A BENCHMARK HIGH-RESOLUTION MULTI-TEMPORAL REMOTE SENSING DATASET FOR URBAN TREE SEMANTIC SEGMENTATION

Hongli Li

Shenzhen University fe

ABSTRACT

Trees are vital to urban ecosystems and the global carbon cycle. Accurate tree identification and segmentation support urban greening monitoring, ecological assessment, and carbon sink estimation. However, existing datasets mostly rely on single-temporal imagery, limiting studies of long-term tree dynamics. To address this, we construct Tree-RGB-MT, a multi-temporal urban tree segmentation dataset based on high-resolution RGB aerial imagery (2013–2022) from multiple cities and seasons, with fine-grained pixel-level annotations. Experiments with state-of-the-art semantic segmentation models demonstrate that Tree-RGB-MT poses significant challenges in boundary complexity, scene diversity, and temporal consistency. Thus, the dataset provides a robust benchmark for studying tree dynamics, phenological variation, and urban greening evolution.

Index Terms— multi-temporal urban tree dataset, high-resolution remote sensing imagery, semantic segmentation, tree segmentation

1. INTRODUCTION

With the accelerating pace of climate change and urbanization, urban ecosystems are facing unprecedented challenges. As a key component of green infrastructure, trees play a vital role in mitigating the urban heat island effect and supporting biodiversity [1, 2, 3]. Consequently, the accurate and continuous identification and monitoring of urban trees has become a pressing scientific priority. Such efforts provide essential data for urban greening strategies, ecosystem service assessment, and carbon accounting, while also underpinning evidence-based policies for sustainable urban development [4, 5].

In recent years, with the rapid development of high-resolution remote sensing imagery and UAV-based aerial photography, the acquisition of sub-meter resolution urban imagery has become feasible [6, 7]. These data provide critical sources for fine-grained identification and dynamic monitoring of urban trees. On the methodological side, the emergence of deep learning, especially convolutional neural networks (CNNs) and Transformer-based architectures, has demonstrated outstanding performance in image segmentation and object recognition tasks [8, 9, 10, 11]. Compared with traditional methods relying on manual interpretation or rule-based image processing, deep learning-based semantic segmentation can automatically learn multi-level features, achieving a higher degree of automation and robustness in complex urban environments. However, most existing studies and publicly available datasets remain focused on single-temporal imagery, without adequately considering the temporal dynamics of trees [12].

In reality, processes such as tree growth, pruning, disease, and seasonal leaf changes can significantly alter tree morphology and spectral characteristics over time [13, 14]. For example, spring budding, dense summer canopies, autumn leaf color changes and shedding, and bare branches in winter all introduce substantial variability in tree appearance across images. When training data are limited to a single temporal snapshot, models often struggle to generalize across different years and seasons, which restricts their application in long-term monitoring. Furthermore, rapid urban development—including road construction, demolition, and re-greening—continuously reshapes the spatial distribution of trees, imposing even higher demands on the temporal adaptability of segmentation models.

At the same time, publicly available tree datasets exhibit clear limitations in the temporal dimension. Representative datasets (e.g., ForestsNet, TreeSatAI) mainly focus on tree segmentation at a single time point [5, 12]. While these datasets have advanced algorithm benchmarking and comparative studies, they generally neglect cross-year and cross-season variability. This not only limits systematic research on urban greening dynamics but also hinders practical deployment of models for continuous monitoring and long-term forecasting. In other words, the lack of high-quality datasets with both multi-temporal coverage and diverse urban contexts has become a bottleneck restricting further progress in tree semantic segmentation and dynamic monitoring research.

To address this gap, we propose and construct a new multi-temporal urban tree semantic segmentation dataset, Tree-RGB-MT. This dataset spans a decade (2013–2022) and covers multiple representative urban regions in China. It is divided into five temporal stages, each comprising a two-year interval (2013, 2015, 2017, 2019, 2021). Each stage includes diverse urban environments—such as streets, parks, campuses, residential areas, and riverside green spaces—and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

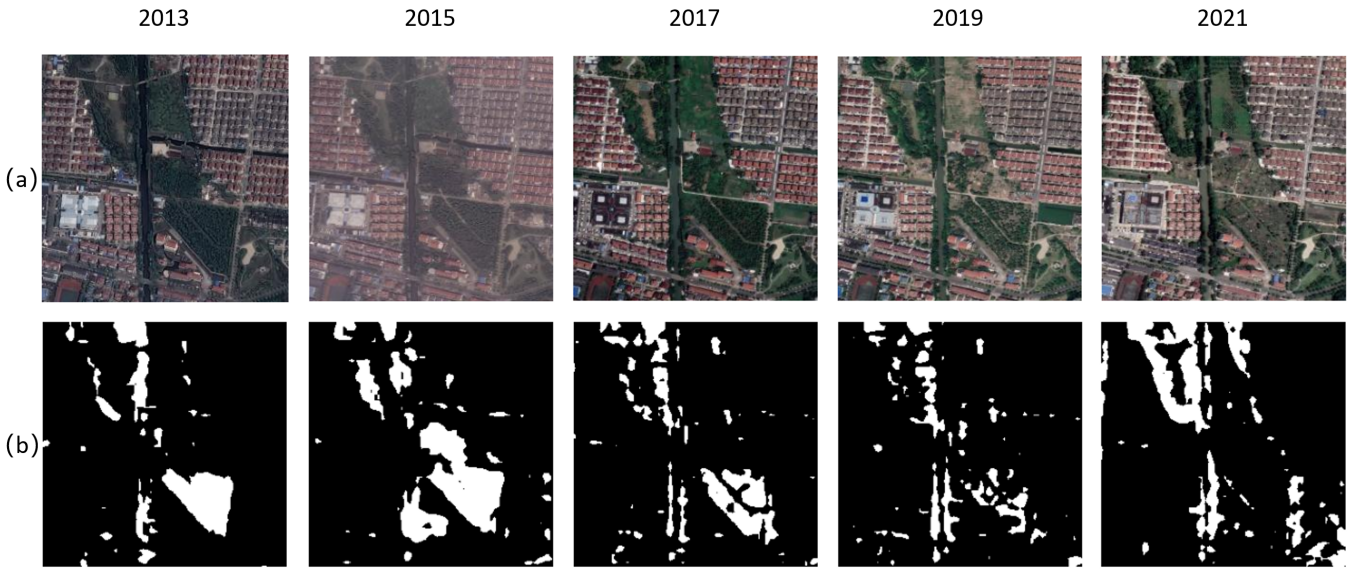

Fig. 1. Visualization of multi-temporal segmentation results on the Tree-RGB-MT dataset. Each column corresponds to a different temporal phase (2013, 2015, 2017, 2019, 2021). (a) High-resolution RGB imagery. (b) Segmentation results produced by our method.

provides high-precision pixel-level annotations. Compared to existing single-temporal datasets, Tree-RGB-MT not only supports standard semantic segmentation tasks but also enables systematic studies of long-term tree dynamics, seasonal variability, and cross-temporal generalization of segmentation models.

2. TREE-RGB-MT DATASET

2.1. Data Collection

The Tree-RGB-MT dataset consists of imagery collected from several representative cities across China, covering coastal, inland, and diverse climatic regions to ensure spatial diversity. The dataset spans the period 2013–2022 and is divided into two-year intervals, forming five temporal phases: 2013, 2015, 2017, 2019, and 2021. The imagery was primarily acquired using UAV platforms and high-resolution aerial photography systems, with a spatial resolution ranging from 0.1 to 0.3 m. This resolution enables clear identification of individual trees, canopy boundaries, and their spatial relationships with surrounding objects. Compared to conventional satellite imagery, such high-resolution aerial data are better suited to the challenges of tree detection in complex urban environments, particularly in scenarios such as roadside green belts, residential courtyards, and small urban woodlands. Moreover, the multi-temporal acquisition strategy ensures coverage across different years and seasons, allowing systematic observation of dynamic changes in trees influenced by growth, pruning, leaf fall, and diseases [15].

2.2. Annotation Strategy

The dataset adopts fine-grained, manually curated pixel-level annotations, with two semantic classes: tree and background. During annotation, the team carefully delineated individual tree crowns, paying close attention to subtle boundaries between trees and background elements such as buildings, roads, and lawns [7]. To guarantee consistency across temporal phases, all annotations followed a unified set of guidelines and standards, and were carried out by annotators with expertise in remote sensing interpretation and ecological knowledge. Each temporal subset underwent multiple rounds of cross-validation and quality control to ensure that tree annotations in the same area were comparable and continuous across different time periods. This rigorous annotation process not only enhances the dataset's reliability but also provides a solid foundation for cross-year model training and evaluation. Compared with automated or semi-automated approaches, fully manual pixel-level annotation significantly reduces errors, particularly in challenging scenarios involving overlapping canopies, shadowed areas, and small-scale tree identification.

2.3. Dataset Scale and Value

The Tree-RGB-MT dataset contains approximately 75,000 high-resolution images, each with a resolution of 1024×1024 pixels. These images are evenly distributed across the five

temporal phases, with about 15,000 images per phase. The dataset is organized into training, validation, and test subsets with a ratio of 70%, 20%, and 10%, respectively. Compared with existing single-temporal tree segmentation datasets, the key strengths of Tree-RGB-MT lie in its temporal depth and diversity. The multi-temporal design enables systematic cross-year monitoring and long-term change detection, facilitating research on how trees evolve across seasons, years, and urban development processes [12]. The high spatial resolution ensures that individual tree boundaries are clearly distinguishable, enabling fine-scale analyses of tree structure and distribution. Additionally, the dataset encompasses a wide range of urban scenarios—including streets, parks, campuses, residential areas, and riverbanks—ensuring strong representativeness and transferability. To further illustrate the characteristics of the Tree-RGB-MT dataset, Figure 1 shows representative samples across five temporal phases, including original imagery and corresponding segmentation results.

3. METHODS

3.1. Baseline Models

To systematically evaluate the challenges and applicability of the Tree-RGB-MT dataset, we conducted experiments on three representative semantic segmentation models: U-Net++, DeepLabV3+, and SegFormer. These models represent mainstream approaches in semantic segmentation, namely convolutional neural networks (CNNs), improved dilated convolution networks, and Transformer-based architectures, respectively. U-Net++ introduces dense skip connections and deep supervision to mitigate semantic gaps across different feature scales, thereby improving boundary recovery of trees across temporal phases. This makes it particularly suitable for handling scenes where tree crowns exhibit complex morphology or are adjacent to background objects. DeepLabV3+ leverages atrous convolution to capture multi-scale contextual features and demonstrates strong robustness when tree morphology, size, and density vary significantly across years. Its decoder structure further enhances boundary delineation in fine-detail regions [10]. SegFormer, based on the Transformer architecture, benefits from global modeling capabilities, which allow it to better capture long-range dependencies and adapt to spectral and textural variations across seasons and cities. This advantage enables it to effectively address the heterogeneous distribution of trees in complex temporal imagery. By comparing these three categories of models, we aim to comprehensively reflect the dataset's challenges while also providing a methodological benchmark for future research.

3.2. Data Augmentation

Given the pronounced seasonal and interannual variations in multi-temporal urban tree imagery—such as lush canopies in summer versus leaf-off conditions in winter, or structural changes caused by pruning and disease—we adopted a diverse set of data augmentation strategies to enhance the temporal generalization ability of models. The augmentations include random rotations and horizontal/vertical flips to simulate variations in aerial view perspectives; illumination perturbations and color jittering to replicate spectral changes under different seasonal or weather conditions; and cropping and scaling operations to mimic tree structures under varying resolutions and growth stages. These strategies effectively enrich training data diversity and prevent models from overfitting to specific temporal phases or static tree characteristics, thereby improving robustness in real-world dynamic scenarios. Compared with traditional augmentation approaches limited to geometric transformations, the multidimensional strategy employed in this study better reflects the temporal variability of urban trees, ensuring more stable model performance on cross-year imagery.

3.3. Training and Optimization

All experiments were conducted using the PyTorch 2.0 framework on an NVIDIA A40 GPU platform to enable efficient training on large-scale, high-resolution data. We adopted the AdamW optimizer, which effectively reduces overfitting through weight decay, with an initial learning rate set to 2×10^{-4}. A cosine annealing learning rate scheduler was applied to gradually decrease the learning rate, ensuring stable convergence in later training stages. The loss function was a combination of cross-entropy and Dice loss: cross-entropy ensured overall classification accuracy, while Dice loss was particularly advantageous for handling tree boundaries and small-scale tree instances. Training was carried out for 100 epochs, with an early stopping mechanism employed to terminate training if validation performance failed to improve over several consecutive epochs, thereby avoiding overfitting and enhancing efficiency. Additionally, mixed precision training was applied to reduce memory usage and accelerate computation. This overall training and optimization strategy ensured stable convergence of models on the Tree-RGB-MT dataset while providing a reproducible benchmark pipeline for cross-temporal tree monitoring tasks.

4. EXPERIMENTS

4.1. Experimental Settings

To comprehensively evaluate the applicability of the Tree-RGB-MT dataset in cross-temporal tree segmentation tasks, we designed two experimental settings: single-temporal and cross-temporal. In the single-temporal experiments, models were trained and tested within the same temporal segment (e.g., 2013 or 2017), in order to assess their segmentation capability under relatively static conditions [8, 9]. In the cross-temporal experiments, however, the training and testing sets

were drawn from different temporal segments—for example, training on the 2013 phase and testing on the 2021 phase—to simulate real-world scenarios where models must generalize across multi-year spans. This design enables both intra-phase evaluation (per temporal subset) and inter-phase generalization tests. To ensure fair comparison, training, validation, and test sets were kept mutually exclusive, and all experiments were conducted under identical hardware and optimization settings.

4.2. Evaluation Metrics

To evaluate model performance from different perspectives, we adopted mean Intersection-over-Union (mean IoU) and Overall Accuracy (OA) as the core metrics. Mean IoU, the most widely used metric in semantic segmentation, effectively measures class-wise segmentation accuracy and is especially informative for assessing model performance on tree boundaries and small-scale regions. In the multi-temporal setting, mean IoU also reflects how consistently models can delineate tree crowns across different years. OA, by contrast, evaluates the overall correctness of pixel-level classification, providing an intuitive measure of model applicability on large-scale temporal datasets and their stability over time. Together, these two metrics capture both fine-grained segmentation precision and large-scale temporal robustness, thereby offering a comprehensive basis for performance comparison across different architectures.

4.3. Experimental Results

Table 1 presents the performance of U-Net++ and SegFormer across the five temporal phases of the Tree-RGB-MT dataset. The results indicate that SegFormer consistently achieved higher mean IoU and OA values across both validation and test sets, confirming its advantage in handling multi-temporal variations with its Transformer-based global modeling capability. In particular, SegFormer demonstrated stable performance across different phases, with test mean IoU values around 0.60–0.61 and OA up to 0.96, showing strong adaptability to variations in illumination, texture, and seasonal changes. U-Net++, while slightly weaker in temporal generalization (test mean IoU around 0.57–0.60), still maintained competitive OA scores (≈0.92–0.93) and exhibited notable strengths in boundary delineation, especially for trees adjacent to roads and buildings. These results underscore both the challenges inherent in cross-temporal tree segmentation and the complementary strengths of CNN-based and Transformer-based models, providing a comprehensive baseline for future research on dynamic tree monitoring.

4.4. Qualitative Analysis

The experimental results validate both the challenges and research value of the Tree-RGB-MT dataset for evaluating cross-temporal tree segmentation tasks. Compared with single-temporal tasks, cross-temporal segmentation substantially increases the difficulty, underscoring the significant impact of temporal variations in trees across years, seasons, and environmental conditions on semantic segmentation models. In terms of model performance, Transformer-based architectures demonstrated stronger robustness in handling long-term temporal spans and complex illumination conditions, while CNN-based models still proved indispensable for detailed boundary recovery and small-scale object segmentation [9, 10]. These findings suggest that future research could explore hybrid architectures that combine the local feature strengths of CNNs with the global modeling power of Transformers [11], thereby building more effective solutions for multi-temporal urban tree monitoring. Moreover, the experimental framework and results presented in this study provide a reproducible benchmark for the community, contributing to further advancements in cross-temporal tree segmentation research.

Table 1. Validation and test performance of U-Net++ and SegFormer on each temporal phase of Tree-RGB-MT.

Phase	Model	Params (M)	Validation		Test	
			mean IoU	OA	mean IoU	OA
2013	U-Net++	90.3	0.66	0.93	0.58	0.92
	SegFormer	44.6	0.68	0.93	0.61	0.95
2015	U-Net++	90.3	0.65	0.92	0.57	0.92
	SegFormer	44.6	0.67	0.93	0.60	0.95
2017	U-Net++	90.3	0.64	0.92	0.59	0.93
	SegFormer	44.6	0.66	0.92	0.60	0.96
2019	U-Net++	90.3	0.65	0.93	0.60	0.93
	SegFormer	44.6	0.67	0.92	0.61	0.96
2021	U-Net++	90.3	0.66	0.93	0.59	0.93
	SegFormer	44.6	0.68	0.93	0.60	0.96

5. CONCLUSION

The paper presents the Tree-RGB-MT dataset, a multi-temporal, high-resolution benchmark for urban tree semantic segmentation, covering Chinese cities from 2013 to 2022 with fine pixel-level annotations. It fills an important gap in existing datasets by combining spatial detail with temporal depth, and highlights challenges in cross-temporal segmentation such as seasonal variation, tree growth, and environmental disturbances. Baseline experiments demonstrate complementary strengths: transformer models show superior temporal generalization, whereas CNNs excel in boundary recovery, underscoring the value of integrating local and global features.

6. REFERENCES

[1] Yude Pan, Richard A. Birdsey, Jingyun Fang, Richard Houghton, Pekka E. Kauppi, Werner A. Kurz, Oliver L. Phillips, Anatoly Shvidenko, Simon L. Lewis, Josep G. Canadell, Philippe Ciais, Robert B. Jackson, Stephen W. Pacala, A. David McGuire, Shilong Piao, Aapo Rautiainen, Stephen Sitch, and Daniel Hayes, "A large and persistent carbon sink in the world's forests," *Science*, vol. 333, no. 6045, pp. 988–993, 2011.

[2] Martin Brandt, Compton J. Tucker, Ankit Kariryaa, Kjeld Rasmussen, Christin Abel, Jennifer Small, Jerome Chave, Laura Vang Rasmussen, Pierre Hiernaux, Abdoul Aziz Diouf, Laurent Kergoat, Ole Mertz, Christian Igel, Fabian Gieseke, Johannes Schöning, Sizhuo Li, Katherine Melocik, Jesse Meyer, Scott Sinno, Eric Romero, Erin Glennie, Amandine Montagu, Morgane Dendoncker, and Rasmus Fensholt, "An unexpectedly large count of trees in the West African Sahara and Sahel," *Nature*, vol. 587, no. 7832, pp. 78–82, Nov. 2020.

[3] Song Leng, Ranhao Sun, Xiaojun Yang, Mingxin Jin, and Liding Chen, "Diverse types of coupling trends in urban tree and nontree vegetation associated with urbanization levels," *npj Urban Sustainability*, vol. 3, no. 1, pp. 33, June 2023.

[4] M. C. Hansen, P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, A. Kommareddy, A. Egorov, L. Chini, C. O. Justice, and J. R. G. Townshend, "High-resolution global maps of 21st-century forest cover change," *Science*, vol. 342, no. 6160, pp. 850–853, 2013.

[5] Shidong Liu, Jie Zhang, Li Wang, Philippe Ciais, Jianjun Zhang, Josep Penuelas, Biswajit Nath, Iskid Jacquet, Xia Wu, Shengping Ding, Wang Li, Ni Huang, Wanjuan Song, WenJian Ni, and Zheng Niu, "Mapping previously undetected trees reveals overlooked changes in pan-tropical tree cover," *Nature Communications*, vol. 16, no. 1, pp. 5561, July 2025.

[6] Jamie Tolan, Hung-I Yang, Benjamin Nosarzewski, Guillaume Couairon, Huy V. Vo, John Brandt, Justine Spore, Sayantan Majumdar, Daniel Haziza, Janaki Vamaraju, Theo Moutakanni, Piotr Bojanowski, Tracy Johns, Brian White, Tobias Tiecke, and Camille Couprie, "Very high resolution canopy height maps from rgb imagery using self-supervised vision transformer and convolutional decoder trained on aerial lidar," *Remote Sensing of Environment*, vol. 300, pp. 113888, Jan. 2024.

[7] Ben G Weinstein, Sarah J Graves, Sergio Marconi, Aditya Singh, Alina Zare, Dylan Stewart, Stephanie A Bohlman, and Ethan P White, "A benchmark dataset for canopy crown detection and delineation in co-registered airborne rgb, lidar and hyperspectral imagery from the national ecological observation network," *PLoS computational biology*, vol. 17, no. 7, pp. e1009180, 2021.

[8] Jonathan Long, Evan Shelhamer, and Trevor Darrell, " Fully convolutional networks for semantic segmentation ," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA, June 2015, pp. 3431–3440, IEEE Computer Society.

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, Eds., Cham, 2015, pp. 234–241, Springer International Publishing.

[10] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[11] Wei Wang, Chen Tang, Xin Wang, and Bin Zheng, "A vit-based multiscale feature fusion approach for remote sensing image segmentation," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[12] F. Yang and Z. Zeng, "Refined fine-scale mapping of tree cover using time series of Planet-NICFI and Sentinel-1 imagery for Southeast Asia (2016–2021)," *Earth System Science Data*, vol. 15, no. 9, pp. 4011–4021, 2023.

[13] Juan P. Ardila, Wietske Bijker, Valentyn A. Tolpekin, and Alfred Stein, "Context-sensitive extraction of tree crown objects in urban areas using VHR satellite images," *International Journal of Applied Earth Observation and Geoinformation*, vol. 15, pp. 57–69, 2012.

[14] L. Monika Moskal, Diane M. Styers, and Meghan Halabisky, "Monitoring urban tree cover using object-based image analysis and public domain remotely sensed data," *Remote Sensing*, vol. 3, no. 10, pp. 2243–2262, 2011.

[15] Juan Ardila, Valentyn Tolpekin, and Wietske Bijker, "Markov random field based super-resolution mapping for identification of urban trees in vhr images," in *2010 IEEE International Geoscience and Remote Sensing Symposium*, 2010, pp. 1402–1405.