The Auto Arborist Dataset: A Large-Scale Benchmark for Multiview Urban Forest Monitoring Under Domain Shift

Sara Beery*, Guanhang Wu†, Trevor Edwards†, Filip Pavetic†, Bo Majewski†, Shreyasee Mukherjee†, Stanley Chan†, John Morgan†, Vivek Rathod†, Jonathan Huang†
California Institute of Technology*  Google†

Abstract

Generalization to novel domains is a fundamental challenge for computer vision. Near-perfect accuracy on benchmarks is common, but these models do not work as expected when deployed outside of the training distribution. To build computer vision systems that truly solve real-world problems at global scale, we need benchmarks that fully capture real-world complexity, including geographic domain shift, long-tailed distributions, and data noise.

We propose urban forest monitoring as an ideal testbed for studying and improving upon these computer vision challenges, while working towards filling a crucial environmental and societal need. Urban forests provide significant benefits to urban societies. However, planning and maintaining these forests is expensive. One particularly costly aspect of urban forest management is monitoring the existing trees in a city: e.g., tracking tree locations, species, and health. Monitoring efforts are currently based on tree censuses built by human experts, costing cities millions of dollars per census and thus collected infrequently.

Previous investigations into automating urban forest monitoring focused on small datasets from single cities, covering only common categories. To address these shortcomings, we introduce a new large-scale dataset that joins public tree censuses from 23 cities with a large collection of street level and aerial imagery. Our Auto Arborist dataset contains over 2.5M trees and 344 genera and is >2 orders of magnitude larger than the closest dataset in the literature. We introduce baseline results on our dataset across modalities as well as metrics for the detailed analysis of generalization with respect to geographic distribution shifts, vital for such a system to be deployed at-scale.

1. Introduction

Urban forests provide critical benefits to the over 4B people living in urban areas worldwide [107]. They filter air and water, capture stormwater runoff, sequester atmospheric carbon dioxide, limit erosion and drought, and save energy in a variety of ways (e.g., by providing shade and thus reducing cooling costs and urban heat islands [100, 120, 138, 141]). In the US alone, urban forests cover 127M acres and produce ecosystem services valued at >$18B [106]. These forests make up the foundations of our urban ecosystems, and provide habitat for diverse urban wildlife and insect populations [44]. Urban forest monitoring, measuring the size, health and species distribution over time, allows us to (1) quantify ecosystem services including air quality improvement [20, 47], carbon sequestration [92, 106, 121], and benefits to public health [28, 47, 125, 125], (2) track damage from extreme weather events [8, 67, 99], and (3) target planting to improve robustness to climate change, disease and infestation [24, 64, 114, 115]. Further, lack of access to urban greenery is a key aspect of urban social inequality [56, 93, 104], including socioeconomic inequality [52, 73, 86] and racial inequality [21, 129]. Urban forest monitoring enables the quantification of this inequality and...
the pursuit of its improvement [22, 118].

To enable automated urban forest monitoring, we focus on the fundamental task of building a tree census (sometimes called a tree inventory). Due to their significant cost (a recent tree census in Los Angeles cost $2M and took 18 months [2]), tree censuses are typically conducted only by cities with the means and will to invest in these undertakings, and even then they are conducted rarely (e.g., once every 5-10 years). We seek to dramatically lower the cost of a tree census by using computer vision to help find, label, and monitor individual trees using a combination of street level and aerial imagery. An automated system could democratize access to urban forest monitoring, providing this valuable information to under-resourced cities that are already disproportionately affected by climate change [109].

While there have been prior works on urban tree species recognition from aerial [4, 5, 63, 78, 124, 137, 139, 140] or street level [95] imagery (or both, in a limited number of cases [27, 130]), a major limitation has been a lack of large-scale labeled datasets. To our knowledge, all prior works have focused on single or limited numbers of cities, and have included only the most common categories. We introduce the Auto Arborist dataset, a multiview urban tree classification dataset which, at 2.6 million trees is two orders of magnitude larger than those in prior work [27, 130] and contains 344 genera (and many more species). To build Auto Arborist, we draw on public tree censuses from 23 cities in the US and Canada and merge these public records with street level and overhead RGB imagery. As the first urban forest dataset to cover multiple cities, Auto Arborist allows for previously-impossible detailed analysis of generalization with respect to geographic distribution shifts, vital to building systems that scale. We propose a set of metrics to evaluate performance with respect to these geographic distribution shifts and show the strengths and weaknesses of typical deep learning models when applied to the Auto Arborist dataset.

Going beyond its immediate application to sustainability and conservation, Auto Arborist can serve as an important challenge benchmark for computer vision. There has been increasing interest in domain generalization, which is ubiquitous in real-world applications [16, 51, 82, 97]. For example, prior works have observed that a model for self-driving cars that can drive safely in San Diego may not work equally well in Seattle [50, 68, 69]. In terms of number of domains, objects, classes, and images, Auto Arborist presents a scale not seen in previous real-world domain generalization benchmarks [13–17, 82]; it focuses on detailed cross domain analysis, and incorporates multiple views and modalities.

To summarize, our main contributions are as follows:

- We develop a pipeline for combining public tree census data with street level and aerial imagery.
- We introduce the Auto Arborist dataset built using this pipeline — the largest dataset of its kind covering >2.6M trees, >9.1M images and 344 categories and first one of its kind to cover multiple cities (23 cities).
- We show that for Auto Arborist, geographic domain shift and the category imbalance are major factors in performance of typical deep learning models.
- We show that diverse geographic coverage is important for generalization to a novel city, and that both multiple views and multiple data modalities are beneficial.

### 2. Related work

Tree detection, localization, and taxonomic identification have been studied in multispectral aerial imagery [49, 78, 140], ground-level imagery [95, 101], and LIDAR [46, 70], with some recent approaches combining data across modalities [6, 27]. Costly high-resolution data collected from low-flying aircraft has been shown to improve performance [18, 110], but this data is not available for much of the world. Though past studies have considered tree detection and categorization, many of these have been limited by perspective (aerial vs ground-level views), geospatial coverage, and taxonomic coverage. Our work seeks to expand upon all three, providing a testbed for urban forest monitoring that is broad in scope and relies on datatypes which are available across cities (aerial and street level RGB imagery) to enable the development of general models and methods which can be used off-the-shelf in novel cities.

**Tree detection and localization from aerial data.** There have been successful, broad-scale studies of tree density, canopy cover, and individual tree delineation from aerial data [9, 25, 38, 40, 46, 62, 66, 77, 94, 108, 110, 112, 122, 132], including tree crown detection across sites from the US National Ecological Observatory Network (NEON) [132, 133], tree canopy mapping in urban forests in cities across the US.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Ground-level images</th>
<th>Aerial Images</th>
<th>Sites</th>
<th>Classes</th>
<th>Forest Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Registree [27]</td>
<td>46,321</td>
<td>28,678</td>
<td>1</td>
<td>40</td>
<td>Urban</td>
</tr>
<tr>
<td>Pasadena Re-ID [102]</td>
<td>6,141 (panoramas)</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>Urban</td>
</tr>
<tr>
<td>NEON Tree Evaluation [132]</td>
<td>0</td>
<td>25,949 (boxes)</td>
<td>22</td>
<td>1</td>
<td>Wilderness</td>
</tr>
<tr>
<td>IDTrees Classification 2017 [98]</td>
<td>0</td>
<td>613 (boxes)</td>
<td>1</td>
<td>9</td>
<td>Wilderness</td>
</tr>
<tr>
<td>IDTrees Classification 2020 [56]</td>
<td>0</td>
<td>452 (boxes)</td>
<td>2</td>
<td>34</td>
<td>Wilderness</td>
</tr>
<tr>
<td>Auto Arborist (Ours)</td>
<td>6,479,077</td>
<td>2,637,208</td>
<td>23</td>
<td>344</td>
<td>Urban</td>
</tr>
</tbody>
</table>

Table 1. Comparison of our dataset to previous tree detection and identification datasets. Note that for Pasadena Re-ID, there is only one class (“tree”) but the task is re-identification as opposed to categorization.
Table 2. Cities by region. The holdout city for each region is in bold.

<table>
<thead>
<tr>
<th>Region W (West)</th>
<th>Region C (Central)</th>
<th>Region E (East)</th>
</tr>
</thead>
<tbody>
<tr>
<td>City</td>
<td>Trees</td>
<td>Genera</td>
</tr>
<tr>
<td>San Francisco, CA</td>
<td>154,698</td>
<td>195</td>
</tr>
<tr>
<td>San Jose, CA</td>
<td>225,655</td>
<td>201</td>
</tr>
<tr>
<td>Cupertino, CA</td>
<td>15,300</td>
<td>104</td>
</tr>
<tr>
<td>Vancouver, BC</td>
<td>121,249</td>
<td>93</td>
</tr>
<tr>
<td>Seattle, WA</td>
<td>150,983</td>
<td>142</td>
</tr>
<tr>
<td>Surrey, BC</td>
<td>62,251</td>
<td>72</td>
</tr>
<tr>
<td>Santa Monica, CA</td>
<td>25,381</td>
<td>126</td>
</tr>
<tr>
<td>Los Angeles, CA</td>
<td>391,788</td>
<td>202</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total West</td>
<td>1,147,305</td>
<td>328</td>
</tr>
</tbody>
</table>

[93], and counting individual trees in Sub-Saharan Africa [25]. These methods rely on a diverse set of aerial data modalities, from low-resolution RGB or hyperspectral satellite data to high-resolution RGB, hyperspectral, and LiDAR data collected from low-flying aircraft and UAVs [18]. However, there are still open challenges in maintaining performance of methods in novel regions [98, 131, 134], and methods must be well-validated and possibly adapted for any novel region before use. Tree crown delineation in dense forests remains a challenge, leading to several studies (e.g., in sub-Saharan Africa [25]) focusing on low-tree-density regions or trees outside forests [43, 117]. Further, there is only so much that can be understood from an aerial view alone. A large amount of the woody vegetation in a forest is hidden under the tree canopy. Understory trees have been mapped with very-high-resolution UAV-collected data [32, 60, 61, 87], but this data is rarely available. Our combined approach allows us to use available ground level imagery to see under the canopy.

Tree taxonomic identification from aerial data. Automated tree identification in aerial data from satellite or low-flying aircraft, including RGB, hyperspectral, LiDAR, or some combination thereof, is well-studied in the remote sensing community. [49] is a thorough review of species classification from remote sensing data which notes the lack of studies considering large spatial extents. Many studies focus on predicting species occurrence, presence/absence, or abundance for a limited set of species [3, 23, 29–31, 33, 34, 41]. Detecting and categorizing a diverse individual trees presents further complexity [4, 26, 35, 36, 41, 42, 48, 48, 53, 55, 59, 65, 70–72, 74, 75, 80, 81, 83, 111, 113, 116, 128], and recently deep learning approaches have been shown to outperform more traditional methods on this task [39, 57, 63, 90, 114, 127, 142]. Generalization to novel regions is a known challenge with many of the proposed methods [110]. The IDTrees challenges [56, 98] were the first to propose a public benchmark for cross-site individual tree categorization, but provided limited labeled data (<1000 labeled trees from <=34 tree categories at 3 NEON wilderness forest plots). Further challenge arises when predicting species in an urban environment, where human intervention leads to a much higher diversity of tree species, with a much longer tail, than is seen in the wild [135]. For this reason, many studies of urban tree categorization focus only on common species [4, 5, 10, 11, 63, 78, 124, 137, 139, 140].

Tree detection and localization in the urban forest from ground-level data. Ground-level data, (e.g., from Google Street View [7], Mapillary [96], and iNaturalist [1]), have been identified as an important source of information for urban monitoring applications [19]. Automated measures of urban “greenness” and tree cover mapping from ground-level data have been proposed, with implications in social justice and public health [45, 88, 89, 119, 123]. Datasets such as Mapillary Vistas [105] and Cityscapes [37] facilitate semantic segmentation of urban categories, including vegetation, but do not provide instance-level information or fine-grained taxonomic labels. Similarly, most current computer vision studies of the urban forest focus on species-agnostic individual tree detection [76, 126] and localization [84, 95, 101–103] across multiple ground-level views of the same tree.

Tree taxonomic identification from combined aerial and ground-level data. Previous large-scale datasets that combine aerial and ground-level data, such as CVUSA [136], were designed for alternative tasks such as image geocalization. Several methods exist for combining aerial + ground-level data, with tree identification as a key application [85, 117]. Here, ground-level data can include RGB imagery, LiDAR, and even physical measurements such as tree diameter or hyperspectral signature [54], [130] and [27] proposed a system for identifying street trees using paired aerial and ground-level RGB imagery for urban forests and released a dataset of paired imagery for Pasadena. [6] proposed a class-agnostic tree detection method from aerial imagery and ground-level LiDAR. Recently, [79] used GNNs to map individual trees across aerial and ground-level community science imagery in forests. All of these prior works trained on a single city and could benefit from a much larger dataset such as ours.
3. The Auto Arborist Dataset

We have generated the largest, and most geographically diverse, computer-vision-ready multi-view dataset of urban trees to date. The Auto Arborist dataset contains 2,637,208 trees across 23 cities. Each tree is represented by a $512 \times 512$ pixel aerial image where each pixel is $5cm \times 5cm$, as well as up to three $768 \times 1152$ pixel street level images \cite{7} of the tree (for a total of 9,116,285 images in the dataset)\footnote{We are publishing all tree records (after curation/merging c.f. 3.1) and a subset of the imagery (verified to obtain consistent results to the full dataset) with personally identifiable information removed. For more information please visit https://google.github.io/auto-arborist}. To avoid taxonomic complexity arising from hybrid and sub-species when developing methods, we have chosen to focus on genus prediction (instead of species-level prediction) as our primary task and have confirmed with ecologists and city planners that a genus-level map would be highly useful as a first step. Our dataset includes 344 unique genera, with a real-world long-tailed class imbalance and unique class distribution for each city on the dataset (Figure 2).

3.1. Dataset curation

To curate Auto Arborist, we started from existing tree censuses which are provided by many cities online. For each tree census considered, we verified that the data contained GPS locations and genus/species labels, and was available for public use. This resulted in data from 23 cities which we then parsed into a common format, fixing common data entry errors (such as flipped latitude/longitude) and mapping groundtruth genus names (and their common misspellings) to a universal label map consisting of 344 categories. We also removed records with invalid genus names, such as “unidentified”. Aggregated into a single dataset, this process yielded localized records for $\sim$5M trees.

Figure 1 shows a map of the 23 selected cities as well as example imagery from the dataset. We partition the cities into three separate regions for evaluation purposes (discussed further in Section 4). Table 2 summarizes the contribution from each of the cities to the Auto Arborist dataset organized by these regions. For this “v1” version of Auto Arborist, we restrict our focus to the US and Canada, with a single genus prediction task. There is room for Auto Arborist to grow in tasks and geographic area: many public tree censuses contain additional metadata (e.g. tree age, health, and trunk diameter), and there are many more cities we might include both in the US and Canada, and globally.

We place our dataset in context with previously published tree classification datasets in Table 1, and emphasize the significantly enhanced scope in number of images across modalities, number of regions, and number of categories.

Extracting street level and aerial imagery. For each city, starting from the parsed tree census, we associate each tree census record to both street level and aerial images. For each tree in our dataset, we sample a $15m \times 15m$, $300 \times 300$ pixel RGB aerial image centered on the tree’s latitude and longitude. We consider all street level images taken within 2-10 meters of the record’s latitude and longitude, filtering out any images which do not meet all of the following criteria:

- Taken on or after Jan 1, 2018.
- Contains the base of the tree near the horizontal center of the image based on the projection of the tree’s latitude/longitude onto the image, based on estimated camera pose generated by the API.
- Contains a significant number of “tree” pixels based on a semantic segmentation model (to avoid cases when the tree has died or been removed, when possible) \textit{and} does not contain any “person” or “bike rider” pixels based on a semantic segmentation model (to remove personally identifiable information).

After filtering, we have 2.6M tree records, each of which is
associated with one aerial image and 1-3 street level images, along with a date and GPS location.

3.2. Challenging aspects of the Auto Arborist data

By matching street level images from existing public records rather than collecting groundtruth labels from scratch, we have been able to achieve a scale much larger than any previous datasets. As we show, scale is important for generalizing to novel cities (which is the ultimate goal). But using public records to generate data across cities also introduces a number of challenges.

Sources of noise and ambiguity. First, we address several known sources of noise and ambiguity in our dataset. See Figure 3 for examples illustrating the following points.

- **Label noise**: There is a known discrepancy between label accuracy of volunteer citizen scientists vs. experts (e.g. with a PhD in Ecology) [12], and there is also no ecologically-agreed-upon definition of tree vs. bush. Cities differ in their labeling protocol.
- **Presence noise**: Tree records in censuses can often be outdated. Specifically, depending on the amount of time since the data was originally collected, there is increasing possibility that trees will have been removed or have died, and new trees planted.
- **Location noise**: Different cities use different data collection protocols and different sensors, leading to discrepancy in the accuracy of the position readings (e.g. by GPS). We estimate visually that they are usually accurate within ∼3 meters.
- **Image quality**: Quality of aerial imagery varies for different cities. The primary tree in a street level image can sometimes be occluded — though we try to guard against this by removing images that are too far from the tree, sometimes vehicles block the tree from view. Qualitatively, access to multiple views frequently helps mitigate occlusion issues. Finally, deciduous trees vary in appearance across seasons, with leaves turning color and then dropping in the winter.
- **Unlabeled visible trees**: Trees on private property (e.g. yards) are not labeled in public censuses, but are visible in the background. While the tree of interest is often the most prominent, the presence of trees of other genera can create classification confusion.

Distribution shift and the long tail. One of our primary challenges is to be able to do well on novel cities that were not part of the training set, but in order for a model to do so, it will have to contend with distribution shift, where the training distribution of cities differs from the novel test distribution on some new city. We remark that there are two kinds of shift that we observe in our data — what we might call “label shift”, and “appearance shift”. Label shift refers to when the marginal distribution $P(y)$ of labels (genera) differs from city to city even if the appearance distribution of image $x$ conditioned on a particular label $P(x|y)$ does not change (e.g., [91]). In our setting this simply can mean that species distributions vary geographically (e.g. we tend to see Palm trees in Southern California and less in Canada) but can also come from cities having different sizes (for example, Los Angeles is much larger than Santa Monica and thus contains many more species).

Figure 4 visualizes the distribution shift between every pair of cities (using $L_1$ distance between normalized genus distributions). In some cases we can see little overlap be-

![Figure 4](image-url)
between genera from two cities, and for cities with similar location, i.e. Denver and Boulder, we tend to see high overlap in genus distribution. However even when two cities are very similar both in size and location, it is still generally the case that one city will contain a number of genera not found in the other due to the long tailed genus distribution. In the extreme setting of “train on one city, test on one city” we thus always have many test genera for which there are no training examples. And even in the regime of training on many cities and testing on a single holdout city we still typically have classes for which there are no training examples, implying value in expanding the dataset in future.

Beyond label shift, we also see “appearance shift” — the images of a particular genus can look different depending on the city. This is partly due to different backgrounds (which can in principle be handled by masking out the background pixels, but is out of scope for this work), but it can also be due to other external factors such as weather conditions (for example, we are likely to see more leafless trees from images in Edmonton than we are to see them in LA) or even “terroir” related factors like soil composition.

4. Evaluation Protocol

Since distribution shift is such a big factor in performance, we have chosen to set up our evaluation protocol to explicitly evaluate distribution shift based on 3 unique types of train/test splits, defined hierarchically:

1. **Per-city splits**: At the first level, we are interested in how well a city generalizes to itself. Here, each city has a defined training region and a defined test region, split geographically (usually based on latitude or longitude) to avoid overfitting on background characteristics. The test sets for each city are never used for training.

2. **Regional splits**: Next, we are interested in generalization within and across larger regions (e.g., how would we fare in cities on the East coast if we trained on West coast trees?) — for this level of evaluation, we split the cities into three regions, Region W (West), Region C (Central), and Region E (East) (Table 2). We build our regional training sets from the per-city training sets for that region. We hold out one city from each region (which we call “holdout cities”) to capture performance on an in-region novel city, and also show results on all out-of-region cities.

3. **Full dataset**: For the final and largest split, we combine training data across the three regions. We maintain the same holdout cities as the regional splits for training, and test on the test sets of all cities (including the holdouts).

**Evaluation metrics**. Due to the long-tailed distribution of the data across genera, a pure accuracy measure is insufficient to capture performance, as it is highly biased towards frequent species. Thus, we report accuracy alongside class-averaged recall (AR), calculated as average over all classes of the proportion of correct predictions for the set of examples of that class (this is sometimes also called class-averaged accuracy). To capture performance in a more nuanced way, we also introduce an LVIS [58]-inspired breakdown of class-averaged recall for frequent \((n \geq 20,000)\) examples, common \((100 \leq n < 20,000)\) examples and rare \((n < 100)\) examples subsets of our data. This results in 29 frequent, 150 common, and 165 rare genera, and we denote these metrics as FAR, CAR, and RAR respectively.

5. Experiments

We now demonstrate the benefits of having a multi-city, multiview dataset by training models on Auto Arborist. In this section we train separate aerial and street level baseline ResNet 101 models for each training split described in Section 4, including the training sets for each individual city, the regional splits, and the full dataset. Training details can be found in the supplementary.

**Single city vs. regional vs. full dataset training**. We begin by experimenting with single-view street level models (as the street level modality gives the most accurate results in isolation). In Figure 5, we compare performance on a city’s test set when training on that city’s training set (city), the aggregation of training sets from that city’s region (region), and all available training data (full). Unsurprisingly, we find that more data is better — we see an average improvement of 21.3% AR across cities when going from training on a single city to the full dataset. However we note that training on a region also gives strong performance gains over training on a city itself (average improvement of 18.3% AR), and for some test cities regional training can be on par with (or even slightly better than) training on the full set.

**Cross-city generalization**. Next we examine cross-city generalization, where we are interested in how effective it would be to train on a certain city A if we are interested in testing on B. For this analysis, we first perform all possible cross training combinations, training on every train split (including per-city, regional and full) and testing each model on the test set for each city. Results for these pairwise combinations are visualized in Figure 4 (bottom). Here we see regional “blocks” of strong generalization, reflecting that cities generalize well to cities in the same geographical area. For example, we tend to get good performance training on one of the Pacific Northwest cities (Seattle, Vancouver, Surrey) and testing on another.

We can also see that some cities tend to generalize quite well to other cities on average whereas some cities tend to generalize poorly to other cities. Figure 6 shows this effect in more detail — here we use a given city as a training set and report the spread of performance when applied to other cities’ test sets. In this plot, a larger gap between “self-test” (red stars) and the box implies less generalizeability. Here, to remove confounding factors due to test genera not seen or
Figure 5. Performance growth when adding regional and continental diversity. For each city, we show test performance from a model trained on that city, trained on the respective region for that city, and trained on the full dataset. Note that performance improves on our holdout cities as well, despite the regional and full training sets not including data from those cities. Average performance from models trained on per-city, regional, and full are shown as horizontal lines.

rarely seen during training, we restrict computation of AR for train city \( A \) and test city \( B \) to only the “frequent” genera seen in the train split of \( A \) and the test split of \( B \).

We observe that cities that are poor “training cities” (on the left side of Fig. 6) tend to be smaller and have poor performance overall, though this is not universally true (consider San Francisco). On the other hand, large cities (e.g. NYC) tend to generalize well on average. But we also see that there are no cities which generalize optimally to all others, and optimal generalization performance is only reached by training on the full dataset. Even restricting our attention to frequent, shared classes, we find that generalization ability continues to be highly correlated with label distribution similarity. In Figure 7, we compare AR across these shared, frequent genera with the \( L_1 \) distance between the genus distribution of train vs. test, as seen in Fig. 4. We see that frequently these are anti-correlated, but training sets for some cities (like Buffalo) struggle to perform well across the board.

where \( x_{SL} \) and \( x_A \) are street level and aerial logits, \( w \in \mathbb{R}^n \) are learned parameters, and \( n \) is the number of classes.

Combining modalities in this way yields an average \( \sim 1\% \) boost for each regional model, compared to average-pooling logits across multiple street level views, and a 3-5% boost over predicting from a single street level image. For the full training set, we find that preserving the regional variations in learned MoE weights \( w \) is important — thus our best model (which achieves 49.96% AR) uses street level and aerial models trained on the full dataset but MoE weights specialized for the region to which a city belongs.

We conjecture that this regional dependence is due primarily to regional variations in aerial image quality/availability. In Fig. 8 we visualize the per-genus weights learned by the MoE per region. Looking more closely at the MoE weights, we find that our models only assigns nonzero weights to aerial data for classes that have \( \geq 400 \) training examples. Moreover, we see that we are able to rely on aerial images more in Region W compared to the other two regions.

We show results from our best model in Table 4, report-
genera with more than ∼Each regional MoE learns to use aerial information only for Figure 8.

formance using full model features paired with region-specific Mixture of the features from the full model are more discriminative, we see best per-

adding the aerial imagery when making predictions. We find that while combining multiple street level images gives a further boost. However, strategies. Street level imagery is much more informative than aerial, and Table 3.

City-averaged percent AR for different regions and ensembling ing accuracy and AR for the full dataset, and broken down by frequent and common genera. Notably, the distribution of the three is quite different, and there are certain genera that are more “aerially distinctive” (we have highlighted one for each region).

6. Limitations and future work

We have presented a baseline modeling approach meant to highlight the performance of a typical CNN and present simple methods for combining signals from multiple views — there is much room for improvement, particularly on rare classes. In future, to predict on cities with no past census, we would need to first localize and geocode the trees to be classified. We also hope to expand our dataset to include more cities, both in North America and worldwide, and include species level predictions and additional features such as tree size and health.

Auto Arborist represents an important first step towards global-scale urban forest monitoring. This has implications for environmental justice: given that marginalized communities have less access to urban greenery, systems trained on Auto Arborist could help equitize access to urban forests by empowering quantifiable analysis and targeted replanting. However we must be responsible with our technology — to this end, we protect the privacy of residents of these urban and suburban areas by explicitly filtering out any imagery containing humans, and blur vehicle plates. Secondly we will need efficient human-in-the-loop validation protocols before such a system could be trusted, to ensure science policy is not based on poorly-generalized ML predictions.

7. Seeing the forest for the trees (Conclusions)

Climate change and loss of ecological diversity are among the most pressing issues of our time. Monitoring is a first crucial step to understanding and mitigating the effects of global warming on urban forests, but many cities cannot afford regular tree censuses. Towards the goal of broad, accessible, and affordable urban forest monitoring, we have introduced the Auto Arborist dataset. This dataset is the first of its kind to expand beyond a single city and common categories: Auto Arborist contains 2.6 million trees across 23 cities, covering 344 unique genera. This dataset will enable the computer vision community to tackle urban forest monitoring at scale, and our evaluation protocols help us measure performance without data poisoning, and to evaluate generalization to novel cities.

Acknowledgements We would like to thank R. Alcantara, T. Birch, V. Birdokar, T. Duerig, K. Gill, J. Hea, A. Kundu, P. Lall, M. Manolides, D. Morris, S. Marquardt, P. Perona, J. Quintero, D. Ross, T. Taylor, J. Thamkul, and X. Yin for their support. This work was supported in part by the Resnick Sustainability Institute and undertaken while S. Beery was a Student Researcher at Google.

<table>
<thead>
<tr>
<th>City</th>
<th>Acc</th>
<th>AR</th>
<th>FAR</th>
<th>CAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vancouver, BC</td>
<td>93.28</td>
<td>67.51</td>
<td>82.76</td>
<td>63.35</td>
</tr>
<tr>
<td>Surrey, BC</td>
<td>82.35</td>
<td>58.96</td>
<td>75.82</td>
<td>48.80</td>
</tr>
<tr>
<td>Seattle, WA</td>
<td>79.68</td>
<td>46.55</td>
<td>74.65</td>
<td>43.08</td>
</tr>
<tr>
<td>San Francisco, CA</td>
<td>58.71</td>
<td>26.39</td>
<td>37.87</td>
<td>31.37</td>
</tr>
<tr>
<td>San Jose, CA</td>
<td>77.71</td>
<td>40.07</td>
<td>63.35</td>
<td>41.13</td>
</tr>
<tr>
<td>Cupertino, CA</td>
<td>74.14</td>
<td>56.86</td>
<td>65.28</td>
<td>55.40</td>
</tr>
<tr>
<td>Santa Monica, CA</td>
<td>56.26</td>
<td>43.29</td>
<td>64.93</td>
<td>44.09</td>
</tr>
<tr>
<td>Los Angeles, CA</td>
<td>76.24</td>
<td>32.62</td>
<td>52.36</td>
<td>35.80</td>
</tr>
<tr>
<td>Boulder City, CO</td>
<td>73.23</td>
<td>42.23</td>
<td>58.61</td>
<td>32.88</td>
</tr>
<tr>
<td>Denver, CO</td>
<td>76.46</td>
<td>29.72</td>
<td>57.16</td>
<td>22.02</td>
</tr>
<tr>
<td>Sioux Falls, SD</td>
<td>93.78</td>
<td>76.76</td>
<td>81.52</td>
<td>62.50</td>
</tr>
<tr>
<td>Calgary, AB</td>
<td>88.81</td>
<td>62.18</td>
<td>70.92</td>
<td>52.32</td>
</tr>
<tr>
<td>Edmonton, AB</td>
<td>87.55</td>
<td>56.67</td>
<td>62.58</td>
<td>43.99</td>
</tr>
<tr>
<td>Washington, DC</td>
<td>77.44</td>
<td>44.49</td>
<td>67.31</td>
<td>30.05</td>
</tr>
<tr>
<td>Charlotteville, VA</td>
<td>73.52</td>
<td>57.77</td>
<td>73.38</td>
<td>42.16</td>
</tr>
<tr>
<td>Pittsburgh, PA</td>
<td>78.84</td>
<td>54.93</td>
<td>71.83</td>
<td>43.97</td>
</tr>
<tr>
<td>Montreal, QC</td>
<td>85.51</td>
<td>49.49</td>
<td>64.99</td>
<td>39.08</td>
</tr>
<tr>
<td>New York, NY</td>
<td>82.54</td>
<td>42.77</td>
<td>66.38</td>
<td>28.16</td>
</tr>
<tr>
<td>Buffalo, NY</td>
<td>86.03</td>
<td>54.01</td>
<td>71.92</td>
<td>43.41</td>
</tr>
<tr>
<td>Kitchener, ON</td>
<td>33.96</td>
<td>17.94</td>
<td>21.31</td>
<td>4.49</td>
</tr>
<tr>
<td>Cambridge, ON</td>
<td>72.16</td>
<td>47.69</td>
<td>65.84</td>
<td>34.38</td>
</tr>
<tr>
<td>Columbus, OH</td>
<td>69.28</td>
<td>55.71</td>
<td>68.29</td>
<td>47.32</td>
</tr>
<tr>
<td>Bloomington, IN</td>
<td>85.50</td>
<td>73.52</td>
<td>79.82</td>
<td>64.46</td>
</tr>
</tbody>
</table>

Table 3. City-averaged percent AR for different regions and ensembling strategies. Street level imagery is much more informative than aerial, and combining multiple street level images gives a further boost. However, even though aerial performance on its own is quite low, we see benefit in adding the aerial imagery when making predictions. We find that while the features from the full model are more discriminative, we see best performance using full model features paired with region-specific Mixture of Experts to combine aerial and street level predictions.

Figure 8. Each regional MoE learns to use aerial information only for genera with more than ~400 training examples. Notably, the distribution of the three is quite different, and there are certain genera that are more “aerially distinctive” (we have highlighted one for each region).

Table 4. Per-city performance (%) with our best model trained on the full dataset combining aerial and multiview street level modalities. AR is class-averaged recall for each city, averaged over the test classes for that city. FAR is “Frequent” AR, CAR is “Common” AR, which serve to further disentangle the commonality of a species in the training data with its per-city performance. Holdout cities in bold.

<table>
<thead>
<tr>
<th>City</th>
<th>Acc</th>
<th>AR</th>
<th>FAR</th>
<th>CAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bloomington, IN</td>
<td>77.52</td>
<td>44.49</td>
<td>67.31</td>
<td>30.05</td>
</tr>
<tr>
<td>Charlotteville, VA</td>
<td>73.52</td>
<td>57.77</td>
<td>73.38</td>
<td>42.16</td>
</tr>
<tr>
<td>Pittsburgh, PA</td>
<td>78.84</td>
<td>54.01</td>
<td>71.83</td>
<td>43.97</td>
</tr>
<tr>
<td>Montreal, QC</td>
<td>85.51</td>
<td>49.49</td>
<td>64.99</td>
<td>39.08</td>
</tr>
<tr>
<td>New York, NY</td>
<td>82.54</td>
<td>42.77</td>
<td>66.38</td>
<td>28.16</td>
</tr>
<tr>
<td>Buffalo, NY</td>
<td>86.03</td>
<td>54.01</td>
<td>71.92</td>
<td>43.41</td>
</tr>
<tr>
<td>Kitchener, ON</td>
<td>33.96</td>
<td>17.94</td>
<td>21.31</td>
<td>4.49</td>
</tr>
<tr>
<td>Cambridge, ON</td>
<td>72.16</td>
<td>47.69</td>
<td>65.84</td>
<td>34.38</td>
</tr>
<tr>
<td>Columbus, OH</td>
<td>69.28</td>
<td>55.71</td>
<td>68.29</td>
<td>47.32</td>
</tr>
<tr>
<td>Bloomington, IN</td>
<td>85.50</td>
<td>73.52</td>
<td>79.82</td>
<td>64.46</td>
</tr>
</tbody>
</table>
References

[2] La begins massive street tree census. 2


[37] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 3


[93] Dexter H Locke, Billy Hall, J Morgan Grove, Steward TA Pickett, Laura A Ogden, Carissa Aoki, Christopher G Boone, and Jarlath PM O’Neil-Dunne. Residential housing segregation and urban tree canopy in 37 us cities. npj Urban Sustainability, 1(1):1–9, 2021. 1, 3


[99] Kendra Marshman. The eye of the storm: extreme weather events and sustainable urban forest management. Dall’housie Journal of Interdisciplinary Management, 14, 2018. 1


Kurt Shickman and Global Cool Cities Alliance. Cool policies for cool cities: Best practices for mitigating urban heat islands in north american cities. 2014. 1


Ben G Weinstein, Sarah J Graves, Sergio Marconi, Aditya Singh, Alina Zare, Dylan Stewart, Stephanie A Bohlman,


