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
[2] DIRO, Université de Montréal venkatesh.ramesh@mila.quebec
[3] School of Computer Science, McGill University,

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 semantic segmentation of trees 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 performances. 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 findings demonstrate the superiority of our methodology in exploiting the time series modality and confirm that enriching labels using taxonomic information improves the semantic segmentation performance.

Keywords: Forest monitoring · Remote sensing · Deep learning · Time series

1 Introduction

Climate change and biodiversity loss in forests are closely intertwined, with each potentially exacerbating the other. As climate patterns shift, the suitable habitats for many tree species change, leading to alterations in forest composition and potential biodiversity loss [1,35]. Conversely, reduced forest biodiversity can diminish carbon absorption and storage capacity, further contributing to climate change. Different tree species exhibit varying tolerances to environmental changes, resulting in diverse phenological responses [65], shifts in species distribution [3], and differential growth patterns [2, 6]. Understanding these species-specific responses is crucial for effective forest monitoring and management.

Increasingly, deep learning-based methods, alongside remote sensing applications (*e.g.* land-use and land-cover mapping [25,26,28,63], change detection [33]), have helped with advancing the field of forest monitoring in tree species classification [21], biomass estimation [70] and tree crown segmentation [57, 67].

The use of temporal data as inputs for these methods has shown promise in various applications, including crop mapping [9, 56, 60] and forest health mapping [25]. However, the potential of leveraging time series data for tree crown segmentation, particularly to capture phenological changes, remains largely unexplored.

In this work, we address this gap by evaluating multiple models on the task of tree crown segmentation using a rich dataset from the Laurentides region of Québec, Canada [16]. This dataset offers unique characteristics, including high-resolution time series data and a number of closely-related tree species classes, allowing us to investigate the impact of phenological changes on tree species identification.

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. To address the challenge of mixed label granularity in our dataset, we propose a custom hierarchical loss function that incorporates species, genus, and family-level labels. Our key contributions are:

- 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.
- Our results demonstrate the importance of time series data in identifying tree species, particularly when considering phenological changes.
- We evaluate models that perform well across taxonomic hierarchies by leveraging a custom hierarchical 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) [41], U-Net [53], and DeepLab [12].

The 'dilated' (also named 'atrous') convolution [12, 69] has been introduced to increase the receptive field of CNNs, while attention mechanisms [22, 48] have been incorporated to focus on relevant regions. Multi-scale and pyramid pooling approaches, such as PSPNet [71] and DeepLabV3+ [13], 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 [15] and V-Net [45].

Recently, transformer-based models have gained popularity in semantic segmentation, showing impressive results, *e.g.* Mask2Former [14] combining 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 [72], TransUNet [11], and SegFormer [68], 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 performances 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 [43] and crop classification [55]. Attention-based methods have been used for encoding time series, which have shown to be well-suited for satellite imagery [23, 54, 56]. More recently, transformer-based methods have proven their merit using SITS with self-supervised learning exploiting unlabelled data to improve performance on downstream tasks [17, 51, 60, 61].

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 [9].

2.3 Forest monitoring

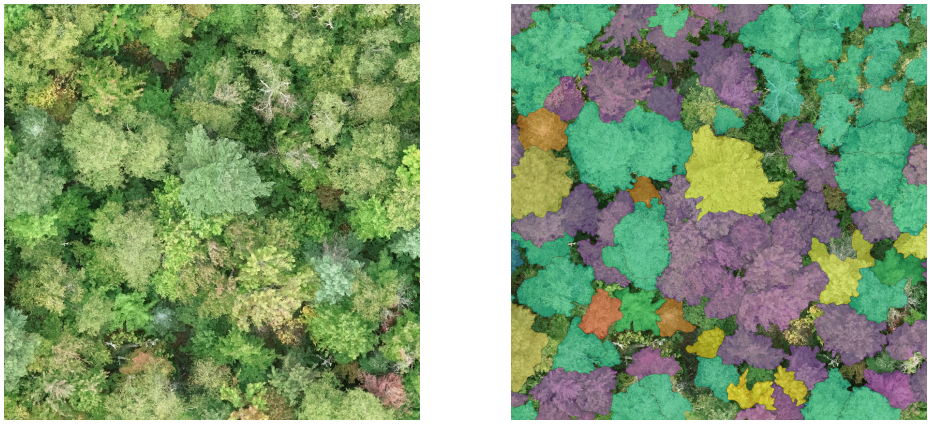
Deep learning methods have helped advance the field of vegetation monitoring using remote sensing, including both satellite and aerial imagery [32], enabling progress in forest monitoring for accurate and efficient analysis at scale [4,47,52]. Such models have achieved state-of-the-art performance in classifying tree species from high-resolution remote sensing imagery [21, 49].

Mapping deforestation at large scale using satellite imagery has also been explored [44, 50]. Computer vision and remote sensing have also been leveraged in applications to plant phenology [29]. Global vegetation phenology has been modelled with satellite imagery alongside meteorological variables as inputs of a 1D CNN [73]. Automated monitoring of forests have also been investigated to accurately identify key phenological events [10, 59, 66].

Deep learning-based segmentation methods have been applied to automatically delineate individual tree crowns from high-resolution remote sensing imagery [8, 37, 57, 67]. In a similar vein, a U-Net architecture has been used for fine-grained segmentation of plant species using aerial imagery [31]. 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 [7].

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



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

Fig. 1: Example of an annotated sample from the studied dataset. The image 1a 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.

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 [24]. Similarly, a loss function evaluated at multiple operating points within the class hierarchy has helped to capture information at various levels of this hierarchy [64]. 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 [5].

Hierarchical loss functions have also been applied to object detection [30, 74] and semantic segmentation [36, 46, 58] demonstrating the effectiveness of incorporating a more structured and informative signal during the learning process.

3 Dataset

The dataset used in our work [16] 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 colour 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 lever-

age 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 create splits which are separated spatially for training, validation, and testing while ensuring an equal distribution of the selected classes in each split. The spatial splits are used to evaluate the performances of each model on geographically distinct areas and to simulate real-world scenarios, *e.g.* applications to unseen locations. The splits that we used are illustrated in Appendix A. A tile is skipped between the three sets to prevent data leakage, ensuring the model avoids spatial autocorrelation between adjacent areas. 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, we ignore classes with less than 50 occurrences throughout the dataset, 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 2.

The dataset is split into train, validation and testing sets with 63%, 16%, and 21% of the samples, respectively. Given that this dataset has a mix of coarse (genus) and fine-grained (species) labels, we leverage this information to create a complete taxonomy of the classes used, as seen in Figure 3.

This taxonomic hierarchy is incorporated in our proposed loss function as detailed in Sec. 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 hierarichal loss used to exploit the tree label taxonomy.

4.1 Single image semantic segmentation

For single image semantic segmentation, we evaluate three state-of-the-art architectures: U-Net [53], DeepLabv3+ [13], and Mask2Former [14]. U-Net utilizes an encoder-decoder structure with skip connections, allowing for precise localization. DeepLabv3+ employs atrous convolutions and atrous spatial pyramid pooling to capture multi-scale context. Mask2Former combines a CNN backbone with a transformer decoder, using masked attention to focus on relevant image regions. We experiment with various backbone networks, including ResNet-34, ResNet-50, and ResNet-101 for U-Net and DeepLabv3+, and Swin-T and Swin-S

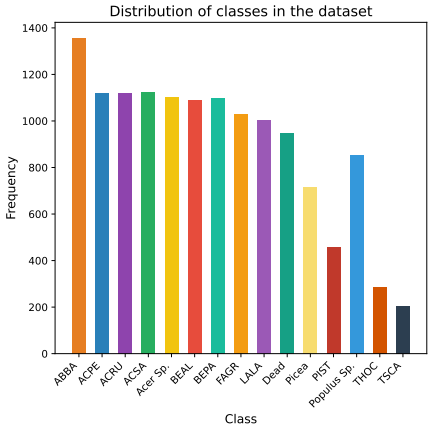


Fig. 2: Distribution of the selected classes in the dataset. We observe that there is a substantial difference in the frequency of occurrence of each tree species. The common and scientific names used for the abbreviations are detailed in Table 1.

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

Table 1: Tree species names and their abbreviations. The color we use to depict each species is highlighted in the second column and is consistent for all the plots and figures.

for Mask2Former. Detailed architecture descriptions are provided in Appendix B.1.

4.2 Time series semantic segmentation

We compare various methods for semantic segmentation with time series data, including 3D-UNet [15] and U-TAE [23] specialized for SITS. 3D-UNet extends the U-Net architecture to 3D, making it suitable for volumetric data and U-TAE employs a temporal attention encoder to capture temporal dependencies in satellite image time series. Detailed descriptions of these methods are provided in Appendix B.2.

Processor module Our proposed Processor module is composed of 3D convolutions and 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 [23, 60].

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

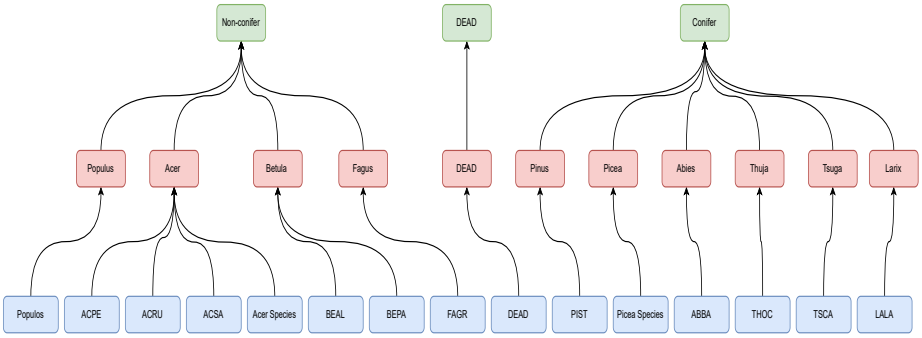


Fig. 3: Taxonomic hierarchy of tree species. The hierarchical structure is visually represented using a tree diagram. Blue nodes represent the species level, the most fine-grained classification in the hierarchy. Red nodes denote the genus level, which groups together closely related species. Finally, 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.

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, let $\mathbf{x} \in \mathbb{R}^{T \times C \times H \times W}$ 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.

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 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 3 and a visual example of these labels is illustrated in Appendix A.

During training, the model predicts only the species level labels for each pixel. These softmax probabilities at species level are then aggregated according to our knowledge of the label taxonomies (see Figure 3) to generate first the genus level predictions (see Equation 3) and second the higher-level predictions (see Equation 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 [62].

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:

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

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:

$$\mathcal{L}_{\text{genus}} := -\frac{1}{G} \sum_{g=1}^{G} \sum_{(h,w)\in\Omega} \mathbf{y}_G[h,w,g] \log\, \mathbf{p}_G[h,w,g] \tag{2}$$

$$= -\frac{1}{G} \sum_{g=1}^{G} \sum_{(h,w)\in\Omega} \Big[\sum_{s=1}^{S_g} \mathbf{y}_S[h,w,s]\Big] \log\Big[\sum_{s=1}^{S_g} \mathbf{p}_S[h,w,s]\Big], \tag{3}$$

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:

$$\mathcal{L}_{\text{taxon}} := -\frac{1}{T} \sum_{t=1}^{T} \sum_{(h,w)\in\Omega} \mathbf{y}_T[h,w,t] \log\, \mathbf{p}_T[h,w,t] \tag{4}$$

$$= -\frac{1}{T} \sum_{t=1}^{T} \sum_{(h,w)\in\Omega} \Big[\sum_{g=1}^{G_t} \sum_{s=1}^{S_g} \mathbf{y}_S[h,w,s]\Big] \log\Big[\sum_{g=1}^{G_t} \sum_{s=1}^{S_g} \mathbf{p}_S[h,w,s]\Big], \tag{5}$$

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 given as:

$$\mathcal{L}_{\text{HLoss}} = \lambda_1 \cdot \mathcal{L}_{\text{species}} + \lambda_2 \cdot \mathcal{L}_{\text{genus}} + \lambda_3 \cdot \mathcal{L}_{\text{taxon}}, \tag{6}$$

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

Model	Backbone	Dice+CE	HLoss	Mask2Former Loss
DeepLabv3+	ResNet34	52.30 ± 0.40	**53.36 ± 0.09**	—
	ResNet50	53.14 ± 0.08	**53.63 ± 0.20**	—
	ResNet101	53.87 ± 0.45	**54.16 ± 0.38**	—
	ResNet50†	43.19 ± 0.45	**43.20 ± 0.06**	—
U-Net	ResNet34	53.00 ± 0.10	**53.10 ± 0.16**	—
	ResNet50	53.30 ± 0.16	**53.53 ± 0.46**	—
	ResNet101	53.90 ± 0.18	**54.31 ± 0.48**	—
	ResNet50†	**42.66 ± 0.38**	42.43 ± 0.96	—
Mask2Former	Swin-t††	—	—	47.41 ± 0.50
	Swin-s††	—	—	46.61 ± 0.10

Table 2: Comparison of single image methods with different losses and backbones. Performances are compared with IoU averaged over all the classes of the dataset (mIoU) for single image models. The † indicates models trained from scratch without using ImageNet weights [18]. The †† indicates Swin-based models using weights from MS-COCO dataset [39]. All the results are averaged over three seeds and the best results for a particular backbone is shown in bold text. The best model overall is highlighted in red.

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

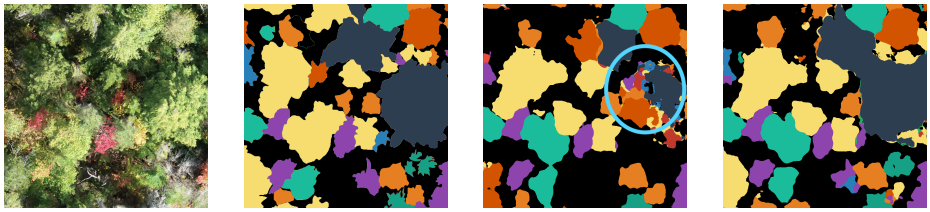
5.1 Experimental setup

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.

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. The performances of our models are evaluated with the Intersection over Union (IoU) metric, also known as the Jaccard index, measuring the overlap between the predicted and ground truth masks. The mean IoU (mIoU) is computed by averaging the IoU scores across all classes. A more detailed explanation of our training process and hyperparamter choices is give in Appendix C.

5.2 Experiment Configuration

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



(a) A sample image. **(b)** Annotation. **(c)** Dice+CE **(d)** HLoss

Fig. 4: 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). First, 4a shows a sample image from the sequence, while 4b displays the corresponding ground truth annotation. Then, 4c depicts the segmentation output obtained by the model trained with the Dice+CE loss, and finally 4d 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.

- 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 are 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 effectiveness 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 performances. 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 [40].

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 performances 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.

Model	Backbone	Dice+CE	HLoss
DeepLabv3+ + Processor	ResNet34	53.32 ± 0.04	**53.43 ± 0.34**
	ResNet50	53.12 ± 0.34	**54.05 ± 0.31**
	ResNet101	53.46 ± 0.53	**54.13 ± 0.11**
	ResNet50†	48.01 ± 0.37	**50.30 ± 2.82**
U-Net + Processor	ResNet34	53.37 ± 0.53	**53.60 ± 0.31**
	ResNet50	53.80 ± 0.15	**54.12 ± 0.15**
	ResNet101	54.46 ± 0.39	**54.88 ± 0.20**
	ResNet50†	49.00 ± 0.19	**49.35 ± 0.25**
UNet 3D†	—	37.74 ± 0.28	**41.38 ± 0.14**
U-TAE†	—	35.59 ± 1.03	**39.78 ± 2.58**

Table 3: Comparison of time-series methods with different losses and backbones. Performances are compared with IoU averaged over all the classes of the dataset (mIoU) for single image models. The † indicates models trained from scratch. All the results are averaged over three seeds and the best results for a particular backbone is shown in bold text. The best model overall is highlighted in red.

As seen in Table 2, both DeepLabv3+ and U-Net architectures show consistent increase in performances with increasing backbone size where the ResNet-101 model achieving the highest mIoU score. Moreover, the proposed HLoss consistently outperforms the Dice+CE loss across all backbones and architectures, demonstrating the effectiveness of leveraging taxonomic information.

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. The best performing single-image model is the U-Net with ResNet101 backbone. 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.

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 performances with DeepLabv3+ and U-Net architectures. Amongst the time series models, we observe a similar pattern in Table 3 as the single-image ones with the benefical impact of using HLoss during training outperforming models trained with the Dice+CE loss. Qualitative results comparing HLoss with Dice+CE loss are illustrated in Figure 4 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 amongst classes that do not belong in the same genera 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 amongst 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.

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 5 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 ACPE, ACRU, or ACSA that have not been assigned a fine-grained label by the annotators due to low confidence.

Class	Processor + U-Net	U-Net
Non-Coniferous Trees		
Populus	**78.49**	74.89
ACPE	29.64	**29.74**
ACRU	**57.51**	55.84
ACSA	**46.56**	44.30
BEAL	**63.00**	62.06
BEPA	**72.55**	71.16
FAGR	55.69	**56.76**
Coniferous Trees		
PIST	75.50	**77.89**
Picea	60.57	**60.82**
ABBA	62.35	**64.15**
THOC	**59.73**	59.68
TSCA	**69.84**	57.75
LALA	**76.68**	76.60
Others		
DEAD	43.38	**46.57**
Overall results	**54.88 ± 0.20**	54.31 ± 0.48

Fig. 5: The table shows the IoU for the individual classes for our best-performing Processor + U-Net and U-Net models, both with ResNet-101 as backbone. 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.

The results show that the time series model consistently outperforms the single-image model across nearly all non-coniferous classes. Even in the few instances where the single-image model achieves a slightly higher IoU, the performance gap is minimal. This finding aligns with our hypothesis that incorporating time series data allows the models to better capture and exploit the phenological changes exhibited by different tree species, particularly those that undergo distinct color changes during the fall season. By leveraging this temporal information, the

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

Fig. 6: 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. The Processor+UNet (ResNet101) architecture, trained with the proposed Hierarchical Loss (HLoss), achieves the highest mIoU score according to Table 3. First, 6a shows a sample image from the sequence, while 6b displays the corresponding ground truth annotation. Then 6c depicts the segmentation output obtained by the single-image model, and finally 6d 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 the time series model consistently 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.

time series model is able to more accurately identify and distinguish between the various non-coniferous species.

In contrast, the performance differences between the single-image and time series models are less pronounced for the coniferous classes. Both models demonstrate strong performances in this category, with the most significant improvement for the time series model observed in the TSCA class. The single image model suffers from confusion between the classes Eastern hemlock (TSCA) and Eastern white pine (PIST) which does not affect the time series model as it takes in multiple input images with different lightning and acquisition angles. This suggests that even though phenological changes in coniferous trees may be less informative for species identification compared to their non-coniferous counterparts, the information from the image time series still helps the model identify the classes better.

An example of the results comparing single-image and time series models is illustrated in Figure 6, 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 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.

7 Conclusion

In this work, we addressed tree species segmentation using aerial image time series, demonstrating the advantages of incorporating temporal information for accurate species identification. We introduced a lightweight Processor module for extracting spatio-temporal features and a hierarchical loss function leveraging taxonomic structure, both enhancing existing segmentation architectures. While our Processor module is designed for a fixed number of time steps, it offers a simple yet effective approach to leveraging temporal information in tree species segmentation. Future work could explore increasing its flexibility for varying temporal resolutions. Our methods have significant implications for forest monitoring and biodiversity conservation, enabling accurate mapping of tree species composition. This work demonstrates the potential of deep learning and time series analysis in advancing forest ecosystem understanding and preservation. Future research could explore additional data modalities and extend these methods to other applications in forest ecology and management, further contributing to global environmental efforts.

Acknowledgements

References

1. Allen, C.D., Macalady, A.K., Chenchouni, H., Bachelet, D., McDowell, N., Vennetier, M., Kitzberger, T., Rigling, A., Breshears, D.D., Hogg, E.T., Gonzalez, P., Fensham, R., Zhang, Z., Castro, J., Demidova, N., Lim, J.H., Allard, G., Running, S.W., Semerci, A., Cobb, N.: 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 (Feb 2010)
2. Anderegg, W.R.L., Kane, J.M., Anderegg, L.D.L.: Consequences of widespread tree mortality triggered by drought and temperature stress. Nature Climate Change **3**(1), 30–36 (Sep 2012)
3. Babst, F., Bouriaud, O., Poulter, B., Trouet, V., Girardin, M.P., Frank, D.C.: Twentieth century redistribution in climatic drivers of global tree growth. Science Advances **5**(1), eaat4313 (2019)
4. Bae, S., Levick, S.R., Heidrich, L., Magdon, P., Leutner, B.F., Wöllauer, S., Serebryanyk, A., Nauss, T., Krzystek, P., Gossner, M.M., 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, W.W., Müller, J.: Radar vision in the mapping of forest biodiversity from space. Nature Communications **10**(1) (Oct 2019)
5. Bertinetto, L., Mueller, R., Tertikas, K., Samangooei, S., Lord, N.A.: Making better mistakes: Leveraging class hierarchies with deep networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (Jun 2020)
6. Bonan, G.B.: Forests and climate change: Forcings, feedbacks, and the climate benefits of forests. Science **320**(5882), 1444–1449 (Jun 2008)
7. Bountos, N.I., Ouaknine, A., Rolnick, D.: Fomo-bench: a multi-modal, multi-scale and multi-task forest monitoring benchmark for remote sensing foundation models (2023)
8. Brandt, M., Tucker, C.J., Kariryaa, A., Rasmussen, K., Abel, C., Small, J., Chave, J., Rasmussen, L.V., Hiernaux, P., Diouf, A.A., Kergoat, L., Mertz, O., Igel, C., Gieseke, F., Schöning, J., Li, S., Melocik, K., Meyer, J., Sinno, S., Romero, E., Glennie, E., Montagu, A., Dendoncker, M., Fensholt, R.: An unexpectedly large count of trees in the west african sahara and sahel. Nature **587**(7832), 78–82 (Oct 2020)
9. Cai, X., Bi, Y., Nicholl, P., Sterritt, R.: Revisiting the encoding of satellite image time series (2023)
10. Cao, M., Sun, Y., Jiang, X., Li, Z., Xin, Q.: Identifying leaf phenology of deciduous broadleaf forests from phenocam images using a convolutional neural network regression method. Remote Sensing **13**(12), 2331 (Jun 2021)
11. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. CoRR **abs/2102.04306** (2021)
12. Chen, L., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. CoRR **abs/1706.05587** (2017)
13. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, p. 833–851. Springer International Publishing (2018)

14. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1280–1289 (2022)
15. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016, pp. 424–432. Lecture notes in computer science, Springer International Publishing, Cham (2016)
16. Cloutier, M., Germain, M., Laliberté, E.: Influence of temperate forest autumn leaf phenology on segmentation of tree species from UAV imagery using deep learning (Aug 2023)
17. Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D.B., Ermon, S.: Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery (2023)
18. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)
19. Dong, H., Yang, G., Liu, F., Mo, Y., Guo, Y.: Automatic Brain Tumor Detection and Segmentation Using U-Net Based Fully Convolutional Networks, p. 506–517. Springer International Publishing (2017)
20. Falk, T., Mai, D., Bensch, R., undefinediçek, O., Abdulkadir, A., Marrakchi, Y., Böhm, A., Deubner, J., Jäckel, Z., Seiwald, K., Dovzhenko, A., Tietz, O., Dal Bosco, C., Walsh, S., Saltukoglu, D., Tay, T.L., Prinz, M., Palme, K., Simons, M., Diester, I., Brox, T., Ronneberger, O.: U-net: deep learning for cell counting, detection, and morphometry. Nature Methods **16**(1), 67–70 (Dec 2018)
21. Fricker, G.A., Ventura, J.D., Wolf, J.A., North, M.P., Davis, F.W., Franklin, J.: A convolutional neural network classifier identifies tree species in mixed-conifer forest from hyperspectral imagery. Remote Sensing **11**(19), 2326 (Oct 2019)
22. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3141–3149 (2019)
23. Garnot, V.S.F., Landrieu, L.: Panoptic segmentation of satellite image time series with convolutional temporal attention networks (2021)
24. Goyal, P., Ghosh, S.: Hierarchical class-based curriculum loss (2020)
25. Hamdi, Z.M., Brandmeier, M., Straub, C.: Forest damage assessment using deep learning on high resolution remote sensing data. Remote Sensing **11**(17), 1976 (Aug 2019)
26. Hamedianfar, A., Mohamedou, C., Kangas, A., Vauhkonen, J.: 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 (Feb 2022)
27. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
28. Helber, P., Bischke, B., Dengel, A., Borth, D.: 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 (2019)
29. Katal, N., Rzanny, M., Mäder, P., Wäldchen, J.: Deep learning in plant phenological research: A systematic literature review. Frontiers in Plant Science **13** (Mar 2022)

30. Katole, A.L., Yellapragada, K.P., Bedi, A.K., Kalra, S.S., Siva Chaitanya, M.: Hierarchical deep learning architecture for 10k objects classification. In: Computer Science & Information Technology (CS & IT). Academy & Industry Research Collaboration Center (AIRCC) (Aug 2015)
31. Kattenborn, T., Eichel, J., Fassnacht, F.E.: Convolutional neural networks enable efficient, accurate and fine-grained segmentation of plant species and communities from high-resolution uav imagery. Scientific Reports **9**(1) (Nov 2019)
32. Kattenborn, T., Leitloff, J., Schiefer, F., Hinz, S.: Review on convolutional neural networks (CNN) in vegetation remote sensing. ISPRS Journal of Photogrammetry and Remote Sensing **173**, 24–49 (Mar 2021)
33. Khelifi, L., Mignotte, M.: Deep learning for change detection in remote sensing images: Comprehensive review and meta-analysis. IEEE Access **8**, 126385–126400 (2020)
34. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2017)
35. Lenoir, J., Gégout, J.C., Marquet, P.A., de Ruffray, P., Brisse, H.: A significant upward shift in plant species optimum elevation during the 20th century. Science **320**(5884), 1768–1771 (Jun 2008)
36. Li, L., Zhou, T., Wang, W., Li, J., Yang, Y.: Deep hierarchical semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1246–1257 (June 2022)
37. Li, S., Brandt, M., Fensholt, R., Kariryaa, A., Igel, C., Gieseke, F., Nord-Larsen, T., Oehmcke, S., Carlsen, A.H., Junttila, S., Tong, X., d'Aspremont, A., Ciais, P.: Deep learning enables image-based tree counting, crown segmentation, and height prediction at national scale. PNAS Nexus **2**(4) (Mar 2023)
38. Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.W., Heng, P.A.: H-denseunet: Hybrid densely connected UNet for liver and tumor segmentation from ct volumes. IEEE Transactions on Medical Imaging **37**(12), 2663–2674 (Dec 2018)
39. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context, p. 740–755. Springer International Publishing (2014)
40. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. CoRR **abs/2103.14030** (2021)
41. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3431–3440. IEEE Computer Society, Los Alamitos, CA, USA (jun 2015)
42. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization (2019)
43. Lucas, B., Pelletier, C., Schmidt, D., Webb, G.I., Petitjean, F.: A bayesian-inspired, deep learning-based, semi-supervised domain adaptation technique for land cover mapping. Machine Learning **112**(6), 1941–1973 (Mar 2021)
44. Maretto, R.V., Fonseca, L.M.G., Jacobs, N., Körting, T.S., Bendini, H.N., Parente, L.L.: Spatio-temporal deep learning approach to map deforestation in amazon rainforest. IEEE Geoscience and Remote Sensing Letters **18**(5), 771–775 (2021)
45. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). pp. 565–571 (2016)
46. Muller, B.R., Smith, W.: A hierarchical loss for semantic segmentation. In: VISIGRAPP (2020)

47. Nguyen, T.A., Rußwurm, M., Lenczner, G., Tuia, D.: Multi-temporal forest monitoring in the swiss alps with knowledge-guided deep learning. Remote Sensing of Environment **305**, 114109 (May 2024)
48. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M.C.H., Heinrich, M.P., Misawa, K., Mori, K., McDonagh, S.G., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D.: Attention U-Net: Learning where to look for the pancreas. CoRR **abs/1804.03999** (2018)
49. Onishi, M., Ise, T.: Explainable identification and mapping of trees using UAV RGB image and deep learning. Scientific Reports **11**(1) (Jan 2021)
50. Ortega Adarme, M., Queiroz Feitosa, R., Nigri Happ, P., Aparecido De Almeida, C., Rodrigues Gomes, A.: Evaluation of deep learning techniques for deforestation detection in the brazilian amazon and cerrado biomes from remote sensing imagery. Remote Sensing **12**(6), 910 (Mar 2020)
51. Reed, C.J., Gupta, R., Li, S., Brockman, S., Funk, C., Clipp, B., Keutzer, K., Candido, S., Uyttendaele, M., Darrell, T.: Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4065–4076 (2023)
52. Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat: Deep learning and process understanding for data-driven earth system science. Nature **566**(7743), 195–204 (Feb 2019)
53. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, p. 234–241. Springer International Publishing (2015)
54. Rußwurm, M., Courty, N., Emonet, R., Lefèvre, S., Tuia, D., Tavenard, R.: 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 (Feb 2023)
55. Rußwurm, M., Körner, M.: Multi-temporal land cover classification with sequential recurrent encoders. ISPRS International Journal of Geo-Information **7**(4), 129 (Mar 2018)
56. Sainte Fare Garnot, V., Landrieu, L., Giordano, S., Chehata, N.: 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). pp. 12322–12331 (2020)
57. Schiefer, F., Kattenborn, T., Frick, A., Frey, J., Schall, P., Koch, B., Schmidtlein, S.: 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 (Dec 2020)
58. Sharma, A., Tuzel, O., Jacobs, D.W.: Deep hierarchical parsing for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015)
59. Song, G., Wu, S., Lee, C.K., Serbin, S.P., Wolfe, B.T., Ng, M.K., Ely, K.S., Bogonovich, M., Wang, J., Lin, Z., Saleska, S., Nelson, B.W., Rogers, A., Wu, J.: 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 (Jan 2022)
60. Tarasiou, M., Chavez, E., Zafeiriou, S.: ViTs for SITS: Vision transformers for satellite image time series (2023)
61. Tseng, G., Cartuyvels, R., Zvonkov, I., Purohit, M., Rolnick, D., Kerner, H.: Lightweight, pre-trained transformers for remote sensing timeseries (2024)
62. Turkoglu, M.O., D'Aronco, S., Perich, G., Liebisch, F., Streit, C., Schindler, K., Wegner, J.D.: Crop mapping from image time series: Deep learning with multi-scale label hierarchies. Remote Sensing of Environment **264**, 112603 (Oct 2021)

63. Vali, A., Comai, S., Matteucci, M.: Deep learning for land use and land cover classification based on hyperspectral and multispectral earth observation data: A review. Remote Sensing **12**(15), 2495 (Aug 2020)
64. Valmadre, J.: Hierarchical classification at multiple operating points. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. NIPS '22, Curran Associates Inc., Red Hook, NY, USA (2024)
65. Visser, M.E., Gienapp, P.: Evolutionary and demographic consequences of phenological mismatches. Nature Ecology & Evolution **3**(6), 879–885 (Apr 2019)
66. Wang, J., Song, G., Liddell, M., Morellato, P., Lee, C.K., Yang, D., Alberton, B., Detto, M., Ma, X., Zhao, Y., Yeung, H.C., Zhang, H., Ng, M., Nelson, B.W., Huete, A., Wu, J.: An ecologically-constrained deep learning model for tropical leaf phenology monitoring using planetscope satellites. Remote Sensing of Environment **286**, 113429 (Mar 2023)
67. Weinstein, B.G., Marconi, S., Aubry-Kientz, M., Vincent, G., Senyondo, H., White, E.P.: Deepforest: A python package for rgb deep learning tree crown delineation. Methods in Ecology and Evolution **11**(12), 1743–1751 (Oct 2020)
68. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 12077–12090. Curran Associates, Inc. (2021)
69. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. In: Bengio, Y., LeCun, Y. (eds.) 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings (2016)
70. Zhang, L., Shao, Z., Liu, J., Cheng, Q.: Deep learning based retrieval of forest aboveground biomass from combined LiDAR and landsat 8 data. Remote Sensing **11**(12), 1459 (Jun 2019)
71. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6230–6239. IEEE Computer Society, Los Alamitos, CA, USA (jul 2017)
72. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.S., Zhang, L.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6877–6886. IEEE Computer Society, Los Alamitos, CA, USA (jun 2021)
73. Zhou, X., Xin, Q., Dai, Y., Li, W.: A deep-learning-based experiment for benchmarking the performance of global terrestrial vegetation phenology models. Global Ecology and Biogeography **30**(11), 2178–2199 (Aug 2021)
74. Zwemer, M.H., Wijnhoven, R.G.J., de With, P.H.N.: Hierarchical Object Detection and Classification Using SSD Multi-Loss, p. 268–296. Springer International Publishing (2022)