Auto Arborist CVPR 2022

Dataset: https://google.github.io/auto-arborist/

The Auto Arborist dataset is a multiview fine-grained visual categorization dataset that contains over 2 million trees belonging to 344 genus-level categories in 23 cities across the US and Canada built to foster the development of robust methods for large-scale urban forest monitoring.

Data Card

DATASET TEAM(S)

The Auto Arborist team

DATASET CONTACT

- Group Email: auto-arborist+managers@googlegroups.com
- Website: https://google.github.io/auto-arborist

DATASET AUTHORS

- Sara Beery, PhD Candidate, Caltech, Google
- Guanhang Wu, Software Engineer, Google
- Trevor Edwards, Software Engineer, Google
- Filip Pavetic, Software Engineer, Google
- Bo Majewski, Software Engineer, Google
- Shreyasee Mukherjee, Software Engineer, Google
- Stanley Chan, Software Engineer, Google
- John Morgan, (formerly at Google)
- Vivek Rathod, Software Engineer, Google
- Jonathan Huang, Research Scientist, Google

PRIMARY DATA MODALITY

Image Data
Text Data
Tabular Data
Audio Data
Video Data
Time Series
Graph Data
Geospatial Data
Multimodal (Image, Geospatial)
Others (please specify)
Unknown

DATASET SNAPSHOT

Size of dataset	24 GB (v1.0. Significant size increase expected)
Number of tfrecords Instances	IN PROGRESS (eventually, 1 million instances with images)
Number of Fields in tfrecords	15
Number of tree_locations instances	4,615,907
Number of Fields in tree_locations	5 (or 6 including implicit city per file)
Labeled Classes	1 (Genus)
Number of Labels	322 Genera
Average labels per instance	N/A
Algorithmic Labels	2 (tree bounding boxes; tree horizontal position

DESCRIPTION OF CONTENT

The content consists of two datasets.

The tree_locations dataset contains basic tree location and genus information derived from 24 cities.

The tfrecords dataset represents a subset of the tree_locations dataset, with additional fields available including an encoded aerial and street level image for each instance.

		in streetlevel image)
	Human Labels	~1 (Genus) *(streetlevel images blurred with help from human labels)
	Other	N/A

DATASET SUBJECT	EXAMPLE: DATA POINT	DATA FIELDS

Sensitive Data about people
Non-Sensitive Data about people
Data about natural phenomena
Data about places and objects
Synthetically generated data
Data about systems or products and their behaviors
Unknown
Others*
(*please specify)
	Example tree_locations datapoint:
	
	Datapoint below is slightly modified (e.g. fake location), but otherwise represents a typical example.
	
	E.g. of Data Point:
	`IDX,SHAPE_LNG,SHAPE_LAT,GENUS,TAXONOMY_ID`
	`057eeab4-1f14-11ec-93z5-eb8801c6f8d0,-85.82556863`
	`059999,37.5307416556,pyrus,246`
	
	Example tfrecords datapoint:
	
	Datapoint below is slightly modified (e.g. fake location, redacted image bytes), but otherwise represents a typical example. It corresponds to the instance above (to show how tfrecords can be matched to tree_locations).
	
	E.g. of Data Point:
	`features: {`
	`  feature: {`
	`  key  : "streetlevel/object/bbox/xmax"`
	`  value: {`
	`  float_list: {`
	`  value: 0.99882144`
	`  value: 0.23218618`
	`  }`
	`  }`
	`  }`
	`  feature: {`
	`  key  : "streetlevel/center/x"`
	`  value: {`
		Tree_locations: Contains the parsed and cleaned up tree inventories of each city in the dataset, along with train/test splits per city.
		• IDX: An identifier for the row which is unique to the city.
		• SHAPE_LNG: The longitude of the tree.
		• SHAPE_LAT: The latitude of the tree.
		• GENUS: The lowercase genus of the tree.
		• TAXONOMY_ID: A unique integer ID corresponding to the genus. Note that this is indexed from 0.
		
		Tfrecords: Contains train/test TFRecord files with one aerial and blurred street level image per available tree for all cities available in this release version. The trees are a subset of the trees in tree_locations.
		• tree/
		  ◦ id: bytes. An ID for the tree that is unique across the release dataset.
		  ◦ tree_locations_idx: bytes. An ID which links to the tree_locations/ CSV IDX for the tree.
		  ◦ city: bytes. The city where the tree is located.
		  ◦ latitude: float. The ground truth latitude of the tree.
		  ◦ longitude: float. The ground truth longitude of the tree.
		  ◦ genus/
		  ▪ label: int64. Holds the ground truth label number representing the tree's genus.
		  ▪ genus: bytes. The ground truth genus of the tree.
		• image/
		  ◦ aerial/
		  ▪ encoded: bytes. An encoded aerial JPEG image of the tree approximately centered on its trunk.
		  ◦ streetlevel/
		  ▪ encoded: bytes. An encoded street level JPEG image of the tree. Non-vegetation pixels are blurred.

```
    int64_list: {
      value: 410
    }
  }
}
feature: {
  key  : "streetlevel/object/bbox/xmin"
  value: {
    float_list: {
      value: 0.90714884
      value: 0.0040578335
    }
  }
}
feature: {
  key  : "tree/genus/label"
  value: {
    int64_list: {
      value: 246
    }
  }
}
feature: {
  key  : "tree/city"
  value: {
    bytes_list: {
      value: "Bloomington"
    }
  }
}
feature: {
  key  : "tree/genus/genus"
  value: {
    bytes_list: {
      value: "pyrus"
    }
  }
}
feature: {
  key  : "streetlevel/object/bbox/ymax"
```


- capturetime: bytes. The month and year that the street level image was captured.
- bbox/: float_lists. Represent tree detection bounding boxes (based on Open Images) as regions scaled from [0, 1], with (0,0) representing the top-left corner of the image.
  - xmin
  - xmax
  - ymin
  - ymax
- center/:
  - x: int64. Represents an approximate (but noisy) location for the horizontal center pixel of the tree in the image.
  - y: int64. This is always set to half of the image height. It is provided for convenience.

```
  value: {
    float_list: {
      value: 0.59795773
      value: 0.54920584
    }
  }
}
feature: {
  key  : "tree/id"
  value: {
    bytes_list: {
      value: "6560451631306680540"
    }
  }
}
feature: {
  key  : "streetlevel/object/bbox/ymin"
  value: {
    float_list: {
      value: 0.29351664
      value: 0.05435237
    }
  }
}
feature: {
  key  : "tree/latitude"
  value: {
    float_list: {
      value: 37.53074
    }
  }
}
feature: {
  key  : "streetlevel/encoded"
  value: {
    bytes_list: {
      value: "JPEGGoesHere"
    }
  }
}
```

```
  feature: {
    key  : "streetlevel/capturetime"
    value: {
      bytes_list: {
        value: "July 2019"
      }
    }
  }
  feature: {
    key  : "streetlevel/center/y"
    value: {
      int64_list: {
        value: 576
      }
    }
  }
  feature: {
    key  : "aerial/encoded"
    value: {
      bytes_list: {
        value: "JPEGGoesHere"
      }
    }
  }
  feature: {
    key  : "tree/idx"
    value: {
      bytes_list: {
        value:
"057eeab4-1f14-11ec-93z5-eb8801c6f8d0"
      }
    }
  }
  feature: {
    key  : "tree/longitude"
    value: {
      float_list: {
        value: -85.8255
      }
    }
```

```
  }
}
```

DATASET PURPOSE(S)	KEY DOMAINS OR APPLICATION(S)	PRIMARY MOTIVATION(S)
Monitoring **Research** Production Others (please specify)	Domains Machine Learning, Object Recognition, Computer Vision, Computing for the Environment, Environmental Monitoring, Biodiversity Monitoring, Urban Planning Problem Space Multiview Recognition, Fine-Grained Visual Categorization, Out-of-domain Recognition, Automated Urban Forest Monitoring	• Enable the computer vision community to tackle impactful environmental challenges • Provide a real-world benchmark for tree categorization in cities from multiview data with spatiotemporal structure • Advocate for robust out-of-domain generalization analysis for SOTA computer vision architectures via cross-domain data splits • Provide the largest ever fine-grained visual categorization benchmark to the computer vision community
DATASET USAGE	**INTENDED AND/OR SUITABLE USE CASE(S)**	**UNSUITABLE USE CASE(S)**
Safe for production use **Safe for research use** Conditional use- some unsafe applications Only approved use Others (please specify)	• Developing a model to predict genera and reporting its architecture and results against the Auto Arborist benchmark • Running a large scale analysis of urban ecology in North America and sharing conclusions from the analysis	• Republishing the Auto Arborist dataset or any data derived from it (such as processed images or examples of images) without authorization
SAFETY OF USE WITH OTHER DATA	**ACCEPTABLE TRANSFORMATIONS**	**BEST PRACTICES FOR JOINING OR AGGREGATING WITH DATASET**
Safe to use with other data Conditionally safe to use with other data Should not be used with other data Unknown Others* (Please specify)	**Joining with other datasets** **Subsampling and splitting** **Filtering** **Joining input sources** **Cleaning missing values** **Anomaly detection** **Grouping and summarizing** **Scaling and reducing** **Statistical transformations** Redaction or Anonymization Others (please specify)	The dataset comes with train/test splits. For benchmarks against the dataset, we recommend following these strictly in order for the benchmark to be comparable to others.
VERSION STATUS	**DATASET VERSION**	**MAINTENANCE PLAN**
Regularly Updated New versions of the dataset have been or will continue to be made available.	**Current Version** 1.0 **Last Updated** 06/2022 (IN PROGRESS) **Release Date** 06/2022 (IN PROGRESS)	The Auto Arborist team intends to update the dataset with new instances until 1 million tfrecords have been released. After this, it expects to shift to Limited Maintenance, and the number of instances may decrease over time for error reasons mentioned below.

Actively Maintained

No new versions will be made available, but this dataset will be actively maintained, including but not limited to updates to the data.

Limited Maintenance

The data will not be updated, but any technical issues will be addressed.

Deprecated

This dataset is obsolete or is no longer being maintained.

- **Versioning:** Versions will be a M.m (Major.minor). Major updates will usually add significant new instances or perform significant error corrections, while minor updates may be error corrections or removal of ~1-10 instances. An example progression may be: 1.0,2.0,2.1,3.0,...
- **Update:** Major updates will normally occur whenever all data which we intend to release for a city is ready. Minor updates may happen at any time as needed.
- **Errors:** The dataset is expected to have errors and noise as described in the paper. These will generally not be corrected unless there is a significant reason to do so.
- **Feedback:** The Auto Arborist team welcomes feedback. Please see the website for feedback instructions.

ACCESS POLICY

Here are the Terms and Conditions

- **Access Prerequisites:** Sign the Terms and Conditions of use, seen at the link above.
- **Data Usage Policy:** Non-commercial, non-exclusive, worldwide, royalty-free, non-transferable and non-sublicenseable license to use (including reproducing and creating derivative works of)
- **Access Control Lists:** Users who have signed the ToC
- **Exemptions & Exceptions:** None

RETENTION POLICY

The retention policy is included in the ToC

There are no retention restrictions for the Auto Arborist dataset

- **Retention Duration:** None
- **Retention Steps:** None
- **Retention Policy:** None
- **Exemptions & Exceptions:** None

WIPEOUT POLICY

The wipeout policy is included in the ToC

Google may receive third party requests to take down or blur a specific panorama on the Google Street View website. Google may forward this request to Organization and provide Organization with an updated version that complies with the takedown request. Organization must delete the Licensed Content originally delivered and replace it with the updated Licensed Content provided by Google.

- **Wipeout Duration:** <summarize here>
- **Deletion Event Steps:** If a third party requests street level imagery to be blurred or taken down, a new, compliant version of the dataset will be sent to all users. They will be asked to delete the prior version of the dataset and work with the new one going forward.
- **Post-deletion Obligations:** <summarize here>
- **Exemptions & Exceptions:** None

DATA COLLECTION METHODS

DATA SOURCES

DATA COLLECTION

API
Artificially Generated
Crowdsourced - Paid
Crowdsourced - Volunteer
Vendor Collection Efforts
Scraped or Crawled
Survey, forms or polls
Taken from other existing datasets
Unknown
To be determined
Others (Proprietary APIs)

City-generated Arboreal Censuses

Arboreal Censuses: These censuses are used by cities to monitor their urban trees and are collected infrequently. We have used censuses that are published publicly, all licensing information is available in the supplementary material of our CVPR 2002 paper.
Date of Collection: Nov 2020 - Nov 2021
Instrumentation: Human-generated GPS locations and categories of trees
Data Modality: Geospatial Data / Text Data

API

Google Street View API: An internal API used to access Google Street View images.
Date of Collection: [April 2009 - June 2021]
Instrumentation: Street View Cameras
Data Modality: Image Data

API

Google Aerial API: An internal API used to access aerial imagery in cities.
Date of Collection: Approximately [Jan 2019-May 2022] (we are unable to fully verify this range)
Instrumentation: Low-flying aircraft and Satellite Imagery
Data Modality: Image Data

API

Google Semantic Segmentation API: An internal API that provides semantic segmentation for Street View data. We used the results from this API to blur PII for our data.
Date of Collection: [Jan 2022 - May 2022]
Instrumentation: Computer Vision Model
Data Modality: Image Masks

Crowdsourced

Collected and included
- none

Collected and excluded
- Boxes around PII that was not blurred by our automated blurring based on internal APIs, used to create the final human-verified, PII-obscured images but not released

INCLUSION CRITERIA

Per-City Tree Instance records

Cities were selected based on availability of tree inventory, the inventory's usage restrictions, quality of the inventory, etc. Cities were restricted to North America. Records that were labeled with a genus that was not mappable into the Catalog of Life taxonomy were removed

Aerial Imagery

Images were obtained by querying a proprietary API with the locations available for each instance.

EXCLUSION CRITERIA

Per-City Tree Instance records

- **Quality:** Instances with invalid lat/lng or lat/lng that are outside of the expected city boundaries. Instances which cannot be mapped to a genus, e.g. because "palm" is not a genus; common typos (e.g. "ginkgo" vs. "ginko") were corrected instead of excluded.
- **Content:** None?

Aerial Imagery

- **Quality:** None.
- **Content:** Instances without available imagery (extremely rare).

DATA PROCESSING

Per-City instance records were processed into a common format. Each instance was supplemented with Aerial and Street Level imagery.

Street Level Imagery

Street Level imagery was processed to generate tree bounding box data and to blur pixels.

Street Level Imagery Street level images taken Jan 1, 2008 or later which are expected to show the instance tree within 10 meters based on a proprietary geolocation API.	Street Level Imagery • **Quality:** Images which are too blurry. • **Content:** Instances without available imagery. Images which contain pixels with people or other S/PII. Images which do not contain a minimum number of pixels associated with trees based on a proprietary pixel segmentation.	
SENSITIVE DATA	**FIELDS WITH SENSITIVE DATA**	**SECURITY AND PRIVACY HANDLING**
User Content User Metadata User Activity Data Identifiable Data **S/PII** Business Data Employee Data Pseudonymous Data Anonymous Data Health Data Children's Data None Others* (*please specify)	Intentionally Collected Sensitive Data none Unintentionally Collected Sensitive Data `streetlevel/encoded`: Identifiable houses, cars, or people in street level imagery	Blurring S/PII in street level imagery • Filtering images that contain people or license plates: We first used an internal privacy API to filter out any images that had visible human pixels • Automated blurring with internal Semantic Segmentation API: Next we blur all pixels that are not "tree", "sky", "paved_road", "dirt_road", "sidewalk", "crosswalk", "water", or "mountain" using an internal semantic segmentation API • Human verified S/PII removal: Finally, we used Crowd Compute to detect and draw boxes around any S/PII that was still visible after our automated method and blurred the interior of those boxes.
TRANSFORMATIONS APPLIED	**FIELDS TRANSFORMED**	**LIBRARIES AND METHODS USED**
Anomaly Detection **Cleaning Mismatched Values** Cleaning Missing Values **Converting Data Types** Data Aggregation Dimensionality Reduction Joining Input Sources **Redaction or Anonymization** Others* (*Please specify)	Cleaning Mismatched Values • Genus (fixed common typos) Converting Data Types • Genus: Label • tree_locations.SHAPE_LNG : tfrecords.tree/longitude (downcast to 32 bit float) • tree_locations.SHAPE_LAT : tfrecords.tree/latitude (downcast to 32 bit float) Redaction or Anonymization • streetlevel/encoded (blurring) • streetlevel/capturetime (reduced timestamp granularity) • Various internal fields are removed for the release	• Blurring S/PII for anonymization: Internal Google Semantic Segmentation API, pixels potentially containing S/PII are blurred using a gaussian kernel
SAMPLING METHOD(S)	**SAMPLING CHARACTERISTIC(S)**	• **SAMPLING CRITERIA**

Cluster Sampling Haphazard Sampling Multi-stage Sampling **Random Sampling** Retrospective Sampling **Stratified Sampling** Systematic Sampling Weighted Sampling Unknown Unsampled Others* (*Please specify) TODO - Maybe Sara can fill this part out? Or whatever details you can easily provide	Stratified Sampling Upstream Source: All tree locations Total data sampled: 2.5M tree records Sample size: 1M tree records Threshold applied: Per city/per genera stratified count Sampling Rate: specific to each city/genera pair	We sampled the total set of trees to determine which imagery to release, with the goal of releasing 1M trees with imagery • Stratified Sampling: We stratified our samples across cities and genera. Our stratification method looped through each city and each genus in a round robin and incremented a count for that city/genus pair that would determine the number of trees to be sampled for that strata. Once the total available number of images for that city/genus pair was reached, the count for that strata was no longer incremented. This was to ensure that the images were stratified across the genera and cities, instead of biased towards large cities and common genera. We used this city/genera count to randomly sample the tree IDs to include with imagery.

ANNOTATION WORKFORCE TYPE	ANNOTATION CHARACTERISTICS	ANNOTATION DESCRIPTION
Annotation Target in Data Machine-generated Annotations **Human Annotations - Expert** Human Annotations - Non-expert Human Annotations - Employees **Human Annotations - Contractors** Human Annotations - Crowdsourcing Human Annotations - Outsourced / Managed Teams Unlabeled Others* (*Please specify)	Human Annotations - Expert Number of annotations (based on tree_locations): 4,615,907 Human Annotations - Contractors Total number of street-level images annotated: 1,000,000	Annotations were used to label the genera of trees and to blur S/PII. Human Annotations - Expert We assume per-city tree inventories were produced by experts and are of high quality. Human Annotations - Contractors Contractors annotated street level imagery for S/PII blurring / removal. A confidential platform is used for collecting these annotations.

Reflections on Data

Trees are non-offensive	- Trees are not considered to be offensive or insulting, and images of trees should not cause anxiety. However, we do not have control over what humans may place on trees (i.e. offensive signs) and have not explicitly removed such objects. If any offensive material is found in the dataset please email

	auto-arborist+managers@googlegroups.com to have it removed.