Cross-view Object Geolocalization for Tree Species Mapping

Anonymous ICCV submission

Paper ID *****

Abstract

Accurate, large-scale tree-species identification is a missing link for many high-impact forestry and climate applications. Though species identification is feasible using ground-level imagery, globally scaleable tree species mapping requires identifying species using remote sensing data. We bridge this gap by formulating and solving the cross-view object geolocalization task: matching every object detected in a ground-level panorama to its counterpart in a precisely geo-referenced satellite image. Our end-to-end pipeline matches each individual tree in a satellite image to a corresponding streetview image of that tree, while filtering out occluded trees using ray-tracing. Evaluated on 7.8 k trees, the method achieves 83.6 percent matching accuracy, rising to 84.3 percent for trees close to the camera. By linking ground-level species labels to satellite-detected crowns, our approach unlocks scalable, multimodal reference data for global tree-species mapping and sets a new direction for fine-grained cross-view localization tasks beyond forestry.

1. Introduction

Forestry is one of the most important natural climate solutions. Forests store over 200 gigatons of carbon [1], and adding trees to cropland has the potential to sequester 5 gigatons of carbon per year [2]. Trees can also provide other key benefits, including provision of additional food and revenue streams, improved soil health, and increased habitat for biodiversity. But research on forestry contains a significant contraint: lack of data.

Recent advances have shown that it is possible to segment individual tree crowns at continental scales with high-resolution satellite imagery [3] and tree count and tree cover using freely available satellite data with global coverage [4]. This represents a transformative opportunity to increase the data available on trees around the world. However, without data on the species of individual trees, the most useful forest monitoring applications are impossible. Tree species is required to accurately measure carbon sequestration from trees as different species sequester different amounts of carbon. Tree species is required for disease monitoring as different species have different diseases. Trees provide significant provisioning services such as timber, food, and other non-timber forest products, but yield monitoring requires tree species, as different species produce different types and quantities of different commodities. [5] Hanan et. al., 2020 note that detecting tree "species will probably remain at the top of the Earth-observation research community's wish list. . . ." (Draper et al., 2020).

Beery et. al. 2022 [6] demonstrated that accurate tree species classification can be achieved with google streetview imagery. However, streetview imagery is not available everywhere, and it is constrained to detecting species only for trees visible from the road. Ahlswede et. al. 2023 [7] demonstrated that tree species identification is also possible from multispectral time-series imagery from sentinel-2 using a large-scale georeferenced dataset in Europe. However, this approach cannot be extended globally without groundtruth georeference data for every geography or species. If one could link the species of individual trees identified using streetview imagery with the geolocation of those trees, one could scalebly generate georeferenced species labels for individual trees around the world. As a result, tree matching and geolocalization across ground and satellite views becomes critical. Each individual tree needs to be accurately geolocated for ground view based species labels to be used to supervise models for scalable tree species mapping from remote sensing data. As a result, scaleable tree species mapping requires object-level geolocalization from ground and satellite views.

Our key contributions in this paper are to 1) Define the cross-view object geolocalization task, 2) Develop a pipeline that can be used for cross-view individual tree geolocalization and achieve strong performance in our human-annotated evaluations.

2. Related Work

Recent work has demonstrated that streetview imagery enables significantly more accurate tree species identification than satellite imagery [6]. Moreover, streetview imagery has also been used to scale up reference data for crop type

mapping, enabling significantly more accurate crop species classification from remote sensing data than was possible solely with field-based ground truthing [8]. This approach of combining ground-level and satellite-based views to improve crop type mapping can also work with geotagged photos from mobile phones or motorbike helmet based cameras [9].

There is a rich literature on image geolocalization as a set of many distinct but related computer vision tasks, which are summarized in this survey from Wilson et. al. 2024 [10], including: 1) single view geolocalization in which one predicts the geolocation of a single image, 2) cross-view geolocalization where one matches a ground-level image to a georeferenced aerial or satellite image, and 3) object geolocalization where one or multiple geolocated ground images are used to geolocate an object. Berton et. al. 2022 [11] release a benchmark for deep visual geolocalization from a single image. Deuser et. al. 2023 [12] is the first paper we find exploring the issue of cross-view geolocalization, and they develop an architecture that addresses distortions from polar transformations to align views. Li et al. 2024 [13] explore unsupervised learning for cross-view geolocalization. As far as we are aware, ours is the first paper exploring cross-view object-level geo-localization, which we apply to matching individual trees across views.

3. Methods

As noted by Wilson et. al 2024 [10], "the concept of geolocalization broadly refers to the process of determining an entity's geographical location, typically in the form of Global Positioning System (GPS) coordinates. The entity of interest may be an image, a sequence of images, a video, a satellite image, or even objects visible within the image." Our problem—matching each individual tree visible in a ground-level photo to the exact tree crown detected from space—falls is an example of cross-view retrieval which incidentally enables object geo-localisation. We therefore introduce cross-view object geo-localisation.

3.1. Task Definition: Cross-view Object geo-localisation

Our task is as follows: Given a street-view panorama and the list of objects detected in a co-located satellite chip, identify the objects within the panorama that correspond to each detected object in the satellite chip. Because the satellite image is precisely georeferenced, solving this task is equivalent to geo-locating each detected object in the ground view. In our context of individual tree species identification, solving this problem enables one to match ground-view images in which tree species are identifiable to exact geo-locations, which can then be linked with satellite remote sensing data to enable scaleable tree species identification. This approach can also be extended to enable large-scale supervision any remote sensing object classification tasks.

3.2. Cross-view Geolocalization Data Pipeline

Stage	Description
Road sampling	Sample candidate points every 25 m along OpenStreetMap® roads inside the study AOIs.
Street-view acquisition	Query the GSV API to check whether imagery exists at these points, snap to the panorama's true capture point to correct the ± 5 m API noise.
Satellite Image Check	Check whether Planet Skysat API contains imagery at these points
Tree-presence filter	Use the 3 m global canopy-cover layer of Brandt *et al.* (2024) to require $> 50\%$ canopy within a 15 m radius.
Individual tree detection	Download Planet SkySat (0.5 m) imagery on this location; run a Faster-rcnn detector trained on the individual tree detection dataset from Sachdeva et. al. 2024 to obtain bounding boxes
Candidate pruning using Ray-tracing	We use the detected bounding boxes to apply a ray-tracing approach out from the location of the streetview image to filter out trees that are occluded by other trees
Compute Streetview Camera Angle	For the un-occluded trees, we compute the location of the detected bounding box using the georeference information of the satellite image. By computing the angle of this point relative to the location of the google streetview image, we can point the streetview camera directly at the tree.
Retrieve Streetview image centered on object	Retrieve a crop of the panorama centered on the object of interest using the generated angle.

We use cross-view geo-localization to assemble a multi-modal dataset of trees combining street-view imagery, high-res satellite imagery, and other data products. We use existing data products from Brandt. et. al. 2024 [14] and OpenStreetmap along with the google street view API to identify lat/long points where trees are present near the road and streetview imagery is available. Then we

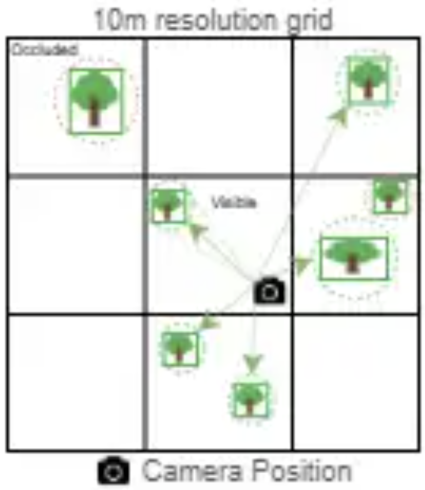


Figure 1. Our ray-tracing algorithm filters out trees that are occluded by other trees from the perspective of the streetview camera

check whether high-resolution satellite imagery from Planet Skysat is available in that area, and if it is, we apply a simple tree detector trained on a dataset from Sachdeva et. al. 2024 [15] that outputs bounding boxes for each individual tree. Because the satellite image is georeferenced, we can convert the bounding boxes to lat/long coordinates for each detected tree.

We use ray-tracing to filter out trees that are occluded by other trees from the perspective of the camera, as shown in figure 1. The data pipeline is also made compatible with remote sensing modalities such as sentinel 2. We utilize the the remote sensing grid to isolate trees that are within the vicinity of the camera point allowing us to construct supervision labels for each remote sensing pixel. This is particularly useful for the species classification task by allowing the construction of a multi-model dataset.

By combining the lat/long coordinates of the streetview image with the lat/long coordinates of each individual tree, we can compute the angle of the tree relative to the streetview image. We use this angle to query a streetview image pointing at the detected tree for all the unoccluded trees in a satellite image. By doing this, we have linked each georeference remote sensing tree detection to a streetview image of the same tree. This pipeline is described in table 1.

3.3. Annotations and Evaluation

To evaluate whether we have correctly matched detected trees from satellite images to images centered on that tree, we construct an annotated dataset. Our dataset consists of 7380 trees, with each tree being comprised of a pair of satellite and street view images generated by our cross-view geo-localization pipeline. The satellite image is centered on and a detected tree with a drawn bounding box, along with a point and line indicating the location and view of the streetview camera. If the match is correct, the street view image should be pointed at and centered on that same tree. We manually annotate 7,860 images as either "correct" or "incorrect" for whether the tree centered in the streetview image is indeed the tree in the satellite image. This task requires using the context of both images and the relative location of the tree to that context to interpret whether the object is correctly matched. For example, we have shown in figure 2 a correct match on the left, and an incorrect match on the right.

Category	**Count**
OSM points sampled with SkySat Coverage	21,301
With Google Street View Image	17,275
With Valid Tree Cover and Farmland Cover	4,840
With non-occluded trees	3,683
Images Captured	7,380

Table 1. Summary of Data Points and Image Statistics

4. Results

We observe strong performance of our cross-view geolocalization pipeline in matching trees detected in the satellite view to the same tree detected in the ground view. In our dataset, 83.6 percent of trees retrieved using our pipeline have been correctly matched, as shown in table 1.

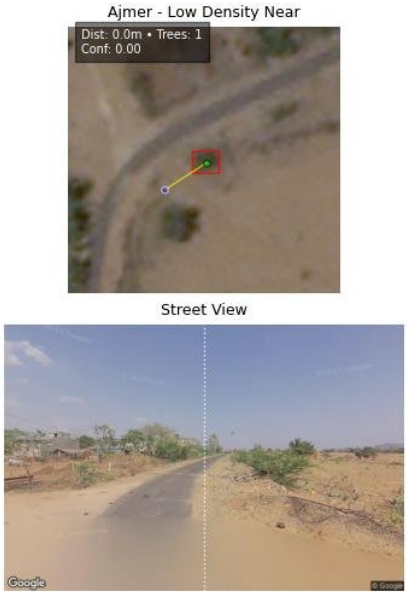
We report accuracy stratified by the distance of the tree from the road. As one might expect, the accuracy for trees near the road at 84.3 percent is slightly higher than for trees far from the road at 81.4 percent, but the relatively high accuracy under both conditions indicating that this approach is robust to the distance of the tree from the street.

Category	Correct	Incorrect	Accuracy	Total
All	6167	1213	0.836	7 380
Near	4580	851	0.843	5 431
Far	1587	362	0.814	1 949

Table 2. Matching accuracy by distance.

We also report accuracy stratified by district in table 2 to ensure that our pipeline is accurate under a wide range of conditions. We observe despite slightly lower accuracy in Ajmer, matching accuracy remains largely stable across districts, indicating our pipeline is relatively robust to geographic variation as well.

(a) This is clearly a correct match, as the location of the tree relative to the road is the same in the both views



(b) This is clearly an Incorrect Match, as the detected tree in the satellite image is not centered in the streetview image

Figure 2. Example cross-view pairs that we annotate as correct matches of trees or not

District	Accuracy
Ajmer	0.600
Alwar	0.827
Bundi	0.917
Jaipur	0.900
Jhunjhunu	0.850
Nagaur	0.866
Pali	0.790
Sikar	0.792

Table 3. Accuracy values by district.

5. Conclusion

To enable remote sensing species identification, we develop an approach to matching individual trees detected in satellite imagery, where resolution is too coarse for visual species identification, to trees in google streetview imagery, where the images are sharp enough to distinguish tree species. We link google street view and planet skysat image modalities together, apply object detection to individual trees from satellite imagery, and retrieve a streetview image pointed at each tree in the satellite image. We then conduct manual annotations to check whether we have correctly matched ground images of the tree to the remotely detected tree.

We observe high accuracy of our pipeline in matching streetview imagery to trees in satellite imagery. The accuracy of this pipeline is robust to whether the tree is near or far from the road, and it is also robust to varying geographies with different environments.

Current approaches for collecting groundtruth species data through field work and mobile phone based geotagging are extremely resource intensive, preventing machine learning applications to tree species mapping to scaling across large geographies. We have ensured that trees are correctly matched from street view and satellite imagery. As a result, we can now scaleably generate geo-located reference species labels for training machine learning models to identify tree species from remote sensing data which is also geo-referenced. Cross-view object matching and geolocalization can dramatically reduce the cost and increase the scaleably of data collection for training remote sensing models to identify tree species. We will leverage this capability in forthcoming work.

Our approach to cross-view object mapping and geolocalization can be extended in many different ways. One could leverage ground view data from open-source databases like Mapillary, or to leverage ground image data taken from cameras attached to cars or helmets driving in rural areas. One could also increase the fidelity of the matching task by increasing the number and type of objects to match between satellite and ground view. A significant limitation of the approach we took is that our pipeline filters out most cases of dense trees because of occlusion. Future work should improve on our approach by enabling dense matching between every point in ground view imagery with

every point in satellite imagery rather than matching every object, as these sorts of approaches will be necessary for cross-view geolocalization in more complex forest landscapes.

Another note worth discussion is that this is clearly a dual use technology. Though we were inspired to develop this method for the application of tree species mapping, the same approach described here could be used to match any type of object between ground view and satellite views to enable previously infeasible fine-grained classification tasks from remote sensing data. This has significant ethical implications as this technology could be used for mass surveillance and military applications, for example by identifying individual buildings or vehicles belonging to specific people. As capabilities combining geospatial data, computer vision, and artificial intelligence improve, there is a need to strike the right balance between realizing the potential of positive applications of these tools while increasing awareness and mitigating risks around the potentially malicious applications of the same technology.

References

[1] Mo L, Zohner CM, Reich PB, et al. Integrated global assessment of the natural forest carbon potential. Nature. 2023;624(7990):92-101. 1

[2] Roe S, Streck C, Obersteiner M, et al. Contribution of the land sector to a 1.5 C world. Nature Climate Change. 2019;9(11):817-28. 1

[3] Brandt M, Tucker CJ, Kariryaa A, et al. An unexpectedly large count of trees in the West African Sahara and Sahel. Nature. 2020;587(7832):78-82. 1

[4] Brandt J, Ertel J, Spore J, Stolle F. Wall-to-wall mapping of tree extent in the tropics with Sentinel-1 and Sentinel-2. Remote Sensing of Environment. 2023;292:113574. 1

[5] Pu R. Mapping Tree Species Using Advanced Remote Sensing Technologies: A State-of-the-Art Review and Perspective. Journal of Remote Sensing. 2021;2021(4):1-26. 1

[6] Beery S, Wu G, Edwards T, Pavetic F, Majewski B, et al. 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 (CVPR); 2022. p. 21294-307. 1

[7] Ahlswede S, Schulz C, Gava C, Helber P, Bischke B, Förster M, et al. TreeSatAI Benchmark Archive: A Multi-Sensor, Multi-Label Dataset for Tree Species Classification in Remote Sensing. Earth System Science Data. 2023;15(2):681-95. 1

[8] Soler JL, Friedel T, Wang S. Combining Deep Learning and Street View Imagery to Map Smallholder Crop Types. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2024. Special Track on AI for Social Impact. 2

[9] Nakalembe C, Zvonkov I, Kerner H, et al.. Helmets Labeling Crops: Kenya Crop Type Dataset Created via Helmet-Mounted Cameras and Deep Learning; 2025. EarthArXiv preprint. 2

[10] Wilson D, Zhang X, Sultani W, Wshah S. Image and Object Geo-Localization. International Journal of Computer Vision. 2024;132:1350-92. 2

[11] Berton G, Mereu R, Trivigno G, Masone C, Csurka G, Sattler T, et al. Deep Visual Geo-Localization Benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022. p. 5396-407. 2

[12] Deuser F, Habel K, Oswald N. Sample4Geo: Hard Negative Sampling For Cross-View Geo-Localisation. In: ICCV; 2023. p. 16847-56. 2

[13] Li G, Qian M, Xia G. Unleashing Unlabeled Data: A Paradigm for Cross-View Geo-Localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024. p. 16719-29. 2

[14] Brandt M, Gominski D, Reiner F, et al. Severe decline in large farmland trees in India over the past decade. Nature Sustainability. 2024;7:860-8. 2

[15] Sachdeva S, Lopez I, Biradar C, Lobell D. A Distribution Shift Benchmark for Smallholder Agroforestry: Do Foundation Models Improve Geographic Generalization? In: Proceedings of the Machine Learning for Remote Sensing (ML4RS) Workshop at ICLR; 2024. p. 1-8. 3