TreeFinder: A US-Scale Benchmark Dataset for Individual Tree Mortality Monitoring Using High-Resolution Aerial Imagery

Zhihao Wang[1], **Cooper Li**[1], **Ruichen Wang**[1], **Lei Ma**[1]
George Hurtt[1], **Xiaowei Jia**[2], **Gengchen Mai**[3], **Zhili Li**[1], **Yiqun Xie**[1]*
[1]University of Maryland, [2]University of Pittsburgh, [3]University of Texas at Austin
`{zhwang1, cligeog25, ruichenw, lma6, gchurtt, lizhili, xie}@umd.edu,`
`xiaowei@pitt.edu, gengchen.mai@austin.utexas.edu`

Abstract

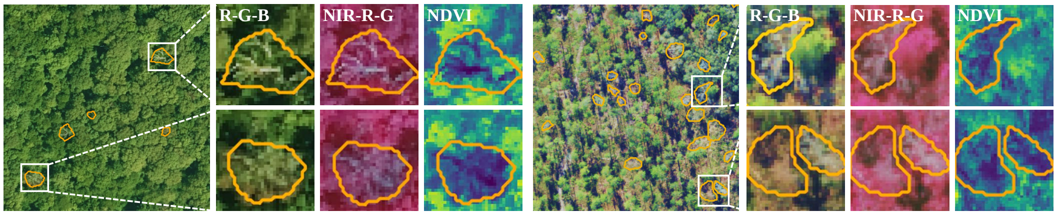
Monitoring individual tree mortality at scale has been found to be crucial for understanding forest loss, ecosystem resilience, carbon fluxes, and climate-induced impacts. However, the fine-granularity monitoring faces major challenges on both the data and methodology sides because: (1) finding isolated individual-level tree deaths requires high-resolution remote sensing images with broad coverage, and (2) compared to regular geo-objects (e.g., buildings), dead trees often exhibit weaker contrast and high variability across tree types, landscapes and ecosystems. Existing datasets on tree mortality primarily rely on moderate-resolution satellite imagery (e.g., 30m resolution), which aims to detect large-patch wipeouts but is unable to recognize individual-level tree mortality events. Several efforts have explored alternatives via very-high-resolution drone imagery. However, drone images are highly expensive and can only be collected at local scales, which are therefore not suitable for national-scale applications and beyond. To bridge the gaps, we introduce TreeFinder, the first high-resolution remote sensing benchmark dataset designed for individual-level tree mortality mapping across the Contiguous United States (CONUS). Specifically, the dataset uses NAIP imagery at 0.6m resolution that provides wall-to-wall coverage of the entire CONUS. TreeFinder contains images with pixel-level labels generated via extensive manual annotation that covers forested areas in 48 states with over 23,000 hectares. All annotations are rigorously validated using multi-temporal NAIP images and auxiliary vegetation indices from remote sensing imagery. Moreover, TreeFinder includes multiple evaluation scenarios to test the models' ability in generalizing across different geographic regions, climate zones, and forests with different plant function types. Finally, we develop benchmarks using a suite of semantic segmentation models, including both convolutional architectures and more recent foundation models based on vision transformers for general and remote sensing images. Our dataset and code are publicly available on Kaggle and GitHub: `https://www.kaggle.com/datasets/zhihaow/tree-finder` and `https://github.com/zhwang0/treefinder`.

1 Introduction

Forests play a critical role in the ecological balance of the Earth, significantly influencing global carbon cycles [20, 31], biodiversity conservation [8, 7], climate regulation [35, 29], and water

[*] Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.



(a) An example of uniform forest in Illinois.

(b) An example of complex forest in Florida.

Figure 1: Examples of scattered tree mortality visualized with three spectral representations: (1) true-color (R-G-B), (2) false-color (NIR-R-G), commonly used in remote sensing to highlight vegetation in red, and (3) vegetation index (NDVI), where brighter colors indicate higher vegetation activities.

resources [9, 18]. The health and stability of forests are increasingly threatened by widespread tree mortality, which can substantially alter carbon storage, disrupt local ecosystems, and increase wildfire risks [47, 14, 32]. While tree wipe-outs as contiguous large patches (e.g. due to wildfire) can be monitored using traditional moderate-resolution remote sensing platforms such as Landsat-8/9 at 30m resolution, scattered (i.e., not contiguous) but widespread tree deaths at the individual level have been largely unmonitored. However, according to recent studies including a *Nature Communications* article [13], such scattered tree deaths have a substantial impact on forest loss and significantly affect carbon budget and sequestration capacity, and serve as critical catalysts for future wildfires. Therefore, accurate mapping and quantification of tree mortality in fine granularity are essential for better ecological monitoring, precise carbon accounting, high-resolution fire risk assessment, and effective forest management policies.

Despite its importance, identifying tree mortality at the individual level remains challenging due to limitations in both data availability and methodological approaches. First, visual signatures of scattered tree deaths are only available in high-resolution images, making traditional remote sensing platforms unsuitable for this detection task. In addition, the high-resolution imagery must have broad-scale geographical coverage (e.g., national level) to answer major carbon cycle questions and inform critical policy and management decisions. Second, compared to traditional geospatial objects or phenomena with sharp contrast and geometric patterns (e.g., buildings), dead trees often present weaker contrast or higher similarity with the surrounding context that makes their pixels harder to separate from the background environment, which may contain varying sunlight, shadows, or landscapes. More importantly, the visual patterns and background contrasts of dead trees can vary significantly across geographic regions due to different climate conditions, forest density, and tree types [3, 33]. This makes it challenging to generalize learned AI models at scale in real-world applications. Finally, the lack of labels that are widely distributed over geographic regions remains a key bottleneck for developing generalizable models for large-scale monitoring.

While efforts have attempted to solve the tree mortality mapping problem, existing datasets are limited in their applicability for monitoring individual-level tree deaths across broad geographical scales. Most current products and monitoring systems are based on low or moderate resolution satellite platforms, such as the 1km-resolution AVHRR imagery [19] or 30m-resolution Landsat imagery [30, 22], which lack sufficient spatial details needed to map fine-granularity tree mortality, including scattered tree deaths. Recent studies have also explored the use of drone imagery at very high resolution (VHR), which is capable of identifying individual tree deaths in localized study sites [43, 27]. However, drone-based monitoring is highly expensive, which significantly constrains its applicability at large scale. Note that even though the labels can be collected at different geographic sites, in practice, it is cost-prohibitive to generate wall-to-wall maps (i.e., spatially contiguous full maps) using these VHR images for applications beyond small local scales. These wall-to-wall maps, though, are often required for scientific research and forest management due to heterogeneity [25].

To address these gaps, we introduce **TreeFinder**, the first high-resolution benchmark dataset designed for individual-level tree mortality mapping across the Contiguous United States (CONUS). Specifically, the dataset uses NAIP imagery at 0.6m-resolution that provides wall-to-wall coverage for the entire CONUS. TreeFinder contains images with pixel-level labels generated via extensive manual annotation that covers forested regions in 48 different states in CONUS with a total area of over 23,000 hectares. The high spatial resolution, combined with broad geographic coverage, offers opportunities to enable accurate identification and delineation of individual tree deaths. Our dead tree annotations are rigorously validated using multi-temporal data from NAIP imagery, ensuring the accuracy and reliability of labeled dead trees. In addition to the dataset, we further implement a suite of machine learning (ML) methods covering both segmentation models based on traditional convolutional neural networks (CNNs) and more recent foundation models to establish performance benchmarks for individual-level tree mortality monitoring. Finally, to facilitate the evaluation of ML models' generalizability under different scenarios, we associate additional metadata with each image patch to provide information about its geographic location, climate zone, and primary tree type. Considering the large degree of spatial variability, these scenarios are necessary to understand if an ML model is able to maintain robust performance in different conditions, especially those not seen during training. Overall, our TreeFinder dataset and benchmarking initiative not only address major gaps in existing datasets, but also offer opportunities to advance machine learning methods for challenging ecological and environmental science problems at a large scale with cross-region variability. Our open-source dataset and code are available on Kaggle `https://www.kaggle.com/datasets/zhihaow/tree-finder` and GitHub `https://github.com/zhwang0/treefinder`. Our key contributions are summarized as follows:

- We create a large-scale, high-resolution dataset covering 1,000 sites over 48 states in CONUS, with a total area of 23,000 hectares. The 0.6m high-resolution NAIP images at each site are manually annotated for dead trees at the pixel level using both visual features from single NAIP images and temporal differences between multi-temporal NAIP images.
- We develop performance benchmarks using a suite of ML models, including traditional CNN-based segmentation methods and more recent foundation models for general and remote sensing images.
- We integrate metadata on geographic locations, climate zones, and primary tree types to each image patch to enable performance evaluation and model comparison under different scenarios and the spatial variability challenge.

2 Related Work

Existing benchmark datasets on remote sensing semantic segmentation. Semantic segmentation has gained increasing attention in the remote sensing domain, as large-scale, pixel-level classification from satellite or aerial imagery provides important and detailed information for the monitoring of diverse Earth surface conditions such as land cover types [37, 44], urban infrastructure [16, 24], and crop growth [45, 5]. Several datasets have been developed for this purpose, including DeepGlobe for land cover segmentation [15], CropHarvest for global crop type mapping using both optical and SAR satellite imagery [45], LoveDA for domain-adaptive segmentation across urban and rural scenes [50], and SAMRS leveraging SAM and existing datasets [48]. These datasets typically focus on well-structured geospatial objects such as buildings, roads, and crop fields, which exhibit strong spatial regularity and clear boundaries. Segmentation of geospatial objects with lower contrast (e.g., individual-level tree deaths) on a large national scale using high-resolution images has been underexplored in existing datasets, as well as generalization across different climate zones and ecological conditions. Our experiments in Sec. 4.2 show that such tasks indeed remain challenging for current segmentation models.

Existing datasets on tree mortality monitoring. Existing datasets remain limited along several critical dimensions, including spatial resolution, geographic coverage, and label availability. While **drone-based datasets** offer very high spatial resolution (e.g., Almorox Crown Dataset [2], FOR-instance [36]), their spatial coverage is highly constrained due to high operational cost, often restricted to localized study areas (e.g., tens or a few hundred hectares). A recent work, *deadtrees.earth* [34], is an encouraging platform effort aiming to support collaborations for tree mortality label collection. However, the dataset relies on drone imagery, which offers centimeter-level resolution but is limited to sample sites due to high cost, constraining its practical applicability for large-scale monitoring tasks. Moreover, according to the paper, its images are biased toward forests located near human settlements as they were originally collected for other purposes. As a result, the data may not be representative of forest ecosystems. On the other hand, coarse-to-moderate resolution imagery from **satellite platforms** makes large-scale coverage possible [40, 41, 38], but the resolution only supports detecting dead trees that form large and contiguous patches and the images lack necessary spatial details to capture fine-granularity tree mortality patterns. Finally, aerial images offer new opportunities to consider both the geographic coverage and resolution [1, 28]. For example, the National Agriculture Imagery

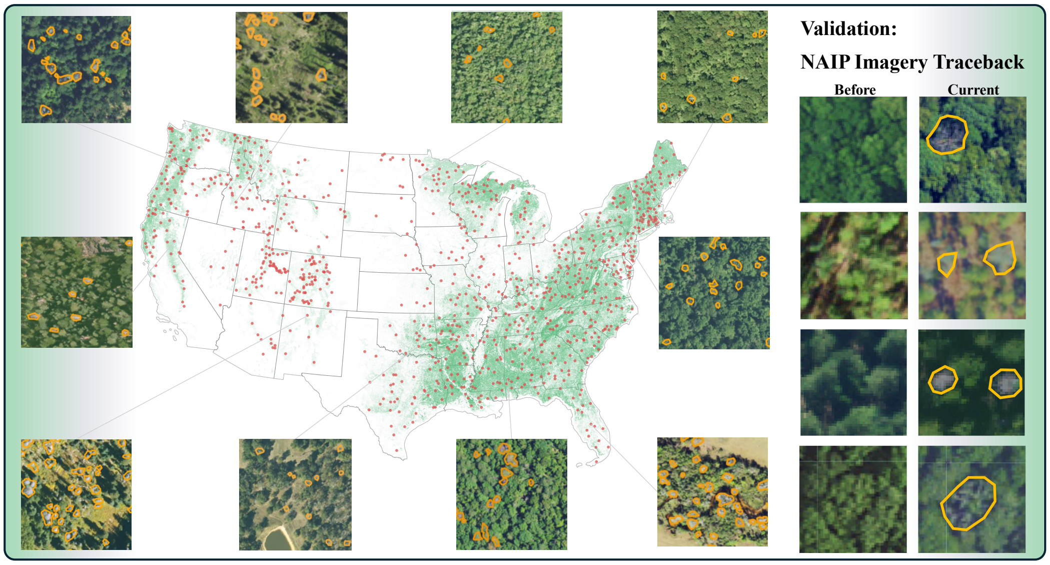


Figure 2: Left: Distribution of the 1000 sites from the Contiguous US and example visualizations of labeled tree deaths. Right: Illustrations of the validation process using multi-temporal images.

Program (NAIP) provides 0.6m resolution images over the entire CONUS region. However, existing datasets have only considered local areas (e.g., the Sierra National Forest in California). In addition, the samples are predominantly generated by ML models, and the manual labels are very limited and not publicly shared. Finally, from the ML model evaluation and benchmarking perspective, existing datasets lack ecological, climate, and geographic diversity, limiting their generalizability across forest types, climatic conditions and locations. As a result, they are not practically suitable for developing models for large-scale monitoring, e.g., national scales.

ML methods for satellite-based segmentation. Deep learning has driven major advances in semantic segmentation. CNN architectures such as U-Net [39], DeepLabV3+ [10], and HRNet [49] are widely adopted in satellite-based segmentation tasks thanks to their stable performance. Recent advances in vision transformers (ViTs) have enabled more flexible and scalable modeling of long-range dependencies. Models such as SegFormer [52] introduced hierarchical transformer encoders with efficient multi-scale fusion, while Mask2Former [11] and Segmenter [42] extend transformers to class-agnostic and class-aware segmentation frameworks. In the geospatial domain, foundation models like NASA-IBM Prithvi [26], SpectralGPT [23] and DOFA [53] are pretrained on large-scale satellite datasets and have demonstrated competitive performance. However, many of them are pretrained using specific types of satellite images and the characteristics may not generalize to other platforms. For example, Prithvi and SpectralGPT were pretrained using moderate-resolution multispectral imagery (e.g., Sentinel-2, or Harmonized Landsat and Sentinel-2), where the special designs along the spectral dimensions may not be suitable for platforms with high-resolution and only a few bands (e.g., NAIP). With that said, these developments provide a diverse set of models for our candidate model selection.

3 TreeFinder: Dataset Construction

3.1 Data Collection, Annotation, and Validation

TreeFinder aims to provide a CONUS-scale, publicly available, and ML-ready dataset using NAIP imagery to support the development of ML-based segmentation methods that have strong generalizability over geographic regions and beneficial for fine-granularity geo-event monitoring at large scale. Specifically, NAIP provides 4-band aerial images including RGB and near-infrared (NIR) channels, with a contiguous coverage over the entire CONUS area. As NAIP has undergone continued enhancements over time, it has historically produced imagery at varying spatial resolutions, including 2 m, 1 m, 0.6 m, and 0.3 m. The varying resolutions and different acquisition conditions (e.g., dates, viewing angles, sunlight angles) make it challenging to construct a consistent multi-temporal dataset. Thus, we only use the most recent high-resolution imagery collected after 2021, ensuring sufficient spatial granularity required for high-quality, individual-level tree mortality mapping. A small subset of post-2021 imagery is available at 0.3 m in a few regions, and we resampled these images to 0.6 m to maintain a consistent resolution across all samples.

As the scope of TreeFinder focuses on forested areas, where individual-level tree mortality has been shown to have substantial impact on forest health, carbon cycles and wildfire risks, we first use a well-established, national-scale forest cover basemap [21] to define the geographic mask for the sampling. From this mask, we randomly sample sites across the CONUS to generate manually annotated labels for the dead trees. Specifically, we delineate the individual dead trees using polygon-based tools through Google Earth Engine, a cloud-based platform hosting the full NAIP imagery archive. One challenge during the labeling process is that the individual dead trees could have similarity to bare ground when there is weak visual contrast, especially in cases of isolated, stand-alone mortality or when shadows from neighboring trees partially obscure the crown. To mitigate this, we utilize several strategies to confirm the class belongings as shown in Fig. 1: (1) We use multiple spectral representations including the true-color composition using the default visible R-G-B channels, the false-color composition with NIR-R-G channels that are commonly used in remote sensing to highlight vegetation distributions, and indices such as the Normalized Difference Vegetation Index (NDVI) to better observe vegetation activities. In false-color composition, vegetation is shown as red, and in NDVI maps, pixels with brighter colors indicate higher vegetation activities. A tree is delineated and labeled as dead if its crown is structurally distinct from neighboring trees and exhibits complete canopy de-saturation (e.g., a consistent gray or brown color with no visible greenness) in different multi-spectral visualizations. Moreover, we also validate the labels using multi-temporal combinations of historical NAIP images. While the images can have a lower spatial resolution (1m) in earlier years, they can still provide some visual cues (e.g. color shifts or structural changes) to inform the annotations. Finally, the annotations undergo a cross-annotator assessment on a randomly selected 20% subset of the sites, achieving an agreement rate of over 95%. More details on annotations and validations are available in the appendix. In total, we annotate NAIP images at 1,000 unique sites across CONUS as shown in Fig. 2.

To facilitate the evaluation of ML models' generalizability under diverse conditions (detailed in Sec. 3.3), we enrich each labeled NAIP image with metadata on **geographic locations**, **climate conditions** and **primary tree types**. Specifically, we include information about latitude, longitude, and state for each image to indicate its location. We also assign a Köppen–Geiger climate classification label to each image using the latest gridded global product, which captures present climatic regimes based on temperature and precipitation seasonality [4]. For tree type information, we overlay each NAIP image with the Individual Tree Species Parameter Maps from the U.S. Department of Agriculture's Forest Service [46], which provide estimates of different tree type composition across forested areas in the U.S. Each image is assigned its primary tree type based on the most frequent tree type. The primary type is often used to reduce uncertainty in tree species mapping. Here we did not include the detailed proportions of tree types for the same reason, and also because we are only using it for later generalization tests across primary tree types.

3.2 ML-Ready Dataset Preparation

In this paper, the ML-ready dataset means that the data is preprocessed into standard input and output formats for convenient use of ML model training and evaluation. Specifically, the raw NAIP scenes and annotated polygons are converted to fixed-size, model-compatible image patches with consistent spatial dimensions. Specifically, each annotated image is split into non-overlapping patches of 224×224 pixels. Based on our visual inspection, a patch size of 224×224 is sufficiently large to capture full details that are needed to identify individual-level dead trees. This leads to a total of $N = 15,489$ image patches, and the input images form a tensor $\mathbf{X} \in \mathbb{R}^{N \times 224 \times 224 \times 4}$ with 4 bands in each image, and the output labels form a tensor $\mathbf{Y} \in \{0,1\}^{N \times 224 \times 224}$, where 1 indicates a pixel of a dead tree. Finally, since aerial images do not always align with the orthogonal directions of the geographic reference systems and their shapes may change after projection and ortho-rectification, there are often areas with empty values in an image patch. Thus, we also provide a binary masking layer $\mathbf{M} \in \{0,1\}^{N \times 224 \times 224}$ to help exclude the null pixels during evaluation. We provide the dataset in multiple formats for convenience. First, we provide the original GeoTIFF format, which preserves all the spatial referencing information (e.g., geographic coordinate system, projection) of each image for visualization or further integration with other spatial data. Second, we develop and share a Python library to load, filter, and batch the dataset based on user-defined criteria. This can convert the data into other more direct formats for ML models: (1) Numpy array format, where the tensors are stored as .npy files; and (2) TFRecords format, which supports easier usage with ML models.

3.3 Scenarios for Generalization Test

To more comprehensively assess the generalizability of ML models under diverse geographical, climatic, and ecological conditions, we define four benchmarking scenarios: (1) baseline (easiest) scenario with random sampling, (2) generalization over geographic regions, (3) generalization over climate zones, and (4) generalization over different types of forests. There are certain connections between generalization over geographic regions and generalization over climates and forest types, as different climates or forest types are also in different regions. We include geographic generalization as a separate scenario as it can be considered as an integration of many factors (e.g., different regional management practices on forest recovery efforts), and it is common in practice to have labels highly localized in certain regions. Each scenario represents a practical deployment situation, and these are important to understand for applications, as the labeled set often only covers a small fraction of the entire study area (e.g., CONUS) in large-scale monitoring tasks.

- **Random split with incremental training sizes.** In this baseline scenario, we evaluate the overall model performance using a standard random split, where 20% of the labeled patches are held out as a fixed test set. The remaining 80% of the dataset is used for training and validation, in which we denote the training data ratio as α. To examine the influence of data size on model performance, we subsample the training set as incremental proportions by varying α from 10% to 80%, while keeping the test set fixed. The validation set is randomly sampled as 10% of the training set. While the baseline scenario is the easiest among the four scenarios, the additional evaluations with different training data proportions reflect real-world settings where labeled data may be limited and help understand the model's sensitivity to sample size.
- **Cross-region scenarios.** To test the spatial generalizability of different ML models on our benchmark dataset, we consider the following splits: (1) Western-eastern split: The dataset is divided into western and eastern regions of the CONUS, using the Mississippi River as a natural boundary. Models are trained in one region and evaluated on the other. As explained earlier, the variability across locations can be considered as an aggregation of factors including climates, forest types, and others. (2) One state vs. all: This is a challenging scenario where the training samples come from one single state and the evaluation is performed on all remaining states. To set up a concrete example, we use Colorado as the single state for training, as the state is well-known for its forested mountains, and the tree mortality problems have been widely observed and studied in the area [51, 6]. This setup also reflects practical scenarios where certain states start the monitoring programs earlier on tree mortality events and thus contribute disproportionately to the training data than others.
- **Cross-climate scenarios.** These scenarios evaluate model generalization across climate regimes by training on data from one set of climate conditions and then test it on the others. First, we build one group of climate zones that include Mediterranean, humid subtropical zone, humid continental zone, etc., comprising approximately 50% of our dataset, and the remaining, such as the humid subtropical climate zone, as the other group. Second, we consider a more challenging scenario, where the training is performed on a climate zone with significantly smaller number of samples and then tested on the rest of the zones. Specifically, we use the humid continental zone as the training climate zone and the rest for testing.
- **Cross-forest-type scenarios.** Finally, we design scenarios to test model generalization across different primary forest types. First, we construct a relatively easier case with a broad training set consisting of the top five most frequent primary tree types–maple, pine, oak, Douglas fir, and cottonwood–which together account for approximately 50% of all labeled samples. Models are trained on this subset and evaluated on all other primary forest types. Second, we define a more challenging generalization scenario, in which the model is trained using only one primary tree type, maple, and evaluated on all others.

Metrics. Performance is evaluated using standard segmentation metrics, including precision, recall, F1 score, intersection-over-union (IoU), and overall accuracy. The segmentation statistics across all image patches are first aggregated together and then used to compute the final metrics, rather than averaging per-patch values. Except for accuracy, the other metrics need to be calculated per class to better reflect the model's performance. Thus, we include both the metrics for the target class (dead trees) and for both (average of the target class and background class) in the result tables.

4 Experiments

4.1 Candidate methods

To benchmark model performance, we consider a set of segmentation architectures covering both CNN-based and transformer-based designs. This selection captures convolutional methods focusing on localized feature extraction as well as more recent transformer-based foundation models pretrained using both general images and remote sensing images. Specifically, we consider the following candidate models:

- **U-Net**: An encoder–decoder network with skip connections. Our U-Net is trained from scratch to provide a baseline with localized spatial modeling and no reliance on pretraining [39].
- **DeepLabV3+**: A CNN-based model with a ResNet-50 backbone, leveraging atrous spatial pyramid pooling to aggregate multi-scale contextual features. The model was pretrained on ImageNet [10] and we customized it with input and output modifications.
- **Vision Transformer (ViT)**: A patch-based transformer with a lightweight transposed convolution decoder that upsamples hidden features back to full resolution [17]. The model is pretrained on ImageNet. Following common strategies, we added a segmentation head to customize it for semantic segmentation [52].
- **SegFormer**: A hierarchical transformer architecture that is designed for semantic segmentation [52]. It uses multi-scale feature encoding with a lightweight decoder, enabling better spatial representation and hierarchical feature extraction. SegFormer was pretrained on pretrained on ADE20k.
- **Mask2Former**: A transformer-based framework that combines a Swin-Tiny backbone with multi-scale deformable attention and a class-agnostic mask prediction head [12]. It models segmentation as a set prediction task using masked attention, and we used the pretrained weights on ADE20K.
- **DOFA**: A multimodal foundation model specifically designed for remote sensing images [53]. It uses wavelengths to embed different spectral bands into a unified feature space, enabling the learning of shared representations across channels. DOFA is pretrained on multi-sensor remote sensing imagery, including Sentinel 1/2 and NAIP. As it is not limited to specific remote sensing sensors and considers NAIP in pretraining, we included it as part of the evaluation. We did not include Prithvi and SpectralGPT as they are specifically designed for multispectral images (e.g., 10+ bands) and pretrained using moderate-resolution remote sensing images, which are largely distinct from NAIP.

All models are trained using the training set of TreeFinder (varying by evaluation scenarios), or fine-tuned if pretrained weights are available. We use a batch size of 32, an initial learning rate of e^{-4} with the Adam optimizer. All models are trained for up to 100 epochs using a combined loss function from binary cross-entropy loss and dice loss to mitigate class imbalance issues. We also applied early stopping based on validation loss to prevent overfitting. More details on training are available in the appendix.

4.2 Results

Random split performance and impact of training size. Table 1 shows model performance under a standard 80-20 random train-test split, using 10% of the training samples for validation. Among all models, it is interesting to see that Mask2Former achieves the highest F1, precision, and IoU, while U-Net has the highest recall. The differences between U-Net, DeepLabV3+, and SegFormer are within 2-3% in this baseline scenario. DOFA did not perform well on the metrics compared to the others, potentially due to the trade-off between its goal to cover broader sensing platforms with different sets of spectral bands, and the performance on specific types of platforms. The accuracy for all models remains very high because the problem has imbalanced class distribution where dead trees

Table 1: Results for the random split scenario (numbers shown as %), with standard deviations across three runs. Best results are bolded.

Model	**F1**		**Precision**		**Recall**		**IoU**		**Accuracy**
	Dead Tree	All	Dead Tree	All	Dead Tree	All	Dead Tree	All	All
U-Net	46.1 ± 7.1	72.9 ± 3.5	35.7 ± 6.8	67.8 ± 3.4	**65.6 ± 5.2**	**82.6 ± 2.6**	30.1 ± 5.9	64.8 ± 2.9	99.4 ± 0.0
DeepLabV3+	49.7 ± 8.7	74.7 ± 4.4	49.6 ± 7.1	74.7 ± 3.5	49.9 ± 10.4	74.8 ± 5.2	33.4 ± 7.5	66.5 ± 3.8	**99.6 ± 0.0**
ViT	43.6 ± 9.0	71.7 ± 4.5	45.1 ± 8.2	72.4 ± 4.1	42.9 ± 11.9	71.3 ± 5.9	28.2 ± 7.4	63.9 ± 3.7	**99.6 ± 0.0**
SegFormer	47.8 ± 8.3	73.8 ± 4.2	49.2 ± 5.4	74.5 ± 2.7	47.0 ± 11.7	73.4 ± 5.8	31.7 ± 7.4	65.7 ± 3.7	99.6 ± 0.1
Mask2Former	**51.9 ± 5.5**	**75.8 ± 2.8**	**55.5 ± 8.7**	**77.6 ± 4.3**	49.4 ± 6.7	74.6 ± 3.3	**35.1 ± 5.1**	**67.4 ± 2.6**	**99.6 ± 0.0**
DOFA	29.2 ± 5.6	64.5 ± 2.8	31.1 ± 2.9	65.4 ± 1.4	28.6 ± 9.1	64.2 ± 4.5	17.2 ± 3.9	58.3 ± 1.9	99.5 ± 0.1

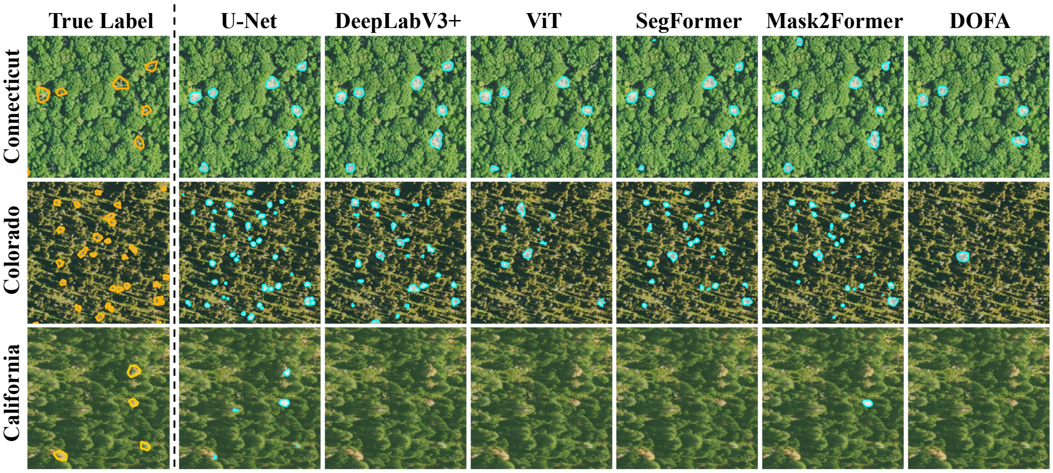


Figure 4: Visualization of example segmentation results from models trained with 80% of the dataset under the random split scenario.

account for a small proportion of the total number of trees. However, their impact on forest health and carbon stock potential is substantial [13]. For example, if a tree dies, not only it will no longer contribute to continued carbon sequestration, but also the existing carbon stock will be taken away, turning to emissions. In Fig. 3, we show the F1 score results by incrementally increasing the training set size from 10% to 80%, while keeping the test set fixed. Results with unstable or poor performance at very small sample size are not included. In general, all models show performance improvements with more training data. DOFA follows the same trend but its overall performance remains lower than the other models. Fig. 4 visualizes several examples of results using models with 80% training data.

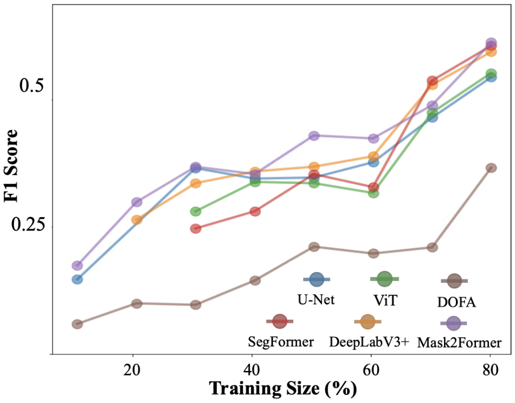


Figure 3: F1 score vs. training data size.

Cross-region generalizability. Table 2 presents the results on model generalizability across different spatial regions. Compared to the model performance in the random split setting, the overall performance in this test setting drops significantly. For example, the F1 score has up to 50% relative decrease in the W-E scenario and 70% in the single-state scenario (Colorado - Others) , which confirms the increased difficulty of generalizing across geographic domains. Among these three scenarios, SegFormer and Mask2Former take most of the top-ranking positions for F1 score, precision, and IoU. U-Net shows the best performance in recall. It is worth noting that in the single-state scenario, most models such as SegFormer and Mask2Former have higher precision than recall, whereas the pattern is the opposite for U-Net, showing different model tendencies. DOFA and ViT in this case still show relatively lower performance. Comparing the scenarios, W-E shows slightly better results

Table 2: Results for cross-region scenarios: **E - W**: train on eastern and test on western states; **W - E**: train on western and test on eastern states; **CO**: train on Colorado and test on all other states. All values are shown as %, and best values are bolded.

Model	**Scenario**	**F1**		**Precision**		**Recall**		**IoU**		**Accuracy**
		Dead Tree	All	Dead Tree	All	Dead Tree	All	Dead Tree	All	
U-Net	E - W	16.1	57.9	11.1	55.5	**29.4**	**64.4**	8.8	54.0	99.3
DeepLabV3+	E - W	20.5	60.2	20.8	60.3	20.3	60.0	11.4	55.5	99.6
ViT	E - W	3.8	51.4	2.1	51.0	16.9	57.5	1.9	49.9	97.9
SegFormer	E - W	**21.8**	**60.8**	18.1	59.0	27.3	63.5	**12.2**	**55.9**	99.5
Mask2Former	E - W	21.6	60.7	**23.8**	**61.8**	19.8	59.8	12.1	**55.9**	**99.7**
DOFA	E - W	4.8	52.0	2.8	51.3	18.4	58.4	2.4	50.3	98.2
U-Net	W - E	13.9	56.8	40.8	70.2	8.4	54.2	7.5	53.5	**99.5**
DeepLabV3+	W - E	15.0	57.4	**41.2**	**70.4**	9.1	54.5	8.1	53.8	**99.5**
ViT	W - E	13.4	56.6	24.9	62.2	9.2	54.5	7.2	53.3	**99.5**
SegFormer	W - E	**22.8**	**61.3**	38.7	69.2	**16.1**	**58.0**	**12.8**	**56.2**	**99.5**
Mask2Former	W - E	21.4	60.6	38.5	69.1	14.8	57.4	12.0	55.8	**99.5**
DOFA	W - E	4.4	52.0	7.0	53.3	3.2	51.5	2.2	50.8	99.4
U-Net	CO	10.2	54.4	5.8	52.8	**40.1**	**68.8**	5.4	51.3	97.3
DeepLabV3+	CO	15.4	57.6	20.2	59.9	12.5	56.2	8.4	53.9	99.5
ViT	CO	6.3	53.0	13.1	56.4	4.1	52.0	3.2	51.4	99.5
SegFormer	CO	**17.5**	**58.7**	27.8	63.7	12.8	56.3	**9.6**	**54.6**	99.5
Mask2Former	CO	11.6	55.7	**29.7**	**64.7**	7.2	53.6	6.1	52.9	**99.6**
DOFA	CO	10.2	54.9	10.2	55.0	10.2	54.9	5.4	52.4	99.3

Table 3: Performance across shifted domains in climate zones and primary tree types. All values are percentages %, and best values are bolded.

Model	**Scenario**	**F1**		**Precision**		**Recall**		**IoU**		**Accuracy**
		Dead Tree	All	Dead Tree	All	Dead Tree	All	Dead Tree	All	
U-Net	Climate	18.6	59.1	17.3	58.5	20.1	59.8	10.3	54.8	99.3
DeepLabV3+	Climate	25.8	62.8	31.4	65.5	21.9	60.9	14.8	57.2	**99.5**
ViT	Climate	21.6	60.7	23.4	61.5	20.1	59.9	12.1	55.8	99.4
SegFormer	Climate	28.0	63.9	**33.7**	**66.7**	24.0	61.9	16.3	57.9	**99.5**
Mask2Former	Climate	**29.8**	**64.7**	25.0	62.4	**36.8**	**68.2**	**17.5**	**58.4**	99.3
DOFA	Climate	12.9	56.3	15.7	57.6	10.9	55.3	6.9	53.1	99.4
U-Net	Climate-hard	19.9	59.8	26.2	62.9	16.1	58.0	11.1	55.3	99.5
DeepLabV3+	Climate-hard	18.0	58.9	36.2	67.9	12.0	56.0	9.9	54.7	**99.6**
ViT	Climate-hard	11.4	55.6	22.7	61.2	7.6	53.7	6.0	52.8	99.5
SegFormer	Climate-hard	**22.6**	**61.2**	33.1	66.4	**17.1**	**58.5**	**12.7**	**56.1**	99.5
Mask2Former	Climate-hard	15.4	57.6	**45.8**	**72.7**	9.2	54.6	8.3	54.0	**99.6**
DOFA	Climate-hard	4.7	52.1	3.9	51.8	6.0	52.7	2.4	50.7	99.0
U-Net	Forest	35.3	67.5	34.3	67.0	36.3	68.0	21.4	60.4	99.5
DeepLabV3+	Forest	36.1	67.9	36.9	68.3	35.3	67.5	22.0	60.8	99.5
ViT	Forest	26.5	63.1	27.3	63.5	25.8	62.8	15.3	57.4	99.4
SegFormer	Forest	**40.1**	**69.9**	**41.5**	**70.7**	38.7	69.2	**25.1**	**62.3**	**99.5**
Mask2Former	Forest	38.6	69.2	36.8	68.3	**40.6**	**70.1**	23.9	61.7	**99.5**
DOFA	Forest	13.9	56.8	14.4	57.0	13.3	56.5	7.5	53.4	99.4
U-Net	Forest-hard	14.7	57.0	9.6	54.7	31.5	65.3	7.9	53.3	98.8
DeepLabV3+	Forest-hard	21.5	60.7	27.5	63.6	17.7	58.8	12.1	55.8	**99.6**
ViT	Forest-hard	19.9	59.7	14.5	57.1	31.7	65.5	11.1	55.1	99.1
SegFormer	Forest-hard	**37.7**	**68.7**	**38.3**	**69.0**	**37.1**	**68.5**	**23.2**	**61.4**	**99.6**
Mask2Former	Forest-hard	11.1	55.4	10.4	55.0	11.9	55.8	5.9	52.6	99.4
DOFA	Forest-hard	14.3	56.9	11.2	55.4	20.0	59.7	7.7	53.4	99.1

compared to E-W on average. The reason could be that the west side covers more conditions (e.g., local climates) or the task there is more challenging with less contrast to the background landscape. The single state case shows significantly reduced scores due to the lack of sufficient representative samples.

Cross-climate and cross-forest-type generalizability. The results for cross-climate and cross-forest-type generalization are shown in Table 3. We skipped the results for train-test group swaps (i.e., similarly like the 2nd row "W-E" in Table 2) due to the space limit, and the full tables are available in the appendix. In Table 3, "Climate" and "Forest" represent the relatively easier scenarios where about half of the data from certain climate zones or forest types are used for training and the rest for testing. The "-hard" modes are the cases where only one climate zone or forest type is used as training, as described in Sec. 3.3. The general trend is similar as before, where major decreases in scores are observed compared to the random-split case due to the intrinsically higher difficulty. Overall, SegFormer tends to have the top-tier performances (i.e., top or very close to top) in F1 score and IoU, demonstrating its consistency. Mask2Former also received the top F1 score, Recall, and IoU in two cases, though followed tightly by SegFormer. As expected, performance drops again notably in the single-vs-all scenarios (i.e., the "-hard" modes in the table), reflecting the difficulty of current models' generalizability in unseen conditions. However, this is a frequently encountered situation in large-scale mapping that need new developments. In comparison to Table 2, the results are slightly better, likely because in cross-region situations, there are more factors contributing to the variability, further reducing the representativeness of highly localized samples.

5 Conclusion and Limitations

TreeFinder offers a high-resolution, large-scale benchmark dataset for individual-level tree mortality mapping with extensive manual labels. Spanning 1,000 sites over 48 states in the CONUS with a 23,000-hectare coverage, TreeFinder supports the development of ML models capable of identifying these fine-granularity events with less contrast over different geographic regions. The dataset is enriched with metadata on climate zones and primary forest types to facilitate generalization tests under various scenarios. We consider a suite of baseline models including both convolutional and ViT-based foundational models across a wide range of generalization scenarios. Our benchmarking experiments highlight the challenges of model generalization across geographic, climate, and forest type conditions and the needs for further model developments. The dataset and corresponding Python libraries are shared to support convenient data usage.

Limitations and future directions. Despite its scale and scope, TreeFinder has several limitations. First, NAIP offers near wall-to-wall coverage across CONUS, but is not available at the global scale. Future expansions may include other regions with wall-to-wall coverage of high-resolution images at national-scale (e.g., from Switzerland) or commercial satellites such as WorldView-3, which offer similar spatial resolution to NAIP over the globe, though the data may not be publicly available for free. Second, the dataset has not yet considered challenges related to the changes in NAIP dataset itself, including the change of resolution over time. Currently, we only included recent years' images at 0.6m resolution, and future extensions are needed to include 1m resolution data to support cross-resolution model development. Third, our evaluation has not considered cases for active learning, meta-learning, etc. The presented scenarios are most commonly encountered situations in practical applications, but future extensions should develop standard testing cases for different types of training strategies as well. We may also explore the applications of emerging general-purpose vision foundation models (e.g., SAM2 and DINOv2) for this challenging segmentation task. Finally, TreeFinder has not considered integration of multiple data sources (e.g., NAIP in combination with other lower-resolution platforms with richer multispectral infomration).

Acknowledgments

Zhihao Wang, Yiqun Xie, Ruichen Wang, and Zhili Li are supported in part by the NSF under Grant No. 2126474, 2147195, 2425844, and 2530610; NASA under grant 80NSSC25K0013 and 80NSSC25K7221; Google's AI for Social Good Impact Scholars program; and the Zaratan cluster at the University of Maryland. George Hurtt is supported by the NASA CMS under Grant No. 80NSSC25K7221 and NASA EIS under Grant No. 80NSSC22K1733. Lei Ma is supported by the NASA ECIPES under Grant No. 80NSSC24K1632. Xiaowei Jia was partially supported by the NSF under Grant No. 2239175, 2147195, 2316305, 2425845, 2530609, 2203581; NASA under Grant No. 80NSSC24K1061 and 80NSSC25K0013; the USGS awards G21AC10564 and G22AC00266; and Pitt Momentum Funds and CRC at the University of Pittsburgh. Gengchen Mai is supported by the NSF under Grant No. 2521631. We would like to acknowledge high-performance computing support from the Derecho system (doi:10.5065/qx9a-pg09) provided by the NSF National Center for Atmospheric Research (NCAR), sponsored by the National Science Foundation.

References

[1] Mete Ahishali, Anis Ur Rahman, Einari Heinaro, and Samuli Junttila. Ada-net: Attention-guided domain adaptation network with contrastive learning for standing dead tree segmentation using aerial imagery. *arXiv preprint arXiv:2504.04271*, 2025.

[2] Matthew J Allen, Daniel Moreno-Fernández, Paloma Ruiz-Benito, Stuart WD Grieve, and Emily R Lines. Low-cost tree crown dieback estimation using deep learning-based segmentation. *Environmental Data Science*, 3:e18, 2024.

[3] William RL Anderegg, Jeffrey M Kane, and Leander DL Anderegg. Consequences of widespread tree mortality triggered by drought and temperature stress. *Nature climate change*, 3(1):30–36, 2013.

[4] Hylke E Beck, Niklaus E Zimmermann, Tim R McVicar, Noemi Vergopolan, Alexis Berg, and Eric F Wood. Present and future köppen-geiger climate classification maps at 1-km resolution. *Scientific data*, 5(1):1–12, 2018.

[5] Vitus Benson, Claire Robin, Christian Requena-Mesa, Lazaro Alonso, Nuno Carvalhais, José Cortés, Zhihan Gao, Nora Linscheid, Mélanie Weynants, and Markus Reichstein. Multi-modal learning for geospatial vegetation forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27788–27799, 2024.

[6] Christof Bigler and Thomas T Veblen. Changes in litter and dead wood loads following tree death beneath subalpine conifer species in northern colorado. *Canadian Journal of Forest Research*, 41(2):331–340, 2011.

[7] Eckehard G Brockerhoff, Luc Barbaro, Bastien Castagneyrol, David I Forrester, Barry Gardiner, José Ramón González-Olabarria, Phil O'B Lyver, Nicolas Meurisse, Anne Oxbrough, Hisatomo Taki, et al. Forest biodiversity, ecosystem functioning and the provision of ecosystem services, 2017.

[8] Eckehard G Brockerhoff, Hervé Jactel, John A Parrotta, and Silvio FB Ferraz. Role of eucalypt and other planted forests in biodiversity conservation and the provision of biodiversity-related ecosystem services. *Forest Ecology and Management*, 301:43–50, 2013.

[9] Ian R Calder. Forests and water—ensuring forest benefits outweigh water costs. *Forest ecology and management*, 251(1-2):110–120, 2007.

[10] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[11] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G Schwing. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*, 2021.

[12] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022.

[13] Yan Cheng, Stefan Oehmcke, Martin Brandt, Lisa Rosenthal, Adrian Das, Anton Vrieling, Sassan Saatchi, Fabien Wagner, Maurice Mugabowindekwe, Wim Verbruggen, et al. Scattered tree death contributes to substantial forest loss in california. *Nature communications*, 15(1):641, 2024.

[14] Brendan Choat, Timothy J Brodribb, Craig R Brodersen, Remko A Duursma, Rosana López, and Belinda E Medlyn. Triggers of tree mortality under drought. *Nature*, 558(7711):531–539, 2018.

[15] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 172–181, 2018.

[16] Foivos I Diakogiannis, François Waldner, Peter Caccetta, and Chen Wu. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:94–114, 2020.

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[18] Solange Filoso, Maíra Ometto Bezerra, Katherine CB Weiss, and Margaret A Palmer. Impacts of forest restoration on water yield: A systematic review. *PloS one*, 12(8):e0183210, 2017.

[19] RH Fraser and R Latifovic. Mapping insect-induced tree defoliation and mortality using coarse spatial resolution satellite imagery. *International Journal of Remote Sensing*, 26(1):193–200, 2005.

[20] Pierre Friedlingstein, Michael O'sullivan, Matthew W Jones, Robbie M Andrew, Luke Gregor, Judith Hauck, Corinne Le Quéré, Ingrid T Luijkx, Are Olsen, Glen P Peters, et al. Global carbon budget 2022. *Earth System Science Data*, 14(11):4811–4900, 2022.

[21] Matthew C Hansen, Peter V Potapov, Rebecca Moore, Matt Hancher, Svetlana A Turubanova, Alexandra Tyukavina, David Thau, Stephen V Stehman, Scott J Goetz, Thomas R Loveland, et al. High-resolution global maps of 21st-century forest cover change. *science*, 342(6160):850–853, 2013.

[22] Sarah J Hart and Thomas T Veblen. Detection of spruce beetle-induced tree mortality using high- and medium-resolution remotely sensed imagery. *Remote Sensing of Environment*, 168:134–145, 2015.

[23] Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, et al. Spectralgpt: Spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[24] Xingliang Huang, Libo Ren, Chenglong Liu, Yixuan Wang, Hongfeng Yu, Michael Schmitt, Ronny Hänsch, Xian Sun, Hai Huang, and Helmut Mayer. Urban building classification (ubc)-a dataset for individual building detection and classification from satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1413–1421, 2022.

[25] G Hurtt, M Zhao, R Sahajpal, A Armstrong, R Birdsey, E Campbell, Katelyn Dolan, R Dubayah, JP Fisk, S Flanagan, et al. Beyond mrv: high-resolution forest carbon modeling for climate mitigation planning over maryland, usa. *Environmental Research Letters*, 14(4):045013, 2019.

[26] Johannes Jakubik, Sujit Roy, CE Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, et al. Foundation models for generalist geospatial artificial intelligence. *arXiv preprint arXiv:2310.18660*, 2023.

[27] Thani Jintasuttisak, Eran Edirisinghe, and Ali Elbattay. Deep neural network based date palm tree detection in drone imagery. *Computers and Electronics in Agriculture*, 192:106560, 2022.

[28] Pratima Khatri-Chhetri, Liz van Wagtendonk, Sean M Hendryx, and Van R Kane. Enhancing individual tree mortality mapping: The impact of models, data modalities, and classification taxonomy. *Remote Sensing of Environment*, 300:113914, 2024.

[29] Yang Li, Yanlan Liu, Gil Bohrer, Yongyang Cai, Aaron Wilson, Tongxi Hu, Zhihao Wang, and Kaiguang Zhao. Impacts of forest loss on local climate across the conterminous united states: Evidence from satellite time-series observations. *Science of the Total Environment*, 802:149651, 2022.

[30] John A Long and Rick L Lawrence. Mapping percent tree mortality due to mountain pine beetle damage. *Forest Science*, 62(4):392–402, 2016.

[31] Lei Ma, George Hurtt, Lesley Ott, Ritvik Sahajpal, Justin Fisk, Rachel Lamb, Hao Tang, Steve Flanagan, Louise Chini, Abhishek Chatterjee, et al. Global evaluation of the ecosystem demography model (ed v3. 0). *Geoscientific Model Development*, 15(5):1971–1994, 2022.

[32] Qin Ma, Yanjun Su, Chunyue Niu, Qin Ma, Tianyu Hu, Xiangzhong Luo, Xiaonan Tai, Tong Qiu, Yao Zhang, Roger C Bales, et al. Tree mortality during long-term droughts is lower in structurally complex forest stands. *Nature communications*, 14(1):7467, 2023.

[33] Nate G McDowell, Gerard Sapes, Alexandria Pivovaroff, Henry D Adams, Craig D Allen, William RL Anderegg, Matthias Arend, David D Breshears, Tim Brodribb, Brendan Choat, et al. Mechanisms of woody-plant mortality under rising drought, co2 and vapour pressure deficit. *Nature Reviews Earth & Environment*, 3(5):294–308, 2022.

[34] Clemens Mosig, Janusch Vajna-Jehle, Miguel D Mahecha, Yan Cheng, Henrik Hartmann, David Montero, Samuli Junttila, Stephanie Horion, Mirela Beloiu Schwenke, Stephen Adu-Bredu, et al. deadtrees. earth-an open-access and interactive database for centimeter-scale aerial imagery to uncover global tree mortality dynamics. *bioRxiv*, pages 2024–10, 2024.

[35] Yude Pan, Richard A Birdsey, Oliver L Phillips, and Robert B Jackson. The structure, distribution, and biomass of the world's forests. *Annual Review of Ecology, Evolution, and Systematics*, 44:593–622, 2013.

[36] Stefano Puliti, Grant Pearse, Peter Surovỳ, Luke Wallace, Markus Hollaus, Maciej Wielgosz, and Rasmus Astrup. For-instance: a uav laser scanning benchmark dataset for semantic and instance segmentation of individual trees. *arXiv preprint arXiv:2309.01279*, 2023.

[37] Christian Requena-Mesa, Vitus Benson, Markus Reichstein, Jakob Runge, and Joachim Denzler. Earthnet2021: A large-scale dataset and challenge for earth surface forecasting as a guided video prediction task. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1132–1142, 2021.

[38] Kyle C Rodman, Robert A Andrus, Cori L Butkiewicz, Teresa B Chapman, Nathan S Gill, Brian J Harvey, Dominik Kulakowski, Niko J Tutland, Thomas T Veblen, and Sarah J Hart. Effects of bark beetle outbreaks on forest landscape pattern in the southern rocky mountains, usa. *Remote Sensing*, 13(6):1089, 2021.

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[40] Felix Schiefer, Sebastian Schmidtlein, Annett Frick, Julian Frey, Randolf Klinke, Katarzyna Zielewska-Büttner, Samuli Junttila, Andreas Uhl, and Teja Kattenborn. Uav-based reference data for the prediction of fractional cover of standing deadwood from sentinel time series. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 8:100034, 2023.

[41] Felix Schiefer, Sebastian Schmidtlein, Henrik Hartmann, Florian Schnabel, and Teja Kattenborn. Large-scale remote sensing reveals that tree mortality in germany appears to be greater than previously expected. *Forestry: An International Journal of Forest Research*, page cpae062, 2024.

[42] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021.

[43] Huan Tao, Cunjun Li, Dan Zhao, Shiqing Deng, Haitang Hu, Xinluo Xu, and Weibin Jing. Deep learning-based dead pine tree detection from unmanned aerial vehicle images. *International Journal of Remote Sensing*, 41(21):8238–8255, 2020.

[44] Aysim Toker, Lukas Kondmann, Mark Weber, Marvin Eisenberger, Andrés Camero, Jingliang Hu, Ariadna Pregel Hoderlein, Çağlar Şenaras, Timothy Davis, Daniel Cremers, Giovanni Marchisio, Xiao Xiang Zhu, and Laura Leal-Taixé. Dynamicearthnet: Daily multi-spectral satellite dataset for semantic change segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21158–21167, June 2022.

[45] Gabriel Tseng, Ivan Zvonkov, Catherine Nakalembe, and Hannah Kerner. Cropharvest: a global satellite dataset for crop type classification. *Neural Information Processing Systems (NeurIPS)*, 2021.

[46] U.S. Department of Agriculture, Forest Service. Individual tree species parameter maps. `https://www.fs.usda.gov/science-technology/data-tools-products/fhp-mapping-reporting/individual-tree-species-parameter-maps`, 2025. Accessed: 2025-05-13.

[47] Phillip J Van Mantgem, Nathan L Stephenson, John C Byrne, Lori D Daniels, Jerry F Franklin, Peter Z Fulé, Mark E Harmon, Andrew J Larson, Jeremy M Smith, Alan H Taylor, et al. Widespread increase of tree mortality rates in the western united states. *Science*, 323(5913):521–524, 2009.

[48] Di Wang, Jing Zhang, Bo Du, Minqiang Xu, Lin Liu, Dacheng Tao, and Liangpei Zhang. Samrs: Scaling-up remote sensing segmentation dataset with segment anything model. *Advances in Neural Information Processing Systems*, 36:8815–8827, 2023.

[49] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.

[50] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021.

[51] James J Worrall, Leanne Egeland, Thomas Eager, Roy A Mask, Erik W Johnson, Philip A Kemp, and Wayne D Shepperd. Rapid mortality of populus tremuloides in southwestern colorado, usa. *Forest Ecology and Management*, 255(3-4):686–696, 2008.

[52] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

[53] Zhitong Xiong, Yi Wang, Fahong Zhang, Adam J Stewart, Joëlle Hanna, Damian Borth, Ioannis Papoutsis, Bertrand Le Saux, Gustau Camps-Valls, and Xiao Xiang Zhu. Neural plasticity-inspired multimodal foundation model for earth observation. *arXiv preprint arXiv:2403.15356*, 2024.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading “NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

   Answer: [Yes]

   Justification: The claims in the abstract are detailed in the paper with evidence.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We explicitly discuss the limitations in the conclusion and future work.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This dataset and evaluation does not cover theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We share the dataset and codes, and all experiment settings are included into the code.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code is open-sourced on GitHub and dataset is shared with sufficient details.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The splits and evaluation scenarios are explicitly discussed with reasons explained.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Our primary goal is to establish baseline performance and benchmark generalization behavior across ecological and spatial domains using a fixed dataset split and deterministic training pipeline. Due to computational cost and the focus on large-scale coverage, we report single-run results without variance estimates. However, the performance differences between methods—especially across generalization settings—are large and consistent across scenarios, which qualitatively supports the main findings.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We describe the computing resource information is details.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, we follow the Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Yes, when discussing future work to mitigate spatial bias using heterogeneity-aware learning.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our tree mortality dataset is rigorously valid with strict criteria and aims to reflect the potential dead tree events. It does not have risk information for misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite all products and models used in this paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new asset is well documented in the paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: We do not have research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Same as above.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: We only used ChatGPT for editing and proofreading our manuscript.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Appendix

A Dataset Generation

A.1 Annotation Details

The labeling of dead trees was performed using the Google Earth Engine (GEE) platform with NAIP imagery (0.6m resolution). Imagery for each state was loaded using GEE scripts that filtered the NAIP collection by date and the state boundary. Visualization layers were configured with both true color (RGB) and false color (NIR–R–G) composites to aid visual interpretation. The false color combination enhances the spectral contrast between healthy and dead vegetation by emphasizing near-infrared reflectance differences.

The tree mortality labeling followed a strict and consistent set of criteria to ensure consistency and accuracy. For visualization, annotators first inspected a false-color composite to detect potential dead trees and then double-checked the observations using a true-color image. In terms of crown condition, a tree was considered for labeling only if at least half of its crown appeared visibly dead, indicated by discoloration toward the crown's edge or trunk, fuzziness, or exposed branches. The certainty threshold required that a tree appear clearly dead in both visualization modes before a label was assigned. To be more consistent in labeling and have higher certainty, annotators avoided over-labeling for trees that were brown but did not have structural indicators of mortality. Additionally, discoloration or standing branches that might have been attributable to seasonal senescence rather than mortality were further investigated through historical imagery to verify the tree's living status before a label was finalized.

Annotations were drawn manually using GEE's geometry tools, with one polygon per dead crown (or in some cases, connected dead patch). Each polygon contained at least ten pixels to maintain consistency and avoid mislabeling tiny ambiguous regions. Polygons were stored as feature collections in GEE, rasterized to binary masks (dead = 1, background = 0), and exported with the corresponding NAIP tiles. A custom function handled rasterization and reprojection at a spatial resolution of 0.6m.

A.2 Validation Details

We validated our annotations following a standard protocol. Specifically, a stratified random sampling approach was used for validation sample collection, with sample size proportional to the total labeled area in each state. Each annotator provided 20% random samples from their labeled tiles for validation. Both commission (false positives) and omission (false negatives) were counted as disagreements between the annotator and validator. The final count of disaggreements were recorded for a consensus review. During the consensus review, each validator presented each disagreement and provided supporting visual evidence, while annotators were given the opportunity to explain their interpretation, including showing the historical images if necessary. All participants then voted on whether each disputed sample should be retained or removed from the final dataset. The resulting dataset represents a majority-voted consensus designed to minimize individual bias and ensure consistent labeling quality. The final agreement score was calculated as the number of correctly assigned labels divided by the total number of labels, leading to about 97% cross-annotator agreement.

B Training details

All benchmark models were trained or fine-tuned (when pretrained weights were available) using the TreeFinder dataset. The training of all models were performed with a batch size of 32 and a maximum of 100 epochs with early stopping. We used a composite loss function combining Binary Cross-Entropy (BCE) loss and Dice loss to address class imbalance and improve segmentation performance. Although we experimented with Focal Loss, commonly used for imbalanced classification, we found it yielded similar performance to BCE and therefore did not include it in the final benchmarks. All models were optimized using the AdamW optimizer with a weight decay of 0.01 and an initial learning rate of 1×10^{-4}, decayed over time using an ExponentialLR scheduler. Early stopping was applied based on validation loss with a patience of 5 epochs to prevent overfitting. Fig. 5 shows the loss changes in training and validation dataset for the random split experiment, where 80% of the dataset is used for training and 10% of the training set is reserved for validation. Both training and

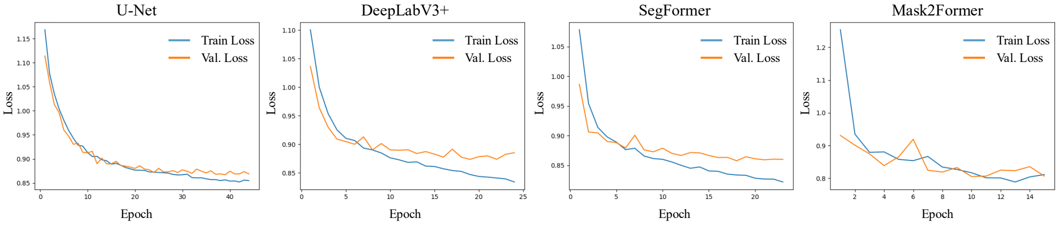


Figure 5: Training and validation loss curves for the random split experiment, where 80% of the dataset is used for training and 10% of the training set is reserved for validation.

Table 4: Performance across shifted domains in climate zones. The climate-swap scenario is added as a complementary part of Table 3 in the main paper. All values are percentages %, and best values are bolded.

Model	**Scenario**	**F1**		**Precision**		**Recall**		**IoU**		**Accuracy**
		Dead Tree	All	Dead Tree	All	Dead Tree	All	Dead Tree	All	
U-Net	Climate	18.6	59.1	17.3	58.5	20.1	59.8	10.3	54.8	99.3
DeepLabV3+	Climate	25.8	62.8	31.4	65.5	21.9	60.9	14.8	57.2	**99.5**
ViT	Climate	21.6	60.7	23.4	61.5	20.1	59.9	12.1	55.8	99.4
SegFormer	Climate	28.0	63.9	**33.7**	**66.7**	24.0	61.9	16.3	57.9	**99.5**
Mask2Former	Climate	**29.8**	**64.7**	25.0	62.4	**36.8**	**68.2**	**17.5**	**58.4**	99.3
DOFA	Climate	12.9	56.3	15.7	57.6	10.9	55.3	6.9	53.1	99.4
U-Net	Climate-swap	29.9	64.9	44.9	72.3	22.4	61.2	17.6	58.6	**99.7**
DeepLabV3+	Climate-swap	31.2	65.5	42.7	71.2	24.6	62.2	18.5	59.1	**99.7**
ViT	Climate-swap	23.8	61.8	30.1	64.9	19.7	59.8	13.5	56.6	99.6
SegFormer	Climate-swap	34.9	67.4	**46.5**	**73.1**	28.0	63.9	21.1	60.4	**99.7**
Mask2Former	Climate-swap	**35.5**	**67.7**	45.6	72.7	**29.0**	**64.5**	**21.6**	**60.6**	**99.7**
DOFA	Climate-swap	17.9	58.9	22.5	61.1	14.9	57.4	9.9	54.7	99.6
U-Net	Climate-hard	19.9	59.8	26.2	62.9	16.1	58.0	11.1	55.3	99.5
DeepLabV3+	Climate-hard	18.0	58.9	36.2	67.9	12.0	56.0	9.9	54.7	**99.6**
ViT	Climate-hard	11.4	55.6	22.7	61.2	7.6	53.7	6.0	52.8	99.5
SegFormer	Climate-hard	**22.6**	**61.2**	33.1	66.4	**17.1**	**58.5**	**12.7**	**56.1**	99.5
Mask2Former	Climate-hard	15.4	57.6	**45.8**	**72.7**	9.2	54.6	8.3	54.0	**99.6**
DOFA	Climate-hard	4.7	52.1	3.9	51.8	6.0	52.7	2.4	50.7	99.0

validation curves gradually decrease and converge without significant divergence, indicating no clear overfitting. For each model, the checkpoint achieving the lowest validation loss was selected for final testing.

C Additional Results

Cross-climate and cross-forest-type generalizability. This appendix provides additional results in Tables 4 and 5 to show a more complete evaluation of model generalization under shifted domains in both climate zones and primary tree types. Specifically, in this main paper we provided the results for the "Climate" and "Forest" scenarios and skipped the train-test swapped versions that were shown for the cross-region generalization test (i.e., "W-E" as a swapped version for "E-W"). Here we provide the full results where in the "Climate-swap" scenario the data in the training climate zones of "Climate" are used as testing and those in the testing climate zones are used for training. The same swapping is done for "Forest-swap" as well where forest types are used instead of climate zones. We did not do the swapping for the hard scenarios, i.e., single state as training in cross-region generalization ("CO"), single climate zone as training in cross-climate generalization ("Climate-hard"), and single forest type as training in the cross-forest-type generalization ("Forest-hard"), because their evaluation goal is to use limited samples for training and see how the model behaves in the more challenge situations. Thus, swapping them will not no longer serve this specific purpose. Looking at the results, we observe the similar performance drops relative to random splits for both scenarios, consistent with expectations due to cross-climate and cross-forest-type variability. SegFormer and Mask2Former consistently rank among the top-performing models across most metrics. In the climate-swap scenario, SegFormer achieves the highest precision, while Mask2Former achieves the highest recall, F1 score, and IoU. In the forest-swap scenario, Mask2Former outperform the other models, tighly followed by SegFormer.

Table 5: Performance across shifted domains in primary tree types. The forest-swap scenario is added as a complementary part of Table 3 in the main paper. All values are percentages %, and best values are bolded.

Model	**Scenario**	**F1**		**Precision**		**Recall**		**IoU**		**Accuracy**
		Dead Tree	All	Dead Tree	All	Dead Tree	All	Dead Tree	All	
U-Net	Forest	35.3	67.5	34.3	67.0	36.3	68.0	21.4	60.4	99.5
DeepLabV3+	Forest	36.1	67.9	36.9	68.3	35.3	67.5	22.0	60.8	99.5
ViT	Forest	26.5	63.1	27.3	63.5	25.8	62.8	15.3	57.4	99.4
SegFormer	Forest	**40.1**	**69.9**	**41.5**	**70.7**	38.7	69.2	**25.1**	**62.3**	**99.5**
Mask2Former	Forest	38.6	69.2	36.8	68.3	**40.6**	**70.1**	23.9	61.7	**99.5**
DOFA	Forest	13.9	56.8	14.4	57.0	13.3	56.5	7.5	53.4	99.4
U-Net	Forest-swap	28.9	64.4	36.1	67.9	24.1	62.0	16.9	58.2	**99.6**
DeepLabV3+	Forest-swap	26.5	63.1	35.4	67.6	21.2	60.5	15.3	57.4	**99.6**
ViT	Forest-swap	10.7	55.2	13.9	56.8	8.7	54.3	5.7	52.6	99.5
SegFormer	Forest-swap	32.1	66.0	39.7	69.7	27.0	63.4	19.1	59.4	**99.6**
Mask2Former	Forest-swap	**33.8**	**66.8**	**44.1**	**71.9**	**27.4**	**63.6**	**20.3**	**60.0**	**99.6**
DOFA	Forest-swap	14.9	57.3	15.1	57.4	14.6	57.1	8.0	53.7	99.4
U-Net	Forest-hard	14.7	57.0	9.6	54.7	31.5	65.3	7.9	53.3	98.8
DeepLabV3+	Forest-hard	21.5	60.7	27.5	63.6	17.7	58.8	12.1	55.8	**99.6**
ViT	Forest-hard	19.9	59.7	14.5	57.1	31.7	65.5	11.1	55.1	99.1
SegFormer	Forest-hard	**37.7**	**68.7**	**38.3**	**69.0**	**37.1**	**68.5**	**23.2**	**61.4**	**99.6**
Mask2Former	Forest-hard	11.1	55.4	10.4	55.0	11.9	55.8	5.9	52.6	99.4
DOFA	Forest-hard	14.3	56.9	11.2	55.4	20.0	59.7	7.7	53.4	99.1

DOFA did not perform very well in both scenarios, likely due to its focus on cross-wavelength applicability and limited specialization for specific bands and tasks. The overall patterns are similar to those from the main paper's results.