https://doi.org/10.5194/essd-2026-380
Preprint. Discussion started: 21 October 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

ArborVerse: Scaling Fine-Grained Urban Tree Species Classification to Millions of Trees Worldwide

Jiapeng Li[1], Yingjing Huang[1,2], Fan Zhang[1], Shengli Tao[3], Zhen Dong[4], Jun Yang[5], and Yu Liu[1]

[1]School of Earth and Space Sciences, Peking University, Beijing, 100871, China
[2]Department of Geography and Regional Research, University of Vienna, Vienna, 1010, Austria
[3]College of Urban and Environmental Sciences, Peking University, Beijing, 100871, China
[4]State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, 430079, China
[5]Department of Earth System Science, Ministry of Education Key Laboratory for Earth System Modeling, Institute for Global Change Studies, Tsinghua University, Beijing, 100084, China

Correspondence: Fan Zhang (fanzhanggis@pku.edu.cn)

Abstract. The fine-grained classification of street trees is a crucial task for urban planning, streetscape management, and the assessment of urban ecosystem services. However, progress in this field has been hindered by the lack of large-scale, geographically diverse, and publicly available benchmark datasets specifically designed for street trees. To address this critical gap, we introduce ArborVerse, the world's first large-scale benchmark dataset dedicated to fine-grained street tree classification. The dataset contains over 12 million images covering more than 8,300 common street tree species, collected from urban streetscapes across 132 countries spanning five continents, and supplemented with expert-verified observational data. ArborVerse poses challenges for pretrained vision models under complex urban environments: high inter-species visual similarity, long-tailed natural distributions, significant intra-class variations caused by seasonal changes, and diverse imaging conditions such as lighting, occlusions from buildings, and varying camera angles. In addition, we provide a hierarchical taxonomy (order–family–genus–species) to support research in hierarchical classification and representation learning. Through extensive experiments with various vision models, we establish solid baselines and reveal the limitations of existing methods in handling such real-world complexities. We believe that ArborVerse will serve as a key resource for driving new advancements at the intersection of computer vision and urban science. The data is now publicly available at https://doi.org/10.6084/m9.figshare.32172486 (Li et al., 2026).

Keywords. Urban forestry, Fine-grained image classification, Street view imagery, Hierarchical taxonomy, Long-tailed distribution

1 Introduction

Urban street trees form an integral component of the global forest system, contributing substantially to carbon sequestration, biodiversity maintenance and local climate regulation (Lian et al., 2025; Wu et al., 2025). Amid accelerating urbanization, these trees constitute the most visible and proximate form of urban greenery, directly influencing residents' daily environments. They

https://doi.org/10.5194/essd-2026-380
Preprint. Discussion started: 21 October 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

play critical roles in mitigating urban heat, improving air quality, and managing stormwater, thereby enhancing urban resilience and livability (Smith et al., 2025; Schwaab et al., 2021). Establishing large-scale, high-quality, and scalable databases of street trees is thus fundamental for advancing evidence-based urban governance and supporting the planning and management of green infrastructure in cities worldwide.

However, constructing and maintaining such databases is both time- and labor-intensive, often relying on expert-led field surveys (Beery et al., 2022). In recent years, remote sensing imagery and unmanned aerial vehicle (UAV) photography have been utilized to analyze tree species (Bolyn et al., 2022; Thapa et al., 2024). Yet, these approaches are primarily suited to large suburban or forested areas, as their spatial resolution is typically too coarse to capture individual trees. Moreover, their top-down perspective provides only limited information on tree canopies, making fine-grained, species-level classification of individual trees particularly challenging.

The rapidly increasing amount of street view imagery has captured detailed visual information of urban streets (Huang et al., 2023, 2025b). It provides a comprehensive and in-depth description of urban environments, making it an excellent data source for observing street trees (Wegner et al., 2016). Drawing upon this rich data source, a growing body of research has used street view imagery to study urban street trees, including measuring urban greenery, identifying tree locations, and establishing tree inventories (Huang et al., 2025a; Arevalo-Ramirez et al., 2024; Oliveira et al., 2022). While the importance of street view–based tree datasets has been increasingly recognized, several limitations remain. Existing datasets are still scarce, with limited geographic coverage and data scale, incomplete taxonomic annotations across hierarchical levels, noisy labeling quality, and often only single-time-point observations for each tree. These constraints continue to hinder the fine-scale management of urban street trees.

To address these limitations, we present **ArborVerse**, a large-scale, fine-grained global dataset of urban street trees (Figure 1). The dataset includes samples from 132 countries across five continents, containing more than 12 million observations of 3,365,485 individual trees annotated with a four-level taxonomic hierarchy. Each tree in ArborVerse is linked to its geographic coordinates, time and season of observation, and hierarchical taxonomic labels spanning order, family, genus, and species. Together, these components provide a comprehensive and structured representation of urban trees worldwide.

Overall, the ArborVerse dataset offers five key contributions:

- **Global coverage.** Over 12 million samples from 132 countries across five continents ensure broad spatial representativeness.
- **Hierarchical taxonomy.** Each tree is annotated with a four-level taxonomy: order, family, genus and species.
- **Street view imagery.** Ground-level images enable detailed observation of taxonomic features such as crown morphology, leaf shape, and bark texture.
- **Seasonal information.** Each tree is labeled with the corresponding season, supporting analysis of phenological variation.
- **Temporal continuity.** Spanning 2015–2025, the dataset provides long-term records for tracking urban vegetation change.

https://doi.org/10.5194/essd-2026-380
Preprint. Discussion started: 21 October 2026
© Author(s) 2026. CC BY 4.0 License.

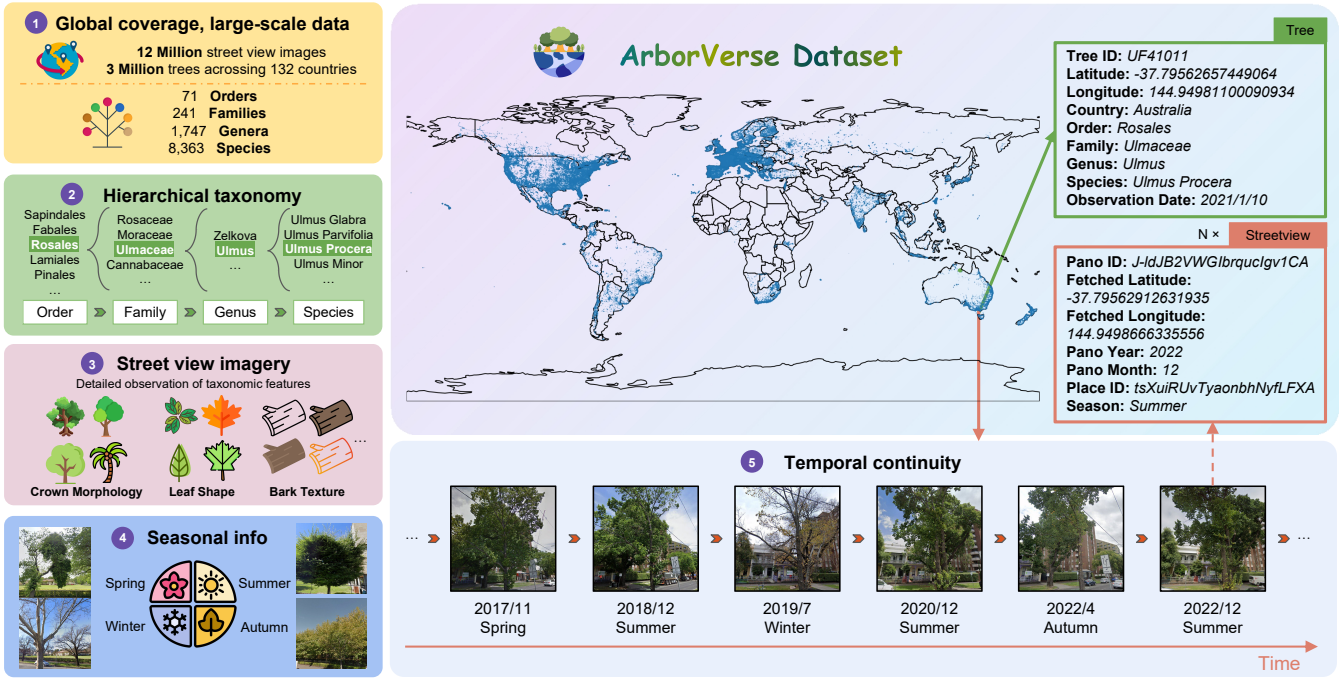


Figure 1. Overview of ArborVerse dataset. The dataset comprises 12,235,152 street view images representing 3,365,485 individual urban trees across 132 countries and five continents, spanning 71 orders, 241 families, 1,747 genera, and 8,363 species. Each tree may correspond to one or multiple street view images. The map illustrates the global geographic distribution of tree samples, where darker shades indicate higher densities. Most individual trees are associated with long-term temporal records, capturing seasonal and interannual variations. Image source: © Google Street View 2026.

1.1 Publicly available tree species datasets

Existing publicly available tree species datasets exhibit considerable diversity. While many are image-based, spanning from satellite to ground-level street view imagery, other tree inventories consist only of georeferenced and species data. Table 1 provides several representative datasets, detailing their geographic coverage, size, taxonomic granularity, temporal span, individual tree information, seasonal coverage, image type, and region type.

Despite their considerable value, these existing datasets still present several shortcomings that remain to be addressed. First, most datasets are limited to a single country or city, resulting in extremely narrow geographic coverage that restricts data diversity and hampers the transferability of models across regions. Second, few datasets provide a complete taxonomic hierarchy, making it difficult to satisfy taxonomic query requirements in a unified framework. Third, some datasets only offer coarse geographic information (e.g. EUForest (Mauri et al., 2017), Pureforest (Gaydon and Roche, 2025), TreeSatAI (Ahlswede et al., 2023) and NEON Vegetation Structure Data (Kampe et al., 2010)), such as approximate locations within $1km \times 1km$ grids, rather than coordinates of individual trees, which prevents accurate spatial localization of species. Last, all these datasets lack

https://doi.org/10.5194/essd-2026-380
Preprint. Discussion started: 21 October 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

Table 1. Overview of publicly available datasets for tree species classification.

Dataset	Geographic Scope	Size	Taxonomic Granularity	Temporal Span	Individual Tree	Seasonal Coverage	Image Type	Region Type
GlobalGeoTree (Mu et al., 2025)	Global	6,263,345	Family / Genus / Species	2015-2024	✓	✗	Satellite Image	Forest
Auto Arborist (Beery et al., 2022)	USA	2,637,208	Genus	2018-2022	✓	✗	Satellite Image and Street View Image	Urban
U.S. 5 Million (McCoy et al., 2022)	USA	5,660,237	Genus / Species	-	✓	✗	Not available	Forest / Urban
Urbanforest (City of Melbourne Open Data Team, 2023)	Australia	76,928	Family / Genus / Species	-	✓	✗	Not available	Urban
EU Forest (Mauri et al., 2017)	Europe	588,983	Genus / Species	-	✗	✗	Satellite Image	Forest
PureForest (Gaydon and Roche, 2025)	France	135,569	Species	-	✗	✗	Satellite Image	Forest
TreeSatAI (Ahlswede et al., 2023)	Germany	50,381	Genus / Species	2011-2020	✗	✗	Satellite Image	Forest
NEON Vegetation Structure Data (Kampe et al., 2010)	USA	-	Family / Genus / Species	-	✗	✗	Satellite Image	Forest
ArborVerse	**Global**	**12,235,152**	**Order / Family / Genus / Species**	**2015-2025**	✓	✓	**Street View Image**	**Urban**

long-term observations and seasonal information at the individual-tree level. Although certain datasets span multiple years, each tree is typically recorded only once, limiting the ability to continuously monitor tree growth and phenological dynamics. Therefore, we introduce the ArborVerse dataset to fill these gaps.

1.2 Applications of street view imagery in urban forestry

Owing to its ability to provide fine-grained, eye-level information on trees at the street scale, street view imagery has increasingly emerged as a crucial data source for urban forestry research. Specifically, researchers use this direct street-level view by applying computer vision techniques such as semantic segmentation and object detection. These methods facilitate various applications, including quantifying urban greenery (Huang et al., 2025a), assessing biodiversity (Velasquez-Camacho et al., 2024), and detecting pests and diseases (Kagan et al., 2021). These results have also been utilized to further examine the role of vegetation in mitigating the urban heat island effect and improving street-level thermal environments (Huang et al., 2025c; Li et al., 2022).

Fine-grained tree species classification using street view imagery is a particularly complex task. Although pioneering work has demonstrated the feasibility of this approach (Branson et al., 2018; Wang et al., 2025), the majority of existing studies are confined to a limited set of species or small geographic regions, hindering the generalizability of their findings. While some efforts have aimed to develop larger-scale datasets to propel this field forward (Beery et al., 2022; McCoy et al., 2022), these resources still lack the necessary scale and comprehensiveness to adequately support the fine-grained classification and management demands of urban forestry.

https://doi.org/10.5194/essd-2026-380
Preprint. Discussion started: 21 October 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

2 ArborVerse Dataset

2.1 Dataset construction

The ArborVerse dataset provides a comprehensive collection of street view imagery alongside multi-level taxonomic annotations for individual urban street trees across the globe. The dataset construction follows a multi-stage pipeline comprising street view retrieval and viewpoint alignment, tree-mask extraction, image filtering, and large-scale screening with a lightweight binary classifier, followed by instance-level dataset partitioning. The complete workflow is illustrated in Figure 2.

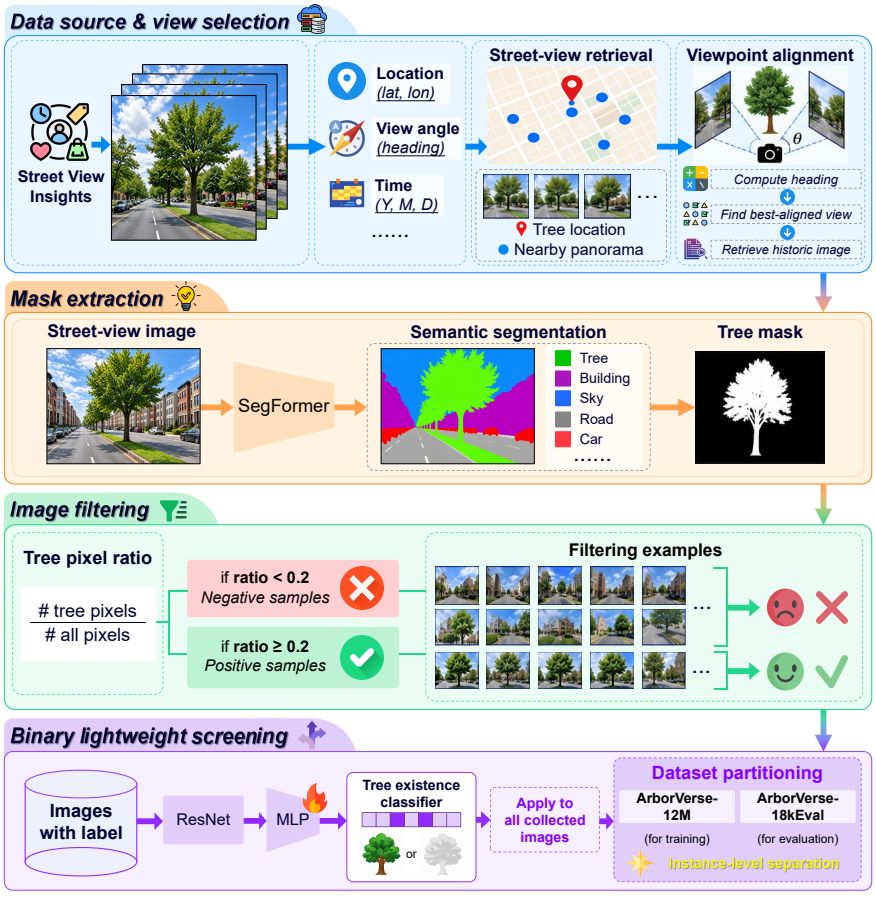


Figure 2. Overview of the dataset construction and filtering pipeline. Street view images are retrieved from Google Street View Insights using tree locations, viewing angles, and temporal information, followed by viewpoint alignment. SegFormer is employed to extract tree masks, and a lightweight binary classifier further removes irrelevant images, followed by instance-level separation to construct the training and evaluation sets.

https://doi.org/10.5194/essd-2026-380
Preprint. Discussion started: 21 October 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

2.1.1 Taxonomic and geolocation data

To achieve broad geographic and taxonomic coverage, we aligned and integrated data from multiple sources. These sources included recently published academic datasets (e.g., TreeML (Yazdi et al., 2024), U.S. 5Million (McCoy et al., 2022), and UrbanForest (City of Melbourne Open Data Team, 2023)), online resources (e.g., Singapore Trees (National Parks Board Singapore and ExploreTrees.SG project, 2025)), data applied from governmental agencies (e.g., Cambridge Tree Dataset), and the Global Biodiversity Information Facility (GBIF (Global Biodiversity Information Facility (GBIF), 2025)).

The GBIF dataset, in particular, required a specialized collection and filtering process due to its comprehensive global scale. To effectively query this extensive dataset, we first constructed a global tree species catalog based on two major inventories, TreeGOER (Kindt, 2023) and GlobalTreeSearch v1.9 (Beech et al., 2017), which includes 59,417 unique species. Subsequently, we used this catalog to query the GBIF database. To ensure the quality and reliability of the retrieved GBIF records, we applied a series of filtering criteria. We restricted the temporal range to observations recorded between 2015 and 2025 and limited the data source to human observations, minimizing interference from automated detections or specimen-based records. Furthermore, we excluded records with geospatial issues or inconsistencies, such as missing coordinates or mismatched country information, and retained only those records with an occurrence status of "present", excluding any trees marked as dead or absent.

After compiling the filtered data from all sources, we refined the consolidated dataset to focus exclusively on urban street trees. This was achieved through an overlay analysis using the Global Human Settlement Layer (GHSL) (ghs, 2021) data, which delineates built-up areas globally. We identified and retained only those trees located within these urbanized zones, ensuring the dataset focuses on urban environments. Finally, for records with incomplete taxonomic information, we further utilized our tree species catalog to impute the missing multi-level taxonomic labels. To resolve taxonomic synonyms and ensure Darwin Core compliance, we aligned all raw labels with standard GBIF species entries via the GBIF Taxonomic Backbone. Our labels were also verified against government data and institutional archives (e.g., museums), and we manually validated a random subset to guarantee high fidelity. For trees with a missing observation date, the publication date of the corresponding dataset was used instead.

2.1.2 Street view imagery

Using the geographic coordinates from the tree records, we queried Google Street View Insights to identify the nearest sampling point for each tree. We then retrieved all available temporal images associated with that sampling point, as a single location often provides imagery from multiple collection dates. Note that the temporal restriction applies only to the source tree records, whereas historical Street View imagery was retrieved without this restriction. To ensure that each view was centered on the target tree, we computed the azimuth angle from the capture point toward the tree location and used this orientation to select the street view perspective.

Uncertainties during the image collection process can lead to invalid images, which may suffer from issues such as poor illumination (low or excessive), occlusion by buildings or vehicles, or camera angle deviations. These issues often result in the target tree being missing or blurred in the image. To filter out such invalid samples, we designed a two-stage filtering procedure.

https://doi.org/10.5194/essd-2026-380
Preprint. Discussion started: 21 October 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

In the first stage, we randomly sampled 10,000 images from the entire street view dataset and applied the SegFormer (Xie et al., 2021) for semantic segmentation to compute the pixel percentage of the "tree" class. An image was classified as a negative sample (tree absent or unclear) if this percentage was below 20%, and as a positive sample otherwise (see Figure 3).

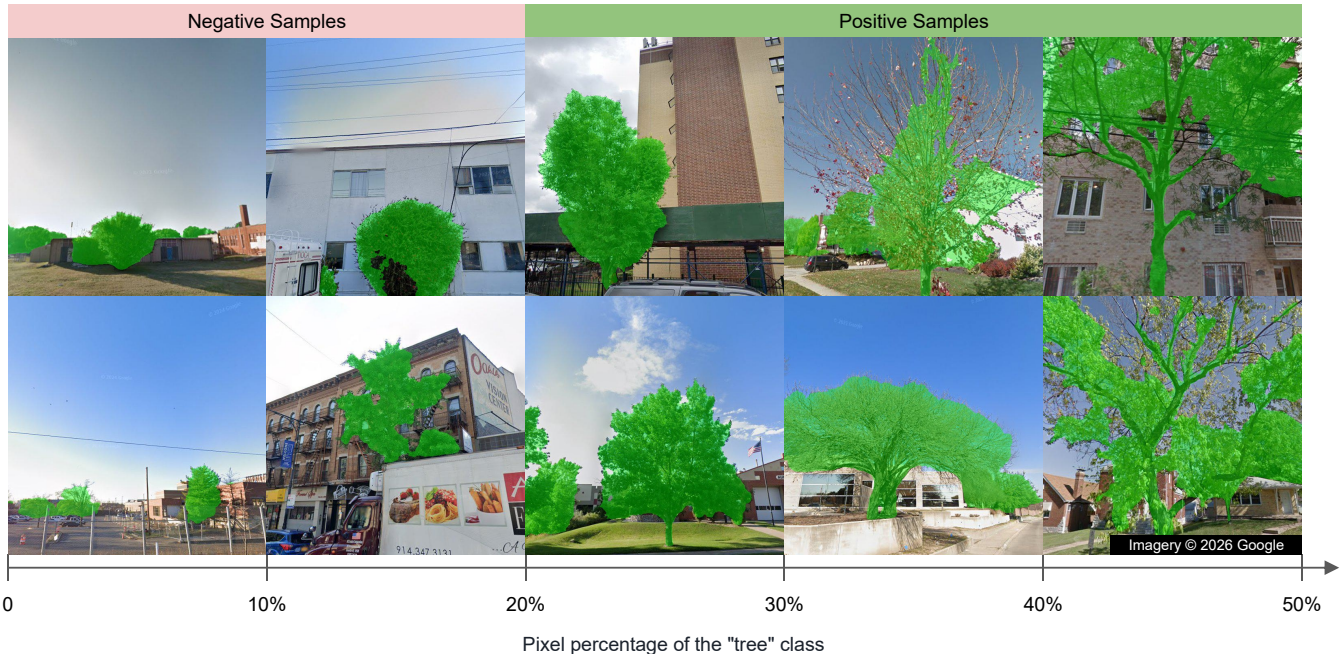


Figure 3. Examples of the binary classification of street view images. Images with less than 20% "tree" class pixels are labeled as negative samples, whereas those exceeding 20% are labeled as positive samples. Segmentation masks of "tree" class are highlighted with green color. Image source: © Google Street View 2026.

In the second stage, we trained an efficient binary classifier to automate this filtering process for the entire dataset. To enhance classification efficiency, we did not use the raw images; instead, we first extracted deep features using a ResNet18 (He et al., 2016) model pretrained on the Places365 (Zhou et al., 2018) dataset. These features were then used to train the classifier, which consisted of a two-layer fully connected multilayer perceptron (MLP). The model was trained for 30 epochs with a batch size of 4096 and a learning rate of 1e-3, using a 9:1 training-validation split. On the validation set, it achieved accuracy of 0.9590 and recall of 0.8214. Finally, this binary classifier was applied to predict and filter the remaining street view images, retaining only the positive samples. Manual inspection confirmed that sub-threshold images suffer from severe occlusion or lack discriminative texture details for classification; excluding these noisy samples prevents feature learning degradation and ensures high training efficacy.

https://doi.org/10.5194/essd-2026-380
Preprint. Discussion started: 21 October 2026
© Author(s) 2026. CC BY 4.0 License.

Open Access Earth System Science Data Discussions

2.2 Dataset description

The ArborVerse dataset contains 12,235,152 samples representing 3,365,485 individual urban street trees across 132 countries. Each record consists of three key components: (1) the taxonomic and spatial attributes of the tree; (2) the corresponding street view image; and (3) the metadata associated with that image. Specifically, the taxonomic and spatial attributes of the tree include its geographic coordinates (latitude and longitude), its location (country and city), the observation date, and a four-level taxonomic hierarchy (order, family, genus, and species; e.g., Order: Sapindales; Family: Sapindaceae; Genus: Acer; Species: Acer rubrum). The street view image metadata includes the latitude and longitude of the sample point, the year and month of image capture, and the corresponding season.

We analyze the key distributional characteristics of the ArborVerse dataset, focusing on its taxonomic composition, seasonal distribution, and temporal coverage. As can be seen from Figure 4 (a)-(d), the sample distribution exhibits a pronounced long-tailed structure across all four taxonomic levels (order, family, genus, and species). Statistical indicators show that at the genus level, skewness is 12.27 and kurtosis is 198.46; at the species level, skewness reaches 14.71 with a kurtosis of 292.03. These values reflect extremely strong positively-skewed and heavy-tailed distributions, in which a small number of high-frequency classes account for a large proportion of samples, while the vast majority of classes appear only rarely. This structure mirrors the inherent unevenness of real-world urban tree distributions and implies that models may naturally bias toward frequent classes while facing generalization challenges on rare categories.

For the seasonal distribution, Figure 4 (e) shows a relatively balanced yet slightly uneven seasonal composition. Spring, summer, autumn, and winter account for 25.3%, 31.6%, 28.0%, and 15.1%, respectively. This distribution reflects the phenological cycles throughout the year and ensures that the dataset captures substantial seasonal variation in tree appearance, including leaf color, bark texture, and growth status. Overall, the diversity of seasonal observations enables models to learn season-dependent visual patterns more effectively and enhances their ability to adapt to appearance variation in real-world environments.

Regarding the temporal distribution, ArborVerse demonstrates substantial long-term observational depth in Figure 4 (f). A total of 921,723 trees have observation records spanning from 1 to 10 years, and 1,411,615 trees have records spanning from 11 to 18 years. Even trees without multi-year histories may include observations from different seasons within a year. This temporal span not only covers a wide range of longitudinal time scales but also provides valuable historical records for studying long-term patterns of urban tree growth dynamics.

Together, these three complementary dimensions reveal the richness and complexity of the ArborVerse dataset's statistical structure: a naturally long-tailed distribution in the species dimension, pronounced intra-species morphological variation across seasons, and extensive long-term observational depth over time. These characteristics capture the complexity of urban ecological systems and offer diverse and meaningful signals for visual learning across multiple scales, while simultaneously imposing higher demands on model design, class-balancing strategies, and generalization ability.

https://doi.org/10.5194/essd-2026-380
Preprint. Discussion started: 21 October 2026
© Author(s) 2026. CC BY 4.0 License.

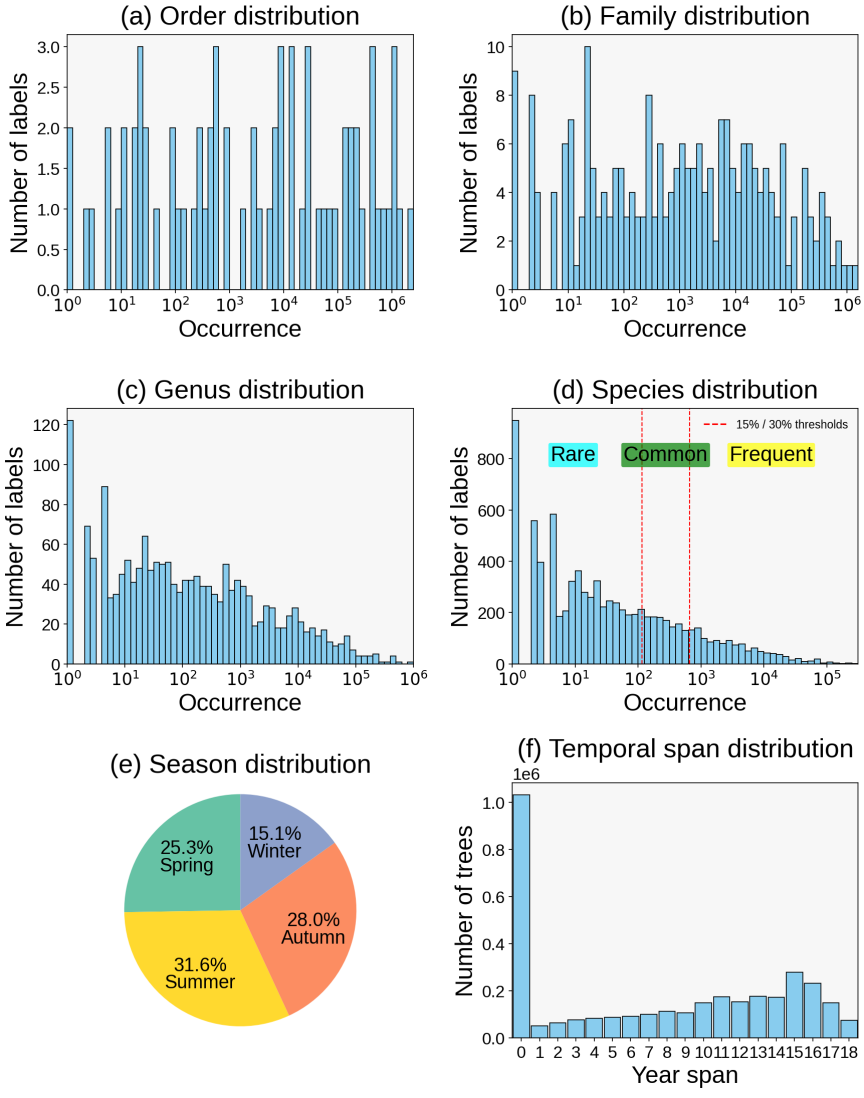


Figure 4. Statistics of ArborVerse dataset. **(a)–(d)** show distributions across the four taxonomic levels: Order, Family, Genus, and Species. The horizontal axis uses a logarithmic scale to represent the magnitude of occurrence counts; **(e)** presents the seasonal distribution; **(f)** illustrates the temporal span distribution of individual trees.

https://doi.org/10.5194/essd-2026-380
Preprint. Discussion started: 21 October 2026
© Author(s) 2026. CC BY 4.0 License.



2.3 Dataset partitioning

To enable effective model training and evaluation, the ArborVerse dataset was divided into two parts: the training set **ArborVerse-12M** and the evaluation set **ArborVerse-18kEval**.

ArborVerse-12M contains the vast majority of samples and is used for fine-tuning large-scale vision models such as Vision Transformer (ViT) (Dosovitskiy et al., 2020), CLIP (Radford et al., 2021), SigLIP (Zhai et al., 2023) and BioCLIP (Stevens et al., 2024). The extensive volume of data allows these models to learn rich visual and contextual representations of tree species and their surrounding urban environments.

ArborVerse-18kEval serves as a benchmark designed to assess model performance in a fair and robust manner across both taxonomic levels and species-frequency categories. To address the long-tailed distribution of species, all species were ranked by their frequency of occurrence and grouped into three categories: frequent, common, and rare. Specifically, frequent species correspond to the top 15% of species by occurrence frequency, common species represent the next 15–30%, and rare species include the remaining 70% (see Figure 4 (d)). Rather than exhaustively covering all species, we adopt frequency-stratified sampling to capture representative challenges across frequent, common, and rare species. This design enables controlled evaluation of model performance under varying data-frequency regimes while maintaining a compact and computationally manageable benchmark. To prevent potential data leakage where different images of the same individual tree might accidentally appear in both the training and evaluation sets, leading the model to memorize specific instances rather than learn generalizable species features, we enforced a strict instance-level separation. For models trained on varying data proportions, we dynamically adjusted the evaluation set. Any test images belonging to individual trees present in the training subset were systematically excluded. To preserve the size and balanced distribution, these excluded samples were replaced with images of different individual trees from the corresponding species. Its geographic distribution closely aligns with that of the full training dataset, broadly spanning multiple regions around the globe and ensuring strong global representativeness. The detailed taxonomic composition of the evaluation set is provided in Table B1.

3 Benchmarks

3.1 Baseline models and experimental setup

To evaluate the performance of current mainstream model architectures in large-scale tree species classification, we compared five models that differ in both training set and architectural design: ViT fine-tuned on ArborVerse-12M, CLIP pretrained by OpenAI, CLIP fine-tuned on ArborVerse-12M, SigLIP fine-tuned on ArborVerse-12M and BioCLIP fine-tuned on ArborVerse-12M. We aim to establish a benchmark for global urban street tree species classification in order to reveal the limitations and potential improvements of existing architectures, thereby enhancing our understanding of urban tree biodiversity across the world.

https://doi.org/10.5194/essd-2026-380
Preprint. Discussion started: 21 October 2026
© Author(s) 2026. CC BY 4.0 License.



3.1.1 Fine-tuned ViT

We initialized the ViT model with weights pretrained on ImageNet (Deng et al., 2009) and then fine-tuned it on the ArborVerse-12M dataset. Specifically, all parameters of the backbone were unfrozen during this process, and the model was trained to simultaneously predict the four hierarchical taxonomic levels: order, family, genus, and species.

3.1.2 CLIP

We used the CLIP visual encoder as a feature extractor for the ArborVerse-12M dataset. The entire CLIP backbone was kept frozen during training. Only the classification heads were trained to predict the four hierarchical taxonomic levels: order, family, genus, and species.

3.1.3 Fine-tuned CLIP / Fine-tuned SigLIP / Fine-tuned BioCLIP

To explore the efficacy of cross-modal alignment, we fine-tuned several CLIP-based architectures. First, we fine-tuned both the visual and text encoders on the ArborVerse-12M dataset, using the images and their corresponding hierarchical textual labels. Then, these fine-tuned encoders were frozen. The visual encoder was then used as a feature extractor to train four classification heads for the hierarchical taxonomic levels (order, family, genus, and species).

3.1.4 Data efficiency experiment

To evaluate the data utilization efficiency of different models, we designed a data efficiency experiment. While the total size of the ArborVerse-12M set reaches tens of millions of samples, we hypothesize that competitive performance can still be achieved even with a small fraction of the data. Specifically, we constructed four training subsets containing 0.1% (12K samples), 1% (120K samples), 10% (1.2M samples), and 100% of the ArborVerse-12M dataset, respectively. To account for sampling variability, each of the three smaller subsets (0.1%, 1%, and 10%) was independently sampled five times using different random seeds. This experiment analyzes how model performance varies with the amount of available training set under identical conditions, thereby allowing us to assess the robustness and generalization capability of different models under data-scarce scenarios.

3.1.5 Training details

All models were trained using PyTorch (Paszke et al., 2019). The ViT model was fine-tuned for 10 epochs on four NVIDIA A100 GPUs (40 GB each). We used the Adam optimizer (Loshchilov and Hutter, 2019) with a learning rate of 5×10^{-4} and a total batch size of 256. For the CLIP, SigLIP, and BioCLIP fine-tuning, we used four NVIDIA L4 GPUs (24 GB each) and the AdamW optimizer (weight decay 0.01) with a per-GPU batch size of 128. The first stage (contrastive learning) ran for 5 epochs with a learning rate warm-up from 1×10^{-5} to 2.5×10^{-4}. The second stage (classifier training) ran for 10 epochs at a constant learning rate of 2.5×10^{-4}. Both full training runs required approximately two days to complete.

https://doi.org/10.5194/essd-2026-380
Preprint. Discussion started: 21 October 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

3.2 Experimental results and evaluation

For all models, predictions were performed on the ArborVerse-18kEval benchmark, dynamically adjusted for each training subset to maintain strict instance-level separation. Model performance was assessed using Top-1 and Top-5 accuracy across the four taxonomic levels (order, family, genus, and species). To offer a granular understanding of how class imbalance affects learning, the mean accuracies are systematically grouped by species-frequency categories (*frequent*, *common*, and *rare*) and detailed in Table 2. Additionally, the overall mean accuracies and standard deviations across 5 random runs are summarized in Table 3 to quantify general model performance and stability. Note that only a single run was performed for models trained on the full training set.

Table 2. Model evaluation on ArborVerse-18kEval. Results are presented as mean accuracy (%) over 5 random runs.

Model	Frequent				Common				Rare			
	Order	Family	Genus	Species	Order	Family	Genus	Species	Order	Family	Genus	Species
0.1% of training set												
Fine-tuned ViT	17.52 (52.31)	13.26 (34.56)	6.33 (20.16)	1.81 (5.76)	12.34 (38.11)	9.47 (24.22)	4.16 (12.35)	1.43 (3.28)	7.82 (25.33)	5.14 (15.29)	2.13 (7.41)	1.16 (1.82)
CLIP	20.35 (60.62)	16.58 (41.25)	5.21 (21.28)	1.54 (5.56)	14.28 (45.18)	11.31 (30.12)	3.14 (14.22)	1.27 (3.18)	9.17 (30.16)	7.18 (18.11)	1.56 (8.12)	1.08 (1.52)
Fine-tuned CLIP	20.06 (58.76)	15.16 (39.84)	6.08 (22.95)	1.76 (5.82)	13.82 (43.21)	10.43 (28.17)	4.07 (15.11)	1.41 (3.59)	8.87 (28.19)	6.42 (16.21)	2.18 (9.14)	1.19 (1.97)
Fine-tuned SigLIP	23.41 (63.26)	18.28 (44.31)	8.42 (25.18)	1.52 (7.82)	16.56 (48.11)	12.87 (33.22)	5.86 (18.81)	1.93 (5.16)	11.18 (33.65)	8.41 (20.17)	3.27 (11.18)	1.46 (2.81)
Fine-tuned BioCLIP	**26.17 (66.42)**	**21.18 (48.21)**	**11.23 (28.14)**	**2.87 (9.11)**	**19.12 (52.19)**	**15.18 (37.51)**	**7.82 (21.18)**	**1.81 (6.87)**	**13.48 (37.11)**	**10.56 (24.12)**	**4.83 (14.16)**	**1.87 (3.92)**
1% of training set												
Fine-tuned ViT	21.85 (58.41)	17.31 (42.49)	10.45 (27.42)	3.32 (12.05)	15.48 (42.17)	11.82 (29.51)	6.81 (18.13)	1.87 (7.82)	10.72 (28.18)	7.46 (18.68)	4.13 (11.19)	1.81 (4.13)
CLIP	20.18 (60.45)	16.52 (41.81)	6.98 (24.42)	1.78 (8.71)	14.07 (44.18)	11.17 (29.35)	4.19 (16.43)	0.93 (5.11)	9.16 (29.78)	6.82 (18.16)	2.84 (9.81)	1.38 (2.87)
Fine-tuned CLIP	22.21 (61.22)	16.65 (43.02)	8.92 (26.18)	3.98 (11.65)	16.18 (45.11)	11.52 (30.18)	5.82 (17.81)	1.72 (7.58)	10.82 (30.34)	7.18 (19.12)	3.19 (10.82)	1.86 (3.81)
Fine-tuned SigLIP	25.28 (64.18)	20.22 (47.48)	12.31 (30.49)	5.08 (14.83)	18.33 (48.68)	14.07 (34.47)	8.16 (21.84)	2.82 (9.98)	12.16 (33.67)	9.19 (22.53)	5.33 (13.07)	1.48 (5.92)
Fine-tuned BioCLIP	**28.17 (67.14)**	**23.94 (51.26)**	**15.81 (34.01)**	**7.40 (17.63)**	**21.04 (52.35)**	**16.32 (38.44)**	**10.66 (25.55)**	**4.56 (12.81)**	**15.67 (37.93)**	**11.64 (26.30)**	**6.58 (16.22)**	**2.50 (7.02)**
10% of training set												
Fine-tuned ViT	30.71 (69.45)	24.82 (53.05)	17.25 (38.28)	10.85 (24.95)	22.41 (52.18)	17.08 (38.37)	11.39 (26.33)	6.41 (16.23)	15.16 (36.24)	11.05 (25.36)	6.82 (16.30)	3.18 (9.76)
CLIP	22.01 (62.31)	17.52 (44.38)	8.68 (27.14)	4.32 (13.98)	15.18 (45.12)	12.07 (31.18)	5.14 (18.27)	2.76 (8.54)	10.93 (30.25)	7.82 (20.40)	2.86 (11.01)	1.82 (4.31)
Fine-tuned CLIP	27.88 (67.22)	22.05 (50.25)	14.72 (35.81)	10.35 (23.75)	19.34 (49.30)	15.38 (36.18)	9.12 (24.16)	5.46 (15.80)	12.86 (34.35)	9.82 (23.87)	5.05 (14.77)	2.82 (8.44)
Fine-tuned SigLIP	34.76 (72.36)	28.16 (56.80)	20.52 (42.00)	13.99 (28.32)	25.76 (55.43)	20.67 (41.06)	14.91 (30.47)	8.01 (19.24)	18.35 (40.14)	14.16 (28.64)	9.02 (19.82)	4.67 (11.49)
Fine-tuned BioCLIP	**38.16 (76.96)**	**32.31 (60.24)**	**24.85 (47.70)**	**16.13 (32.86)**	**29.23 (59.27)**	**24.02 (45.37)**	**17.14 (34.82)**	**11.31 (23.38)**	**21.32 (44.06)**	**17.62 (32.82)**	**11.31 (23.50)**	**6.19 (14.57)**
100% of training set												
Fine-tuned ViT	41.25 (75.46)	33.15 (60.25)	25.60 (45.84)	17.16 (31.77)	28.09 (58.24)	22.51 (44.87)	16.64 (31.23)	9.32 (21.62)	20.85 (42.88)	15.16 (31.10)	10.59 (21.22)	5.14 (13.79)
CLIP	24.99 (66.21)	20.11 (48.13)	11.88 (31.56)	8.67 (17.16)	18.25 (49.12)	14.24 (34.81)	7.43 (21.66)	3.48 (11.58)	12.40 (33.02)	9.43 (23.15)	4.07 (14.62)	1.28 (6.13)
Fine-tuned CLIP	42.31 (78.77)	35.16 (64.83)	28.25 (50.34)	23.77 (36.02)	31.03 (61.81)	25.27 (48.05)	18.30 (35.16)	11.55 (24.79)	23.35 (45.50)	18.94 (35.54)	12.25 (24.12)	6.16 (15.27)
Fine-tuned SigLIP	42.92 (78.06)	36.86 (66.43)	31.41 (57.05)	28.42 (48.47)	39.19 (73.52)	31.34 (60.94)	25.73 (50.33)	23.17 (41.85)	37.51 (71.59)	23.30 (59.76)	23.43 (46.63)	18.29 (38.37)
Fine-tuned BioCLIP	**46.41 (81.12)**	**40.30 (69.00)**	**37.70 (63.28)**	**30.26 (52.07)**	**41.67 (78.09)**	**38.69 (68.81)**	**30.96 (52.63)**	**27.15 (46.01)**	**40.42 (77.20)**	**31.64 (64.12)**	**28.27 (48.95)**	**20.60 (41.76)**

3.2.1 Comparative analysis of model performance

Our comparative analysis reveals distinct performance dynamics among the models, heavily influenced by both the architecture and data scale. Across all data scales, the specialized vision models, particularly Fine-tuned BioCLIP and Fine-tuned

https://doi.org/10.5194/essd-2026-380
Preprint. Discussion started: 21 October 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

Table 3. Overall model evaluation. Results represent the mean accuracy (%) ± standard deviation (%) across all frequency categories over 5 random runs. The experiments using 100% of training set do not include a standard deviation.

Model	Top-1				Top-5			
	Order	Family	Genus	Species	Order	Family	Genus	Species
0.1% of training set								
Fine-tuned ViT	16.37 ± 0.58	12.39 ± 0.30	5.84 ± 0.88	1.73 ± 0.22	49.15 ± 1.24	32.27 ± 0.22	18.49 ± 0.64	5.23 ± 0.55
CLIP	19.01 ± 0.28	15.43 ± 1.12	4.76 ± 1.28	1.48 ± 0.56	57.15 ± 0.70	38.71 ± 1.03	19.72 ± 1.46	5.05 ± 0.89
Fine-tuned CLIP	18.69 ± 0.44	14.12 ± 1.37	5.63 ± 0.19	1.69 ± 0.42	55.27 ± 0.66	37.20 ± 2.34	21.24 ± 0.87	5.34 ± 0.54
Fine-tuned SigLIP	21.91 ± 1.55	17.09 ± 0.18	7.84 ± 1.42	1.58 ± 0.67	59.86 ± 0.58	41.74 ± 1.46	23.70 ± 1.12	7.23 ± 0.69
Fine-tuned BioCLIP	**24.62** ± 0.79	**19.87** ± 0.94	**10.47** ± 0.38	**2.67** ± 0.91	**63.18** ± 0.85	**45.71** ± 1.06	**26.57** ± 1.32	**8.58** ± 0.44
1% of training set								
Fine-tuned ViT	20.46 ± 0.54	16.11 ± 0.46	9.66 ± 0.67	3.04 ± 0.27	54.82 ± 0.92	39.63 ± 0.74	25.40 ± 0.42	11.11 ± 0.65
CLIP	18.84 ± 0.33	15.34 ± 1.02	6.40 ± 0.63	1.63 ± 0.82	56.84 ± 0.64	39.04 ± 1.35	22.66 ± 1.82	7.94 ± 0.72
Fine-tuned CLIP	20.87 ± 0.66	15.52 ± 0.70	8.24 ± 0.85	3.55 ± 0.36	57.63 ± 0.62	40.18 ± 0.48	24.33 ± 0.34	10.74 ± 1.32
Fine-tuned SigLIP	23.74 ± 0.61	18.87 ± 0.78	11.42 ± 0.49	4.60 ± 0.85	60.70 ± 1.03	44.58 ± 2.31	28.53 ± 1.04	13.76 ± 0.79
Fine-tuned BioCLIP	**26.62** ± 1.22	**22.31** ± 1.32	**14.68** ± 0.44	**6.78** ± 0.96	**63.81** ± 0.76	**48.39** ± 0.68	**32.07** ± 1.18	**16.51** ± 0.58
10% of training set								
Fine-tuned ViT	28.87 ± 0.42	23.13 ± 1.57	15.97 ± 0.68	9.89 ± 0.32	65.59 ± 0.80	49.79 ± 0.56	35.64 ± 1.52	23.05 ± 0.85
CLIP	20.55 ± 0.44	16.33 ± 1.14	7.92 ± 0.76	3.99 ± 0.62	58.50 ± 0.32	41.48 ± 1.33	25.19 ± 0.38	12.79 ± 0.36
Fine-tuned CLIP	26.02 ± 0.55	20.58 ± 0.49	13.50 ± 0.84	9.32 ± 0.31	63.27 ± 2.33	47.13 ± 0.91	33.25 ± 1.55	21.97 ± 0.68
Fine-tuned SigLIP	32.78 ± 0.76	26.50 ± 1.63	19.24 ± 1.20	12.72 ± 0.85	68.59 ± 0.79	53.35 ± 1.43	39.42 ± 0.76	26.31 ± 1.31
Fine-tuned BioCLIP	**36.18** ± 1.34	**30.50** ± 1.06	**23.17** ± 0.84	**15.03** ± 0.61	**73.05** ± 1.42	**56.96** ± 0.76	**44.84** ± 0.39	**30.74** ± 0.57
100% of training set								
Fine-tuned ViT	38.47	30.85	23.67	15.51	71.63	56.83	42.69	29.55
CLIP	23.50	18.82	10.91	7.59	62.38	45.18	29.42	15.90
Fine-tuned CLIP	39.88	33.04	26.13	21.23	74.96	61.18	47.05	33.54
Fine-tuned SigLIP	42.14	35.53	30.24	27.25	77.12	65.33	55.63	47.08
Fine-tuned BioCLIP	**45.45**	**39.76**	**36.31**	**29.45**	**80.50**	**68.81**	**61.10**	**50.76**

SigLIP, consistently demonstrated superior performance compared to general-purpose architectures like Fine-tuned ViT and the zero-shot CLIP. Even in the extremely low-data scenario (0.1% subset), Fine-tuned BioCLIP exhibited powerful few-shot generalization, achieving 26.17% and 21.18% Top-1 accuracy at the order and family levels for the *frequent* class, outperform-

https://doi.org/10.5194/essd-2026-380
Preprint. Discussion started: 21 October 2026
© Author(s) 2026. CC BY 4.0 License.


ing all other models. As the training data scaled up to 100%, finetuned-bioclip maintained its significant lead, achieving the highest accuracies across all taxonomic levels and frequency categories.

3.2.2 Impact of data scale on model performance

Model performance on our task is highly dependent on the training set volume. As the training set size increases, all models show consistent improvement. For instance, the Top-1 species accuracy of Fine-tuned BioCLIP on the *frequent* subset steadily climbs from 2.87% (on the 0.1% subset) to 7.40% (1%), 16.13% (10%) and 30.26% (100%). Crucially, the substantial performance jump from the 10% subset to the 100% subset indicates that model performance is far from saturated. This trend underscores the necessity of large-scale datasets, suggesting that further expanding the training data could continue to yield significant performance gains.

3.2.3 Fine-grained nature of the classification task

The difficulty of the task is highlighted by the large gap between Top-1 and Top-5 accuracies, even when using the full training set. For example, the best-performing Fine-tuned BioCLIP model (at 100% scale, *frequent* class) achieved high Top-5 accuracies of 81.12% (order), 69.00% (family), 63.28% (genus), and 52.07% (species). In sharp contrast, its Top-1 accuracies were significantly lower at 46.41%, 40.30%, 37.70%, and 30.26% respectively. This discrepancy reveals the fine-grained nature of the challenge: models can effectively narrow down the possibilities, but they struggle to make the final, precise decision among visually similar taxonomic groups.

3.2.4 Impact of class frequency on model performance

Our analysis reveals that model efficacy is heavily influenced by the sample distribution, with performance exhibiting a strict hierarchy across the *frequent*, *common*, and *rare* classes. Due to the abundance of training samples, models generally demonstrate the strongest capability to learn representative features from the *frequent* category, which acts as a relatively simpler task. Using the full training set, the Top-1 species accuracy of Fine-tuned BioCLIP strictly follows the *frequent* > *common* > *rare* pattern, dropping from 30.26% (*frequent*) to 27.15% (*common*) and ultimately to 20.60% (*rare*). Similar degradation is consistently observed across other models, including Fine-tuned SigLIP (28.42% > 23.17% > 18.29%) and Fine-tuned ViT (17.16% > 9.32% > 5.14%). This trend highlights a fundamental challenge in fine-grained classification: while models can comfortably master well-represented classes, their generalization ability degrades significantly on *common* and particularly *rare* species, underscoring the bottleneck imposed by long-tailed class imbalance.

4 Discussion

Our expanded evaluation reveals a nuanced performance dynamic contingent on data scale. At lower data scales, the single-modality Fine-tuned ViT remains highly competitive, often performing comparably to or even outperforming Fine-tuned CLIP in several taxonomic categories. This suggests that with limited data, ViT can efficiently extract visual supervisory signals.

https://doi.org/10.5194/essd-2026-380
Preprint. Discussion started: 21 October 2026
© Author(s) 2026. CC BY 4.0 License.



However, a significant shift occurs when scaling to the full training set: the fine-tuned cross-modal models (Fine-tuned CLIP, Fine-tuned SigLIP, and particularly Fine-tuned BioCLIP) comprehensively surpass Fine-tuned ViT. This highlights the latent power of large-scale vision-language pretraining, which seemingly requires massive data to fully bridge the modal gap and adapt to fine-grained tasks. Specifically, Fine-tuned BioCLIP, which incorporates biological and taxonomic knowledge during pretraining, demonstrates the strongest capability across almost all scales, proving that domain-specific cross-modal alignment is highly effective for parsing the complex visual hierarchy of tree species. Although the four-level taxonomic labels used as supervision currently carry limited semantic information, the superiority of cross-modal architectures at scale is evident. In future work, incorporating large language models (LLMs) to generate more structured and semantically rich, human-readable descriptions of tree species, thereby replacing simple label concatenation with stronger semantic signals, may further unlock the potential of these models (Hou et al., 2025; Feng et al., 2025). However, even when trained on the full dataset with the best-performing BioCLIP, the Top-1 accuracy remains insufficient for practical deployment. This limitation arises not only from the models' capacity but also from the inherent complexity of real-world urban tree classification: many species exhibit high inter-species visual similarity, while the same species can show substantial intra-species variation across seasons, cities, and climatic conditions; in addition, street view imagery is strongly affected by lighting, occlusion, camera viewpoint, and scene noise (Huang et al., 2025a), which introduces considerable uncertainty into visual appearance. In fact, even tree experts would find it difficult to make fully accurate species-level judgments from a single street view image in the absence of multi-view information or key details. Therefore, the limited Top-1 accuracy is a faithful reflection of the intrinsic difficulty of the task, and also indicates that current vision models still face clear performance bottlenecks when performing fine-grained classification in diverse urban environments.

The ArborVerse dataset itself plays an important role in advancing urban ecological research and urban governance. ArborVerse provides unprecedented global coverage, rich seasonal information, multi-temporal records, and fine-grained four-level taxonomic labels, offering a high-quality, easily accessible, and well-structured data foundation for building more powerful computer vision and deep learning models, as well as a standardized benchmark for large-scale tree-species classification. Compared with previous tree datasets that were often fragmented, limited in scale, or lacking temporal depth, ArborVerse has higher information density and more comprehensive content, enabling researchers to conduct large-scale experiments more conveniently and systematically improve model performance in complex urban environments. Such data-driven approaches have the potential to substantially reduce the labor and time costs of urban tree inventory, health monitoring, and fine-grained management, and to promote a transition toward more automated, scalable, and intelligent urban forest management.

Despite these advantages, the ArborVerse dataset still has several limitations: the distribution of tree species across global cities is highly uneven, and data from some regions remain sparse. Future work may further enrich data sources, improve the accuracy of taxonomic and geographic annotations, and incorporate multimodal information to enhance the completeness and reliability of the dataset; explore more advanced visual models and multimodal large language models to improve fine-grained feature representation and cross-modal alignment; and develop methods to mitigate the impact of long-tailed distributions and seasonal fluctuations, while strengthening research on model interpretability. Ultimately, applying ArborVerse in real-

https://doi.org/10.5194/essd-2026-380
Preprint. Discussion started: 21 October 2026
© Author(s) 2026. CC BY 4.0 License.



world urban forest management and biodiversity monitoring scenarios may help drive the development of urban ecological governance.

5 Conclusion

This study introduces **ArborVerse**, the largest and most comprehensive global dataset of urban street trees to date, together with the rigorously designed evaluation set **ArborVerse-18kEval**, providing a high-quality foundation for global research on fine-grained tree species classification. The more than 12 million samples in ArborVerse span 132 countries, offering complete four-level taxonomic annotations, rich seasonal information, and long-term temporal records, thereby greatly enhancing the availability and representativeness of data for urban tree research. Through evaluations conducted on ViT, CLIP, SigLIP and BioCLIP models, we identify the key challenges of large-scale, fine-grained, and real-world tree classification, offering valuable insights for future model development and methodological advances. Furthermore, ArborVerse is posed to drive progress in urban forest monitoring: with its high-resolution, multi-season, and spatially aligned global street view imagery, the dataset provides a robust basis for more accurate, scalable, and automated urban tree inventory and ecological analysis. We believe that ArborVerse will serve as a crucial bridge between computer vision, deep learning, and urban ecology, supporting the future development of data-driven urban green management.

Code and data availability. The ArborVerse dataset was constructed using Google Street View Insights under the current terms applicable at the time of submission. Under the Google Street View Insights *Asset Management* use case, Google's permitted-use documentation explicitly includes "green infrastructure" among supported assets and permits "creating reports and datasets summarizing the findings" of the corresponding asset analyses.[1] All access to and processing of these contents, including model training and evaluation, were conducted within Registered GCP Projects using authorized Google Cloud Platform products, services, and environments, consistent with Sections 3.4 and 3.6.1 of the current *Google Maps Platform Analytics Service Specific Terms*.[2] All models trained using Street View Insights content were retained within the authorized Google Cloud Platform environment.

The public release is separated from Google Maps Content. The released taxonomic and geospatial annotations originate from independent sources, including TreeML, U.S. 5 Million, UrbanForest, Singapore Trees, GBIF, and other governmental, institutional, and publicly accessible datasets, rather than from predictions or information inferred from Street View Insights imagery. These source datasets are subject to their respective licensing terms, including CC0 1.0, CC BY 4.0, and CC BY-NC 4.0. The released dataset does not contain personally identifiable or sensitive information. In accordance with the applicable Google Maps Platform terms, ArborVerse does not redistribute street view imagery, embeddings, or model weights.

The ArborVerse dataset is publicly available on Figshare at https://doi.org/10.6084/m9.figshare.32172486.v3 (Li et al., 2026). It is released as a comprehensive metadata index rather than a collection of raw image files. Specifically, the released dataset includes **unique panorama identifiers (PanoIDs), geographical coordinates (latitude and longitude), observation time (year and month), local season, country of origin, complete multi-level taxonomic information (order, family, genus, and species), and species frequency categories (*frequent*,**

[1] Google Street View Insights, *Permitted Use Cases*; last updated October 16, 2026.
[2] Google Maps Platform, *Analytics Service Specific Terms*, Sections 3.4 and 3.6.1; last updated September 2, 2026.

https://doi.org/10.5194/essd-2026-380
Preprint. Discussion started: 21 October 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

common*, or *rare), together with predefined training and evaluation splits. The underlying street view imagery itself is not redistributed. Researchers seeking access to the corresponding imagery must obtain it independently through Google Maps Platform services for which they are authorized and under the terms applicable to their own use. For panorama identifiers specifically, Google's published panorama-ID caching policy states that the panorama ID is "exempt from the caching restriction" and that panorama ID values may be stored indefinitely.[3] By providing these panorama IDs together with the corresponding image retrieval parameters, we enable researchers to retrieve the same street view imagery used in our experiments.

To use ArborVerse, researchers first obtain the released benchmark metadata and accompanying code. With independently authorized access to the applicable Google Maps Platform services, researchers can retrieve the corresponding street view observations using the released metadata, and the retrieved images can then be organized according to the predefined benchmark splits and used with the evaluation code to reproduce the results. This workflow enables researchers to follow a consistent dataset construction and evaluation protocol without redistributing Google Maps Content. The source code for this study is available from the corresponding author upon reasonable request.

Together, the released metadata, fixed benchmark splits, and accompanying code provide a stable specification of the benchmark and its evaluation protocol across releases. The underlying Street View imagery remains accessed through Google Maps Platform under the applicable terms. We are committed to the continued maintenance and expansion of ArborVerse, including addressing user feedback, improving data quality, and integrating additional high-value data sources to further support research on urban trees.

All policy statements here reflect the current Google Maps Platform policies available at the time of this preprint.

Appendix A: ArborVerse dataset statistics

The ArborVerse dataset is a large-scale and fine-grained global tree species classification resource. This chapter presents a systematic and in-depth description of its statistical properties and structural components.

A1 Basic statistics

ArborVerse offers georeferenced tree observations accompanied by a complete four-level taxonomic hierarchy and other structured attributes. The scale and coverage of the dataset are summarized in Table A1.

Table A1. Basic statistics of the ArborVerse dataset.

Attribute	Value
Total samples	12,235,152
Countries coverage	132
Taxonomic span	71 orders, 241 families, 1,747 genera, and 8,363 species
Data source distribution	U.S. 5 Million (44.58%), GBIF (36.95%), SGTrees (16.20%), UrbanForest (1.52%), Cambridge (0.73%), and TreeML (0.01%)

[3] Google Street View Static API, *Policies and Attributions*, "Exceptions from caching restrictions"; last updated October 16, 2026.

https://doi.org/10.5194/essd-2026-380
Preprint. Discussion started: 21 October 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

A2 Long tail distribution analysis

The dataset exhibits a pronounced long-tailed distribution at both the genus and species levels, as illustrated in Figure A1. At the genus level, the top 20% of genera (349 in total) account for 97.85% of all samples. In contrast, 460 genera contain no more than 10 samples each, forming a substantial long-tail of rare categories.

A similar pattern emerges at the species level: the top 20% of species (1,672 in total) contribute 97.70% of the dataset, whereas 3,333 species have 10 samples or fewer. This highlights an even more extreme long-tailed distribution at the species level.

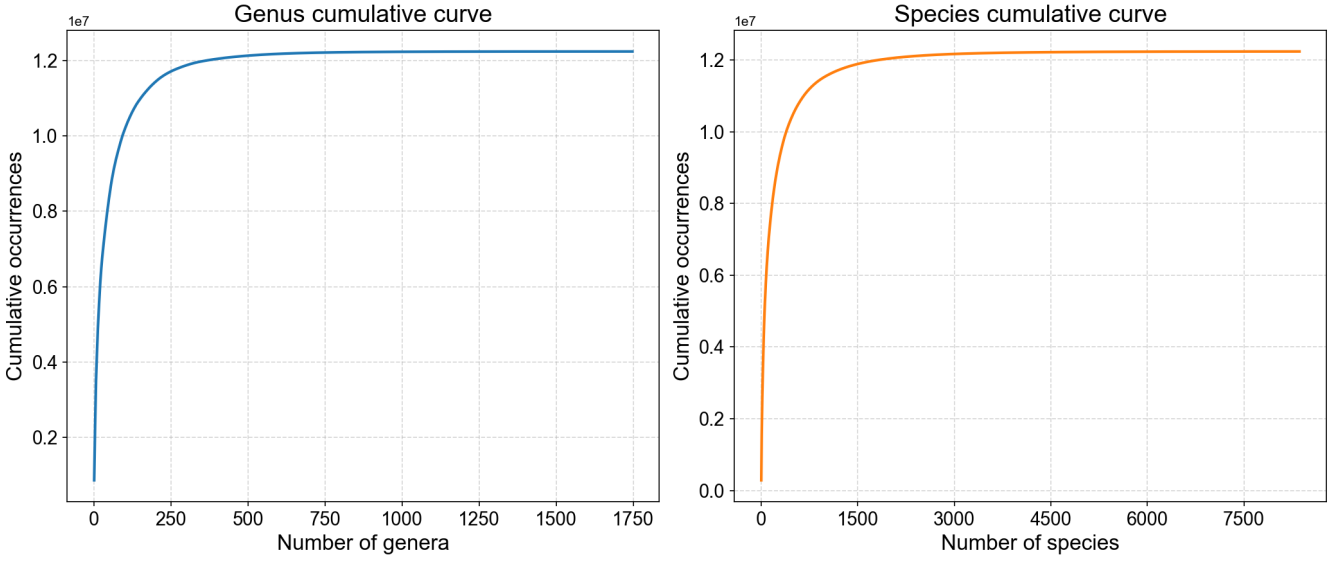


Figure A1. Long-tail distribution across genus and species level in ArborVerse dataset.

This long-tailed distribution reflects high global ecological imbalance in the abundance of urban tree species, implying that models tend to learn features of frequent species more readily while exhibiting limited recognition performance on rare species. ArborVerse-18kEval explicitly accounts for the impact of such uneven species distributions: by stratifying species according to their frequency category and applying balanced sampling, the evaluation set achieves both representativeness and uniformity across frequency categories.

A3 Detailed statistics of category attributes

This chapter consolidates the statistical characteristics of the ArborVerse dataset across several key attributes, encompassing its geographic coverage as well as the composition and diversity of its four-level taxonomic structure. Together, these analyses provide a comprehensive depiction of the dataset's overall distributional landscape.

https://doi.org/10.5194/essd-2026-380
Preprint. Discussion started: 21 October 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

A3.1 Geographic coverage

The 20 most represented countries in the dataset are shown in Table A2.

Table A2. Top 20 countries by sample count in the ArborVerse dataset, ordered by descending sample count.

Country	Samples	Country	Samples
United States of America	7,607,849	United Kingdom of Great Britain and Northern Ireland	88,666
Singapore	1,993,414	South Africa	73,120
France	521,830	Belgium	59,099
Australia	395,828	Portugal	55,300
Italy	222,089	Brazil	52,429
Spain	207,863	Colombia	29,234
Chinese Taipei	173,021	India	27,565
Canada	144,023	Austria	24,243
New Zealand	114,313	Chinese Hong Kong	24,026
Mexico	95,923	Czechia	20,827

A3.2 Distribution of order

The sample distribution of the top 20 most represented orders is visualized in Figure A2.

A3.3 Distribution of family

The sample distribution of the top 20 most represented families is visualized in Figure A3.

A3.4 Distribution of genus

The sample distribution of the top 20 most represented genera is visualized in Figure A4.

A3.5 Distribution of species

The sample distribution of the top 20 most represented species is visualized in Figure A5.

A4 Real-world challenges and complexities in the dataset

Although the ArborVerse dataset has undergone extensive cleaning and quality checks to achieve best performance, it still confronted the substantial complexities inherent to real-world street view images. In the raw street view images, trees are fre-

https://doi.org/10.5194/essd-2026-380
Preprint. Discussion started: 21 October 2026
© Author(s) 2026. CC BY 4.0 License.

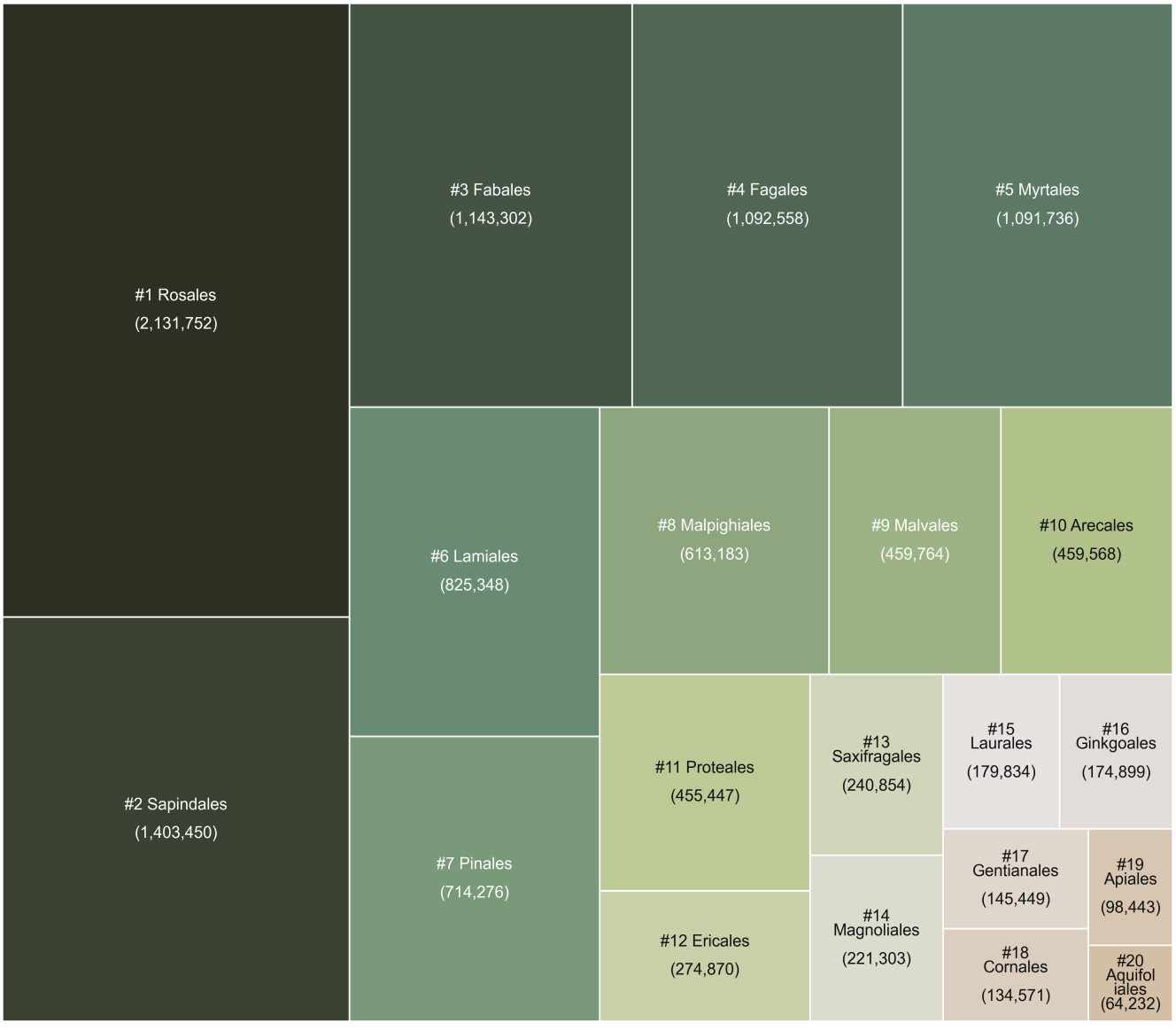


Figure A2. The sample distribution among the 20 most frequent orders. Larger areas indicate higher sample counts, highlighting the dominant orders within the dataset.

https://doi.org/10.5194/essd-2026-380
Preprint. Discussion started: 21 October 2026
© Author(s) 2026. CC BY 4.0 License.

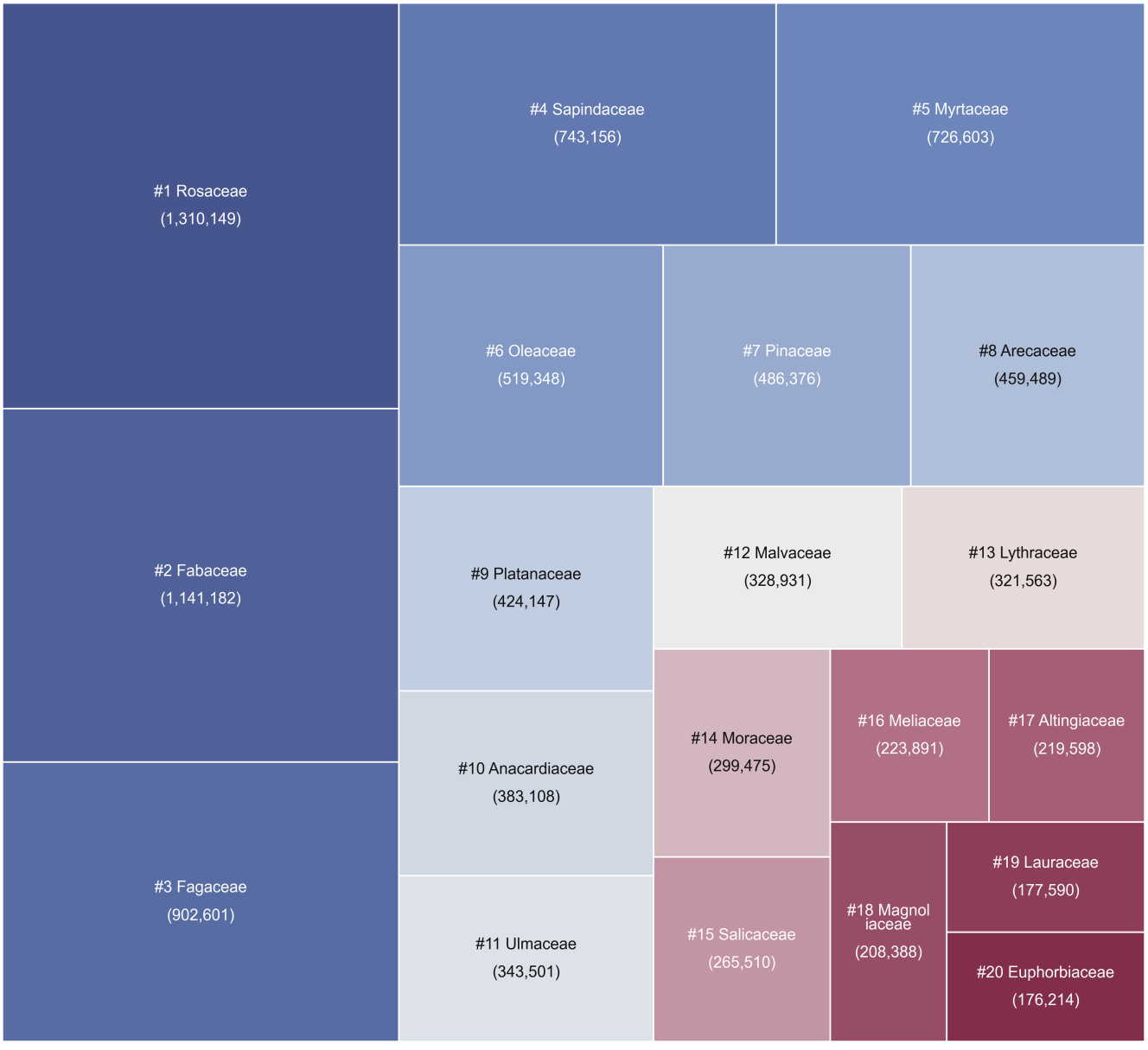


Figure A3. The sample distribution among the 20 most frequent families.

https://doi.org/10.5194/essd-2026-380
Preprint. Discussion started: 21 October 2026
© Author(s) 2026. CC BY 4.0 License.

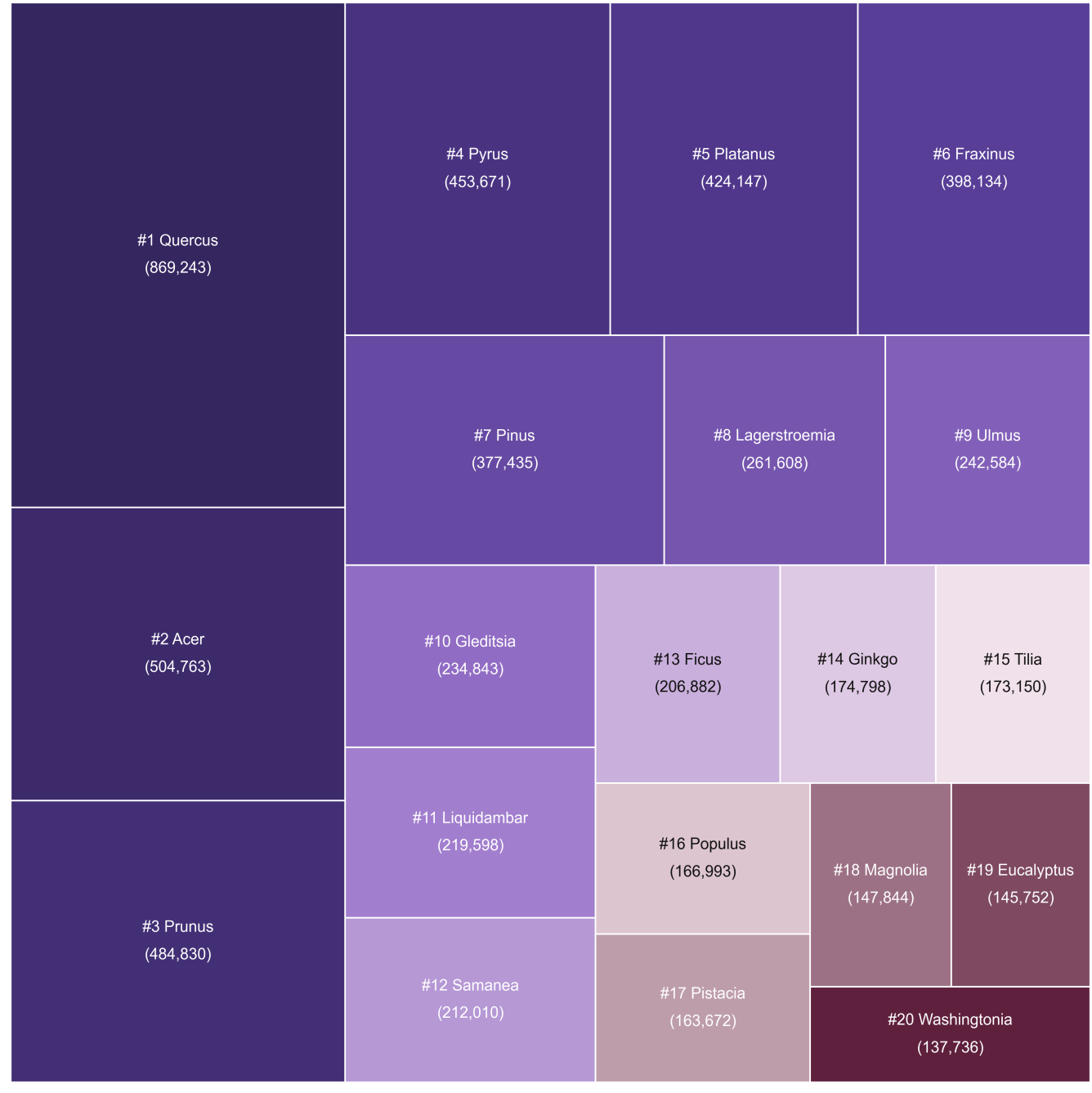


Figure A4. The sample distribution among the 20 most frequent genera.

https://doi.org/10.5194/essd-2026-380
Preprint. Discussion started: 21 October 2026
© Author(s) 2026. CC BY 4.0 License.

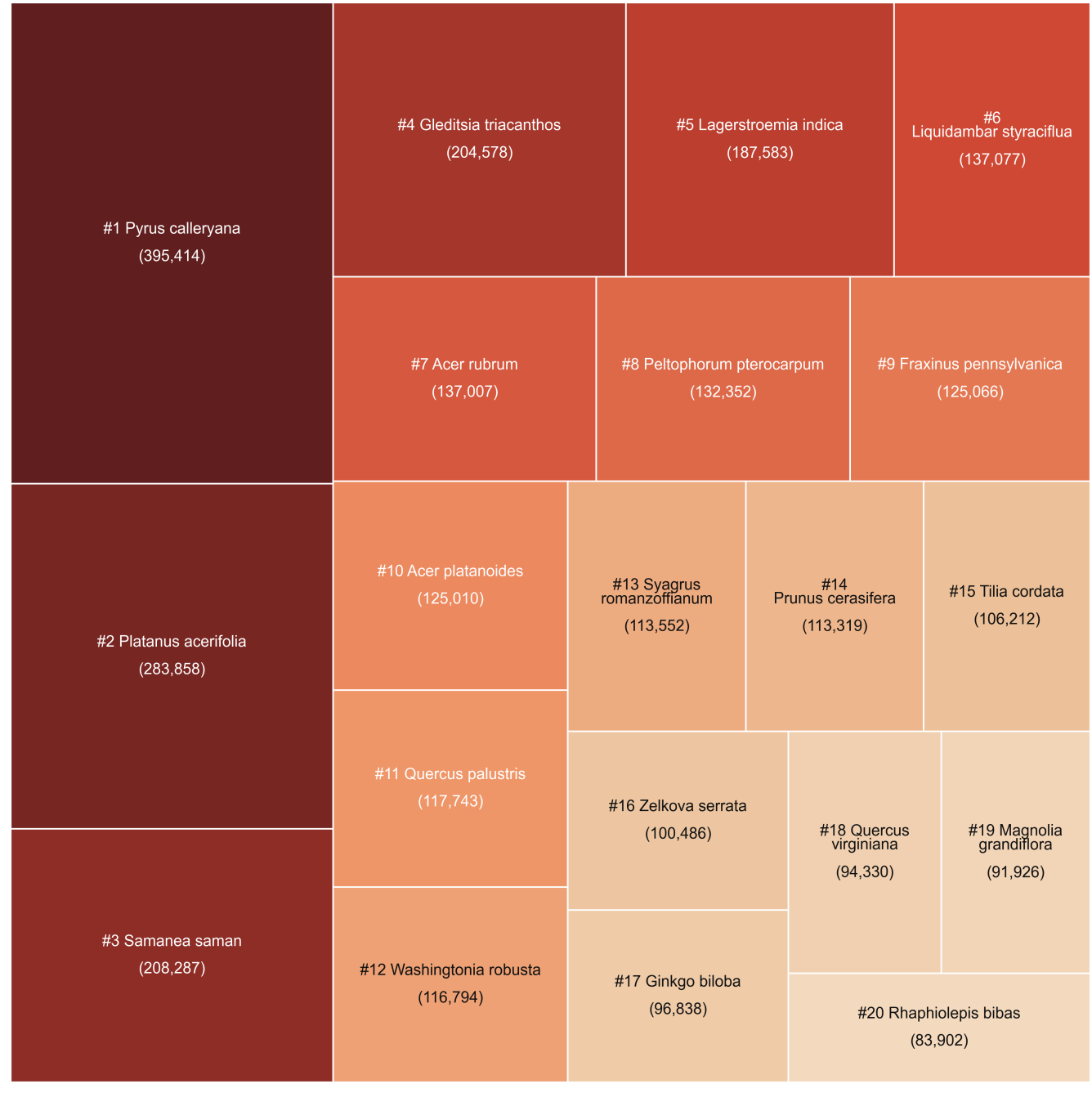


Figure A5. The sample distribution among the 20 most frequent species.

https://doi.org/10.5194/essd-2026-380
Preprint. Discussion started: 21 October 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

quently affected by a range of uncontrolled visual factors, including occlusions caused by street infrastructures, excessively strong or weak lighting conditions, intra-species variability driven by seasonal changes, tree misalignment arising from imprecise camera angles and visually ambiguous boundaries among highly similar species. These issues collectively increase the difficulty of data filtering process and pose new challenges to model training. Below, we provide representative examples of these issues in the raw data to illustrate the underlying challenges involved in the ArborVerse dataset.

A4.1 Occlusions caused by street infrastructures

Tree crowns in street view images are frequently blocked by buildings, traffic signs or other urban structures, resulting in incomplete or visually distorted observations. Such occlusions make it difficult for models to capture the full morphological characteristics of the tree. A portion of such data has already been filtered out during the data cleaning process, and Figure A6 presents several examples.



Figure A6. Example of occlusions caused by street infrastructures. Image source: © Google Street View 2026.

A4.2 Excessively strong or weak lighting conditions

Extreme lighting, either overly bright or insufficient, can obscure fine-grained visual cues such as leaf texture or branch structure. These illumination variations introduce substantial noise into feature extraction and species recognition. Figure A7 presents several examples.

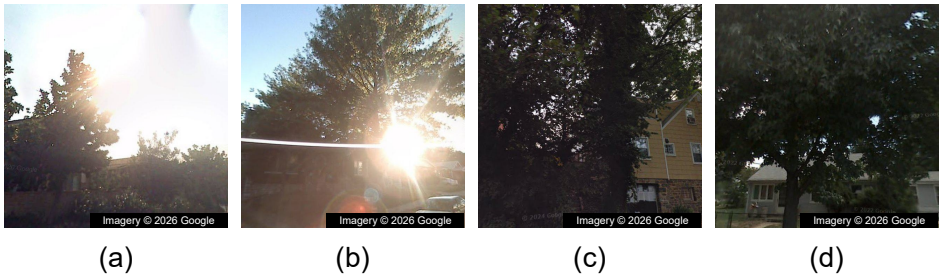


Figure A7. Example of excessively strong or weak lighting conditions. **(a)** and **(b)** illustrate cases of excessively strong lighting, while **(c)** and **(d)** show cases of insufficient illumination. Image source: © Google Street View 2026.

https://doi.org/10.5194/essd-2026-380
Preprint. Discussion started: 21 October 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

A4.3 Intra-Species variability driven by seasonal changes

Seasonal transitions cause dramatic changes, like Figure A8. Differences in crown and branch morphology, leaf color, and overall tree appearance can lead to large intra-species variability, which complicates the learning of stable discriminative features.

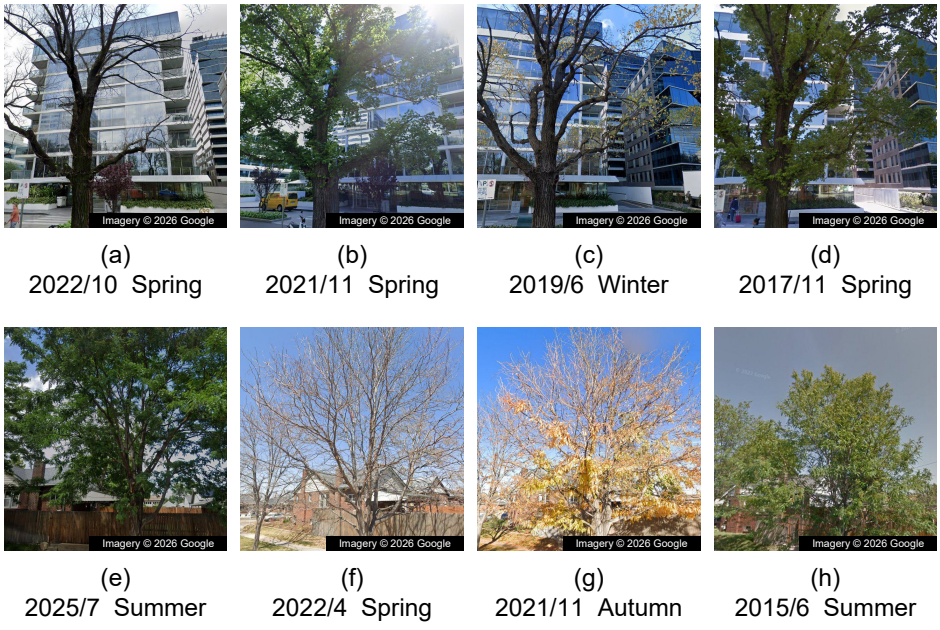
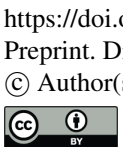


Figure A8. Example of intra-species variability driven by seasonal changes. **(a)-(d)** and **(e)-(h)** illustrate the seasonal progression observed in two individual trees, illustrating their seasonal variability. **(a)** corresponds to early spring, where the tree has recently emerged from dormancy and the branches remain bare. **(b)** and **(d)** show the tree in mid to late spring, characterized by dense and deep green leaves. **(c)** appears to depict early winter or a late autumn state where most leaves have fallen, leaving only a few yellow leaves clinging to the branches. For another tree, a similar pattern is observed. **(e)** and **(h)** were captured in summer, showing full, vibrant green leaves. **(f)** represents spring, during which the branches are completely bare. **(g)** reflects an autumn condition, with abundant yellow leaves signaling seasonal transition. Image source: © Google Street View 2026.

A4.4 Tree misalignment arising from imprecise camera angles

Inaccuracies in camera orientation or GPS alignment may capture the wrong tree or only part of intended target. These misalignments reduce labeling precision and introduce unintended visual context, see Figure A9.

https://doi.org/10.5194/essd-2026-380
Preprint. Discussion started: 21 October 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions



Figure A9. Example of tree misalignment arising from imprecise camera angles. **(a)-(d)** illustrate how inaccurate camera angle cause the target tree to appear too low, too high, too far to the left, or too far to the right in the captured image respectively. Image source: © Google Street View 2026.

A4.5 Visually ambiguous boundaries among highly similar species

Many tree species share extremely similar visual traits, leading to ambiguous boundaries even in high-quality images. Such fine-grained similarities pose inherent challenges for reliable species-level classification. Figure A10 presents several species of visually similar yet taxonomically different.

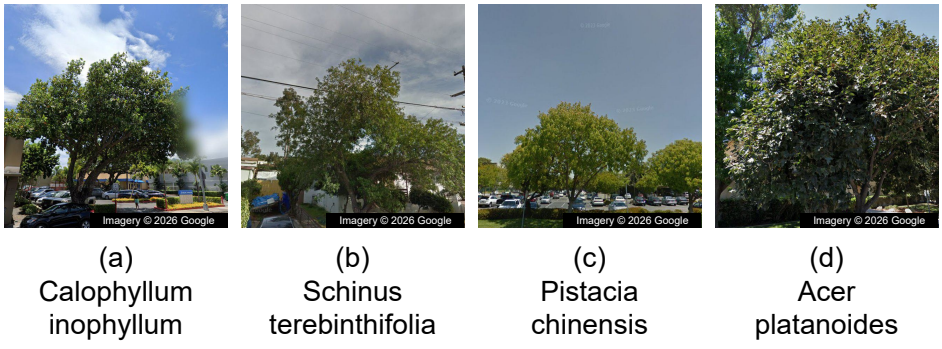


Figure A10. Example of visually ambiguous boundaries among highly similar species. The species name of each tree is shown below corresponding image. Image source: © Google Street View 2026.

Appendix B: Further details of evaluation set

To enable fair and representative benchmarking across different taxonomic levels and species, we constructed an independent evaluation set, ArborVerse-18kEval. We first partitioned all species in the full ArborVerse dataset into three frequency categories: frequent, common, and rare, based on their observed sample counts. From each frequency group, we then randomly selected 30 species. Finally, we randomly sampled 500, 100, and 20 images from the frequent, common, and rare species respectively, thereby forming an evaluation set with a controlled gradient of species frequency. This design enables rigorous

https://doi.org/10.5194/essd-2026-380
Preprint. Discussion started: 21 October 2026
© Author(s) 2026. CC BY 4.0 License.



assessment of a model's generalization ability across species with markedly different frequency categories. Table B1 below summarizes the taxonomic composition of the evaluation set.

Table B1. Taxonomic composition of ArborVerse-18kEval dataset.

Order	Family	Genus	Species	Sample count
Frequent species				
Pinales	Pinaceae	Picea	Picea pungens	500
Rosales	Rosaceae	Malus	Malus sylvestris	500
Laurales	Lauraceae	Machilus	Machilus japonica	500
Fagales	Fagaceae	Quercus	Quercus ilicifolia	500
Myrtales	Myrtaceae	Melaleuca	Melaleuca linariifolia	500
		...		
Total		(A total of 30 frequent species were selected)		15,000
Common species				
Malpighiales	Salicaceae	Salix	Salix bonplandiana	100
Fabales	Fabaceae	Dalbergia	Dalbergia cochinchinensis	100
Malvales	Malvaceae	Durio	Durio zibethinus	100
Magnoliales	Myristicaceae	Knema	Knema globularia	100
Lamiales	Oleaceae	Notelaea	Notelaea microcarpa	100
		...		
Total		(A total of 30 common species were selected)		3,000
Rare species				
Aquifoliales	Aquifoliaceae	Ilex	Ilex micrococca	20
Ericales	Ericaceae	Kalmia	Kalmia latifolia	20
Celastrales	Celastraceae	Euonymus	Euonymus europaeus	20
Laurales	Lauraceae	Nectandra	Nectandra oppositifolia	20
Dipsacales	Viburnaceae	Viburnum	Viburnum opulus	20
		...		
Total		(A total of 30 rare species were selected)		600

https://doi.org/10.5194/essd-2026-380
Preprint. Discussion started: 21 October 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

Appendix C: Algorithm for azimuth angle calculation

After obtaining the geographic coordinates of each tree, we retrieve the corresponding street view images through Google Street View Insights. Since the camera usually captures panoramic imagery centered at the capture point, it is essential to specify an angle that directs the camera toward the target tree. To achieve this, we first search for available capture locations around each tree and compute the azimuth angle from the capture point to the tree, which is then used as the heading parameter α.

Specifically, let the capture point be located at (φ_1, λ_1) and the tree at (φ_2, λ_2), where φ denotes latitude and λ denotes longitude. Let $\Delta\lambda = \lambda_2 - \lambda_1$. The azimuth angle θ from the capture point to the tree is computed as:

$$\theta = \operatorname{atan2}\begin{pmatrix} \sin(\Delta\lambda)\cos\varphi_2, \\ \cos\varphi_1 \sin\varphi_2 - \sin\varphi_1 \cos\varphi_2 \cos(\Delta\lambda) \end{pmatrix} \tag{C1}$$

Since θ lies in the range $(-\pi, \pi]$, we convert it to degrees and map it to $[0, 360)$ to obtain the final heading α:

$$\alpha = \left(\theta \times \frac{180}{\pi} + 360\right) \bmod 360 \tag{C2}$$

By determining an appropriate capture point and its corresponding azimuth angle for each tree, we are able to precisely control the camera orientation and ensure that the tree appears within the main field of image. This procedure enables large-scale acquisition of high-quality street view images and provides the foundation for subsequent data-cleaning steps.

Author contributions. Jiapeng Li: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft preparation. Yingjing Huang: Methodology, Data curation, Validation, Writing – review and editing. Fan Zhang: Conceptualization, Supervision, Project administration, Funding acquisition, Writing – review and editing. Shengli Tao: Resources, Data curation, Writing – review and editing. Zhen Dong: Writing – review and editing. Jun Yang: Resources, Data curation, Writing – review and editing. Yu Liu: Conceptualization, Supervision, Funding acquisition, Writing – review and editing.

Competing interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements. This work was supported by the National Natural Science Foundation of China under Grant 42371468. This work was also supported by the High-performance Computing Platform of Peking University.

https://doi.org/10.5194/essd-2026-380
Preprint. Discussion started: 21 October 2026
© Author(s) 2026. CC BY 4.0 License.



References

Global Human Settlement Layer: Understanding Human Presence on Planet Earth : Science Based Information on Population and Human Settlements to Support Informed Policy Decisions, Publications Office of the European Union, ISBN 978-92-76-42061-3, 2021.

Ahlswede, S., Schulz, C., Gava, C., Helber, P., Bischke, B., Förster, M., Arias, F., Hees, J., Demir, B., and Kleinschmit, B.: *TreeSatAI Benchmark Archive*: a multi-sensor, multi-label dataset for tree species classification in remote sensing, Earth System Science Data, 15, 681–695, 2023.

Arevalo-Ramirez, T., Alfaro, A., Figueroa, J., Ponce-Donoso, M., Saavedra, J. M., Recabarren, M., and Delpiano, J.: Challenges for computer vision as a tool for screening urban trees through street-view images, Urban Forestry & Urban Greening, 95, 128 316, 2024.

Beech, E., Rivers, M., Oldfield, S., and Smith, P.: GlobalTreeSearch: The First Complete Global Database of Tree Species and Country Distributions, Journal of Sustainable Forestry, 2017.

Beery, S., Wu, G., Edwards, T., Pavetic, F., Majewski, B., Mukherjee, S., Chan, S., Morgan, J., Rathod, V., and Huang, J.: The Auto Arborist Dataset: A Large-Scale Benchmark for Multiview Urban Forest Monitoring Under Domain Shift, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21 294–21 307, 2022.

Bolyn, C., Lejeune, P., Michez, A., and Latte, N.: Mapping tree species proportions from satellite imagery using spectral–spatial deep learning, Remote Sensing of Environment, 280, 113 205, 2022.

Branson, S., Wegner, J. D., Hall, D., Lang, N., Schindler, K., and Perona, P.: From Google Maps to a Fine-Grained Catalog of Street Trees, ISPRS Journal of Photogrammetry and Remote Sensing, 135, 13–30, 2018.

City of Melbourne Open Data Team: Trees, with species and dimensions (Urban Forest), Online dataset, accessed: Jan. 16, 2025, 2023.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255, ISSN 1063-6919, 2009.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N.: An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale, in: ICLR, 2020.

Feng, J., Zhang, J., Liu, T., Zhang, X., Ouyang, T., Yan, J., Du, Y., Guo, S., and Li, Y.: CityBench: Evaluating the Capabilities of Large Language Models for Urban Tasks, in: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2, KDD '25, pp. 5413–5424, Association for Computing Machinery, New York, NY, USA, ISBN 979-8-4007-1454-2, 2025.

Gaydon, C. and Roche, F.: PureForest: A Large-Scale Aerial Lidar and Aerial Imagery Dataset for Tree Species Classification in Monospecific Forests, in: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 5895–5904, IEEE Computer Society, ISBN 979-8-3315-1083-1, 2025.

Global Biodiversity Information Facility (GBIF): Occurrence Download API, https://www.gbif.org/occurrence/, 2025.

He, K., Zhang, X., Ren, S., and Sun, J.: Deep Residual Learning for Image Recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, ISSN 1063-6919, 2016.

Hou, C., Zhang, F., Li, Y., Li, H., Mai, G., Kang, Y., Yao, L., Yu, W., Yao, Y., Gao, S., et al.: Urban sensing in the era of large language models, The Innovation, 6, 2025.

Huang, Y., Zhang, F., Gao, Y., Tu, W., Duarte, F., Ratti, C., Guo, D., and Liu, Y.: Comprehensive urban space representation with varying numbers of street-level images, Computers, Environment and Urban Systems, 106, 102 043, 2023.

Huang, Y., Sanatani, R. P., Liu, C., Kang, Y., Zhang, F., Liu, Y., Duarte, F., and Ratti, C.: No "true" greenery: Deciphering the bias of satellite and street view imagery in urban greenery measurement, Building and Environment, 269, 112 395, 2025a.

https://doi.org/10.5194/essd-2026-380
Preprint. Discussion started: 21 October 2026
© Author(s) 2026. CC BY 4.0 License.



Huang, Y., Zhang, F., Wu, L., and Liu, Y.: Measuring urban physical environments using image deep features, Cities, 166, 106 196, 2025b.

Huang, Z., Duan, L., Xu, Y., Yang, S., Lin, Z., Yue, H., and Yang, J.: Exploring the Influence of Urban Green Space and Urban Morphology on Urban Heat Islands Using Street View and Satellite Imagery, Scientific Reports, 15, 23 759, 2025c.

Kagan, D., Fuhrmann Alpert, G., and Fire, M.: Automatic Large Scale Detection of Red Palm Weevil Infestation Using Street View Images, ISPRS Journal of Photogrammetry and Remote Sensing, 182, 122–133, 2021.

Kampe, T. U., Johnson, B. R., Kuester, M. A., and Keller, M.: NEON: The First Continental-Scale Ecological Observatory with Airborne Remote Sensing of Vegetation Canopy Biochemistry and Structure, Journal of Applied Remote Sensing, 4, 043 510, 2010.

Kindt, R.: TreeGOER: A Database with Globally Observed Environmental Ranges for 48,129 Tree Species, Global Change Biology, 29, 6303–6318, 2023.

Li, B., Xing, H., Cao, D., Yang, G., and Zhang, H.: Exploring the Effects of Roadside Vegetation on the Urban Thermal Environment Using Street View Images, International Journal of Environmental Research and Public Health, 19, 1272, 2022.

Li, J., Huang, Y., Tao, S., Dong, Z., Yang, J., Liu, Y., and Zhang, F.: ArborVerse: Scaling Fine-Grained Urban Tree Species Classification to Millions of Trees Worldwide, https://doi.org/10.6084/m9.figshare.32172486.v4, 2026.

Lian, X., Liu, W., Liu, Z., Jiao, L., Shackleton, C. M., Li, B., Zhong, J., Yang, R., Fu, M., Zhang, H., Sumari, N. S., and Liu, Y.: Tree shortages in informal settlements across African cities, Nature Cities, 2025.

Loshchilov, I. and Hutter, F.: Decoupled Weight Decay Regularization, in: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019, 2019.

Mauri, A., Strona, G., and San-Miguel-Ayanz, J.: EU-Forest, a High-Resolution Tree Occurrence Dataset for Europe, Scientific Data, 4, 160 123, 2017.

McCoy, D. E., Goulet-Scott, B., Meng, W., Atahan, B. F., Kiros, H., Nishino, M., and Kartesz, J.: Species Clustering, Climate Effects, and Introduced Species in 5 Million City Trees across 63 US Cities, eLife, 11, e77 891, 2022.

Mu, Y., Xiong, Z., Wang, Y., Shahzad, M., Essl, F., van Kleunen, M., and Zhu, X. X.: GlobalGeoTree: A Multi-Granular Vision-Language Dataset for Global Tree Species Classification, 2025.

National Parks Board Singapore and ExploreTrees.SG project: Explore Trees in Singapore, https://exploretrees.sg/, 2025.

Oliveira, A. A. A. M., Wang, Z., and Hirata, R.: Locating Urban Trees Near Electric Wires Using Google Street View Photos: A New Dataset and a Semi-Supervised Learning Approach in the Wild, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4286–4294, 2022.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library, in: Advances in Neural Information Processing Systems, edited by Wallach, H., Larochelle, H., Beygelzimer, A., dAlché-Buc, F., Fox, E., and Garnett, R., vol. 32, Curran Associates, Inc., 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision, in: Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, edited by Meila, M. and Zhang, T., vol. 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, PMLR, 2021.

Schwaab, J., Meier, R., Mussetti, G., Seneviratne, S., Bürgi, C., and Davin, E. L.: The role of urban trees in reducing land surface temperatures in European cities, Nature Communications, 12, 6763, 2021.

https://doi.org/10.5194/essd-2026-380
Preprint. Discussion started: 21 October 2026
© Author(s) 2026. CC BY 4.0 License.



Smith, I. A., Li, D., Fork, D. K., Wellenius, G. A., and Hutyra, L. R.: Integrated tree canopy expansion and cool roofs can optimize air temperature and heat exposure reductions in Boston, Communications Earth & Environment, 6, 507, 2025.

Stevens, S., Wu, J., Thompson, M. J., Campolongo, E. G., Song, C. H., Carlyn, D. E., Dong, L., Dahdul, W. M., Stewart, C., Berger-Wolf, T., Chao, W.-L., and Su, Y.: BioCLIP: A Vision Foundation Model for the Tree of Life, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19 412–19 424, 2024.

Thapa, B., Darling, L., Choi, D. H., Ardohain, C. M., Firoze, A., Aliaga, D. G., Hardiman, B. S., and Fei, S.: Application of multi-temporal satellite imagery for urban tree species identification, Urban Forestry & Urban Greening, 98, 128 409, 2024.

Velasquez-Camacho, L., Merontausta, E., Etxegarai, M., and de-Miguel, S.: Assessing Urban Forest Biodiversity through Automatic Taxonomic Identification of Street Trees from Citizen Science Applications and Remote-Sensing Imagery, International Journal of Applied Earth Observation and Geoinformation, 128, 103 735, 2024.

Wang, S., Liu, C., Wei, S., and Tang, H.: Urban Tree Species Classification Using Multisource Satellite Remote Sensing Data and Street View Imagery, Geo-spatial Information Science, 2025.

Wegner, J. D., Branson, S., Hall, D., Schindler, K., and Perona, P.: Cataloging Public Objects Using Aerial and Street-Level Images - Urban Trees, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6014–6023, 2016.

Wu, Z., Zohner, C. M., Zhou, Y., Crowther, T. W., Wang, H., Wang, Y., Peñuelas, J., Gong, Y., Zhang, J., Zou, Y., Van Den Hoogen, J., and Fu, Y. H.: Tree Species Composition Governs Urban Phenological Responses to Warming, Nature Communications, 16, 3696, 2025.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P.: SegFormer: Simple and efficient design for semantic segmentation with transformers, in: Proceedings of the 35th International Conference on Neural Information Processing Systems, NIPS '21, Curran Associates Inc., Red Hook, NY, USA, ISBN 978-1-7138-4539-3, 2021.

Yazdi, H., Shu, Q., Rötzer, T., Petzold, F., and Ludwig, F.: A Multilayered Urban Tree Dataset of Point Clouds, Quantitative Structure and Graph Models, Scientific Data, 11, 28, 2024.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L.: Sigmoid Loss for Language Image Pre-Training, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11 975–11 986, 2023.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A.: Places: A 10 Million Image Database for Scene Recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, 40, 1452–1464, 2018.