International Journal of Applied Earth Observation and Geoinformation 152 (2026) 105449

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag

Individual urban street tree detection and species-level mapping from street-view images and GIS data

Weijia Li [a], Xinjie Huo [b], Junyan Ye [b], Jinhua Yu [b], Hanxi Wang [b], Juepeng Zheng [c,*], Xiang Zhang [b,*], Haohuan Fu [a]

[a] Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[b] School of Geospatial Engineering and Science, Sun Yat-sen University, Zhuhai, China
[c] School of Artificial Intelligence, Sun Yat-sen University, Zhuhai, China

ARTICLE INFO

Keywords:
Tree species classification
Street-view image
OpenStreetMap
Object detection
Street tree mapping

ABSTRACT

Accurate urban street tree recognition is crucial for ecological management. Street-view images offer a unique pedestrian perspective, capturing details often missed by remote sensing imagery. Current street tree recognition methods using street-view images are limited by scarce labeled data and high manual annotation costs. To address these challenges, we propose a novel urban street tree recognition and mapping method that enables accurate instance-level mapping of street trees. The method supports species identification where sufficient visual distinction exists. First, we design a two-stage street tree recognition method. In stage one, an object detection model is used to extract individual tree regions from street-view images, decoupling tree detection from species classification. In stage two, tree species are classified using a model pre-trained on open-source tree species datasets and fine-tuned with web-sourced labeled images, achieving annotation cost reduction. Second, StreetTreeMapper reduces duplicate detections by integrating street-view image features with OpenStreetMap's geographic constraints, achieving precise tree instance mapping. To validate our proposed method, we construct a street tree dataset from New York, Zhuhai, and Seoul. Experimental results validate the overall effectiveness of our proposed method in both tree recognition and mapping. Through decoupled detection and classification, our method achieves a 2.7% to 11.4% increase in average precision compared to state-of-the-art models on Zhuhai datasets. The StreetTreeMapper achieves instance-level tree mapping with median errors of 4.5 m (New York), 4.8 m (Seoul), and 2.9 m (Zhuhai), reaching minimum observed localization errors of 0.21 m, 0.13 m, and less than 0.001 m, respectively. Our method enables automated detection and species identification of street trees. These capabilities provide valuable data support and decision-making foundations for smart city development and sustainable ecological maintenance.

1. Introduction

The accurate mapping of individual urban street trees plays a crucial role in advancing urban sustainability. This capability allows for the precise quantification of key ecosystem services, including carbon sequestration (Lind et al., 2023) and climate regulation (Şimşek et al., 2024), while also facilitating the evaluation of species-specific stress responses (Zhang et al., 2024). The potential of deep learning for individual tree detection is demonstrated by its application to high-resolution satellite imagery (Li et al., 2019). When deployed on street-level imagery, these deep learning approaches address biological heterogeneity and environmental noise, providing essential data for biodiversity monitoring and ecosystem assessment. Specifically, systematic species-level mapping serves as technical support for analyzing the impact of atmospheric NO_X concentrations on vegetation health. Furthermore, when combined with emerging structural analysis (Long et al., 2026) and crown extraction techniques (Lian et al., 2024), this approach enables a detailed characterization of urban forests, assisting in quantifying carbon density (Liang et al., 2024) and monitoring allergenic species for public health risk assessment (Li et al., 2024).

While remote sensing provides a viable path for large-scale mapping (Ventura et al., 2024), its application is hindered by a heavy reliance on large-scale, manually annotated datasets. This challenge of annotation dependency is equally prevalent in deep learning-based tree detection and recognition studies (Arevalo-Ramirez et al., 2024). The development of intelligent approaches for instance-level street tree mapping under limited annotation constraints is thus essential for

* Corresponding authors.
E-mail addresses: zhengjp8@mail.sysu.edu.cn (J. Zheng), zhangx795@mail.sysu.edu.cn (X. Zhang).

https://doi.org/10.1016/j.jag.2026.105449
Received 31 October 2025; Received in revised form 16 May 2026; Accepted 28 June 2026
Available online 14 July 2026

progress in urban sustainability and precision forestry. Street-view images have become a cornerstone data source for urban intelligence, with applications ranging from habitat monitoring and infrastructure assessment to the tracking of Sustainable Development Goals (SDGs) and urban land use mapping (Zhang et al., 2021; Yang et al., 2025; Li and Ni, 2025; Che et al., 2025). The geometric and perceptual information in SVIs further enables the modeling of complex human–environment interactions, such as using depth perception to explain driving behavior (Zhang et al., 2025). While Yang et al. (2025) demonstrated the effectiveness of mapping standardized infrastructure like streetlights, our study addresses the higher complexity of urban vegetation. Beyond straightforward spatial localization, the biological heterogeneity and seasonal phenology of trees introduce significant semantic ambiguities, making instance-level correspondence and fine-grained classification a more formidable challenge. The integration of advanced analytics, such as unsupervised learning and perception quantification, enables the establishment of measurable frameworks for monitoring urban sustainability, including the assessment of physical changes and their socio-environmental consequences (Liu et al., 2025). Within this context, the accurate mapping of ubiquitous urban elements like street trees emerges as a valuable and challenging application. Compared to top-down remote sensing data such as satellite imagery and UAV images (Zheng et al., 2024, 2023, 2021), street-view images provide a more intuitive and cost-effective means for individual tree detection and classification (Liu et al., 2023), primarily because the ground-level vantage point enables fine-grained observation of tree morphology (Sarlin et al., 2023; Hu et al., 2023). However, species-level mapping is constrained by the scarcity of detailed annotations, whose costly manual collection limits the scalability of deep learning for this task. On the other hand, open geospatial data (e.g., OpenStreetMap) contains structural information about urban environments, including road networks and spatial layouts. Recent efforts in Bird's-Eye-View (BEV) representation have sought to address such cross-view perspective differences (Ye et al., 2024a, 2025a). However, due to the lack of instance-level tree data and the perspective gap with street-view images, it is challenging to directly utilize this data for accurate tree-level mapping.

Deep learning-based approaches have become the predominant method for detecting and classifying trees from street-level images (Wu and Qian, 2021; Li and Yao, 2020). However, the development of robust deep learning models for street-tree classification is hindered by the scarcity of large-scale (Yang et al., 2023; Beery et al., 2022), finely annotated street-tree datasets.

While this data limitation poses a significant challenge, an even more critical technical problem emerges when integrating recognition results into accurate, instance-level maps, namely, the instance correspondence problem (Arevalo-Ramirez et al., 2024). This problem, which refers to the difficulty of determining whether trees detected across multiple overlapping street-view images belong to the same physical entity, remains a major obstacle. Current mapping methods, which often rely on error-prone GPS metadata or visual features alone, struggle with this issue in complex urban environments (Li and Yao, 2020; Roberts et al., 2019; Laumer et al., 2020), leading to inaccuracies such as double-counting or missing individual trees. Therefore, there is a pressing need for a method that can not only perform well under limited annotation constraints but also effectively resolve the instance correspondence problem to generate reliable, large-scale street tree maps.

To address these challenges, we propose StreetTreeMapper, a novel framework that integrates street-view images with open geospatial data for instance-level street tree mapping. The framework is designed to provide a robust processing solution for large-scale urban forestry surveys, assisting municipal departments in mapping tree distributions more effectively. By providing species-level maps, our approach can be integrated with emerging structural analysis (Long et al., 2026) and crown extraction techniques (Lian et al., 2024) to facilitate a more detailed characterization of urban forests. Such a synergy is essential for downstream applications, including the quantification of carbon density (Liang et al., 2024) and the enhancement of city-scale ecological resilience (Velasquez-Camacho et al., 2024).

The framework adopts a two-stage strategy to reduce reliance on large-scale fine-grained species annotations. By separating tree detection from species recognition, web-sourced imges can be incorporated to facilitate species classification training. In addition, a dedicated mapping module addresses the instance correspondence problem by fusing visual similarity with spatial constraints from open map data, enabling unique identification of individual street trees.

In this paper, our main contributions are:

- We introduce a street tree recognition method that alleviates the heavy reliance on manually annotated street-view images by decoupling detection from species classification,as validated for species classification on the Zhuhai dataset.
- We propose StreetTreeMapper, a street tree mapping method that achieves instance-level results by uniquely combining image similarity computation and OpenStreetMap spatial constraints to avoid double-counting the same tree across multiple street-view images.
- We construct and release street tree datasets from New York, Zhuhai, and Seoul to validate our proposed methods for tree recognition and instance-level mapping, while also providing a valuable resource for the research community for further urban tree analysis.

2. Related work

2.1. Deep learning-based approaches for urban street tree recognition

Research on urban street tree analysis has increasingly relied on deep learning models applied to street-view imagery, with a primary focus on tree recognition tasks, namely detection and species classification (Liu et al., 2023; Jiang et al., 2024). These methods have become the predominant approach due to their strong performance. Object detection algorithms, particularly the efficient YOLO series (Redmon et al., 2016), are widely adopted to locate individual tree instances in street-level photographs. For the more fine-grained task of species classification, convolutional neural networks (CNNs) have demonstrated remarkable capability in distinguishing visual features, achieving robust performance even on close-up imagery such as leaves (Wegner et al., 2016; Branson et al., 2018; Wu and Qian, 2021).

However, the primary challenge hindering the development of more robust and generalizable recognition models is no longer the model architecture, but the scarcity of large-scale, finely annotated street-tree datasets (Yang et al., 2023). The high cost and labor-intensive nature of manual annotation create a significant bottleneck. This data scarcity directly limits the scalability and practical deployment of fully-supervised recognition methods in diverse urban environments. While some studies have explored multi-modal data fusion (e.g., with LiDAR or satellite imagery) to enhance accuracy (Wu et al., 2018; Ye et al., 2024b), these approaches often introduce higher data acquisition costs and complexity, rather than alleviating the core data dependency. Recent geometry-aware semi-supervised approaches offer a promising alternative (Li et al., 2025), though the complex morphology of street trees presents unique challenges compared to buildings.

In summary, while deep learning provides powerful tools for street tree recognition, the field's progress is constrained by its reliance on extensive manual annotations. This limitation highlights the demand for innovative methods capable of accurate recognition without the need for large-scale labeled datasets.

Table 1
Characteristics of the proposed StreetTreeMapper in the context of existing street-level tree detection and mapping literature.

Aspect	Lumnitz et al. (2021)	Liu et al. (2023)	Choi et al. (2022)	This study
Primary objective	Detection and geolocation	Species classification	Urban tree profiling	Instance-level detection and mapping
Study area	Vancouver	Jinan	Seoul	Zhuhai, Seoul, New York
Reported performance	>70% detection, 4-6 m error	0.587 mAP@0.5	0.564 mAP	4-5 m median localization error

2.2. Instance-level mapping of street trees from street-view images

Research on instance-level mapping from street-view imagery focuses on the core challenge of the instance correspondence problem: accurately determining the geographic location of each unique tree instance. Early approaches relied on simple metadata like GPS coordinates for spatial association (Zhou et al., 2022). However, the limitations of these metadata-based methods motivated a shift towards vision-based techniques. The feasibility of this approach was initially demonstrated by CNN-based methods for detecting individual trees in street-level images (Wegner et al., 2016; Branson et al., 2018; Wu and Qian, 2021; Li and Yao, 2020).

Subsequent advancements have focused on deriving precise locations through geometric analysis. Key techniques include using depth estimation and triangulation from image sequences to determine 3D positions (Lumnitz et al., 2021; Liu et al., 2023; Jiang et al., 2024), as well as employing photogrammetry with deep learning for accurate coordinate extraction. These methods have been extended to estimate tree parameters like DBH and height by using reference objects for scale (Wang et al., 2018; Choi et al., 2022).

The method of geometric localization from street-view images is also applied to other urban objects. A pertinent comparison is the work of (Yang et al., 2025) on streetlight assessment, which also follows a “detect-locate-assess” pipeline. A key difference lies in the localization method. Their approach for sparse streetlights is based on multi-view triangulation, whereas our method adopts a map-constrained, single-view geometric localization approach to estimate tree locations. This design choice prioritizes computational efficiency and scalability for city-wide deployment, offering a pragmatic alternative to multi-view solutions that require cross-view matching.

In summary, a core challenge in instance-level street tree mapping is the instance correspondence problem—the need to match tree detections across views while minimizing double-counting. While existing methods excel at detection, they offer limited solutions for reliable cross-view association. StreetTreeMapper presents an improved approach that integrates image similarity with OpenStreetMap constraints to provide more reliable association results for urban tree mapping.

To further contextualize our contribution, Table 1 summarizes the key characteristics of StreetTreeMapper in comparison with other representative domain-specific studies. Unlike previous works that often focus on single-city optimization, our framework emphasizes cross-city scalability and instance-level mapping across heterogeneous urban environments.

3. Study area and data curation

3.1. Study areas and data sources

The study area comprises three regions characterized by distinct urban morphologies, including Brooklyn in New York, Xiangzhou in Zhuhai, and Seongdong-gu in Seoul. Data distribution and acquisition details are shown in Fig. 1. The evaluation leverages three datasets for distinct purposes. The New York dataset, with its manual instance-level location annotations, is used for testing basic mapping capability. The Seoul dataset, featuring a large image collection but limited annotations, is employed to challenge the method’s performance in large-scale environments under the constraint of limited labeled data. Meanwhile, the Zhuhai dataset, with its detailed species labels, is utilized for validating the two-stage recognition method.

Table 2
Dataset statistics and processing pipeline.

Study area	Zhuhai	Seoul	New York
Training set (images)	2078	400	5063
Test set (images)	1016	100	174
Original resolution (pixels)	512 × 512	512 × 512	512 × 512
Cropped resolution	Variable (as detected)		
Label procedure:			
1. Tree instances manually annotated (Labelme) with bounding boxes.			
2. Annotations saved in JSON.			
Key preprocessing steps:			
1. Object detection to locate tree instances.			
2. Cropping based on bounding boxes.			

Table 3
Statistical summary of cropped image dimensions (pixels). Median represents the central value, and IQR represents the interquartile range.

City	Stat	Dimension (pixels)	
		Width	Height
Zhuhai	Median	56	99
	IQR	36–84	65–155
Seoul	Median	78	177
	IQR	39–153	117–274
New York	Median	96	207
	IQR	67–140	126–298

Street-view images for the three datasets are sourced from Google Street View (GSV),[1] Baidu Street View (BSV),[2] and the OmniCity dataset (Li et al., 2023). The New York region primarily uses the OmniCity dataset. For object detection, OmniCity’s panoramic images are divided into four perspective-view images based on fixed viewing angles. For the Zhuhai region, we directly use single-view images provided by BSV without additional processing. To address the missing true north orientation information in street-view images of Zhuhai, we use OrienterNet (Sarlin et al., 2023) to estimate the orientation of image query locations. OrienterNet infers direction by matching neural bird’s-eye views with OpenStreetMap image maps. We further utilize OpenStreetMap (OSM)[3] data to extract the road network geometry of each study area, serving as a geographic reference for tree localization. In addition, forestry survey data from New York[4] and Seoul[5] are used to validate and refine the mapping results. The key statistics of the datasets, including image counts, resolutions, and species distributions, are summarized in Tables 2 and 4.

The cropped individual tree images exhibit variable dimensions, as they are tightly fitted to the detection bounding boxes. A statistical summary of the cropped image sizes across the three cities is provided in Table 3. The median widths range from 56 to 96 pixels, and the median heights from 99 to 207 pixels. The interquartile ranges (IQRs) indicate that the majority of cropped images preserve sufficient pixel-level detail for distinguishing key species-specific features,

[1] https://www.google.com/streetview/
[2] https://map.baidu.com
[3] https://www.openstreetmap.org/
[4] https://climate.cityofnewyork.us/zh/subtopics/green-space/
[5] https://data.seoul.go.kr/dataList/OA-1325/S/1/datasetView.do?tab=F

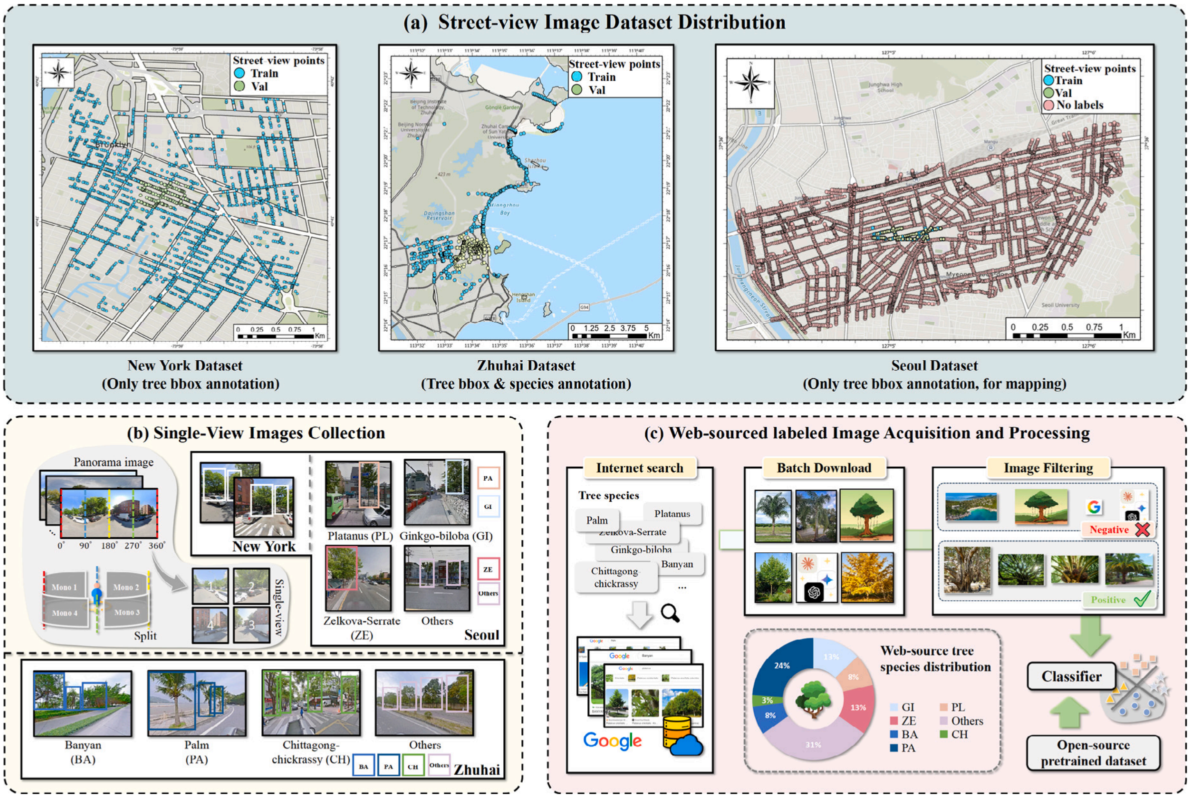


Fig. 1. Overview of the datasets used in this work. (a) Our street-view image dataset covers Brooklyn (New York), Xiangzhou (Zhuhai), and Seongdong-gu (Seoul). (b) Our single-view image collection includes images derived from panoramas in New York and Seoul. (c) Our labeled web image acquisition workflow involved Internet Search, Batch Download, and Image Filtering. This data, combined with a publicly available tree image dataset for pretraining, was then fine-tuned for classification.

Table 4
Tree species distribution.

Data	Species	Train	Test	Total
Zhuhai	BA	1242	586	1828
	CH	432	143	575
	PA	2804	523	3327
	Others	3882	1775	5657
	Total	**8,360**	**3,027**	**11,387**
Seoul	GI	–	12,303	12,303
	PL	–	40,673	40,673
	ZE	–	3322	3322
	Others	–	7381	7381
	Total	–	**63,679**	**63,679**

such as overall tree shape and bark texture, ensuring reliable species identification.

An urban forestry study (Jiang et al., 2013) served as the foundation for classifying tree species in Zhuhai, as it had documented the dominant species through analyses of their morphology and spatial distribution. Four representative categories were adopted: Banyan (BA), Chittagong-chickrassy (CH), Palm (PA), and Other. The Seoul dataset categorizes trees into four dominant species according to historical official census data[6] and existing studies (Choi et al., 2022), specifically Ginkgo biloba (GI), Platanus (PL), Zelkova serrata (ZE), and Other. The annotators determined tree species by examining morphological characteristics in the street-view images.

3.2. Annotation and data collection

The manual annotation of street trees for the three cities was conducted by five volunteers with academic backgrounds in geospatial studies. All annotators received unified labeling guidelines prior to the annotation process. To ensure label consistency, a cross-review protocol was implemented in which each annotation was independently checked by at least one additional annotator. Discrepancies were resolved through consensus discussion before final confirmation. We enriched the training data with multiple sources in order to enhance the classification model's performance. We utilized an open-source tree image dataset (Yang et al., 2023), which provides representative samples across a wide range of tree species, for pretraining the classifier. Then, we fine-tuned it with supplementary internet images sourced through Internet Search. We employed a browser extension named ImageAssistant Batch Image Downloader[7] to download images in batches, filtering out low-quality or irrelevant results. Specifically, we defined "low-quality" images as those with a resolution below 100 × 100 pixels (typically icons), or those that were so blurred as to be unrecognizable. Based on these criteria, approximately 15% of web images were discarded due to irrelevance, taking only a fraction of the total preparation time.

In order to compare the efficiency of acquiring web images with that of annotating street-view images, we estimated the time cost per valid image. It takes 3–5 min (including automated filtering) to procure 60–100 qualified web images for a single species, which equates to about 2–3 s per image. By contrast, labeling tree instances in a street-view image requires around 30 s per image, making it a task that demands sustained focus in complex scenes. Compared to the laborious

6 Seoul Open Data Portal

7 https://www.pullywood.com/ImageAssistant/

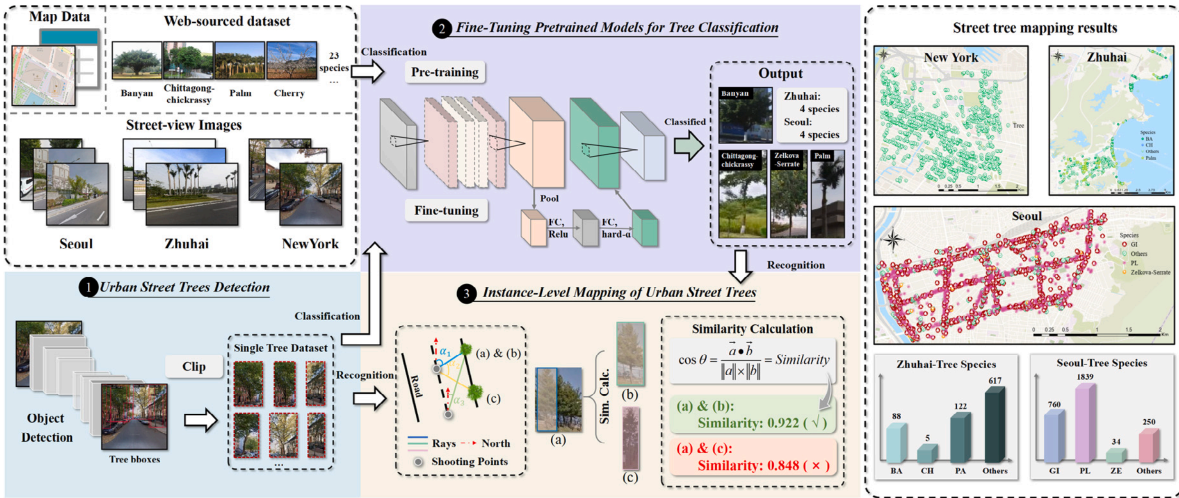


Fig. 2. An overview of the proposed urban street tree recognition and mapping method includes (1) Street tree detection from street-view images, creating a Single Tree Dataset. (2) Tree species classification using a fine-tuned classifier, pre-trained and fine-tuned with web-sourced data and the Single Tree Dataset, respectively. (3) Instance-Level Mapping of Urban Street Trees employs StreetTreeMapper to generate an instance-level map by integrating street-view images, OSM data, the Single Tree Dataset, and species labels for tree instances.

annotation of individual tree instances, acquiring images from the web significantly reduces the time required per training image, which also facilitates the development of larger, more scalable datasets for future research.

4. Methods

4.1. Overview of the proposed method

Fig. 2 illustrates the two-stage street tree recognition method and the StreetTreeMapper tree mapping method. First, we designed a two-stage method for instance-level tree classification. In the first stage, we used an object detection model to detect "tree bounding boxes" and to crop individual tree images based on the bounding box results. In the second stage, we fine-tuned a deep learning classification model for instance-level tree classification. This involved leveraging a model pre-trained on an open-source tree image dataset (Yang et al., 2023) and further adapting it to different urban environments using web-sourced labeled images. This approach enabled accurate species identification of the cropped individual tree images.

Second, we proposed StreetTreeMapper, a tree mapping approach that integrates feature vector similarity with spatial distance constraints. Specifically, the method calculates the feature vector similarity between individual tree images and applies a spatial constraint strategy derived from street-view acquisition points in OpenStreetMap (OSM) to determine whether trees detected in adjacent street-view images correspond to the same object, thereby supporting instance-level tree mapping.

4.2. Two-stage urban street tree recognition

The scarcity of annotated street-view data limits the training of street tree recognition models. Accordingly, we developed a two-stage method aimed at reducing reliance on instance-level annotations and boosting instance-level classification accuracy. In the first stage, we used an object detection model to frame the task as a binary classification problem and extract tree regions from street-view images. In the second stage, we incorporate publicly available web-sourced labeled data to enable instance-level tree classification, thereby alleviating the need for extensive manual annotations.

4.2.1. Urban street tree detection

We utilized a YOLO architecture based on Ultralytics' implementation (Khanam and Hussain, 2024) for street tree detection in street-view images, framing it as a binary classification task. Input images were resized to a resolution of 1280 × 1280 using zero-padding to preserve aspect ratio. We customized anchor box aspect ratios to $[0.35, 1.2, 3.1]$ to better match street tree morphology (slender trunks, dense foliage). We adopted CIoU loss for improved bounding box regression. During inference, detections were refined using confidence thresholding and NMS. For each detection, the tree region was cropped using predicted bounding box coordinates (x_1, y_1, x_2, y_2). All cropped regions, representing the area of interest, are saved as individual images, annotated with the original street-view image filename, bounding box coordinates, center coordinates denoted as x_{center}, and detection confidence score. These outputs serve as inputs for the second-stage classification and the StreetTreeMapper process.

4.2.2. Fine-tuning pretrained models for tree classification

In this stage, we utilized the cropped individual tree images obtained from the first-stage detection and incorporate web-sourced labeled data to achieve low-cost instance-level street tree classification.

To optimize the model for dominant tree species, we created a web-sourced labeled dataset of annotated images collected from the Internet and combined it with the pre-training dataset for fine-tuning. For Seoul, we used 1000 pre-training images and approximately 3000 web-sourced labeled images. For Zhuhai, we used 8303 pre-training images, cropped from our previously annotated dataset, and approximately 2500 web-sourced labeled images. This custom preprocessing was necessary because open-source datasets lacked sufficient images of BA and PA trees, which are prevalent in Zhuhai.

4.3. Instance-level mapping of urban street trees

A challenge in street-level mapping is the instance correspondence problem, which refers to the difficulty of correctly identifying and unifying the same physical tree across multiple overlapping viewpoints. To address this, we proposed StreetTreeMapper, a novel street tree mapping method that integrates image similarity analysis with spatial constraints from OpenStreetMap (OSM) to achieve precise identification and localization of individual trees across street-level images. Based on the bounding box and center coordinates from the detection phase, we introduced a ray-casting-based method to geolocate trees

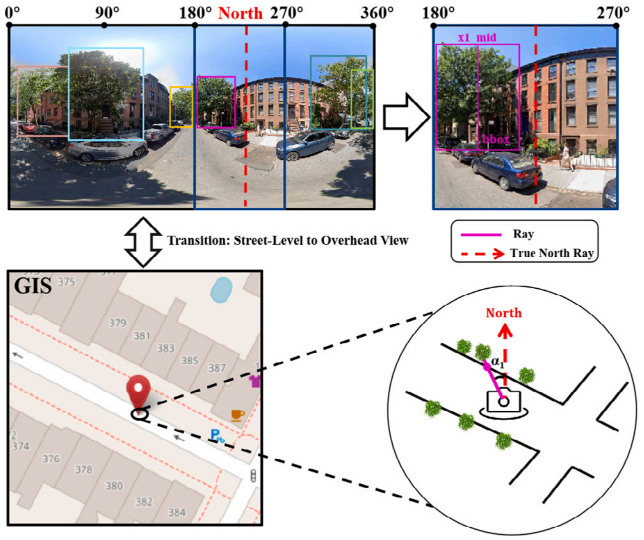
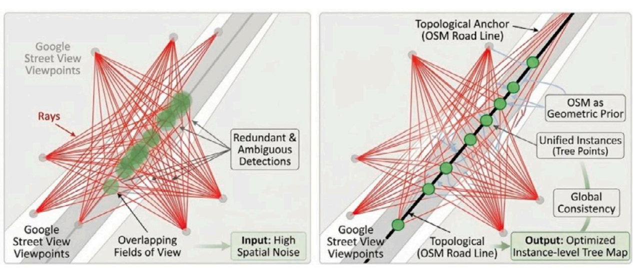


Fig. 3. StreetTreeMapper method leverages spatial geometric relationships between street-view images and OSM road networks for street tree mapping. It identifies trees in images using object detection, estimates their location through ray-casting based on panoramic image projection properties (angle relative to true north), and refines the location with OSM road network constraints.



(A) PROBLEM: Raw Visual Ray Ambiguity (Without Constraints) (B) SOLUTION: Unified Mapping with Dual Constraints (Our Method)

Fig. 4. Conceptual logic of handling the instance correspondence problem in StreetTreeMapper.

from street-view images. This method follows the procedure in Fig. 3, involving the extraction of the true north direction column index (α) from the image metadata, while the bounding box center column index is denoted as d. The azimuth angle θ is computed as:

$$\theta = \alpha - d \tag{1}$$

Using this angle, we cast a ray from the camera location in the direction of θ. The point where the ray intersects with the OSM road network is designated as the predicted tree location.

The integration of OSM road networks serves as a topological anchor to resolve these spatial ambiguities. As illustrated in Fig. 4, by constraining the candidate matching space for the cast rays, we effectively mitigate the instance correspondence problem and avoid redundant detections.

To merge duplicate tree predictions, we used DBSCAN (Deng, 2020) with a custom ϵ-neighborhood that considers both spatial distance denoted as dis and image similarity denoted as sim. Predictions within 2 m are considered the same tree due to potential ray-casting errors. This threshold is based on standard street tree spacing, typically ranging from 4 to 6 m.

For distances between 2 and 6 m, we further introduced image similarity constraints to improve clustering accuracy. Specifically, we used the Dinov2 model (Oquab et al., 2023) to extract a 768-dimensional feature vector and calculate the cosine similarity between tree images. To avoid incorrectly merging different but spatially close trees into the same cluster, we empirically select a similarity threshold of 0.85.

The final eps-neighborhood condition is defined as:

$$dis < 2 \vee (2 \leq dis < 6 \wedge sim > 0.85) \tag{2}$$

Once matched predictions are grouped, we performed weighted merging to determine the final tree location. The weight for each prediction is computed as:

$$W = \text{con} + \exp\left(-\frac{\text{dis}}{10}\right) \tag{3}$$

We define con the detection confidence score and dis the distance of the predicted point from the camera. This formulation accounts for the "large-near, small-far" effect in street-view images, where objects farther from the camera are more sensitive to spatial deviations. To ensure reliability, we exclude predictions with $dis > 40$. This threshold reflects the effective recognition range of street-view imagery, within which trees retain sufficient detail for identification, while beyond this distance, our observations indicate a noticeable decline in image quality and positional accuracy. It also corresponds to a typical coverage segment along urban roads, supporting continuous detection. The exponential decay term models the decreasing trustworthiness of predictions at increasing distances while keeping the confidence and distance components on a comparable scale.

StreetTreeMapper uses majority voting for class assignment within each cluster. This, combined with multi-view detection, spatial reasoning, and image similarity, improves the accuracy and robustness of street tree mapping, especially in complex urban areas.

5. Experiment

5.1. Experiment setting

5.1.1. Experimental design

The experimental design reflects the characteristics and label availability of the three datasets. Tree detection is evaluated on all three cities, as each provides bounding-box annotations. Species identification is primarily evaluated on the Zhuhai and Seoul datasets, which contain structured species categories suitable for classification assessment. For New York, due to its highly diverse species composition without a dominant category, the primary emphasis is placed on instance-level mapping performance. Species-level classification results for New York are also reported for completeness. Ground truth data are obtained from municipal forestry surveys for New York and Seoul, and from manual street-view annotations for Zhuhai. The expected positional accuracy of these reference datasets is at the meter level for municipal inventories and at the sub-meter level for manually aligned annotations.

5.1.2. Implementation details

We used a manually annotated street-view image dataset encompassing three study areas Zhuhai, New York, and Seoul to conduct a comparative experiment on the object detection performance of five mainstream object detection models RTMDet (Lyu et al., 2022), EfficientDet (Tan et al., 2020), Faster R-CNN (Ren et al., 2015), YOLOv8 (Varghese and Sambath, 2024), and YOLOv11 (Khanam and Hussain, 2024). All five models were trained utilizing the computational resources of a single NVIDIA RTX 3090 GPU, employing a batch size of 15, an initial learning rate of 0.00008, and a maximum training duration of 20 epochs.

For the classification stage, we employed MobileNetV3-Small, a network architecture comprising 16 convolutional layers with bottleneck modules and Squeeze-and-Excitation attention, using HSwish activation. We pre-trained this model for 100 epochs on an open-source tree image dataset (Yang et al., 2023), using RMSprop optimization

Table 5
Quantitative results for urban street tree detection in Brooklyn (New York), Xiangzhou (Zhuhai), and Seongdong (Seoul). Performance comparison using AP_{50} (%) and $AP_{[50:95]}$ (%).

Methods	Zhuhai		New York		Seoul		Mean	
	AP_{50}	$AP_{[50:95]}$	AP_{50}	$AP_{[50:95]}$	AP_{50}	$AP_{[50:95]}$	AP_{50}	$AP_{[50:95]}$
RTMDet (Lyu et al., 2022)	55.4	19.8	62.8	23.3	49.9	9.2	56.0	17.4
EfficientDet (Tan et al., 2020)	47.0	13.5	60.5	20.1	59.0	21.9	55.5	18.5
Faster R-CNN (Ren et al., 2015)	51.7	22.0	61.4	21.1	62.8	24.9	58.6	22.7
YOLOv8 (Varghese and Sambath, 2024)	**64.8**	25.8	71.0	28.3	**74.6**	33.5	70.1	29.2
YOLOv11 (Khanam and Hussain, 2024)	64.5	**25.9**	**72.5**	**28.1**	73.5	**34.6**	**70.2**	**29.5**

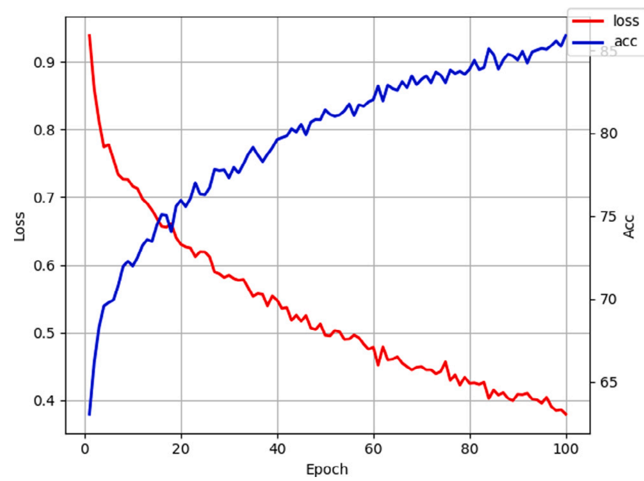


Fig. 5. Training loss (red) and training accuracy (blue) during 100-epoch fine-tuning on the Zhuhai dataset.

with a learning rate of 0.001. We performed this pre-training on an RTX 3090 GPU. Subsequently, to adapt the model to our specific urban environments, we fine-tuned it. For fine-tuning, we combined web-sourced labeled images with cropped single tree images from Zhuhai, initializing the model with the pre-trained weights obtained after 100 epochs. The classifier head consisted of stacked linear layers (576 to 1024 channels) with 0.2 dropout. The fine-tuning process was conducted for 100 epochs. A separate validation set was used for model selection, and the checkpoint corresponding to the lowest training loss was selected for subsequent testing, as shown in Fig. 5. For clarity, the experimental design distinguishes between detection and species classification across cities. Tree detection is trained and evaluated independently within each city, as all datasets provide bounding-box annotations. For species classification, the framework adopts a shared backbone with city-specific classification heads. Because species differ across Zhuhai, Seoul, and New York, a unified cross-city classifier is not assumed. Instead, the backbone network is first pre-trained to learn general visual representations of tree morphology, and then fine-tuned separately for each city's species set. Therefore, cross-city transfer occurs at the feature representation level rather than at the label space level. This design supports cross-domain knowledge transfer while accommodating region-specific ecological differences. To assess the practical feasibility for city-scale applications, we evaluated the end-to-end inference cost, including object detection, DINOv2 feature extraction, similarity matching, and clustering. The lightweight YOLOv11n detector (2.6M parameters) processes a street-view image in approximately 0.18 s for detection alone. When accounting for feature extraction and geometric matching, the average total processing time is approximately 6 s per image. Under this configuration, processing 3000 street-view images required approximately five hours on a single RTX 3090 GPU (24 GB), with peak memory usage of 8–10 GB. Extrapolating under near-linear data parallelism, mapping a mid-sized city with approximately 200,000 images would require about 100 h using four comparable GPUs, demonstrating practical scalability for urban forestry management applications.

5.2. Evaluation metrics

To evaluate model performance, we adopted both detection and classification metrics. For object detection, we used AP_{50} and $AP_{[50:95]}$, following the COCO evaluation protocol (Lin et al., 2014). AP_{50} measures average precision at an IoU threshold of 0.5, while $AP_{[50:95]}$ reports the mean AP across IoU thresholds from 0.5 to 0.95 in steps of 0.05.

For tree species classification using MobileNetV3, model performance was assessed using Accuracy, Precision, as well as Recall and the F1-score.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

$$F1\text{-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

In these equations, TP, TN, FP, and FN correspond to true positives, true negatives, false positives, and false negatives, respectively. For joint evaluation of localization and classification, we defined compound precision and recall metrics that require correct localization and species classification. A prediction is considered a true positive, denoted as $\text{TP}_{\text{loc+cls}}$, if the predicted tree falls within a spatial buffer b of the ground truth location and the species label is correct. Based on this, we computed:

$$\text{Precision}_{\text{loc+cls}} = \frac{\text{TP}_{\text{loc+cls}}}{\text{TP}_{\text{loc+cls}} + \text{FP}_{\text{loc+cls}}} \tag{8}$$

$$\text{Recall}_{\text{loc+cls}} = \frac{\text{TP}_{\text{loc+cls}}}{\text{TP}_{\text{loc+cls}} + \text{FN}_{\text{loc+cls}}} \tag{9}$$

where $\text{FP}_{\text{loc+cls}}$ represents predictions with either incorrect species or localization error beyond b, and $\text{FN}_{\text{loc+cls}}$ includes undetected trees or those with incorrect class or location.

5.3. Experiment results

5.3.1. Street tree detection results of different object detection models

Example street tree detection results from different object detection models were shown in Fig. 6. As in Table 5, the experimental results demonstrated that YOLOv8 (Varghese and Sambath, 2024) and YOLOv11 (Khanam and Hussain, 2024) significantly outperformed other models under the AP_{50}, with both achieving average precision values exceeding 0.6. Specifically, YOLOv11 (Khanam and Hussain, 2024) attained AP_{50} of 0.645 in Zhuhai, 0.725 in New York City, and 0.735 in Seoul, respectively, highlighting its superior and consistent detection capability across different urban environments.

Furthermore, under the $AP_{[50:95]}$, YOLOv11 (Khanam and Hussain, 2024) achieved an average precision of 0.2953, slightly surpassing YOLOv8's (Varghese and Sambath, 2024) 0.292, further validating

Fig. 6. Performance comparison of urban street tree detection models. Results from (a-b) Brooklyn, NewYork, USA, (c-d) Xiangzhou, Zhuhai, China, and (e-f) Seongdong-gu, Seoul, South Korea demonstrate both accurate detections and error cases, facilitating comparative analysis of model accuracy and failure modes.

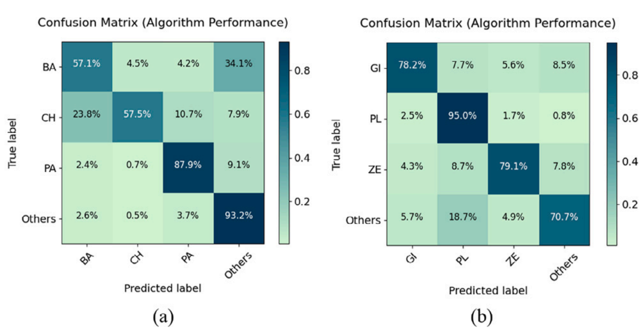


Fig. 7. The normalized confusion matrices, where (a) represents the classification results of species recognition in Zhuhai, and (b) represents the classification results of species recognition in Seoul.

its robustness in multi-scale detection tasks. Considering YOLOv11's (Khanam and Hussain, 2024) high generalizability and detection accuracy, this study ultimately selected it as the object detection model for subsequent tasks of urban street tree species classification and mapping.

5.3.2. Performance evaluation of two-stage street tree species recognition method

Fig. 7(a) illustrated the classification results for BA, CH, PA, and Others. BA achieved 57.1% accuracy and was most frequently misclassified as Others. CH achieved 57.5% accuracy and was most frequently misclassified as BA. PA had the highest accuracy at 87.9%. Others achieved a strong 93.2% accuracy. Fig. 7(b) presents classification results for GI, PL, ZE, and Others, with corresponding accuracies of 78.2%, 95.0%, 79.1%, and 70.7% respectively. PL achieved the highest accuracy. ZE was most often misclassified as PL, suggesting potential visual similarities between these species. Others was also most often misclassified as PL, indicating possible overlap in features with the PL category.

We primarily focus on tree species classification analysis in Zhuhai and Seoul, where dominant species structures allow clearer evaluation.

Table 6
AP_{50} (%) comparison for tree species classification in Zhuhai.

Method	Tree species				
	BA	CH	PA	Other	All
RTMDet	14.5	0.5	51.5	38.4	21.0
EfficientDet	7.6	0.4	35.3	27.2	17.6
Faster R-CNN	11.3	2.2	51.4	42.6	26.9
YOLOv8	11.1	2.4	48.4	39.3	25.3
YOLOv11	12.2	1.8	51.5	40.7	26.6
Ours	**26.2**	**13.8**	**53.3**	**53.1**	**29.3**

Note: RTMDet (Lyu et al., 2022), EfficientDet (Tan et al., 2020), Faster R-CNN (Ren et al., 2015), YOLOv8 (Varghese and Sambath, 2024), YOLOv11 (Khanam and Hussain, 2024).

In New York, species classification is also conducted, although higher inter-species similarity increases discrimination difficulty.

Table 6 shows that our two-stage street tree recognition method outperformed advanced direct detection models across all tree species in terms of AP50. Notably, for the BA category, our method achieved an AP50 of 26.2%, nearly twice that of the best baseline at 14.5%. Even for challenging categories such as CH, where baseline scores were extremely low, our method delivered a substantial improvement of over 10%. For high-performing categories like PA and Other, our approach also excelled, achieving 53.3% and 53.1%, respectively, surpassing all comparison models. Overall, the method achieved an average AP_{50} of 29.3%, representing a 2.4% improvement over the best baseline, demonstrating consistent advantages across different tree species.

Overall, our method achieved an average AP_{50} of 29.3% across all categories, which was superior to all compared models. Fig. 8 provides a visual comparison of various object detection models directly performing multi-class tree detection, highlighting the qualitative improvements offered by our approach. Our method achieves higher classification accuracy through its staged process, particularly evident in the improved performance for BA and CH compared to direct detection methods. These results indicate that our proposed method contributes positively to addressing the urban street tree recognition and classification challenge.

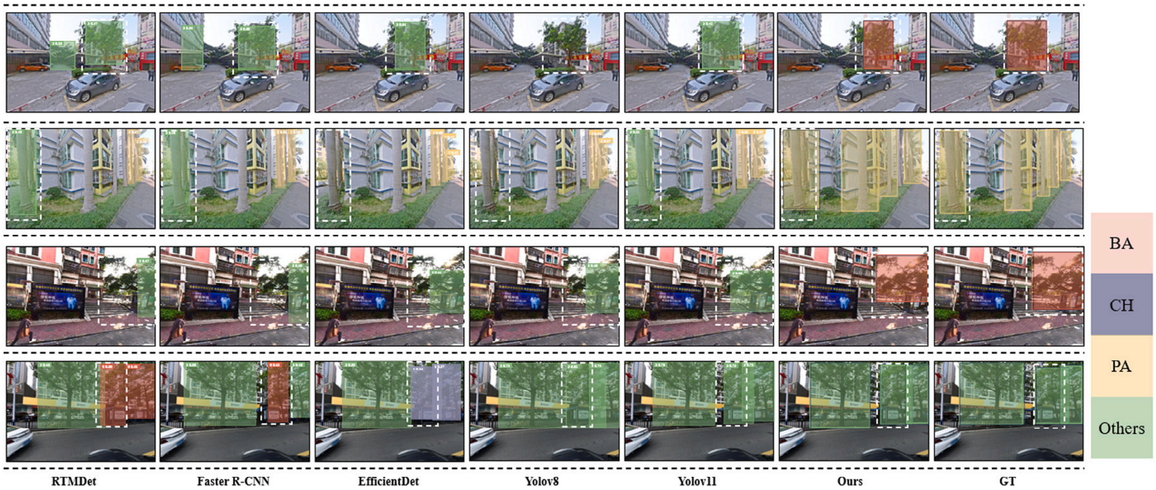


Fig. 8. Visual comparison of object detection models was performed for urban street tree recognition in Zhuhai Xiangzhou street-view images. The white bounding box showcases instances of tree species misclassification in street-view images.

Table 7
Street tree matching: Distance distributions between predicted and observed locations by study area (m). The value <0.001 m in the Zhuhai dataset represents the lower bound of the error distribution, corresponding to samples with near-perfect geometric alignment.

	New York	Seoul	Zhuhai
Median (m)	4.490	4.835	2.901
Min (m)	0.212	0.133	<0.001
Mean (m)	6.703	5.729	4.690

5.3.3. Instance-level mapping results of urban street trees

StreetTreeMapper integrates street-view images with OpenStreetMap data and visual feature comparison, which reduces duplicate tree detections and improves urban tree mapping accuracy. Table 7 summarized distance distributions and matching numbers. Fig. 9 presented instance-level street tree maps for three areas, with detailed views in (a)–(c).

In Brooklyn, New York, the regular street grid yields a uniform spatial pattern of street trees. Although species classification is conducted, the high diversity and relatively balanced species distribution increase fine-grained discrimination difficulty compared to cities with dominant species.

As demonstrated in Table 8, our two-stage recognition method achieved a $\text{Precision}_{loc+cls}$ of 35.5% on the Zhuhai dataset. For the Seoul dataset, the overall $\text{Precision}_{loc+cls}$ reached 29.9% while maintaining a stable $\text{Recall}_{loc+cls}$ of 30.2%. Notably, PL species in Seoul exhibited exceptional performance with $\text{Precision}_{loc+cls}$ at 43.0%. However, the accuracy for ZE in Seoul and CH in Zhuhai was significantly lower at 8.0% and 20.0% respectively, indicating the model's limited discriminative capability for morphologically similar species. For New York, the overall $\text{Precision}_{loc+cls}$ reached 19.0% with a $\text{Recall}_{loc+cls}$ of 14.0%. Compared to Zhuhai and Seoul, the lower classification performance in New York may be associated with higher species heterogeneity and seasonal variability in the dataset.

Beyond localization and classification accuracy, instance-level correspondence was further evaluated to assess whether the multi-view merging process preserves individual tree instances. On a subset of the New York dataset, ground-truth trees within the prediction coverage were analyzed and categorized into one-to-one matches, under-merging (a single tree split into multiple clusters), and over-merging (multiple trees merged into one cluster), as summarized in Table 9. Among the evaluated instances, 91.47% exhibited correct one-to-one correspondence, while under-merging and over-merging accounted for 7.75% and 0.78%, respectively. These results indicate consistent instance-level mapping performance (see Table 9).

6. Discussion

6.1. Impact of web-sourced data on fine-tuning classifier performance

Our analysis of the classifier training strategy within our two-stage street tree recognition method reveals significant regional variations in sensitivity to web-sourced image quantity. Of particular note, the Seoul dataset achieved a substantial improvement in accuracy after adding approximately 3000 web-sourced images, as shown in Table 10. This reflects a marked improvement in the model's capability for recognizing trees in the Seoul region. Owing to the model's high accuracy on this dataset along with the potential for overfitting or diminishing returns, further expansion of the Seoul dataset was deemed unnecessary. In contrast, Zhuhai's initial Top-1 accuracy was 80%, but with diminishing returns; approximately 2500 additional images only improved accuracy by 4 percentage points to 84%. This performance disparity can be attributed to two key factors. First, the urban morphology of the two cities significantly influences the visual domain gap. The street-view dataset for Seoul is relatively consistent, due to the city's characteristic high-density urban form featuring uniform street layouts and greening patterns. This structural regularity provides a more stable and coherent learning foundation for the model. In contrast, the urban environment of Zhuhai exhibits stronger ecological and visual heterogeneity, increasing the difficulty of feature alignment. Secondly, differences in baseline performance also constrain the degree of improvement possible. The Zhuhai model's higher initial accuracy, primarily due to its dataset's more precise category annotations, inherently limits the absolute margin for performance improvement.

These findings highlight that the effectiveness of transfer learning is closely related to region-specific visual coherence and baseline annotation quality. Cross-city deployment may therefore benefit from adaptive sampling or domain-aware training strategies that account for structural and ecological differences across urban environments.

6.2. Parameter sensitivity analysis for street-view tree detection

6.2.1. Impact of buffer distance on detection accuracy

As summarized in Table 11, spatial evaluation buffers of 4 m, 5 m, and 6 m were examined. Increasing the buffer from 4 m to 6 m leads

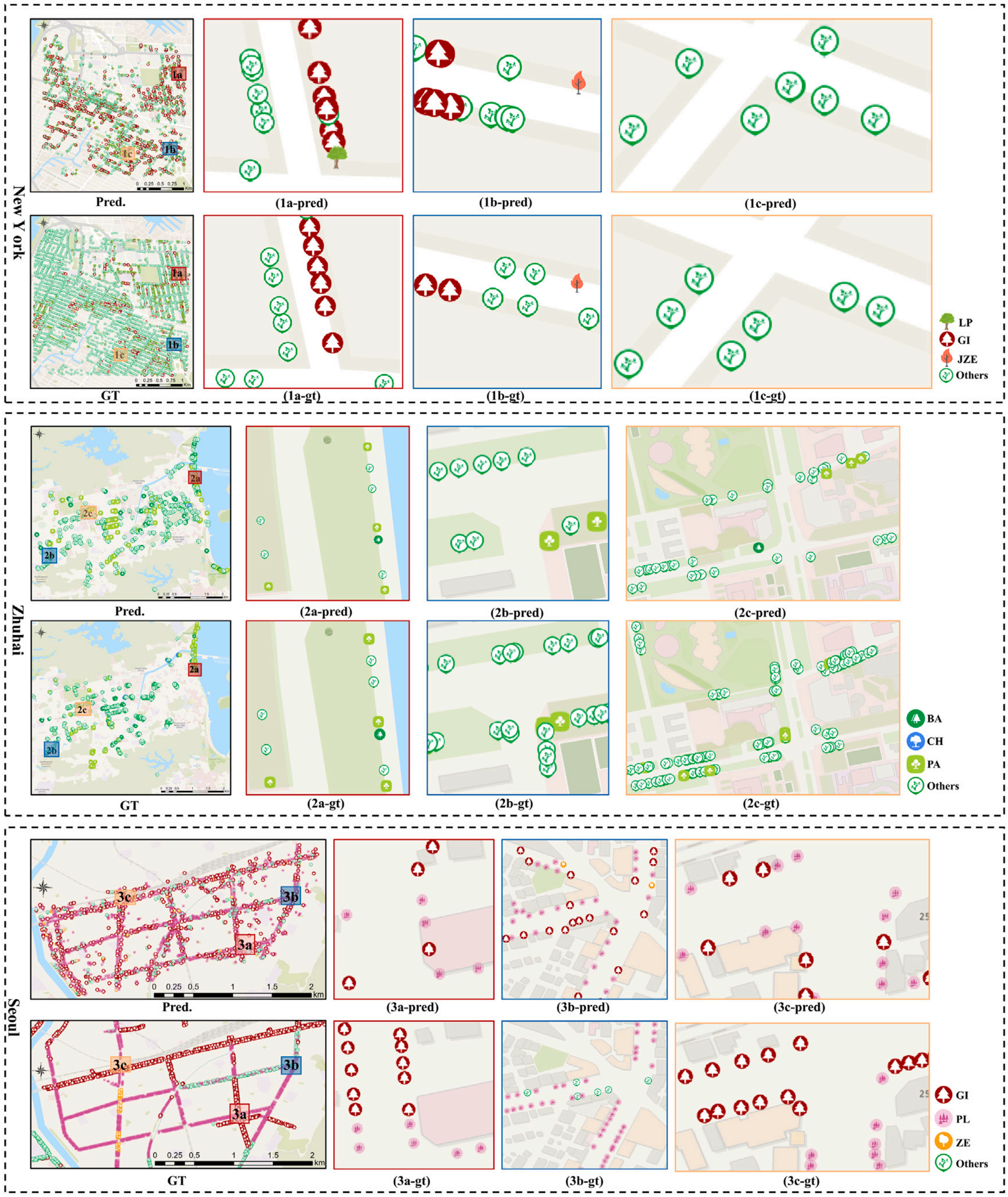


Fig. 9. Instance-level mapping results in New York, Zhuhai, and Seoul. Predictions (Pred) are shown above ground truth (GT), including overall maps and local enlargements.

Table 8
Tree species classification performance across the Zhuhai, Seoul, and New York datasets. PA, BA, and CH correspond to species in Zhuhai, PL, ZE, and GI to species in Seoul, and LP, GI, and JZE to London plane, Ginkgo, and Japanese zelkova in New York.

Metrics	Zhuhai					Seoul					New York				
	PA	BA	CH	Others	Mean	PL	ZE	GI	Others	Mean	LP	GI	JZE	Others	Mean
$Precision_{loc+cls}$	35.5	14.5	20.0	38.4	35.4	43.0	8.0	18.0	3.0	29.9	16.0	22.0	18.0	20.0	19.0
$Recall_{loc+cls}$	22.1	9.2	1.2	33.2	25.6	42.6	6.1	12.3	6.0	30.2	10.0	13.0	14.0	19.0	14.0

Table 9
Instance-level correspondence evaluation on a subset of the New York dataset. Results are computed on ground-truth trees within the spatial coverage of predictions.

Metric	Count	Rate
One-to-one correspondence	118	91.47%
Under-merge (split instances)	10	7.75%
Over-merge (merged distinct trees)	1	0.78%

to noticeable improvements in precision and recall, particularly in New York and Seoul, where F1-score increases by more than 20 percentage points. This sensitivity reflects the interaction between localization uncertainty and the one-to-one spatial matching protocol, under which each reference tree can be matched at most once. When predicted coordinates deviate slightly from reference positions, small adjustments in spatial tolerance can substantially alter matching outcomes.

Importantly, the buffer parameter governs evaluation tolerance rather than the detection or projection mechanism itself. A 5 m threshold provides a balanced tolerance that accommodates meter-level positional uncertainty while limiting excessive spatial merging. In practice, buffer size can be initialized based on local planting intervals or inventory standards and refined through lightweight validation without modifying the trained model.

6.2.2. Effect of street-view distance threshold and weighting

To examine the robustness of the localization framework, we conducted systematic sensitivity analyses on key parameters, including the DINOv2 similarity threshold, the maximum matching distance threshold, the evaluation buffer size, and the weighting decay denominator.

An ablation study was performed using thresholds of 0.80, 0.85, and 0.90 (Table 13). The median localization error remains stable within a narrow range (4.43–4.59 m), indicating that geometric matching accuracy is largely unaffected by moderate variations in feature similarity tolerance. Threshold 0.85 achieves the highest F1-score (59.5%) by balancing recall and precision, and is therefore adopted as the default setting.

We evaluated search radii from 35 m to 100 m (Table 12). The median error varies only slightly (4.45–4.69 m), demonstrating that the core matching mechanism is stable across a wide spatial range. The gradual increase in mean error reflects the inclusion of more distant candidate matches with naturally higher spatial uncertainty rather than degradation of the model itself. Beyond 40 m, improvements in coverage become marginal while computational cost increases, making 40 m a practical trade-off between spatial coverage and efficiency.

We further evaluated the denominator α in the weighting function $\exp(-dis/\alpha)$ using values of 5, 8, 10, 12, and 15 (Table 14). The median localization error remains within 4.47–4.54 m across all tested values, confirming that moderate changes in decay rate do not materially affect localization accuracy. $\alpha = 10$ provides a balanced decay behavior and is adopted as the default.

Overall, these results indicate that the geometric localization framework remains stable across broad parameter ranges. Parameter selection primarily influences evaluation tolerance or search scope rather than the intrinsic matching accuracy, supporting the transferability of the method to new urban environments.

6.2.3. Robustness analysis of orientation estimation and geometric modeling

Our approach utilizes OrienterNet (Sarlin et al., 2023) for camera orientation estimation. To assess the impact of potential inaccuracies in horizontal orientation (azimuth), we conducted a perturbation experiment by introducing angular offsets of $\pm1°$, $\pm3°$, and $\pm5°$. As shown in Table 15, the median localization error exhibits a minor upward trend as the orientation offset increases, rising from 2.900 m (baseline) to 3.260 m at a $\pm5°$ perturbation. This marginal increase (+0.360 m) confirms that while localization accuracy is positively correlated with orientation precision, the overall impact remains within a manageable range for urban-scale mapping.

Although panoramic imagery originates from spherical projection, Baidu Street View images are provided as rectified equirectangular projections. Under this representation, horizontal pixel coordinates correspond linearly to azimuth angles, allowing direct computation of object bearings from pixel positions. As a result, lateral localization primarily depends on azimuth estimation and depth estimation, while higher-order lens distortion effects are largely mitigated by platform rectification. Regarding geometric modeling, we employ a simplified linear projection model for practical reasons. By using single-perspective images rather than full panoramas, the model provides sufficient approximation for tree detection while avoiding the need for complex, often unavailable metadata required by more sophisticated models. This trade-off between precision and scalability is common in street-view analysis (Zhang et al., 2022; Yu et al., 2026).

We also evaluated the impact of pitch angle variations (vertical tilt). According to Baidu Street View API documentation, the default pitch is 0°, offering the widest coverage. Comparative analysis of 0° vs. 5° pitch datasets (1622 vs. 929 points) showed an average deviation of only 1.0–1.75 m in overlapping regions. In practical applications, we can strictly filter for data with a 0° pitch angle through platform settings, thereby minimizing the impact of such errors.

6.3. Seasonal variation in street-level tree classification

Seasonal variation in street-level tree imagery challenges visual classification models, as changes in appearance can affect recognition accuracy (Zhao et al., 2017; Han et al., 2023). The seasonal fluctuation of vegetation also correlates with variations in ecosystem services, such as urban thermal comfort (Chiang et al., 2023), highlighting the need for seasonally robust classification methods.

Our study focuses on tree species that exhibit stable seasonal characteristics. As shown in Fig. 10, dominant species in our primary study area, such as the Banyan and Palm, exhibit minimal variation in canopy structure and leaf morphology across seasons. This stability is typical of subtropical evergreen species, which are commonly selected in cities like Zhuhai for their ability to provide consistent shade and year-round greenery.

In Zhuhai, street-view imagery was collected during a concentrated, phenologically stable period (June–July), which provided a consistent dataset for model development. Similarly, in Seoul, the images used for cross-city validation were captured during another stable phenological window (March–April), further supporting the model’s generalization to stable evergreen-dominated urban forests. In contrast, the public dataset from New York contains imagery from varied seasons; therefore, we deliberately restricted its application to the less seasonally sensitive task of tree detection, avoiding phenological interference in fine-grained species classification.

Future efforts could extend validation to deciduous-dominated urban areas, where stronger phenological shifts would provide a more rigorous assessment of the method’s seasonal robustness.

6.4. Analysis of typical challenges and failure cases

Fig. 11 illustrates representative failure modes. In (a), BA is confused with CH under canopy overlap and partial occlusion, where discriminative trunk and structural cues are degraded and classification increasingly depends on coarse canopy texture. (b)–(c) show boundary ambiguity between Others and BA, particularly at moderate viewing distances where fine-grained leaf morphology is not reliably resolved. These errors indicate that confusions are concentrated near inter-class decision boundaries under feature degradation rather than uniformly across categories. In contrast, (d) demonstrates correct prediction despite severe occlusion, indicating robustness when sufficient structural cues are preserved.

Table 10
Performance comparison of urban tree species classification across study areas, showing baseline pre-training results and location-specific fine-tuning improvements. Metrics include Top-1 Accuracy (Top-1 Acc),Precision, Recall and F1 score (%). For subsequent classification, we employed the Open-source dataset augmented with 3000 images in Seoul and Manual annotations augmented with 2500 images in Zhuhai.

Study areas	Pre-trained	web-sourced images	Performance metrics (%)			
			Top-1 Acc	Precision	Recall	F1 score
Seoul	Open-source dataset (Yang et al., 2023)	✗	67.80	73.43	68.17	67.78
	Open-source dataset (Yang et al., 2023)	~1500	73.00	75.15	73.00	72.66
	Open-source dataset (Yang et al., 2023)	**~3,000**	**88.60**	**88.64**	**88.52**	**88.55**
Zhuhai	Manual annotations	✗	80.26	74.71	64.35	68.27
	Manual annotations	**~2,500**	**83.80**	**77.98**	**68.36**	**72.18**

Table 11
Accuracy of tree prediction at different buffer thresholds b, in terms of Precision, Recall, and F1 score metrics (%). We evaluated the accuracy of tree prediction based on tree detection results obtained using our YOLOv11 model.

Threshold	New York			Seoul			Zhuhai		
	Precision	Recall	F1 score	Precision	Recall	F1 score	Precision	Recall	F1 score
b=4 m	43.2	48.2	45.6	38.3	36.5	37.4	48.3	29.5	38.9
b=5 m	55.0	64.6	59.5	52.5	50.8	51.7	52.5	34.1	43.3
b=6 m	63.1	76.7	69.2	65.9	64.5	65.2	56.8	38.1	47.5

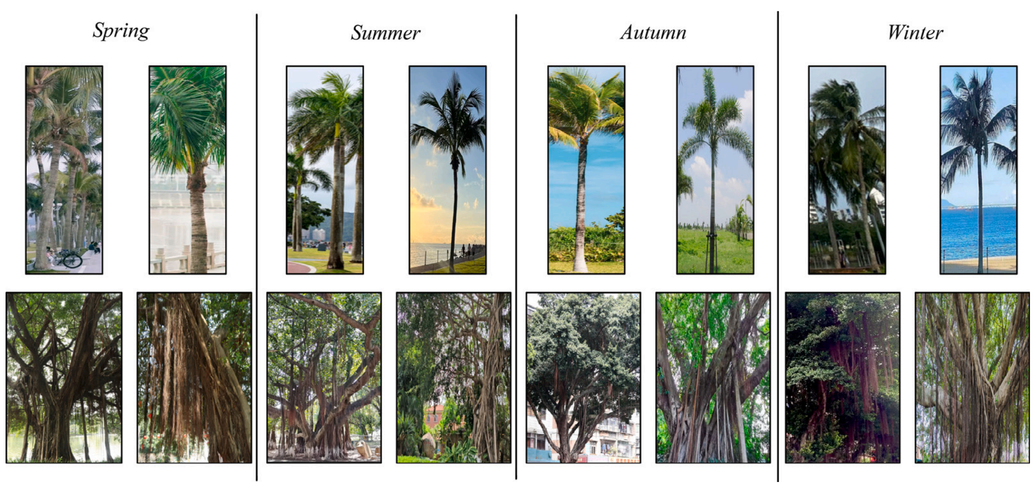


Fig. 10. Seasonal variation of dominant evergreen street trees (Palm and Banyan) in Zhuhai.

Table 12
Sensitivity analysis of distance threshold selection.

Threshold (m)	Median error (m)	Mean error (m)
35	4.45	5.91
40	4.49	6.70
50	4.58	6.74
60	4.62	6.89
80	4.65	7.30
100	4.69	7.46

Table 13
Ablation study on DINOv2 similarity threshold.

Threshold	Precision (%)	Recall (%)	F1 (%)	Median error (m)
0.80	55.5	52.9	54.2	4.59
0.85	55.0	64.6	59.5	4.49
0.90	57.3	54.9	56.1	4.43

Table 14
Sensitivity analysis of weighting scheme denominator (α).

Denominator (α)	Median error (m)
5	4.54
8	4.47
10	**4.49**
12	4.54
15	4.54

Table 15
Impact of Orientation Perturbation on Localization Accuracy.

Orientation offset	Median error (m)
Baseline (0°)	2.900
±1°	2.916
±3°	3.255
±5°	3.260

The pronounced performance gap of the "Other" category between Seoul (3.0% precision) and Zhuhai (38.4% precision) is primarily associated with differences in intra-class diversity. As summarized in Table 16, the Zhuhai "Other" class is dominated by a limited set of visually similar evergreen species, resulting in lower intra-class variance and a more coherent visual representation. In contrast, the Seoul "Other" category aggregates a broader mixture of deciduous and evergreen taxa, increasing structural variability and boundary ambiguity, which directly contributes to the observed disparity in precision.

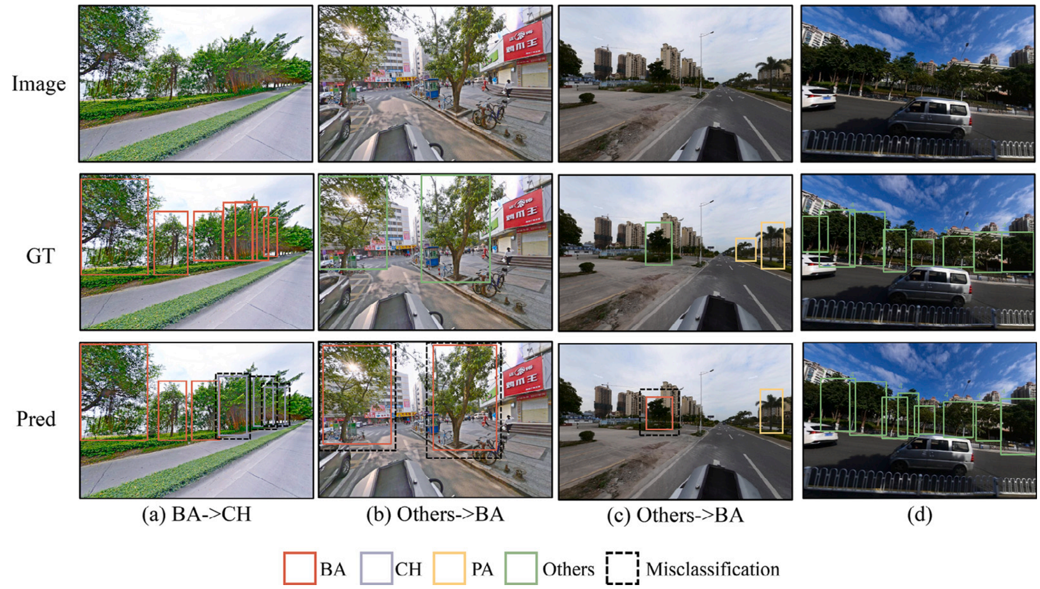


Fig. 11. Qualitative analysis of representative failure cases. Bounding box colors indicate species categories, BA (red), CH (purple), PA (yellow), and Others (green); dashed boxes denote misclassifications.

Table 16
Comparison of the composition and visual characteristics of the “Other” category in Zhuhai and Seoul.

Aspect	Zhuhai “Other”	Seoul “Other”
Representative species	Camphor, Elaeocarpus, Cinnamon	Birch, Cherry, Maple, etc.
Vegetation type	Predominantly evergreen broad-leaved	Mixed deciduous and evergreen
Visual coherence	Relatively homogeneous canopy structure	High morphological variability
Learning implication	Lower intra-class variance	High intra-class variance

It is important to note that the “Other” category represents a residual grouping of non-target taxa rather than a single botanical species, and its performance therefore reflects the challenge of modeling heterogeneous taxa under a unified label.

7. Conclusion

This work introduces a novel urban street tree recognition method to overcome the challenges of limited labeled data and high manual annotation costs, enabling accurate street tree recognition and instance-level mapping. We address the scarcity of labeled tree data through a two-stage recognition method, effectively combining tree detection and species classification. Furthermore, we developed StreetTreeMapper, which uses the geometric relationship between street-view images and OpenStreetMap (OSM) data to achieve improved accuracy and stability in instance-level mapping. Experiments across different cities validate the method’s practicality for large-scale urban vegetation monitoring, resulting in relatively accurate street tree recognition and precise instance-level mapping.

Future work may explore the frontier of model architecture and scale to further enhance performance. At the architectural level, the success of pre-trained Vision Transformers in related complex urban sensing tasks, such as fusing multi-modal data for noise modeling (Zhang et al., 2026), suggests their potential for learning more robust feature representations under street-view occlusions. A comparative study of such advanced backbones for street tree recognition would be valuable. Moreover, incorporating generative cross-view synthesis with spatial-textual guidance (Ye et al., 2025b) could provide synthetic overhead cues to bolster instance mapping robustness. At the model scale level, larger vision foundation models (e.g., DINOv3) could be investigated. However, for city-scale deployment under practical computational constraints, models like DINOv2-B/14 used in this study provide a critical reference point for balancing discriminative performance with inference efficiency. Systematic exploration across this architecture-scale design space could yield optimal solutions for different application scenarios.

CRediT authorship contribution statement

Weijia Li: Writing – original draft, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Xinjie Huo:** Writing – original draft, Software, Methodology, Conceptualization. **Junyan Ye:** Software, Methodology, Conceptualization. **Jinhua Yu:** Writing – original draft, Visualization. **Hanxi Wang:** Investigation, Data curation. **Juepeng Zheng:** Writing – review & editing. **Xiang Zhang:** Writing – review & editing. **Haohuan Fu:** Supervision, Resources.

Data and code availability statement

The processed street-view image patches and corresponding manual annotations for Zhuhai, Seoul, and New York will be made publicly available at: https://github.com/huoxinjie28-code/Urban-Street-Tree-Recognition. The repository will include cropped single-tree image patches, bounding-box annotations, species labels, geographic coordinates, and evaluation scripts necessary to reproduce the experimental results reported in this study.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partially supported by National Natural Science Foundation of China (Grant No. T2125006, 62571560, 42401415), the National Science Foundation of Guangdong Province (Grant No. 2025A1515010400, 2024A1515012083), and the Science and Technology Project of the Department of Natural Resources of Guangdong Province (Grant No. GDZRZYKJ2026011).

Data availability

Data will be made available on request.

References

Arevalo-Ramirez, T., Alfaro, A., Figueroa, J., Ponce-Donoso, M., Saavedra, J.M., Recabarren, M., Delpiano, J., 2024. Challenges for computer vision as a tool for screening urban trees through street-view images. Urban For. Urban Green. 95, 128316. http://dx.doi.org/10.1016/j.ufug.2024.128316.

Beery, S., Wu, G., Edwards, T., Pavetic, F., Majewski, B., Mukherjee, S., Chan, S., Morgan, J., Rathod, V., Huang, J., 2022. 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, pp. 21294–21307. http://dx.doi.org/10.1109/CVPR52688.2022.02061.

Branson, S., Wegner, J.D., Hall, D., Lang, N., Schindler, K., Perona, P., 2018. From google maps to a fine-grained catalog of street trees. ISPRS J. Photogramm. Remote Sens. 135, 13–30. http://dx.doi.org/10.1016/j.isprsjprs.2017.11.008.

Che, L., Chen, Y., Jin, T., Raubal, M., Schindler, K., Kiefer, P., 2025. Unsupervised urban land use mapping with street view contrastive clustering and a geographical prior. In: Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems. pp. 28–38. http://dx.doi.org/10.1145/3748636.3762707.

Chiang, Y.C., Liu, H.H., Li, D., Ho, L.C., 2023. Quantification through deep learning of sky view factor and greenery on urban streets during hot and cool seasons. Landsc. Urban Plan. 232, 104679. http://dx.doi.org/10.1016/j.landurbplan.2022.104679.

Choi, K., Lim, W., Chang, B., Jeong, J., Kim, I., Park, C.R., Ko, D.W., 2022. An automatic approach for tree species detection and profile estimation of urban street trees using deep learning and google street view images. ISPRS J. Photogramm. Remote Sens. 190, 165–180. http://dx.doi.org/10.1016/j.isprsjprs.2022.06.004.

Deng, D., 2020. DBSCAN clustering algorithm based on density. In: 2020 7th International Forum on Electrical Engineering and Automation. IFEEA, IEEE, pp. 949–953. http://dx.doi.org/10.1109/IFEEA51475.2020.00199.

Han, Y., Zhong, T., Yeh, A.G., Zhong, X., Chen, M., Lü, G., 2023. Mapping seasonal changes of street greenery using multi-temporal street-view images. Sustain. Cities Soc. 92, 104498. http://dx.doi.org/10.1016/j.scs.2023.104498.

Hu, Y., Wang, H., Yan, H., Han, Q., Nan, X., Zhao, K., Bao, Z., 2023. Alternative scenarios for urban tree surveys: Investigating the species, structures, and diversities of street trees using street view imagery. Sci. Total Environ. 895, 165157. http://dx.doi.org/10.1016/j.scitotenv.2023.165157.

Jiang, Y., Liu, D., Ren, L., Grekousis, G., Lu, Y., 2024. Tree abundance, species richness, or species mix? Exploring the relationship between features of urban street trees and pedestrian volume in jinan, China. Urban For. Urban Green. 95, 128294. http://dx.doi.org/10.1016/j.ufug.2024.128294.

Jiang, N., Yan, S., Mao, X., 2013. Zhuhai shi chengshi xingdaoshu shuzhong xuanze chutan. Shandong For. Sci. Technol. 43 (3), 75–77, in Chinese.

Khanam, R., Hussain, M., 2024. YOLOv11: An overview of the key architectural enhancements. http://dx.doi.org/10.48550/arXiv.2410.17725, arXiv Preprint, arXiv:2410.17725.

Laumer, D., Lang, N., van Doorn, N., Mac Aodha, O., Perona, P., Wegner, J.D., 2020. Geocoding of trees from street addresses and street-level images. ISPRS J. Photogramm. Remote Sens. 162, 125–136. http://dx.doi.org/10.1016/j.isprsjprs.2020.02.001.

Li, W., Dong, R., Fu, H., Yu, L., 2019. Large-scale oil palm tree detection from high-resolution satellite images using two-stage convolutional neural networks. Remote. Sens. 11 (1), 11. http://dx.doi.org/10.3390/rs11010011.

Li, W., Lai, Y., Xu, L., Xiangli, Y., Yu, J., He, C., Xia, G.S., Lin, D., 2023. OmniCity: Omnipotent city understanding with multi-level and multi-view images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 17397–17407. http://dx.doi.org/10.1109/CVPR52729.2023.01669.

Li, D., Ni, Y., 2025. Assessing street environments for older adults in urban villages using POIs and street view images—A case study of Guangzhou, China. Sustainability 17 (1), 31. http://dx.doi.org/10.3390/su17010031.

Li, Z., Tao, H., Huai, Y., Nie, X., 2024. Urban street tree species identification and factor interpretation model based on natural images. Urban For. Urban Green. 101, 128512. http://dx.doi.org/10.1016/j.ufug.2024.128512.

Li, M., Yao, W., 2020. 3D map system for tree monitoring in Hong Kong using google street view imagery and deep learning. ISPRS Ann. Photogramm. Remote. Sens. Spat. Inf. Sci. V-3-2020, 765–772. http://dx.doi.org/10.5194/isprs-annals-V-3-2020-765-2020.

Li, W., Yu, J., Chen, D., Lin, Y., Dong, R., Zhang, X., He, C., Fu, H., 2025. Fine-grained building function recognition with street-view images and GIS map data via geometry-aware semi-supervised learning. Int. J. Appl. Earth Obs. Geoinf. 137, 104386. http://dx.doi.org/10.1016/j.jag.2025.104386.

Lian, X., Jiao, L., Liu, Z., Jia, Q., Liu, W., Liu, Y., 2024. A detection of street trees and green space: Understanding contribution of urban trees to climate change mitigation. Urban For. Urban Green. 102, 128561. http://dx.doi.org/10.1016/j.ufug.2024.128561.

Liang, C., Jiang, H., Yang, S., Tian, P., Ma, X., Tang, Z., Wang, H., Wang, W., 2024. Characterizing street trees in three metropolises of central China by using street view data: From individual trees to landscape mapping. Ecol. Inf. 80, 102480. http://dx.doi.org/10.1016/j.ecoinf.2024.102480.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: Common objects in context. In: Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V. Springer, pp. 740–755. http://dx.doi.org/10.1007/978-3-319-10602-1_48.

Lind, r., Prade, T., Sjöman Deak, J., Levinsson, A., Sjöman, H., 2023. How green is an urban tree? The impact of species selection in reducing the carbon footprint of park trees in Swedish cities. Front. Sustain. Cities 5, 1182408. http://dx.doi.org/10.3389/frsc.2023.1182408.

Liu, D., Jiang, Y., Wang, R., Lu, Y., 2023. Establishing a citywide street tree inventory with street view images and computer vision techniques. Comput. Environ. Urban Syst. 100, 101924. http://dx.doi.org/10.1016/j.compenvurbsys.2022.101924.

Liu, Y., Wang, Z., Ren, S., Chen, R., Shen, Y., Biljecki, F., 2025. Physical urban change and its socio-environmental impact: Insights from street view imagery. Comput. Environ. Urban Syst. 119, 102284. http://dx.doi.org/10.1016/j.compenvurbsys.2025.102284.

Long, C., Chen, D., Chen, Z., Ding, R., Dong, Z., Yang, B., 2026. TSC-net: A roadside tree structure parameter computation network using street-view images. Geo-Spat. Inf. Sci. 1–21. http://dx.doi.org/10.1080/10095020.2026.2626665.

Lumnitz, S., Devisscher, T., Mayaud, J.R., Radic, V., Coops, N.C., Griess, V.C., 2021. Mapping trees along urban street networks with deep learning and street-level imagery. ISPRS J. Photogramm. Remote Sens. 175, 144–157. http://dx.doi.org/10.1016/j.isprsjprs.2021.01.016.

Lyu, C., Zhang, W., Huang, H., Zhou, Y., Wang, Y., Liu, Y., Zhang, S., Chen, K., 2022. RTMDet: An empirical study of designing real-time object detectors. http://dx.doi.org/10.48550/arXiv.2212.07784, arXiv Preprint, arXiv:2212.07784.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P., 2023. DINOv2: Learning robust visual features without supervision. http://dx.doi.org/10.48550/arXiv.2304.07193, arXiv Preprint, arXiv:2304.07193.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 779–788. http://dx.doi.org/10.1109/CVPR.2016.91.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. NIPS, In: Advances in Neural Information Processing Systems, vol. 28, pp. 91–99. http://dx.doi.org/10.5555/2969239.2969250.

Roberts, J., Koeser, A., Abd-Elrahman, A., Wilkinson, B., Hansen, G., Landry, S., Perez, A., 2019. Mobile terrestrial photogrammetry for street tree mapping and measurements. Forests 10 (8), 701. http://dx.doi.org/10.3390/f10080701.

Sarlin, P.E., DeTone, D., Yang, T.Y., Avetisyan, A., Straub, J., Malisiewicz, T., Bulo, S.R., Newcombe, R., Kontschieder, P., Balntas, V., 2023. Orienternet: Visual localization in 2d public maps with neural matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 21632–21642.

Şimşek, Ç.K., Arabacı, D., Yücel, C., Öztürk, B., 2024. Monitoring the climatic effects of street tree plantation in different urban patterns by synthetic image based BPNN simulations. Build. Env. 250, 111173. http://dx.doi.org/10.1016/j.buildenv.2024.111173.

Tan, M., Pang, R., Le, Q.V., 2020. EfficientDet: Scalable and efficient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 10781–10790. http://dx.doi.org/10.1109/CVPR42600.2020.01079.

Varghese, R., Sambath, M., 2024. YOLOv8: A novel object detection algorithm with enhanced performance and robustness. In: 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems. ADICS, IEEE, pp. 1–6. http://dx.doi.org/10.1109/ADICS58448.2024.10533619.

Velasquez-Camacho, L., Merontausta, E., Etxegarai, M., de Miguel, S., 2024. Assessing urban forest biodiversity through automatic taxonomic identification of street trees from citizen science applications and remote-sensing imagery. Int. J. Appl. Earth Obs. Geoinf. 128, 103735. http://dx.doi.org/10.1016/j.jag.2024.103735.

Ventura, J., Pawlak, C., Honsberger, M., Gonsalves, C., Rice, J., Love, N.L., Han, S., Nguyen, V., Sugano, K., Doremus, J., et al., 2024. Individual tree detection in large-scale urban environments using high-resolution multispectral imagery. Int. J. Appl. Earth Obs. Geoinf. 130, 103848. http://dx.doi.org/10.1016/j.jag.2024.103848.

Wang, W., Xiao, L., Zhang, J., Yang, Y., Tian, P., Wang, H., He, X., 2018. Potential of internet street-view images for measuring tree sizes in roadside forests. Urban For. Urban Green. 35, 211–220. http://dx.doi.org/10.1016/j.ufug.2018.09.008.

Wegner, J.D., Branson, S., Hall, D., Schindler, K., Perona, P., 2016. Cataloging public objects using aerial and street-level images - urban trees. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 6014–6023. http://dx.doi.org/10.1109/CVPR.2016.647.

Wu, P., Qian, Z., 2021. Leaf classification based on convolutional neural network. J. Phys. Conf. Ser. 1820 (1), 012161. http://dx.doi.org/10.1088/1742-6596/1820/1/012161.

Wu, J., Yao, W., Polewski, P., 2018. Mapping individual tree species and vitality along urban road corridors with LiDAR and imaging sensors: Point density versus view perspective. Remote. Sens. 10 (9), 1403. http://dx.doi.org/10.3390/rs10091403.

Yang, W., Wang, A., Zhang, G., Zhang, Y., Xu, T., Liao, M., 2025. Assessment of solar utilization potential for streetlights using street view imagery and deep learning: A case study in Hong Kong. Renew. Energy 258, 124951. http://dx.doi.org/10.1016/j.renene.2025.124951.

Yang, T., Zhou, S., Huang, Z., Xu, A., Ye, J., Yin, J., 2023. Urban street tree dataset for image classification and instance segmentation. Comput. Electron. Agric. 209, 107852. http://dx.doi.org/10.1016/j.compag.2023.107852.

Ye, J., He, J., Li, W., Lv, Z., Lin, Y., Yu, J., Yang, H., He, C., 2025a. Leveraging BEV paradigm for ground-to-aerial image synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. ICCV, pp. 28451–27761. http://dx.doi.org/10.1109/ICCV51701.2025.02642.

Ye, J., He, J., Zhang, X., Lin, Y., Lin, H., He, C., Li, W., 2025b. Satellite image synthesis from street view with fine-grained spatial textual guidance: A novel framework. IEEE Geosci. Remote. Sens. Mag..

Ye, J., Luo, Q., Yu, J., Zhong, H., Zheng, Z., He, C., Li, W., 2024a. Sg-bev: satellite-guided bev fusion for cross-view semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 27748–27757. http://dx.doi.org/10.1109/CVPR52733.2024.02622.

Ye, J., Lv, Z., Li, W., Yu, J., Yang, H., Zhong, H., He, C., 2024b. Cross-view image geo-localization with panorama-BEV co-retrieval network. In: Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XXXVII. In: Lecture Notes in Computer Science, vol. 15095, Springer, pp. 74–90. http://dx.doi.org/10.1007/978-3-031-72913-3_5.

Yu, J., Ye, J., Lin, Y., Li, W., 2026. CRFFNet: A cross-view reprojection based feature fusion network for fine-grained building segmentation using satellite-view and street-view data. Inf. Fusion 127, 103795. http://dx.doi.org/10.1016/j.inffus.2025.103795.

Zhang, Y., Chen, N., Du, W., Li, Y., Zheng, X., 2021. Multi-source sensor based urban habitat and resident health sensing: A case study of Wuhan, China. Build. Environ. 198, 107883.

Zhang, Q., Du, Z., Li, L., Zhao, Z., Liang, J., Li, H., 2024. Urbanized and climatic influences on vegetation changes of urban green space under the increase of atmospheric NOX: A case study in Tianjin. Ecol. Indic. 165, 112222. http://dx.doi.org/10.1016/j.ecolind.2024.112222.

Zhang, Y., Ke, E., Kwan, M.P., Fang, L., Li, M., 2026. Multi-frequency street-level urban noise modeling and mapping through street view and remote sensing image fusion. Comput. Environ. Urban Syst. 126, 102401. http://dx.doi.org/10.1016/j.compenvurbsys.2026.102401.

Zhang, Y., Ma, H., Kwan, M.P., 2025. How drivers' depth perception of environmental features influences traffic speed. Geo-Spat. Inf. Sci. 1–17. http://dx.doi.org/10.1080/10095020.2025.2584862.

Zhang, J., Yang, K., Ma, C., Reiß, S., Peng, K., Stiefelhagen, R., 2022. Bending reality: Distortion-aware transformers for adapting to panoramic semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 16917–16927. http://dx.doi.org/10.1109/CVPR52688.2022.01641.

Zhao, J., Xu, W., Li, R., 2017. Visual preference of trees: The effects of tree attributes and seasons. Urban For. Urban Green. 25, 19–25. http://dx.doi.org/10.1016/j.ufug.2017.04.015.

Zheng, J., Fu, H., Li, W., Wu, W., Yu, L., Yuan, S., Tao, W.Y.W., Pang, T.K., Kanniah, K.D., 2021. Growing status observation for oil palm trees using unmanned aerial vehicle (UAV) images. ISPRS J. Photogramm. Remote Sens. 173, 95–121. http://dx.doi.org/10.1016/j.isprsjprs.2021.01.008.

Zheng, J., Yuan, S., Li, W., Fu, H., Yu, L., Huang, J., 2024. A review of individual tree crown detection and delineation from optical remote sensing images: Current progress and future. IEEE Geosci. Remote. Sens. Mag. 13 (1), 209–236. http://dx.doi.org/10.1109/MGRS.2024.3479871.

Zheng, J., Yuan, S., Wu, W., Li, W., Yu, L., Fu, H., Coomes, D., 2023. Surveying coconut trees using high-resolution satellite imagery in remote atolls of the Pacific ocean. Remote Sens. Environ. 287, 113485. http://dx.doi.org/10.1016/j.rse.2023.113485.

Zhou, H., Wang, J., Wilson, K., 2022. Impacts of perceived safety and beauty of park environments on time spent in parks: Examining the potential of street view imagery and phone-based GPS data. Int. J. Appl. Earth Obs. Geoinf. 115, 103078. http://dx.doi.org/10.1016/j.jag.2022.103078.