, **Ouaknine A**, **Papoutsis I and Rolnick D** (2025) Fomo-bench: A multi-modal, multi-scale and multi-task forest monitoring benchmark for remote sensing foundation models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. *39*. Association for the Advancement of Artificial Intelligence (AAAI), pp. 27858–27868. https://doi.org/10.1609/aaai.v39i27.35002.

Brandt M, **Tucker CJ**, **Kariryaa A**, **Rasmussen K**, **Abel C**, **Small J**, **Chave J**, **Rasmussen LV**, **Hiernaux P**, **Diouf AA**, **Kergoat L**, **Mertz O**, **Igel C**, **Gieseke F**, **Schöning J**, **Li S**, **Melocik K**, **Meyer J**, **Sinno S**, **Romero E**, **Glennie E and Montagu A** (2020) Morgane Dendoncker, and Rasmus Fensholt. An unexpectedly large count of trees in the west african sahara and Sahel. *Nature 587*(7832), 78–82.

Cai X, **Bi Y**, **Nicholl P and Sterritt R** (2023) Revisiting the encoding of satellite image time series. In *34th British Machine Vision Conference 2023*, *{BMVC} 2023*. BMVA. Available at https://papers.bmvc2023.org/0402.pdf.

Cao M, **Sun Y**, **Jiang X**, **Li Z and Xin Q** (2021) Identifying leaf phenology of deciduous broadleaf forests from phenocam images using a convolutional neural network regression method. *Remote Sensing 13*(12), 2331.

Chen J, **Mei J**, **Li X**, **Lu Y**, **Yu Q**, **Wei Q**, **Luo X**, **Xie Y**, **Adeli E**, **Wang Y**, **Lungren MP**, **Zhang S**, **Xing L**, **Lu L**, **Yuille A and Zhou Y** (2024) TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis 97*, 103280. https://doi.org/10.1016/j.media.2024.103280.

Chen LC, **Zhu Y**, **Papandreou G**, **Schroff F and Adam H** (2018a) Encoder-decoder with atrous separable convolution for semantic image segmentation. In Ferrari V, Hebert M, Sminchisescu C and Weiss Y (eds.), *Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science*, Vol. *11211*. Cham: Springer, pp. 801–818. https://doi.org/10.1007/978-3-030-01234-2_49.

Chen L-C, **Zhu Y**, **Papandreou G**, **Schroff F and Adam H** (2018b) Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Computer Vision – ECCV 2018*. Springer International Publishing, pp. 833–851. https://doi.org/10.1007/978-3-030-01234-2_49.

Cheng B, **Misra I**, **Schwing AG**, **Kirillov A and Girdhar R** (2022) Masked-attention mask transformer for universal image segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 1280–1289.

Çiçek Ö, **Abdulkadir A**, **Lienkamp SS**, **Brox T and Ronneberger O** (2016) 3D U-net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 424–432.

Cloutier M, **Germain M and 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. https://doi.org/10.1016/j.rse.2024.114283.

Cong Y, **Khanna S**, **Meng C**, **Liu P**, **Rozi E**, **He Y**, **Burke M**, **Lobell DB and Ermon S** (2022) SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. In Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K and Oh A (eds), *Advances in Neural Information Processing Systems. Curran Associates, Inc.*, pp. 197–211. Available at https://proceedings.neurips.cc/paper_files/paper/2022/file/01c561df365429f33fcd7a7faa44c985-Paper-Conference.pdf.

Deng J, **Dong W**, **Socher R**, **Li L-J**, **Li K and Fei-Fei L** (2009) Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 248–255.

Dong H, **Yang G**, **Liu F**, **Mo Y and Guo Y** (2017) *Automatic Brain Tumor Detection and Segmentation Using U-Net Based Fully Convolutional Networks*. Springer International Publishing, pp. 506–517

Falk T, **Mai D**, **Bensch R**, **Çiçek Ö**, **Abdulkadir A**, **Marrakchi Y**, **Böhm A**, **Deubner J**, **Jäckel Z**, **Seiwald K**, **Dovzhenko A**, **Tietz O**, **Bosco CD**, **Walsh S**, **Saltukoglu D**, **Tay TL**, **Prinz M**, **Palme K**, **Simons M**, **Diester I**, **Brox T and Ronneberger O** (2018) U-net: Deep learning for cell counting, detection, and morphometry. *Nature Methods 16*(1), 67–70.

Fricker GA, **Ventura JD**, **Wolf JA**, **North MP**, **Davis FW and Franklin J** (2019) A convolutional neural network classifier identifies tree species in mixed-conifer forest from hyperspectral imagery. *Remote Sensing 11*(19), 2326.

Fu J, **Liu J**, **Tian H**, **Li Y**, **Bao Y**, **Fang Z and Lu H** (2019) Dual attention network for scene segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 3141–3149.

Garnot VSF and Landrieu L (2021) Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4872–4881.

Goyal P and Ghosh S (2021) Hierarchical class-based curriculum loss. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, pp. 2448–2454. https://doi.org/10.24963/ijcai.2021/337.

Hamdi ZM, **Brandmeier M and Straub C** (2019) Forest damage assessment using deep learning on high resolution remote sensing data. *Remote Sensing 11*(17), 1976.

Hamedianfar A, **Mohamedou C**, **Kangas A and Vauhkonen J** (2022) Deep learning for forest inventory and planning: A critical review on the remote sensing approaches so far and prospects for further applications. *Forestry: An International Journal of Forest Research 95*(4), 451–465.

He K, **Zhang X**, **Ren S and Sun J** (2016) Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 770–778.

Helber P, **Bischke B**, **Dengel A and Borth D** (2019) EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 12*(7), 2217–2226.

Katal N, **Rzanny M**, **Mäder P and Wäldchen J** (2022) Deep learning in plant phenological research: A systematic literature review. *Frontiers in Plant Science 13*. https://doi.org/10.3389/fpls.2022.805738.

Katole AL, **Yellapragada KP**, **Bedi AK**, **Kalra SS and Chaitanya MS** (2015) Hierarchical deep learning architecture for 10k objects classification. In *2nd International Conference on Computer Science & Engineering (CSEN 2015). Computer Science & Information Technology (CS & IT)* Vol. *5*, Dubai, UAE: Academy & Industry Research Collaboration Center (AIRCC), pp. 79–94. https://doi.org/10.5121/csit.2015.51408

Kattenborn T, **Eichel J and Fassnacht FE** (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). https://doi.org/10.1038/s41598-019-53797-9.

Kattenborn T, **Leitloff J**, **Schiefer F and Hinz S** (2021) Review on convolutional neural networks (CNN) in vegetation remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing 173*, 24–49.

Khelifi L and Mignotte M (2020) Deep learning for change detection in remote sensing images: Comprehensive review and meta-analysis. *IEEE Access 8*, 126385–126400.

Kingma DP and Ba J (2015) Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Lenoir J, **Gégout JC**, **Marquet PA**, **de Ruffray P and Brisse H** (2008) A significant upward shift in plant species optimum elevation during the 20th century. *Science 320*(5884), 1768–1771.

Li S, **Brandt M**, **Fensholt R**, **Kariryaa A**, **Igel C**, **Gieseke F**, **Nord-Larsen T**, **Oehmcke S**, **Carlsen AH**, **Junttila S**, **Tong X**, **d'Aspremont A and Ciais P** (2023) Deep learning enables image-based tree counting, crown segmentation, and height prediction at national scale. *PNAS Nexus 2*(4). https://doi.org/10.1093/pnasnexus/pgad076.

Li X, **Chen H**, **Qi X**, **Dou Q**, **Fu C-W and Heng P-A** (2018) H-denseunet: Hybrid densely connected UNet for liver and tumor segmentation from ct volumes. *IEEE Transactions on Medical Imaging 37*(12), 2663–2674.

Li L, **Zhou T**, **Wang W**, **Li J and Yang Y** (2022) Deep hierarchical semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 1246–1257.

Lin T-Y, **Maire M**, **Belongie S**, **Hays J**, **Perona P**, **Ramanan D**, **Dollár P and Zitnick CL** (2014) Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*. Zurich, Switzerland: Springer International Publishing, pp. 740–755. https://doi.org/10.1007/978-3-319-10602-1_48.

Liu Z, **Lin Y**, **Cao Y**, **Hu H**, **Wei Y**, **Zhang Z**, **Lin S and Guo B** (2021) Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986.

Long J, **Shelhamer E and Darrell T** (2015) Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA: IEEE Computer Society, pp. 3431–3440.

Loshchilov I and Hutter F (2019) Decoupled weight decay regularization. In *International Conference on Learning Representations*. Available at https://openreview.net/forum?id=Bkg6RiCqY7.

Lucas B, **Pelletier C**, **Schmidt D**, **Webb GI and Petitjean F** (2021) A bayesian-inspired, deep learning-based, semi-supervised domain adaptation technique for land cover mapping. *Machine Learning 112*(6), 1941–1973.

Mahecha MD, **Bastos A**, **Bohn FJ**, **Eisenhauer N**, **Feilhauer H**, **Hickler T**, **Kalesse-Los H**, **Migliavacca M**, **Otto FEL**, **Peng J**, **Sippel S**, **Tegen I**, **Weigelt A**, **Wendisch M**, **Wirth C**, **Al-Halbouni D**, **Deneke H**, **Doktor D**, **Dunker S**, **Duveiller G**, **Ehrlich A**, **Foth A**, **García-García A**, **Guerra CA**, **Guimarães-Steinicke C**, **Hartmann H**, **Henning S**, **Herrmann H**, **Hu P**, **Ji C**, **Kattenborn T**, **Kolleck N**, **Kretschmer M**, **Kühn I**, **Luttkus ML**, **Maahn M**, **Mönks M**, **Mora K**, **Pöhlker M**, **Reichstein M**, **Rüger N**, **Sánchez-Parra B**, **Schäfer M**, **Stratmann F**, **Tesche M**, **Wehner B**, **Wieneke S**, **Winkler AJ**, **Wolf S**, **Zaehle S**, **Zscheischler J and Quaas J** (2024) Biodiversity and climate extremes: Known interactions and research gaps. *Earth's Future*, *12*(6). https://doi.org/10.1029/2023ef003963.

Maretto RV, **Fonseca LMG**, **Jacobs N**, **Körting TS**, **Bendini HN and Parente LL** (2021) Spatio-temporal deep learning approach to map deforestation in amazon rainforest. *IEEE Geoscience and Remote Sensing Letters 18*(5), 771–775.

Milletari F, **Navab N and Ahmadi S-A** (2016) V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE, pp. 565–571.

Muller BR and Smith W (2020) A hierarchical loss for semantic segmentation. In *VISIGRAPP*, pp. 260–267. https://doi.org/10.5220/0008946002600267

Nguyen T-A, **Rußwurm M**, **Lenczner G and Tuia D** (2024) Multi-temporal forest monitoring in the swiss alps with knowledge-guided deep learning. *Remote Sensing of Environment 305*, 114109.

Oktay O, **Schlemper J**, **Le Folgoc L**, **Matthew CHL**, **Heinrich MP**, **Misawa K**, **Mori K**, **McDonagh SG**, **Hammerla NY**, **Kainz B**, **Glocker B and Rueckert D** (2018) Attention U-net: Learning where to look for the pancreas. *CoRR* **abs/1804.03999**.

Onishi M and Ise T (2021) Explainable identification and mapping of trees using UAV RGB image and deep learning. *Scientific Reports 11*(1).

Ouaknine A, **Kattenborn T**, **Laliberté E and Rolnick D** (2025) OpenForest: A data catalogue for machine learning in forest monitoring. *Environmental Data Science 4*. https://doi.org/10.1017/eds.2024.53.

Reed CJ, **Gupta R**, **Li S**, **Brockman S**, **Funk C**, **Clipp B**, **Keutzer K**, **Candido S**, **Uyttendaele M and Darrell T** (2023) Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, pp. 4065–4076.

Reichstein M, **Camps-Valls G**, **Stevens B**, **Jung M and Denzler J** (2019) Nuno Carvalhais, and Prabhat. Deep learning and process understanding for data-driven earth system science. *Nature 566*(7743), 195–204.

Ronneberger O, **Fischer P and Brox T** (2015) U-Net: Convolutional networks for biomedical image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Munich, Germany: Springer International Publishing, pp. 234–241

Rußwurm M, **Courty N**, **Emonet R**, **Lefèvre S**, **Tuia D and Tavenard R** (2023) End-to-end learned early classification of time series for in-season crop type mapping. *ISPRS Journal of Photogrammetry and Remote Sensing 196*, 445–456.

Rußwurm M and Körner M (2018) Multi-temporal land cover classification with sequential recurrent encoders. *ISPRS International Journal of Geo-Information 7*(4), 129.

Sainte Fare Garnot V, **Landrieu L**, **Giordano S and Chehata N** (2020) Satellite image time series classification with pixel-set encoders and temporal self-attention. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* JEEE, pp. 12322–12331. https://doi.org/10.1109/CVPR42600.2020.01234.

Schiefer F, **Kattenborn T**, **Frick A**, **Frey J**, **Schall P**, **Koch B and 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.

Sharma A, **Tuzel O and Jacobs DW** (2015) Deep hierarchical parsing for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA: IEEE, pp. 530-538. https://doi.org/10.1109/CVPR.2015.7298651.

Song G, **Wu S**, **Calvin KFL**, **Serbin SP**, **Wolfe BT**, **Ng MK**, **Ely KS**, **Bogonovich M**, **Wang J**, **Lin Z**, **Saleska S**, **Nelson BW**, **Rogers A and Wu J** (2022) Monitoring leaf phenology in moist tropical forests by applying a superpixel-based deep learning method to time-series images of tree canopies. *ISPRS Journal of Photogrammetry and Remote Sensing 183*, 19–33.

Tarasiou M, **Chavez E and Zafeiriou S** (2023) ViTs for SITS: Vision transformers for satellite image time series. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10418–10428. https://doi.org/10.1109/CVPR52729.2023.01004.

Tseng G, **Cartuyvels R**, **Zvonkov I**, **Purohit M**, **Rolnick D and Kerner H** (2023) Lightweight, pre-trained transformers for remote sensing timeseries. In *NeurIPS 2023 Workshop on Tackling Climate Change with Machine Learning*. Available at https://www.climatechange.ai/papers/neurips2023/58.

Turkoglu MO, **D'Aronco S**, **Perich G**, **Liebisch F**, **Streit C**, **Schindler K and Wegner JD** (2021) Crop mapping from image time series: Deep learning with multi-scale label hierarchies. *Remote Sensing of Environment 264*, 112603.

Vali A, **Comai S and Matteucci M** (2020) Deep learning for land use and land cover classification based on hyperspectral and multispectral earth observation data: A review. *Remote Sensing 12*(15), 2495.

Valmadre J (2022) Hierarchical classification at multiple operating points. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*. Red Hook, NY: Neural Information Processing Systems Foundation.

Visser ME and Gienapp P (2019) Evolutionary and demographic consequences of phenological mismatches. *Nature Ecology & Evolution 3*(6), 879–885.

Wang J, **Song G**, **Liddell M**, **Morellato P**, **Calvin KFL**, **Yang D**, **Alberton B**, **Detto M**, **Ma X**, **Zhao Y**, **Henry CHY**, **Zhang H**, **Ng M**, **Nelson BW**, **Huete A and Wu J** (2023) An ecologically-constrained deep learning model for tropical leaf phenology monitoring using planetscope satellites. *Remote Sensing of Environment 286*, 113429.

Weinstein BG, **Marconi S**, **Aubry-Kientz M**, **Vincent G**, **Senyondo H and White EP** (2020) Deepforest: A python package for rgb deep learning tree crown delineation. *Methods in Ecology and Evolution 11*(12), 1743–1751.

Xie E, **Wang W**, **Yu Z**, **Anandkumar A**, **Alvarez JM and Luo P** (2021) Segformer: Simple and efficient design for semantic segmentation with transformers. In Ranzato M, Beygelzimer A, Dauphin Y, Liang PS and Vaughan JW (eds), *Advances in Neural Information Processing Systems*, Vol. *34*. Curran Associates, Inc, pp. 12077–12090

Yu F and Koltun V (2016) Multi-scale context aggregation by dilated convolutions. In Bengio Y and LeCun Y (eds), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2–4, 2016, Conference Track Proceedings*

Zhang L, **Shao Z**, **Liu J and Cheng Q** (2019) Deep learning based retrieval of forest aboveground biomass from combined LiDAR and landsat 8 data. *Remote Sensing 11*(12), 1459.

Zhang C, **Xia K**, **Feng H**, **Yang Y and Du X** (2020) Tree species classification using deep learning and rgb optical images obtained by an unmanned aerial vehicle. *Journal of Forestry Research 32*(5), 1879–1888. https://doi.org/10.1007/s11676-020-01245-0.

Zhao H, **Shi J**, **Qi X**, **Wang X and Jia J** (2017) Pyramid scene parsing network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA: IEEE Computer Society, pp. 6230–6239.

Zheng S, **Lu J**, **Zhao H**, **Zhu X**, **Luo Z**, **Wang Y**, **Fu Y**, **Feng J**, **Xiang T**, **Torr PS and Zhang L** (2021) Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA: IEEE Computer Society, pp. 6877–6886.

Zhou X, **Xin Q**, **Dai Y and Li W** (2021) A deep-learning-based experiment for benchmarking the performance of global terrestrial vegetation phenology models. *Global Ecology and Biogeography 30*(11), 2178–2199.

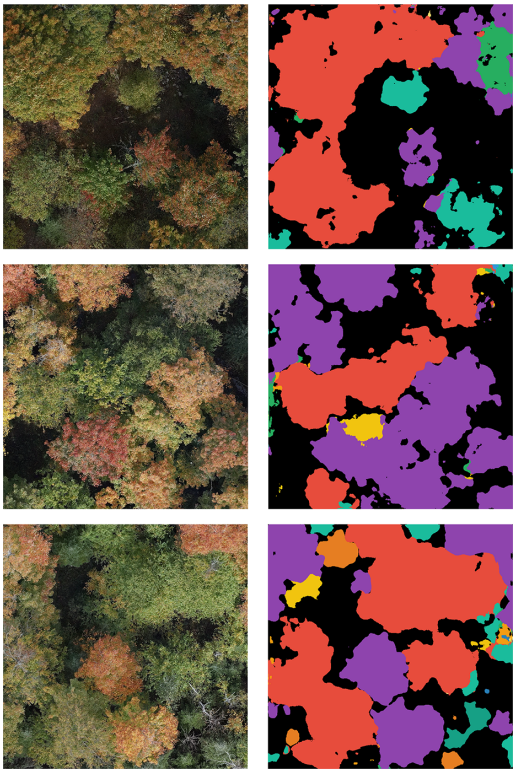
Zwemer MH, **Rob GJW and Peter HN d W** (2022) *Hierarchical Object Detection and Classification Using SSD Multi-Loss*. Springer International Publishing, pp. 268–296

A . Appendix

A.1. Spatial transferability

To assess the geographic generalization capabilities of our model, we conducted additional experiments using aerial imagery from the municipality of Stornoway, located in Quebec's Le Granit regional county municipality within the administrative region of Estrie. This location was specifically chosen for evaluation as it shares similar ecological characteristics with our training site in the Laurentides region, particularly in terms of tree species composition and forest structure typical of Quebec's temperate-mixed forests.

For generating the predictions, we chose the best performing Processer+U-Net model with ResNet-101 backbone. Given the scarcity of high-resolution time-series datasets and the difficulty of collecting such datasets for tree species segmentation, we conducted our experiments by replicating a single image across four time steps in our time-series model. The results of these predictions are shown in Figure A1. While a comprehensive quantitative evaluation was not possible due to the absence of ground truth labels for this region, qualitative assessment of the model's predictions reveals that the model demonstrates robust capabilities in both detecting individual trees and accurately delineating crown boundaries, even in areas with dense canopy cover and varying lighting conditions.



(a) Input image. (b) Model output.

Figure A1. *Evaluation of spatial transferability in Stornoway, Quebec. The figure presents paired comparisons of original input imagery (left) and corresponding model predictions (right) from a geographically distinct test location. Each image represents an area of approximately 20.1m × 20.1m. Here, we used the time-series model with a single image replicated across four time steps due to the scarcity of relevant time-series datasets for tree species segmentation. Tree species are color-coded according to the scheme established in Table 1. The model is effective in segmenting and delineating tree crowns in this new location, particularly for distinguishing between neighboring trees with different species compositions. This transfer to a new geographic location, while within a similar ecological zone, suggests the model's potential for broader regional application in forests of similar tree composition.*

The model's ability to maintain consistent performance in distinguishing between different tree species suggests that the learned features are sufficiently generalizable across similar forest ecosystems within Quebec. However, it is important to acknowledge that this evaluation is limited to regions with comparable ecological conditions.

A.2. Classification metrics

In addition to the mIoU metric presented in the main text, Tables A1 and A2 show F1-score, precision, and recall metrics for our single-image and time-series models, respectively. These metrics provide complementary insights into our models' classification performance.

For single-image models (Table A1), the U-Net architecture with a ResNet101 backbone trained with HLoss achieved the highest mean scores (F1: 71.09 ± 0.24, Precision: 69.43 ± 0.63, Recall: 71.89 ± 0.67). Comparing this model to the same architecture trained with Dice+CE loss (F1: 70.64 ± 0.18, Precision: 69.49 ± 0.68, Recall: 71.83 ± 0.73), the HLoss model showed a likely statistically significant improvement in F1-score. However, the differences in mean Precision and Recall were small relative to their standard deviations, with overlapping error ranges suggesting these metrics were comparable between the two loss functions for this model. The performance degradation observed in models trained from scratch (denoted by [a]) was consistent across all classification metrics, reinforcing the importance of transfer learning.

In time-series approaches (Table A2), the integration of temporal information via our Processor module generally led to strong classification performance, particularly when combined with U-Net and HLoss. The Processor+U-Net model with a ResNet101 backbone and HLoss achieved the highest mean scores across all metrics (F1: 71.77 ± 0.39, Precision: 72.28 ± 1.20, Recall: 71.27 ± 1.17). Compared to the same model trained with Dice+CE (F1: 71.00 ± 0.39, Precision: 71.66 ± 1.00, Recall: 70.35 ± 0.96), the HLoss version showed slight improvements in F1 and Recall, while Precision was comparable. Specialized time-series architectures like UNet 3D and U-TAE, while designed to capture temporal patterns, achieved lower classification scores (F1-scores around 55 ± 0.55 and 55 ± 0.37), likely due to the lack of pretrained weights and the challenge of training such architectures from scratch on limited data.

Overall, the analysis of classification metrics aligns with the mIoU findings. The benefits of the Processor module and HLoss are evident. However, the improvements are not uniformly significant across all metrics or all model comparisons when considering the error margins from the three-fold cross-validation.

Table A1. *Comparison of single image methods with different classification metrics*

Model	Backbone	F1-score	Precision	Recall
DeepLabv3+	ResNet34	**68.96 ± 0.33**	**67.22 ± 0.17**	70.21 ± 0.19
		68.68 ± 0.26	66.90 ± 1.27	**70.56 ± 1.41**
	ResNet50	**69.35 ± 0.27**	66.40 ± 0.49	**71.42 ± 0.57**
		68.82 ± 0.31	**67.08 ± 1.64**	70.65 ± 1.82
	ResNet101	**70.81 ± 0.37**	69.17 ± 0.39	**71.36 ± 0.41**
		70.25 ± 0.43	**69.28 ± 1.37**	71.25 ± 1.45
	ResNet50[a]	**60.05 ± 0.88**	**60.91 ± 0.49**	59.14 ± 0.46
		60.01 ± 0.81	59.80 ± 0.87	**60.22 ± 0.88**
U-Net	ResNet34	**68.90 ± 0.55**	66.79 ± 0.22	**70.49 ± 0.24**
		68.59 ± 0.56	**67.84 ± 1.20**	69.36 ± 1.25
	ResNet50	**69.64 ± 0.37**	67.64 ± 0.35	**70.62 ± 0.38**
		69.10 ± 0.50	**68.02 ± 0.49**	70.22 ± 0.52
	ResNet101	**71.09 ± 0.24**	69.43 ± 0.63	**71.89 ± 0.67**
		70.64 ± 0.18	**69.49 ± 0.68**	71.83 ± 0.73
	ResNet50[a]	**59.32 ± 0.68**	**60.79 ± 3.18**	57.82 ± 2.88
		59.27 ± 0.74	57.65 ± 0.93	**60.98 ± 1.04**
Mask2Former	Swin-t[b]	62.81 ± 1.48	64.28 ± 2.58	61.41 ± 2.35
	Swin-s[b]	62.77 ± 1.49	63.89 ± 1.86	61.69 ± 1.73

Note. Performances are compared using F1-score, Precision, and Recall averaged over all the classes of the dataset. For DeepLabv3+ and U-Net models, each backbone shows two rows of results: HLoss metrics in the first row and Dice Loss metrics in the second row.

[a]Indicates models trained from scratch without using ImageNet weights (Deng et al., 2009).

[b]Indicates Swin-based models using weights from the MS-COCO dataset (Lin et al., 2014). All results are averaged across three-fold cross-validation, and the best result for each backbone will be shown in bold text.

Table A2. *Comparison of time-series methods with different classification metrics*

Model	Backbone	F1-score	Precision	Recall
DeepLabv3+ + Processor	ResNet34	**69.02 ± 0.48**	67.95 ± 1.33	**70.12 ± 1.42**
		69.03 ± 0.59	**70.80 ± 0.21**	67.35 ± 0.19
	ResNet50	**69.63 ± 0.64**	**69.02 ± 0.44**	**70.25 ± 0.46**
		68.89 ± 0.55	67.72 ± 0.09	70.10 ± 0.09
	ResNet101	**70.91 ± 0.26**	**71.40 ± 0.82**	70.43 ± 0.80
		70.68 ± 0.36	71.23 ± 0.43	**70.14 ± 0.42**
	ResNet50[a]	65.17 ± 0.69	**65.00 ± 1.48**	**66.70 ± 1.56**
		65.60 ± 0.28	64.75 ± 0.27	65.84 ± 0.79
U-Net + Processor	ResNet34	**69.20 ± 0.50**	**68.10 ± 0.04**	**70.34 ± 0.04**
		69.19 ± 0.50	68.18 ± 0.55	70.23 ± 0.58
	ResNet50	**70.00 ± 0.72**	**68.86 ± 1.03**	**71.18 ± 1.10**
		69.59 ± 0.71	67.78 ± 1.32	69.42 ± 1.38
	ResNet101	**71.77 ± 0.39**	**72.28 ± 1.20**	71.27 ± 1.17
		71.00 ± 0.39	71.66 ± 1.00	**70.35 ± 0.96**
	ResNet50[a]	**65.83 ± 0.41**	**64.78 ± 0.30**	**66.91 ± 0.32**
		65.74 ± 0.33	63.49 ± 0.05	68.16 ± 0.05
UNet 3D[a]	–	**55.55 ± 0.27**	**47.95 ± 0.49**	**66.02 ± 0.93**
		55.01 ± 0.61	48.02 ± 0.23	64.39 ± 0.34
U-TAE[a]	–	**55.37 ± 0.35**	**50.15 ± 1.52**	**61.81 ± 2.31**
		53.84 ± 0.96	47.93 ± 1.29	61.41 ± 2.10

Note. Performances are compared using F1-score, Precision, and Recall averaged over all the classes of the dataset. For each model and backbone combination, the first row shows results using HLoss metrics and the second row shows results using Dice Loss metrics.
[a]Indicates models trained from scratch. All results are averaged across three-fold cross-validation, and the best result for each backbone will be shown in bold text.

Cite this article: Ramesh V, Ouaknine A and Rolnick D (2025). Tree semantic segmentation from aerial image time series. *Environmental Data Science*, 4: e37. doi:10.1017/eds.2025.10013