International Cartographic Association
Association Cartographique Internationale

Training-free Urban Tree Counting with Multi-modal data

Zhengsen Xu [a], Lanying Wang [a], Hongjie He [a], Haiyan Guan [b], Lingfei Ma [c] Jonathan Li [a, d], Lincoln Xu [e*]

[a] *Department of Geography and Environmental Management, University of Waterloo, {zhengsen.xu, lanying.wang, hongjie.he, junli}@uwaterloo.ca*
[b] *School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, guanhy.nj@nuist.edu.cn*
[c] *School of Statistics and Mathematics, Central University of Finance and Economics, l53ma@cufe.edu.cn*
[d] *Department of Systems Design Engineering, University of Waterloo, junli@uwaterloo.ca*
[e] *Schulich School of Engineering, University of Calgary, lincoln.xu@ucalgary.ca*

* Corresponding author

Keywords: remote sensing, urban trees, foundation model, multi-modal model, unsupervised learning

Abstract:

Trees play a crucial role in urban environments, contributing significantly to both aesthetics and improvements in urban and global ecosystems (Eisenman et al., 2019). However, there is a lack of research on urban tree counting, despite its potential to support forest management decisions, improve the accuracy of carbon sequestration estimates, and refine assessments of vegetation access inequality for urban residents (Leng et al., 2023).

In recent years, advances in artificial intelligence have demonstrated remarkable efficacy in accurate tree counting from 2D remotely sensed imagery (Ammar et al., 2021). However, a notable limitation of data-driven artificial neural networks is their reliance on extensive labeled datasets for parameter training. Given the promising results of recently developed foundation models in remote sensing, this study explores the feasibility of counting urban trees using high-resolution satellite imagery and multimodal information without relying on labeled data or retraining deep learning models.

Specifically, we utilize the Grounding DINO foundation model (Liu et al., 2024). To address the variability in canopy scales, we propose a hierarchical input design. Furthermore, to compensate for the lack of height information in 2D satellite imagery compared to LiDAR data, we incorporate another foundation model, Depth Anything (Yang et al., 2024), which generates pseudo-depth images from 2D inputs. Finally, experiments conducted on the KCL-London dataset (Amirkolaee et al., 2023) demonstrate that the proposed approach achieves promising results.

First, preliminary experiments reveal limitations in handling tree canopies of varying scales, as its fixed feature fusion mechanisms struggle to distinguish overlapping canopies, often segmenting them as a single entity. To address this, we designed a hierarchical input framework. First, large canopies are segmented and counted using the original image. Then, the image is divided into non-overlapping sub-images of size 512×512 with a 50% overlap, which are fed into the model to enhance its focus on smaller canopies and mitigate incomplete detection caused by sub-image boundaries.

After obtaining all segmentation results, a post-processing step is applied to isolate small canopies from large ones and eliminate shadow interference. Specifically, for detected overlapping canopies, overlap ratios are calculated, and if the overlap exceeds an empirical threshold (e.g., 0.5), only the smaller canopy is retained, effectively splitting large canopies into smaller ones. Moreover, Grounding DINO primarily relies on 2D image information and is limited in its sensitivity to height variations, potentially leading to higher false-positive or false-negative rates when depth data is unavailable. Considering the similarity between tree canopy shadows and the canopy itself, high-resolution imagery is processed through Depth Anything to generate pseudo-height information for the entire image. The height information for each detected canopy region is then calculated, and if the canopy's height exceeds the surrounding mean, it is considered a true canopy; otherwise, it is discarded as a shadow. Notably, the proposed model tends to underestimate the results, prompting the use of linear regression to correct the predictions.

The final quantitative and qualitative results are presented in Table 1 and Figure 1. As shown in Table 1, the proposed multi-modal foundation model achieves competitive MAE and RMSE values compared to state-of-the-art models, such as TreeFormer, while maintaining a high R^2 that is nearly identical to TreeFormer and significantly higher than models trained with annotated data. This highlights the strength of our unsupervised approach, which does not rely on labeled training data, yet effectively addresses challenges in accurately capturing complex canopy structures and small targets. As illustrated in Figure 1, our method demonstrates strong performance in bridging the gap with fully supervised models, despite its reliance solely on pre-trained foundation models and unsupervised refinements.

Abstracts of the International Cartographic Association, 10, 307, 2025
32nd International Cartographic Conference (ICC 2025), 17–22 August 2025, Vancouver, Canada.
https://doi.org/10.5194/ica-abs-10-307-2025

Metric	MCNN*	CSRNet*	SwinUnet*	FusionNet*	SASNet*	EDNet*	TreeFormer**	Ours***
MAE	25.87	23.27	36.45	28.45	24.33	26.18	16.7	19.73
RMSE	34.12	29.62	47.56	35.67	30.12	32.02	22.98	25.24
R^2	0.45	0.59	0.24	0.47	0.56	0.52	0.75	0.71

Table 1. Performance comparison of different models across metrics MAE, RMSE, and R^2 on the KCL-London dataset. Note: *, **, and *** denote fully supervised, semi-supervised, and unsupervised methods, respectively.

However, it is worth noting that although this study demonstrates the feasibility of using foundation models for urban tree canopy counting, it represents only an initial exploration. Future research directions include the following: (1) designing more effective data preprocessing methods to enhance the model's attention to multi-scale canopy features, addressing challenges such as overlapping and small canopies; (2) effectively integrating depth information to strengthen canopy edge delineation, reducing misclassification caused by shadows and structural complexity; (3) testing the model on additional datasets to assess its applicability to lower-resolution imagery or diverse ecological regions (Veitch-Michaelis et al., 2024); and (4) exploring the use of segmentation results to advance downstream tasks, such as improving urban tree carbon sequestration estimation and assessing green space equity. These efforts aim to refine the model's performance while broadening its practical applications in urban sustainability and environmental equity.

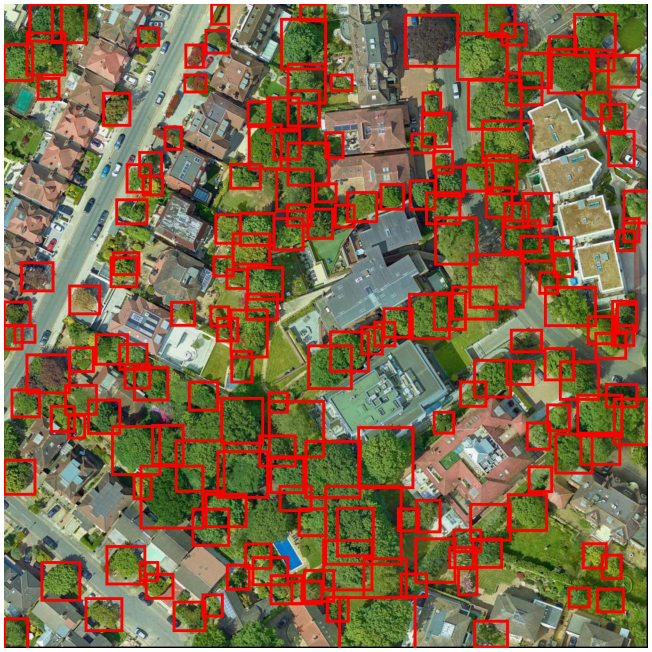

Figure 1. The partial counting results of the proposed model. Green bounding boxes represent the detected canopies.

References

Amirkolaee, H. A., Shi, M. and Mulligan, M., 2023. Treeformer: a semi-supervised transformer-based framework for tree counting from a single high resolution image. *IEEE Transactions on Geoscience and Remote Sensing*.

Ammar, A., Koubaa, A. and Benjdira, B., 2021. Deep-learning-based automated palm tree counting and geolocation in large farms from aerial geotagged images. *Agronomy* 11(8), pp. 1458.

Eisenman, T. S., Churkina, G., Jariwala, S. P., Kumar, P., Lovasi, G. S., Pataki, D. E., Weinberger, K. R. and Whitlow, T. H., 2019. Urban trees, air quality, and asthma: An interdisciplinary review. *Landscape and urban planning* 187, pp. 47–59.

Leng, S., Sun, R., Yang, X. and Chen, L., 2023. Global inequities in population exposure to urban greenspaces increased amidst tree and nontree vegetation cover expansion. *Communications Earth & Environment* 4(1), pp. 464.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J. and Zhang, L., 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection.

Veitch-Michaelis, J., Cottam, A., Schweizer, D., Broadbent, E., Dao, D., Zhang, C., Zambrano, A. A. and Max, S., 2024. OAM-TCD: A globally diverse dataset of high-resolution tree cover maps. In: *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J. and Zhao, H., 2024. Depth anything: Unleashing the power of large-scale unlabeled data. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10371–10381.

https://doi.org/10.5194/ica-abs-10-307-2025