ISPRS Journal of Photogrammetry and Remote Sensing 175 (2021) 144–157

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Mapping trees along urban street networks with deep learning and street-level imagery

Stefanie Lumnitz [a,b,*], Tahia Devisscher [a], Jerome R. Mayaud [c], Valentina Radic [d], Nicholas C. Coops [a], Verena C. Griess [e]

[a] Department of Forest Resources Management, University of British Columbia, Vancouver, British Columbia, Canada
[b] ESRIN, European Space Agency, Frascati, Italy
[c] Spare Labs Inc, Vancouver, British Columbia, Canada
[d] Department of Earth, Ocean, and Atmospheric Sciences, University of British Columbia, Vancouver, British Columbia, Canada
[e] Institute of Terrestrial Ecosystems, Department of Environmental System Sciences, ETH, Zurich, Switzerland

ARTICLE INFO

Keywords:
Deep learning
Instance segmentation
Monocular depth estimation
Street-level images
Urban forest management

ABSTRACT

Planning and managing urban forests for livable cities remains a challenge worldwide owing to sparse information on the spatial distribution, structure and composition of urban trees and forests. National and municipal sources of tree inventory remain limited due to a lack of detailed, consistent and frequent inventory assessments. Despite advancements in research on the automation of urban tree mapping using Light Detection and Ranging (LiDAR) or high-resolution satellite imagery, in practice most municipalities still perform labor-intensive field surveys to collect and update tree inventories. We present a robust, affordable and rapid method for creating tree inventories in any urban region where sufficient street-level imagery is readily available. Our approach is novel in that we use a Mask Regional Convolutional Neural Network (Mask R-CNN) to detect and locate separate tree instances from street-level imagery, thereby successfully creating shape masks around unique fuzzy urban objects like trees. The novelty of this method is enhanced by using monocular depth estimation and triangulation to estimate precise tree location, relying only on photographs and images taken from the street. Experiments across four cities show that our method is transferable to different image sources (Google Street View, Mapillary) and urban ecosystems. We successfully detect >70% of all public and private trees recorded in a ground-truth campaign across Metro Vancouver. The accuracy of geolocation is also promising. We automatically locate public and private trees with a mean error in the absolute position ranging from 4 to 6 m, which is comparable to ground-truth measurements in conventional manual urban tree inventory campaigns.

1. Introduction

Urban forests are gaining global attention as evidence is gathered about the diverse benefits they provide to human health and well-being through various ecosystem services (Nowak et al., 2014; van den Bosch, 2017). Planning and managing urban forests and trees on the basis of urban tree inventories is increasingly coming to the fore in the context of rapid urbanization trends, climate change and increasingly global trade (Padayachee et al., 2017). Unfortunately, urban forest planning and management remains an outstanding challenge worldwide owing to relatively scarce information on the spatial distribution and accessibility of urban forests and trees, as well as their health condition, composition, structure and function (Kelly et al., 2007). In practice, most municipalities still perform labor-intensive field surveys to collect and update inventories of public trees, and lack information on private trees. Despite the importance of urban trees, national and municipal sources of tree inventory lack in detail, consistency and quantity due to the cost associated with mapping and monitoring trees through time and over large areas (Nielsen et al., 2014).

Recent research has focused on remote sensing data and techniques allowing for the remote and automated recognition and characterization of individual trees (Ke and Quackenbush, 2011). Individual tree mapping from remotely-sensed data, termed as Individual Tree Crown Delineation or Detection (ITCD), has gained popularity since the mid-1980s as an alternative to ground-truth measurements (Zhen et al., 2016). However, mapping and monitoring of individual trees in

* Corresponding author at: Department of Forest Resources Management, University of British Columbia, Vancouver, British Columbia, Canada.
E-mail address: Stefanie.Lumnitz@gmail.com (S. Lumnitz).

https://doi.org/10.1016/j.isprsjprs.2021.01.016
Received 22 September 2020; Received in revised form 31 December 2020; Accepted 19 January 2021
Available online 18 March 2021

heterogeneous urban areas using remotely-sensed data and current ITCD methods remains challenging (Aval et al., 2018). The small size of individual tree crowns in urban areas binds the use of most satellite imagery sources to analyzing clusters of urban trees or requires a process for spectral unmixing (Small, 2001). Very high-resolution (VHR) satellite or aerial imagery (<80 cm) can help provide the level of detail required for individual urban tree assessments, but are often impacted by urban shadows (Li et al., 2019; Plowright et al., 2016). Similarly, the use of high-resolution Light Detection and Ranging (LiDAR) data for individual tree assessments is impacted by vertical urban structures such as power lines and lamp posts (Zhen et al., 2016). Datasets such as LiDAR or VHR aerial imagery are usually collected at one-point in time and can be expensive to acquire (Li et al., 2019; Alonzo et al., 2014). Additionally, LiDAR or data fusion approaches often lack processing methods that can be generalized or automated over large areas with high accuracy (Alonzo et al., 2014). Novel, readily-accessible methods and data sources to build standardized tree inventories on a large spatial scale allowing for affordable, seamless and recurrent data collection and rapid processing are still needed (Ke and Quackenbush, 2011).

Two recent trends have gained attention in assessing urban greenery along street networks over large areas, at low cost, and promoting uptake from a larger number of municipalities in recent literature (Stubbings et al., 2019). First, the growing availability of low-cost, detailed and increasingly crowd-sourced street-level imagery (photographs of street scenes taken from the ground) (Li et al., 2016; Berland and Lange, 2017). Second, the success of Convolutional Neural Networks (CNN) out-competing other methods for extracting abstract features and objects in imagery (Ma et al., 2019). Street-level imagery is used to quantify 'perceived urban canopy cover' by estimating the percentage of detected tree canopy cover pixels relative to the total number of pixels in an image (Seiferling et al., 2017). Similarly, Li et al. (2017) assessed the percentage of vegetation in streets by quantifying the amount of green pixels seen in a street view scene. Both of these methodologies calculated a Green View Index (GVI), a metric that quantifies the proportional amount of green pixels in each image. The index serves as a proxy for how urban vegetation is perceived by pedestrians, and has since been applied to a variety of cities all over the world (Li et al., 2015). Other applications of street-level imagery in combination with CNNs include quantifying the benefits provided by trees along urban street networks, such as shade provisioning (Li et al., 2018), public health (Kang et al., 2020), as well as accessibility to greenspaces (Jang et al., 2020). Additionally, street-level imagery and deep learning were used in combination to update tree inventories based on coarse street addresses with precise geographic coordinates Laumer et al. (2020). Laumer et al. (2020) applied the developed method to match 38% of >50000 detected street trees to street-level addresses. Wegner et al. (2016) designed a workflow for automatic detection and geolocation of street trees, by combining Faster Region Convolutional Neural Network (R-CNN) tree detection results from Google Street View (GSV) and aerial imagery, with information retrieved from Google Maps in a probabilistic model. The authors then used street-level and aerial imagery to classify 18 different species among the detected trees. Branson et al. (2018) subsequently built upon methods for object detection in Wegner et al. (2016) by including a Siamese CNN, to verify whether detected trees had changed visibly over time. Most approaches generating urban tree inventory data using GSV and computer vision techniques, however, still rely on a data fusion approach with VHR aerial imagery or other secondary data sources to locate urban trees (Duarte and Ratti, 2017).

Localizing objects that have previously been detected in photographs acquired with smart phones or cameras from the street is a unique challenge for remote sensing practitioners. Satellite or aerial imagery pixels and LiDAR point clouds inherently store either relative or absolute geographic location information and make the need to additionally compute three dimensional geographic pixel coordinates redundant (Plowright et al., 2017). The translation process for features from street-level imagery to a geographical location is usually achieved using one of two principal approaches: (1) objects are either matched to locations using overhead or 3-dimensional data (e.g. passive aerial imagery or active LiDAR) (Lefèvre et al., 2017), or (2) the location is directly retrieved from street-level imagery by reconstructing 3-dimensional space or feature data (e.g. camera-to-object depth) (Agarwal et al., 2010). The latter approach can be achieved through multi-view stereo methods (using multiple images to reconstruct the objects) (Cheng et al., 2018), binocular methods (using two images) (Hirschmuller, 2008) or monocular methods (using only one image) (Godard et al., 2016). Because monocular depth estimation can be made using a single image, it does not require a large amount of images taken from multiple perspectives or additional knowledge about the analysed scene, and allows for the analysis of features from various data sources taken at different points in time (Tippetts et al., 2016). Monocular depth estimation has benefited from recent development of novel deep learning approaches, particularly in the field of self-driving cars (Michels et al., 2005). The potential to retrieve location information of detected objects from a single image through the use of monocular depth estimation enhances the potential to use street-level imagery collected with different sensors over time to match detected objects to specific locations (Wegner et al., 2016).

Our aim is to produce an automatic, affordable and novel method for streamlined tree detection and geolocation along street networks that can be used in any urban region where sufficient street-level imagery is readily available. The novelty of this model is enhanced by using monocular depth estimation and triangulation to estimate precise tree locations without the need to rely on secondary datasets. In addition, we introduce state-of-the-art instance segmentation (object detection and pixel masking) with deep learning frameworks to extract and mask fuzzy features like trees in images. We investigate the generalizability and transferability of our tree detection model by applying it to different geographical locations in three cities in the Metro Vancouver region (Canada) and the city of Pasadena (US). We further test the robustness of our approach on images provided by two different street-level imagery sources, namely GSV, which provides proprietary data, and Mapillary, which offers crowd-sourced data. We validate our model by comparing its output with on-the-ground tree location measurements in the Metro Vancouver region. Ultimately, we aim to fulfill the need for inventory methods that can be automated and generalized over multiple cities in order to develop national and international recommendations and standards for urban tree inventory data collection along street networks (Nielsen et al., 2014).

2. Methodology

In this study, we propose a novel, low-cost method for urban tree detection and geolocation using readily available geo-tagged street-level images. We started by detecting trees and generating tree instance masks in all available panorama imagery prior to mapping detected trees in space using two deep learning architectures (s. Fig. 1). Our approach assesses the potential to map urban trees in four areas of interest and consists of the following steps (s. Fig. 1):

1. Street-level imagery retrieval for areas of interest (s. 3.3)
2. Tree instance segmentation with Mask R-CNN architecture (s. 2.1.1)
3. Geolocation of trees detected in panoramas through monocular depth estimation and panorama metadata (s. 2.2.1)
4. Geolocation correction of trees present in multiple panoramas through triangulation (s. 2.2.2)

2.1. Tree instance segmentation

2.1.1. Tree instance segmentation model

We trained the Mask Regional Convolutional Neural Network (Mask R-CNN) architecture for tree instance segmentation, the task of pixel-

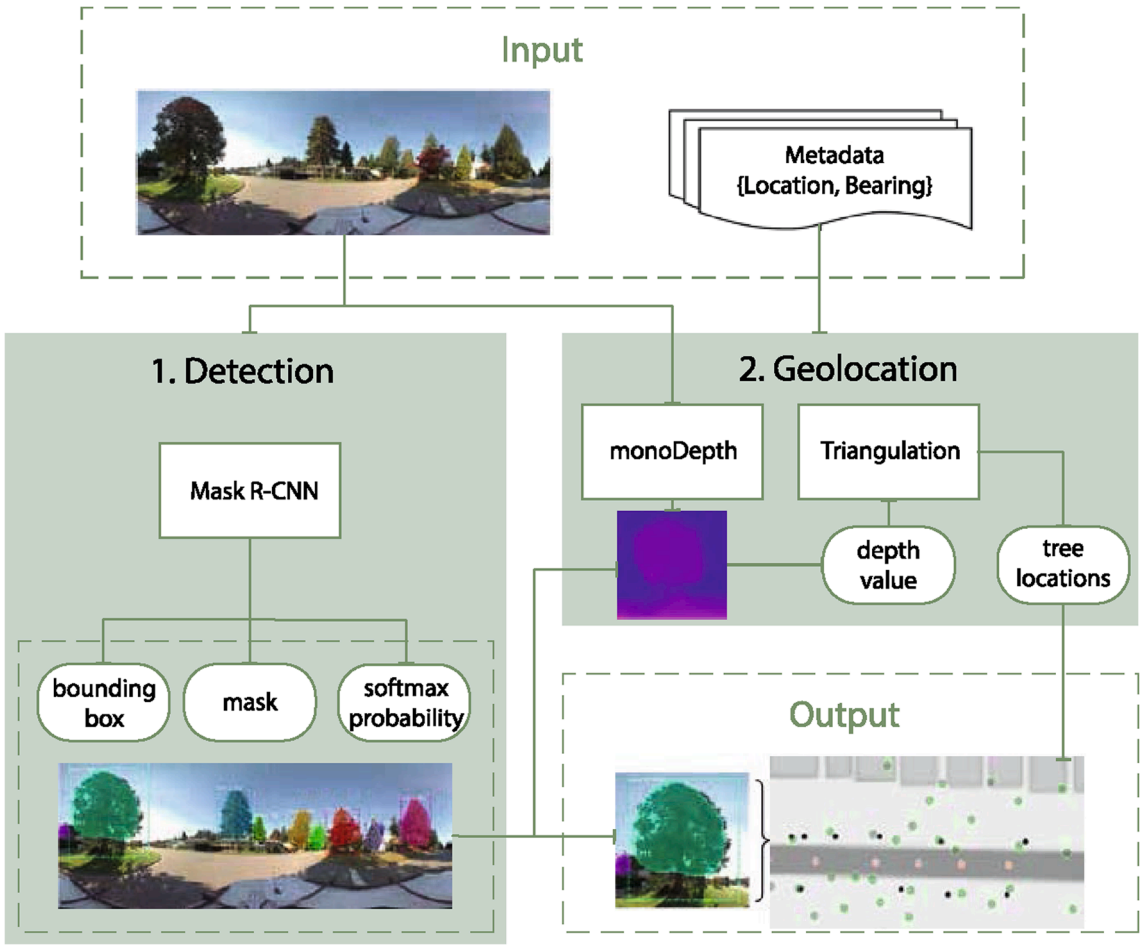

Fig. 1. Urban tree mapping workflow: We first generate a bounding box, a mask and a probability score for all urban trees for the input imagery using our trained Mask R-CNN algorithm. We then compute a dense depth mask with monoDepth for the same input imagery and extract a depth value for every previously generated tree mask. Finally, we use the depth value and imagery metadata in our triangulation pipeline to generate tree locations as geographic coordinates. The output is a map of urban tree positions connected to generated tree masks.

wise detection and delineation of separate objects of interest in an image. Instance segmentation for 'Stuff classes' (fuzzy object classes without clearly delineated shapes, like the sky, trees or other vegetation) is technically challenging and has only recently gained attention in the deep learning community, resulting in a relative scarcity of architectures performing well for this task (Caesar et al., 2016). We chose Mask R-CNN due to its generality, flexibility and the best performing architecture in the recent COCO 2017 Instance Segmentation Challenge (Bolei, 2017). Mask R-CNN is a state-of-the-art architecture that is implemented in the model in a modular way so that it can be replaced, by surpassing instance segmentation algorithms in future. The implementation of Mask R-CNN adopts He et al. (2017) original framework implemented in Python 3, Keras and TensorFlow and can be accessed through Abdulla (2017). Generated outputs include: (1) bounding boxes in pixel coordinates around each detected tree object, (2) a probability score of the class label assigned to a detected object (binary: tree or non-tree), and (3) a pixel mask through the assignment of single pixels to individually detected objects (He et al., 2017). For more detailed information about the architecture of Mask R-CNN c.f. He et al. (2017) and the Appendix.

2.1.2. *Training and inference of the segmentation model*

We used all three, transfer learning, a layered-training approach and fine-tuning to train the Mask R-CNN architecture (Chollet, 2017). In this way, the model first learned to distinguish tree structure through semantic segmentation and was later able to separate single trees more effectively through instance segmentation (Cai et al., 2018).

Through transfer learning, a model pre-trained for one task (e.g. semantic segmentation of 'thing' classes not containing a 'tree' class in the Common Objects in Context (COCO) dataset) is re-trained for another task (e.g. tree instance segmentation on street-level imagery) through the transferal of weights (Bolei, 2017; Pan and Yang, 2010). We transferred weights (i.e. feature representations) of a Mask R-CNN model pre-trained on the COCO dataset to the first (deep) layers of the fresh model, and initialized the training process with these COCO weights. In the subsequent training iteration, defined as the layered-training approach, we used images containing tree objects in the COCO Stuff dataset. For this, we trained the last 5+ top layers of Mask R-CNN (Goodfellow et al., 2016). Finally, we fine-tuned the model by training with the labeled data from Vancouver and Surrey. We therefore trained the model heads (the most shallow or last layers of the model), followed by another iteration training +5 layers of the Residual Learning Network (ResNet101).For more detailed information about the architecture of Mask R-CNN and training strategy used c.f. Goodfellow et al. (2016), He et al. (2017) and the Appendix.

We used heavy data augmentation at train-time during the fine-tuning procedure to avoid over-fitting our model to both, relatively few tree instance samples from Surrey and Vancouver (1000 tree instances), and to the distortions in GSV panorama imagery. In brief, we split the panorama images in half at the start of each training epoch, downscaled the halves to 1024x1024 pixels in size. Each time an image was loaded into memory: (1) we flipped the images left to right 50% of the time; (2) we either re-scaled the image in the x and/or y direction by

a variable factor between 0.8 and 1.2, rotated the image with a random angle between -4 and +4 degrees, or sheared the image with a random angle between -2 and +2 degrees; (3) and performed contrast normalization using a random target factor between 0.9 and 1.1 of the initial contrast of the image.

After finalizing training, the Mask R-CNN model was applied to detect and mask new tree objects in images which were not exposed during the previous training step, which we refer to as the process of inference (Goodfellow et al., 2016). We removed the most heavily distorted top and bottom 150 pixels of all panoramas during inference to further mitigate the impact of panorama distortion.

2.2. Geolocation of trees

2.2.1. Depth estimation

To geolocate individual trees, we first created a dense depth estimate layer for each panorama using monocular depth estimation (Godard et al., 2016). To develop dense depth masks for GSV panorama images, we adapted the monocular depth estimation architecture MonoDepth developed by Godard et al. (2016) because of its applicability to images captured with varying lens types (e.g. panoramic or narrow view) typically used for street-level imagery datasets (Stubbings et al., 2019). MonoDepth is available off-the-shelf as a fully trained unsupervised deep learning model with a depth error margin of less than 20%. Godard et al. (2016) provide multiple trained weights for non-commercial usage, which allows researchers to use the model for inference without having to perform laborious training stages. For this analysis, we followed Godard et al. (2016) recommendations to adopt the best performing weights when MonoDepth was pre-trained on the Cityscapes dataset and fine-tuned using the KITTY vision benchmark dataset (Cordts et al., 2016; Geiger et al., 2013).

The depth estimation model typically computes disparities (D) between objects in each panorama image. In a post processing step, disparities need to be translated into absolute depth ($depth$, s. Eq. (1)) in meters given the focal length (F) and baseline (W_0) of the camera used for capturing training images:

$$depth = \frac{W_0 * F}{D} \tag{1}$$

These parameters are published for datasets such as KITTI and Cityscapes (Godard et al., 2016; Cordts et al., 2016; Geiger et al., 2013). In our work, these parameters needed to be calibrated (C) for the camera model used to capture a GSV image and corrected for differences in image sizes between the input image (W_1) and the original Cityscapes and KITTY datasets (W_0). Depth was then interpreted as per pixel depth estimate ($depth_C$, s. Eq. (2)) in meters between the object in the image and the camera position at the time of image capture:

$$depth_C = \frac{W_0 * F}{(W_1 * D)} * C \tag{2}$$

Based on (Godard et al., 2016), we set F to 0.54 m, W_0 to 721 pixels, W_1 to 6656 pixels and C to 1.5 to account for the use of a camera lens that captures panorama images. The calibration parameter C was determined by minimizing the average error in the absolute monocular depth estimates compared to measured ground-truth tree-to-camera distance.

2.2.2. Triangulation

Once dense depth layers were computed a preliminary geographic location for each tree observation can be calculated (s. Fig. 2 (a)). We extracted a single depth value in meters for each previously detected tree by using the depth pixel values at the center of mass calculated for each instance mask. We then translated each detected tree into geographic coordinates by linking: (1) the calculated depth value from camera to tree, (2) the bearing of each tree in respect to the camera, and (3) the camera's geographic location. Both of the latter are recorded in the imagery's metadata.

In a subsequent processing step, triangulation is used to reduce duplicate observations of individual tree predictions and correct their position estimation where multiple observations of the same tree are recorded (s. Fig. 2 (a)). We assumed that there are no false-positive tree instances (i.e. each tree detected as an object by Mask R-CNN is a real

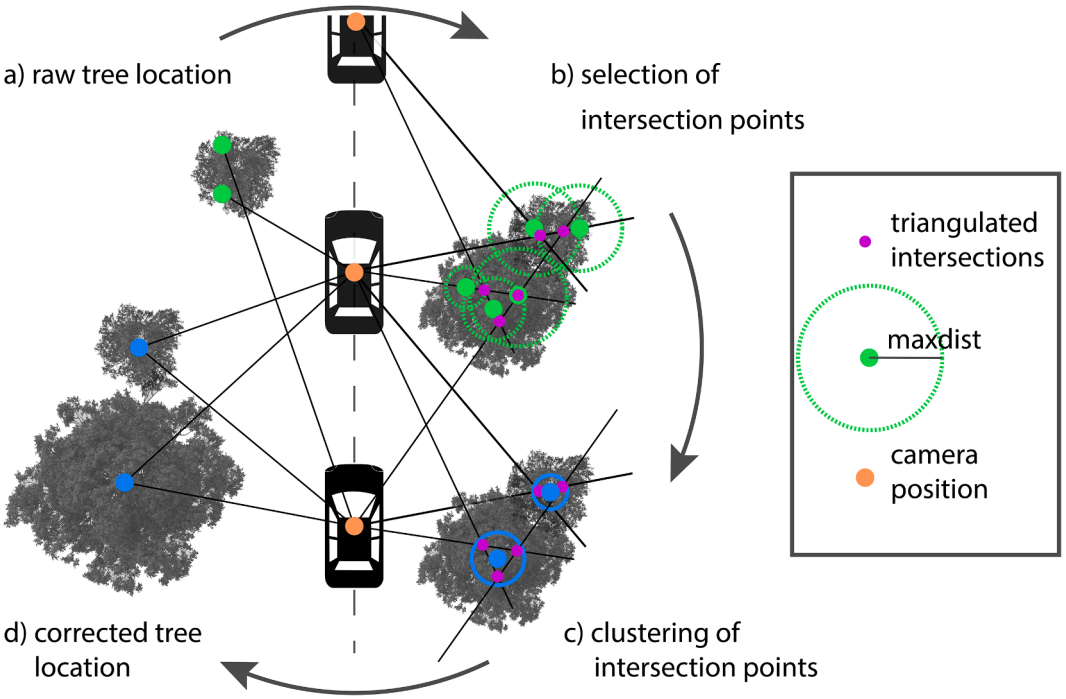

Fig. 2. Correction of predicted tree locations through triangulation. We draw bearings between raw tree locations from monocular depth estimation and camera positions (a). The triangulated intersections of these bearings are selected if they are located within a maximum distance (*maxdist*) from raw tree locations, for a minimum of two raw predictions (b). The closest intersections to the raw tree positions are chosen and clustered (c) to create the final corrected tree position (d).

tree). Using triangulation, we drew edges originating in the camera location and extending beyond the corresponding preliminary tree locations. All crossing edges created nodes of intersections of which some intersections can be considered as triangulated, predicted tree locations. We selected candidate intersections as preliminary tree location estimates according to the following two criteria: First, we picked intersections within a maximum distance of *maxdist* (s. Eq. (3)), to at least two preliminary tree locations (s. Fig. 2 (b)):

$$maxdist = c_0 + c_1 * depth_C \quad (3)$$

c_0 is a constant offset in meters. We chose a value of 3 m for c_0 to account for the average inaccuracy in the GPS positioning of camera locations. c_1 describes the maximum relative error in the depth estimate, calculated as 65%. Second, given that each edge has potentially multiple candidate intersections falling within *maxdist*, we selected only one closest intersection to each preliminary tree position per edge. What remained are candidate intersections that represent potential duplicate locations of detected trees from different panorama images.

Finally, we used hierarchical clustering of all selected candidate intersections to correct the output tree position exploiting multiple detections of the same tree in different panorama images (s. Fig. 2 (c), (d)). To avoid multiple position estimates for the same tree, we chose a 3 m threshold for the clustering as the minimum distance between two trees or cluster centers. We decided to use 3 m as a threshold for clustering as an analysis of the distances between all ground-truth tree measurements revealed that over 99% of all trees were separated by at least 3 m. Final output tree coordinates were the average location of all selected candidate points within a cluster (s. Fig. 2 (c), (d)).

3. Experiments

3.1. Study site

For the majority of model training and assessment, we chose imagery and ground-truth measurement plots distributed over the Metro Vancouver area (49°N, 123°W), specifically in the municipalities of Vancouver, Surrey and Coquitlam (s. Fig. 3). Metro Vancouver is located on Canada's south-west coast, being one of the warmest Canadian regions in winter and experiencing relatively high rainfall rates throughout the year. Mild climatic conditions are favouring growth and survival for tree species from harsher and milder climatic conditions (Stewart and Oke, 2010). One of the largest and most dense cities in Metro Vancouver is the City of Vancouver. Vancouver's urban forest includes many exotic tree species imported from different climatic zones in North America as well as over 60% of Canada's native tree species resulting in one of the most diverse urban forests in Canada (Steele, 2016). Vancouver strives to be one of the world's greenest cities by 2020, resulting in spatially varying types of proactive urban forest management (Isaac et al., 2018). In the following experiments, detected trees encompass both, all public trees planted on streets, referred to as 'street trees', as well as all trees in front yards that could be seen from the street, referred to as 'private trees'. To demonstrate the potential transferability of the model to other urban ecosystems (Alberti, 2008), we evaluate the trained tree instance segmentation model on imagery of the Pasadena Urban Tree dataset (34°N, 118°W), located 2000 km further south in the west coast of the United States (s. Fig. 3). We understand urban ecosystems as described by Alberti et al. (2003) as "human-dominated ecosystems", where human decision making within a city boundary is the "primary driving force behind environmental conditions" (p.1175).

3.2. Ground-truth measurements

To evaluate our model's geolocation performance, we conducted a field campaign in March 2019 to collect ground-truth location measurements of all public and private trees in four areas of interest: Vancouver, Surrey and two in Coquitlam (urban and suburban areas) (s. Fig. 3). We define a tree as any vegetation with a clearly distinguishable stem and crown, which has the potential to grow over 5 m in height in full maturity and has reached a minimum of one meter in height at the

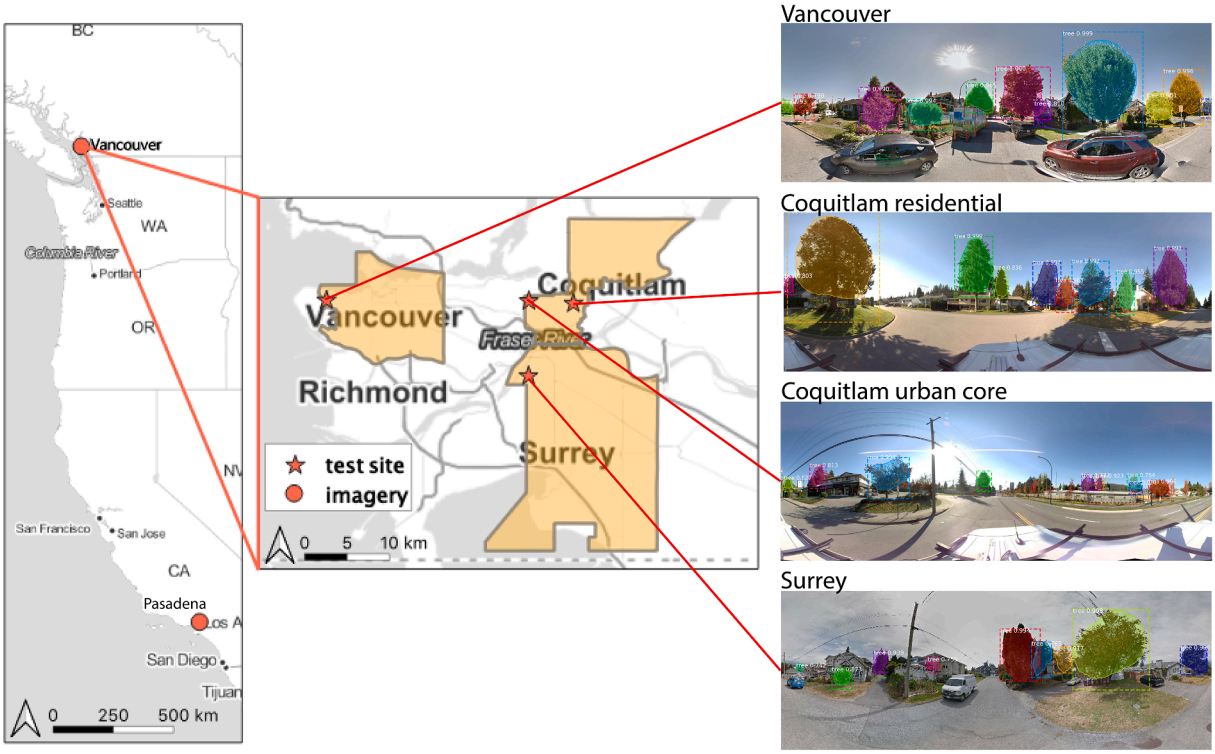

Fig. 3. Location of street-level imagery datasets and ground-truth measurements. Street-level imagery is available for Metro Vancouver and Pasadena. Geolocation test sites are located in the cities of Vancouver, Coquitlam and Surrey, Canada.

time of in-situ measurements (Nitoslawski and Duinker, 2016). We recorded the GPS positions of each visible tree from the sidewalk using a Trimble Geo 7X handheld device in an unobstructed position to maximize GPS signal. We used the rangefinder on the handheld device to measure the offset between the measurement position and the tree, from which a precise tree location could be determined during post-processing. The measured GPS locations were corrected using Base Station Data for Vancouver (BCVC: 491632.73535 N, 123521.58021 W) and Surrey (BCSF: 491131.49655 N, 1225136.24849 W). This provided us with an overall accuracy of under 0.5 m for 294 recorded trees in Vancouver, 152 trees in Surrey, and 336 trees in Coquitlam.

3.3. Street-level imagery

It is common practice in training a CNN to divide a dataset into training, development and testing datasets (Schmidhuber, 2015). This practice ensures that model parameters are adjusted to support the best possible generalization of the model and its applicability to various datasets and tree objects.

Building training and development datasets. We compiled a total of over 36,560 street-level images containing tree labels. We compiled two datasets for two separate training steps described in Section 2.1.1. We split each of the compiled datasets with a 80:20 ratio into training and development datasets. First, we extracted 36,500 images from the openly available COCO Stuff semantic segmentation dataset that contains semantic labels (classified pixels) for trees. The COCO Stuff dataset is, to date, the most expansive collection of images with semantic segmentation labels (~ 164,000) for 'amorphous' Stuff classes (e.g. sky, roads, brick walls, trees, etc.) (Caesar et al., 2016). In contrast, most datasets focus on clearly delineated thing' classes (e.g. people, cars, traffic lights, etc) (Bolei, 2017). Second, we acquired GSV images from Vancouver and Surrey in March 2018 (s. Fig. 3, Table 1). We used the "Labelbox" web tool to create single tree instance labels for combined 60 images for Vancouver and Surrey by manually masking all visible trees, resulting in approximately 1200 tree masks (i.e. instance labels), 453 for Vancouver and 711 for Surrey (Labelbox Inc.).

Building test datasets. To assess the models generalization performance and transferability to imagery of a different urban ecosystem, we evaluated our model's performance on an independent test dataset consisting of imagery from the city of Pasadena (s. Table 1). The dataset, which covers all of Pasadena, was created in March 2016 and is available from Branson et al. (2018). In addition, we used imagery acquired from Mapillary for the city of Coquitlam as a second independent test dataset to demonstrate the robustness of the model applied to different street-level imagery providers (s. Fig. 3, Table 1). For Coquitlam, we downloaded panorama images in February 2019 (Sweden). We randomly sampled 30 test images from both datasets using the NumPy random number generator (assuming a univariate Gaussian distribution) in order to test imagery from different types of city structure. We masked and annotated approximately 360 individual tree masks for Pasadena and 470 for Coquitlam using the same method for labeling Vancouver and Surrey imagery described above. All panorama imagery contained metadata about the camera location of capture, and a 360°bearing reference to true north.

Table 1
Street-level imagery and mask annotations for training and evaluation. Compiled datasets consist on 30 panorama images each.

	dataset	provider	tree masks	green infrastructure
Vancouver	train/dev	Google	453	street and private trees
Surrey	train/dev	Google	711	private trees
Pasadena	test	Google	365	street and private trees
Coquitlam	test	Mapillary	471	street and private trees

3.4. Evaluation of the tree identification model

We evaluated the method on two development datasets (Vancouver and Surrey, using GSV imagery) and two test datasets (Coquitlam using Mapillary, and Pasadena using GSV imagery). To evaluate model performance, we chose three commonly used evaluation metrics for instance segmentation frameworks (Bolei, 2017): (1) mean average precision (*mAP*), (2) average precision over Intersection over Union (IoU) using a threshold of 0.5 (AP_{50}), and (3) average precision over IoU with threshold 0.75 (AP_{75}). To quantify the known, negative influence of small tree mask sizes on instance segmentation performance in detail, we iteratively excluded all smaller masks under a mask size threshold and compared recalculated evaluation metrics (Kisantal et al., 2019). Next, we manually inspect failure cases according to three different tree mask sizes to identify the most frequent tree detection error. For detailed tree instance segmentation error assessment, we defined small masks to be under 3000 pixels in size (approximately 0.3% of total image pixels per small mask), representing detections of very distant trees (>70 m) relative to the camera position at image capture. We defined medium masks to be between 3000 and 30,000 pixels in size, roughly all private trees, found in front yards, and big masks to be over 30,000 pixels in size (approximaely 3% of total image pixels per big mask), repetitive to large front yard trees, or street trees.

Additionally, we calculated precision and recall to compare detection results with annotated images (Davis and Goadrich, 2006):

$$precision = \frac{tree_{pred}}{tree_{pred} + other_{pred}} \tag{4}$$

$$recall = \frac{tree_{pred}}{tree_{annotated}} \tag{5}$$

All final evaluation metrics and precision-recall curves to compare model performance for different datasets were calculated excluding very small masks under a size of 3000 pixels (approximately 0.3% of total image pixels per small mask) (Caesar et al., 2016).

3.5. Evaluation of the tree location model

We used measured public and private tree positions to evaluate the tree location model performance for areas of interest in Vancouver, Surrey, an urban and a suburban area in Coquitlam. We evaluated the absolute positioning error of tree predictions as the euclidean distance between the ground-truth measurement and the tree location predictions (Yin and Wang, 2016). We used a greedy algorithm to assign closest matching trees first, and then took matched trees out of the running process until no ground-truth measurements were left to match (Wegner et al., 2016). A match is kept as a true positive match if the distance between ground-truth measurement and tree prediction does not exceed 15 m. A 15 m threshold was chosen to include as many private trees as possible in the error assessment, typically found in a 15–30 m range from the camera position at time of image capture. Trees located more than 15 m from the measured private ground-truth data represent distant tree detections behind houses without comparable measured ground-truth data and were excluded in the error assessment by the given threshold. We defined a measure of absolute tree positioning accuracy as the mean of absolute positioning errors $mean_{epos}$ (Yin and Wang, 2016):

$$mean_{epos} = \frac{1}{n_{TP}} \sum_{i=1}^{n_{TP}} \sqrt{(x_i - x_{predi})^2 + (y_i - y_{predi})^2} \tag{6}$$

We evaluate absolute tree positioning accuracy and the ratio of matches to non-matches for public, private and all trees respectively.

4. Results

4.1. Tree instance segmentation

4.1.1. Effects of layered training

Performing layered training with COCO Stuff images improves the detection of tree objects from ~ 80% without COCO Stuff training to a slight overcounting of trees with 103% detected trees compared to the labeled tree masks in the combined Vancouver and Surrey datasets (s. Table 2). The overall model accuracy with AP_{50} and *mAP* improves slightly while values for AP_{75} decline. The small decrease in mask accuracy for AP_{75} may be a direct result of the layered model including detections of trees which are more difficult to detect compared to the lower number of detections when layered training is not used.

4.1.2. Transferability between different urban ecosystems and data sources

We find that Mask R-CNN developed on Vancouver and Surrey training datasets was successfully applied to detecting trees across all four datasets (s. Fig. 4, Table 3). AP_{50} values ranging from 0.620 to 0.682 and values of other evaluation metrics are consistent with state-of-the-art tree or plant semantic segmentation performances found in other studies, with the difference that we not only evaluate pixel-based classification results, but also distinguish between different tree objects (Caesar et al., 2016) (s. Table 3). AP_{75} and AP_{50} values are lowest for Surrey (0.157, 0.262) and highest for Pasadena (0.262, 0.316) and Coquitlam (0.261, 0.342) datasets. Overall, at a recall threshold of 0.6, precision is above 0.8 for all four datasets (s. Fig. 4). With a recall of 0.35 or higher, precision-recall curves for both development datasets are slightly higher than for the testing datasets, which is to be expected since assessed features in the development datasets directly influence the training process (s. Fig. 4) (LeCun et al., 2015).

Mask R-CNN performance for the Pasadena and Mappillary test datasets is very similar (*mAP*) to slightly better (AP_{75},AP_{50}) than that of the Vancouver and Surrey dataset (s. Table 3). Slight differences in precision-recall curves and variations in evaluation metrics may be attributed to overall varying tree shapes and sizes found in each dataset. Features learned throughout the trained Mask R-CNN model appear to be sufficient to detect a variety of urban trees in different urban greenspace management settings, i.e. they are not limited to tree species and forms observed in Vancouver and Surrey (i.e. detection of palm trees in Pasadena). The tree detection model is therefore robust to a variety of urban ecosystems and urban green space design without the need for extensive retraining.

Furthermore, performing inference on Coquitlam (Mapillary) test imagery without retraining results in the highest *mAP* value of the four datasets. Model performance appears robust to the different data source and sensors used for street-level photography, i.e. Mapillary, (s. Fig. 4, Table 3). Precision-recall curves for both testing datasets appear to be very similar. This indicates that the presented model has the ability to generalize well for both a city with a very different urban ecosystem (Pasadena) and imagery from different data sources or sensors (Mapillary in the case of Coquitlam). The consistency of model performance with an $AP_{50} > 0.6$ regardless of data and sensor source implies that panoramas acquired from both GSV and Mapillary are suitable for use in the urban tree mapping model.

4.1.3. Instance segmentation performance as a function of mask size

Next, we assessed the influence of tree mask sizes on Mask R-CNN instance segmentation performance. Plotting *AP* metrics values against mask size thresholds shows that larger masks get predicted more accurately (*AP* values over 0.6) (s. Fig. 5). This is expected, since the ratio between the outline of a tree and the tree mass contained by the outline (i.e. outline-mass-ratio) decreases with object size, resulting in a bigger weight of the fuzzy tree outline with decreasing mask size in the calculation of model evaluation metrics (Kisantal et al., 2019). Predicting precise fuzzy tree outlines (tree to sky interface) is often harder than predicting the tree mass. Outlines therefore often differ more from the ground-truth outline than the actual mass of a tree, resulting in declining evaluation metrics numbers with smaller tree size. Notably, the accuracy for large masks are similar for both the Vancouver (GSV) and Coquitlam (Mapillary) datasets, while accuracy for Surrey (GSV) and Pasadena (GSV) datasets are approximately 0.2 lower. This indicates that similarity in tree object structure, to a degree, may have a bigger influence on image segmentation performance than similarity in image quality when instance segmentation is performed on multiple datasets. Values for AP_{50} increase once masks under 3000 pixels (which represent very small or distant trees) are removed. There is a corresponding significant decrease in accuracy for masks smaller than 3000 pixels once all masks over 3000 pixels in size are discarded.

Table 2
Evaluation metrics for training with and without COCO Stuff on combined Vancouver and Surrey dataset.

	AP_{50}	AP_{75}	*mAP*	mask predicted	mask annotations
COCO only	0.608	0.252	0.281	74	90
COCO-Stuff	0.661	0.199	0.290	93	90

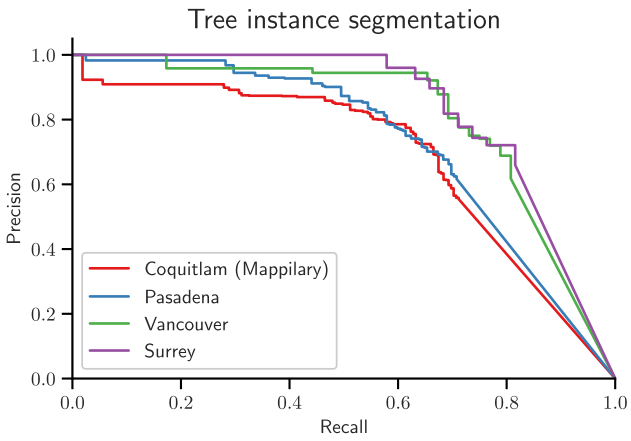

Fig. 4. Precision-Recall curves for development (Surrey, Vancouver) and test (Coquitlam, Pasadena) datasets.

Table 3
Evaluation metrics for Vancouver, Surrey and Pasadena. (Excluding masks under 3000 pixels in size.)

	AP_{50}	AP_{75}	*mAP*	mask predicted	mask annotations
Vancouver	0.682	0.232	0.312	53	52
Surrey	0.634	0.157	0.262	40	38
Pasadena	0.628	0.262	0.316	194	202
Coquitlam	0.620	0.261	0.342	238	215

4.1.4. Error analysis for instance segmentation

We manually inspected 296 failure cases (including small masks) to identify the most frequent tree detection error (s. Fig. 6 and 7). The majority of errors arise from densely planted public and private trees, resulting in two trees being detected as one combined tree, occlusion of trees, or otherwise overlapping trees (s. Fig. 6 (a), (b), (e), (h)). This source of error confirms that distinguishing between visually overlapping amorphous objects is a difficult task (Caesar et al., 2016). Detecting trees using multiple street-level perspectives potentially offsets this error source, as occluded or overlapping trees can either be seen in the foreground or are otherwise distinguishable from another perspective and image in the full model (Krylov et al., 2018). Detecting hedges as false positives, and detecting small trees, trees in shadows of buildings and trees with leaf-off condition on non-sky background as

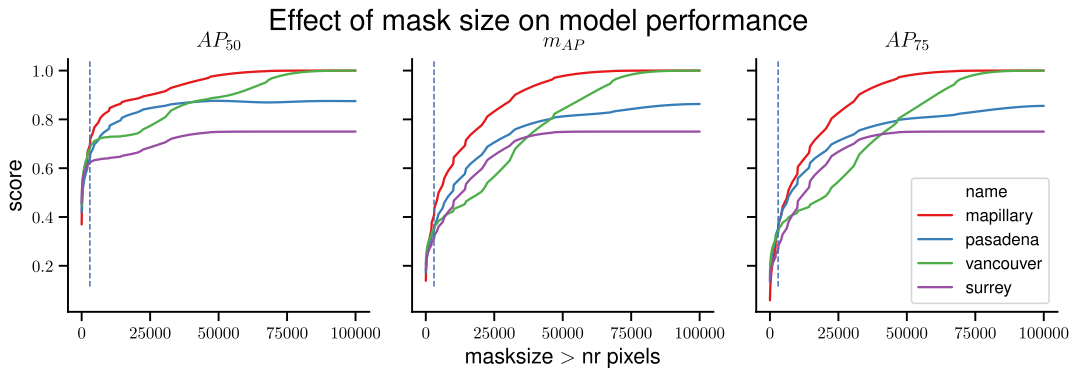

Fig. 5. Effect of mask size on model performance. We iteratively excluded smaller masks, measured by the number in pixels per mask in calculating AP_{50}, mAP and AP_{75}. The dotted vertical line indicates the cut-off size of >3000 pixels for final calculation of the evaluation metrics. The model can predict big tree masks of Coquitlam (Mappilarry) in red and Vancouver in green datasets the most accurate.

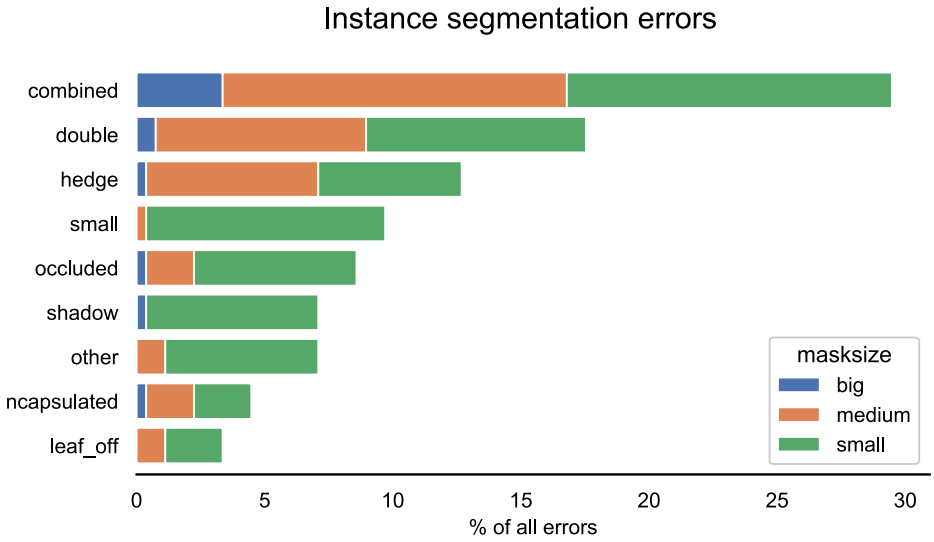

Fig. 6. The most common inference errors of tree instance segmentation with Mask R-CNN (in percent).

false negatives (s. Fig. 7, (c), (d), (f), (g), (i)) is a direct result of having very few training examples of these special cases in our datasets. As expected, most of these errors were detected for small mask sizes (<3000 pixel) (s. Fig. 6).

Trees seen far in the distance were often disregarded in the manual labeling process due to their small mask size and relative distance for the direct camera location at image capture time. We note that 200 of these human labeling errors were recorded, describing instances where the model correctly identified a tree but no corresponding tree label was created. We disregarded human labeling errors for small masks, as masks under a threshold of > 3000 pixels in size were not included in the final evaluation (s. Section 4.1.3). These smaller masks, which represent trees found in backyards or distant trees, could be included with help of additional data augmentation methods mentioned in (Kisantal et al., 2019) in future analyses.

4.2. Tree geolocation

4.2.1. Comparison to ground-truth

We matched 70% of all ground-truth tree measurements with tree predictions after excluding all matches over 15 m in distance as false positive matches (s. Fig. 8). Non-matched ground-truth measurements often result from a tree missing in the tree detection process, through either occlusion by larger trees in the front of an image (s. Fig. 8 (a)), or by the absence of a tree in either the ground-truth measurements or the street-level imagery, due to a two or more year time difference in the ground-truth and imagery datasets (s. Fig. 8 (b)). Localizing trees using monocular depth estimation can potentially help to prevent loss of information since every single tree detection can be localized and is not dependent on many photographs from different views (Krylov et al., 2018). We note that triangulation, in comparison to raw tree location predictions, successfully reduced the mean absolute position error for all trees from 9 to 7 m, by approximately 2 m, and the total count of tree predictions by 45-55%.

Minimum distances between tree location prediction and ground-truth measurements for all areas are 0.26 m or higher (s. Table 4). Tree location prediction in Vancouver is, with a mean of 5.28 and a median position accuracy of 4.36 m approximately 2 m more accurate than all other areas, followed by Coquitlam's urban and suburban areas with a mean and medium slightly above 6 m. With a mean of 7.06 m and a medium of 6.87 m, geolocation performance in Surrey is lower than in all other areas. Overall lower position accuracy in the Surrey dataset could be attributed to the area of interest being located on a slope >15% negatively influencing triangulation, compared to other areas with no or a relatively low slope (<5%). Another source of error may be the overall more spread building and green space structure of the Surrey area. This structural characteristic is leading to trees being located further away from the camera position of image capture with a slight decline in detection and location accuracy with resulting smaller tree masks, discussed in Section 4.2.3 (s. Fig. 8 (c)).

Fig. 7. Examples of masking errors in instance segmentation with Mask R-CNN. Most common errors include: (a) separate trees detected as one, (b) one tree split into multiple detections, (c) hedges or shrubs detected as trees, (d) small trees that were not detected, (e) undetected trees behind detected trees, (f) undetected trees in shadows of buildings, (g) non-tree objects detected as trees and masking errors, (h) undetected small trees in front of large trees, and (i) undetected trees with leaf-off condition and non-sky background.

4.2.2. Location accuracy for private and street trees

After triangulation, 143 street trees (93% of the ground-truth measurement) were successfully located in the Vancouver area. 11 trees (7%) of 154 street trees remained unmatched (<15 m ground-truth to prediction). The majority (9 trees) of ground-truth street trees that were not matched were either newly or re-planted small trees in between the date of capture for street-level images and the collection of ground-truth data (i.e. 2 years) (s. Fig. 8, (c)). The two remaining unmatched trees were not detected in the instance segmentation step due to large vehicles obstructing the trees in respective street-level images. Distances from ground-truth to tree predictions for street trees range from 0.26 to 13.14 m with a mean of 4.31 m, a median of 3.92 m and a standard deviation of 2.76 m (s. Table 4). The mean of street trees (red) can be detected almost 1 m more accurately than the mean of all trees (blue) and 4 m more accurately than the mean of private trees (green) in the Vancouver area (s. Fig. 9). The overall more accurate predictions in Vancouver are possibly a result of the presence of uniformly and separately planted street trees (s. Fig. 9, red, and Fig. 8). Owing to street trees proximity to the camera position, tree masks are bigger and predicted more accurate which has a potential influence on the location prediction of street trees (s. Section 4.1.3).

Absolute location accuracy for private trees with the presence of street trees is 3 m less accurate compared to the previously discussed overall accuracies of all of Vancouver, Surrey and Coquitlam areas (s. 4). Values for Vancouver's private trees are a minimum of 1.56 m, a mean of 8.55 m, a median of 8.38 m and a standard deviation of 3.77 m. 70% (97

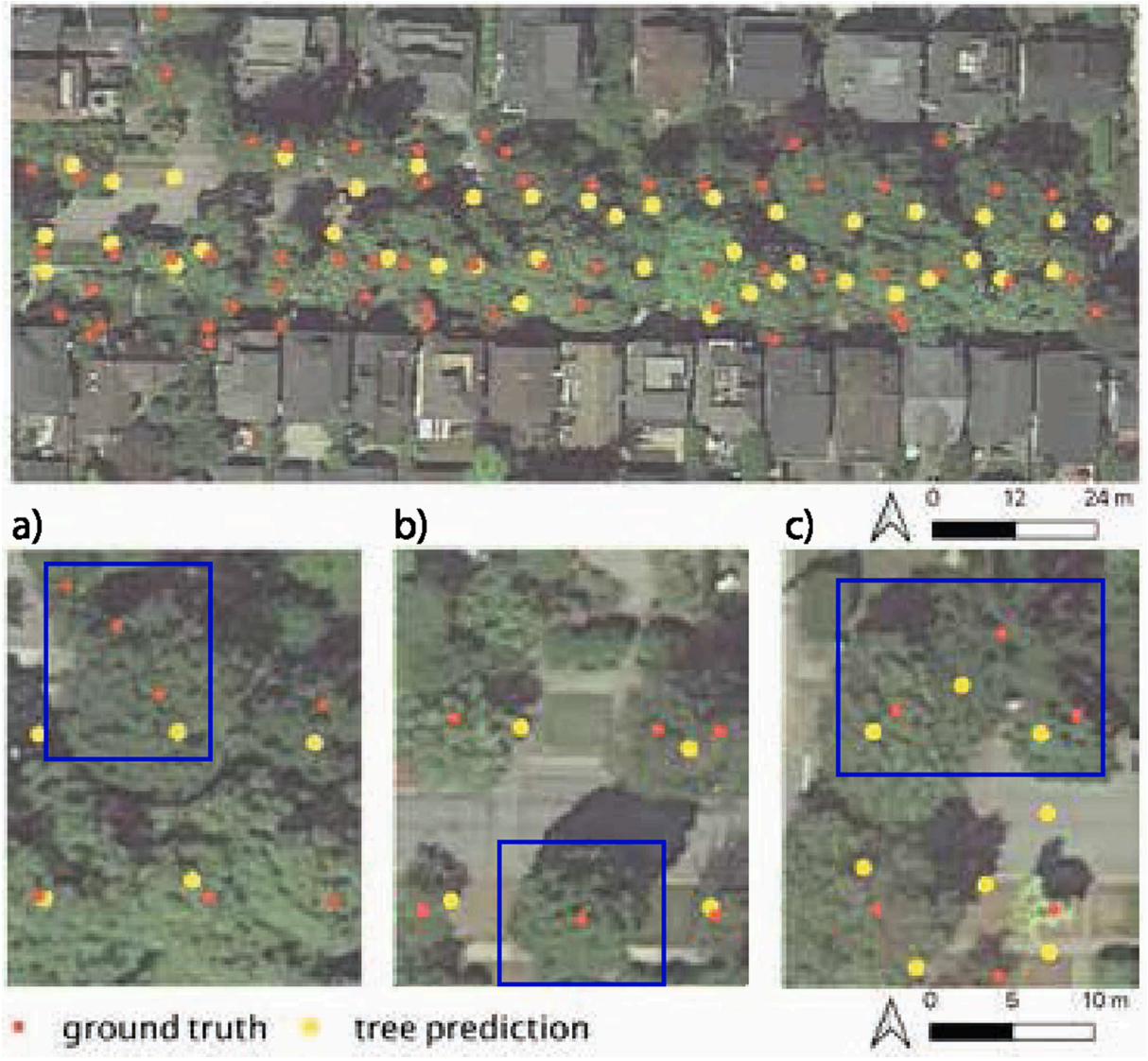

Fig. 8. Location prediction results of trees. Predicted tree positions (yellow), ground-truth measurements (red) and common detection errors (blue). 70% of all measured trees were detected, 30% are missing through occlusion by large trees (a). The time difference between ground-truth measurements and when a street-level image was taken (>=2 years) results in the absence of either a tree prediction or a ground-truth measurement (b). Geolocation accuracy decreases slightly with increasing distance of trees to the car position of image capture (c).

Table 4
Absolute geolocation accuracy (in meters).

	match	min	mean	median	std
Vancouver	235	0.26	5.28	4.36	3.59
street trees	143	0.26	4.31	3.92	2.76
private trees	97	1.56	8.55	8.38	3.77
Surrey	94	0.42	7.06	6.87	3.36
Coquitlam (urban)	64	0.46	6.58	6.26	3.22
Coquitlam (suburban)	159	0.55	6.83	6.07	3.73

trees) of all recorded private trees were matched, 30% (43 trees) were not matched. Both, non-matches and low position accuracy of private trees may be influenced by the more varying spatial pattern of planted private trees. Surrey and Coquitlam areas with similar spatial tree heterogeneity, not influenced by street trees, also recorded approximately 30% of non-matched ground-truth trees. As previously discussed, the combination of two trees detected as a single tree is the most common tree detection error for all mask sizes (s. Section 4.1.4). This error is expected to occur more often for densely planted private trees with overlapping canopy, than uniformly planted street trees (Wegner et al., 2016).

It is also possible that the presence of street trees influences our model's location accuracy. Surrey and Coquitlam's (suburban) private trees (no presence of street trees) show lower positional errors than Vancouver's private trees. Street trees often overlap with private trees in street-level photographs due to their proximity to the camera position. Street trees therefore influence both, monocular depth estimation of private trees and the bearing information of the tree detection bounding box from the camera position, as the center of mass shifts towards the larger part of the mask, the street tree. These detection errors negatively influence our localization process and may result in lower positional accuracy for private trees in areas with street tree presence.

4.2.3. Location accuracy with distance of tree from position of image capture

Distances of street tree measurements to the camera positions range between 6-14 m, distances of private tree measurements to the camera position are typically >15 m away, resulting in a bi-modal distribution of all distances between ground measurements and car positions for all

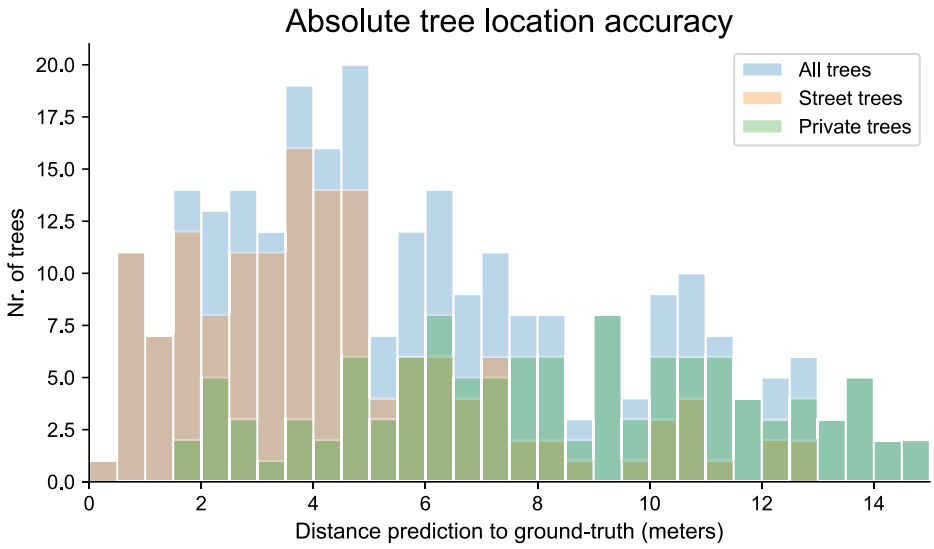

Fig. 9. Absolute geolocation accuracy for street trees (red), private trees (green), and all trees (blue) in the Vancouver area. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

areas (s. Fig. 10). Another reason for more accurate geolocation in Vancouver may be attributed to Vancouver trees being positioned closer to the camera at the time of image capture than in Surrey and Coquitlam image datasets (s. Fig. 10). The range of absolute position error from the tree location prediction to the tree ground-truth measurement is slightly lower for trees closer to the camera position than the error range for trees further away from the camera (s. Fig. 8, (b)), indicated by the shape of the kernel-density estimate in Fig. 10. However, only a low correlation with R-square of 0.17 can be recorded between the distance of the predicted tree to the ground-truth measurement and the distance of the ground-truth measurement to the car position. The mean positional error increases with distance to the camera by approximately 0.23 times the distance between ground-truth tree location and camera, indicated by the slope of Fig. 10. This aligns in magnitude with a Root Mean Squared Log Error (RMSLE) of approximately 0.2 reported by Godard et al. (2016) for the increase in error for monocular depth estimation with distance from the camera position. This suggests that the distance from the camera position at time of image capture to trees of interest should be a consideration when choosing or generating a dataset for urban tree mapping for future application. Random noise of 6.3 m is introduced, likely through different street slopes, described tree detection errors and resulting triangulation errors. We also detect a systematic error, an intercept of 2.7 m with a potential cause through systematic car position GPS inaccuracies in urban landscapes (Falco et al., 2017). Another cause for this systematic error could be the initial tree location prediction using the center of mass for each tree crown, retrieving a depth measurement for the outside of the crown diameter instead of the usually measured stem position.

5. Conclusion

To support decision-making and research that can improve the management of urban forests, cities need more cost-efficient and widely applicable tools that can provide high-resolution spatial information on single urban trees for the entire urban and peri-urban landscape (Kelly et al., 2007). We presented a promising low-cost framework for mapping individual urban trees along street networks over large areas that shows potential to be adopted in different cities around the world. This novel model relies solely on street-level imagery as a data input and does not require any additional, potentially expensive VHR aerial or satellite

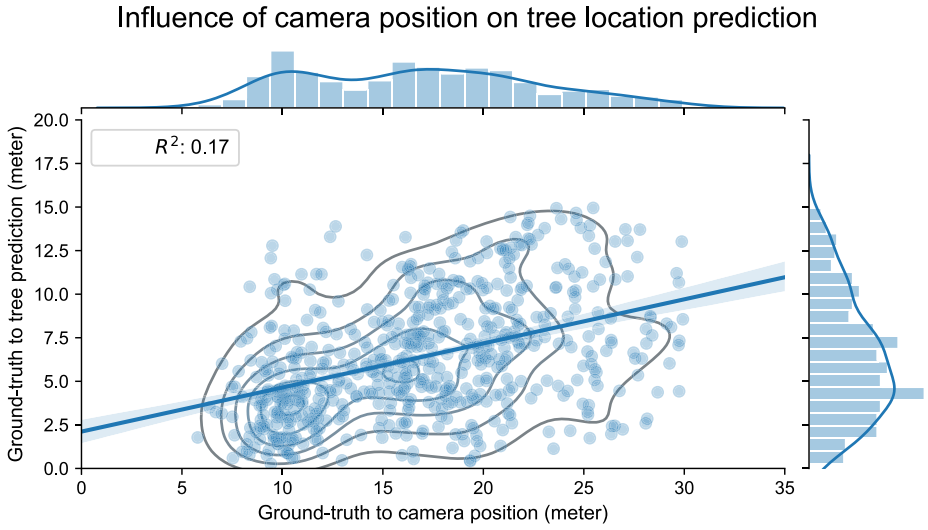

Fig. 10. Influence of camera position at time of image capture on tree location prediction. Comparison of the distance of ground-truth measurements to the camera location at time of image capture vs. ground-truth measurements to predicted tree locations for all data points (Vancouver, Surrey, Coquitlam)).

imagery for the geolocation of trees. Furthermore, it is developed and tested to be transferable over different image sources and geographical regions as evidenced by our experimental results.

The approach can be applied to a diversity of urban trees and forests, both public and private, and could form the basis for urban assessments that require single tree detection. We found that Mask R-CNN can be successfully trained to identify fuzzy objects like trees to a high precision with a minimal amount of training images (48 images) and a layered training approach integrating open source imagery datasets (COCO Stuff). The experimental results of this study demonstrate that a layered training approach resulted in a 23% higher tree detection rate compared to only using transfer learning. AP_{50} values over 0.62 are consistent with state-of-the-art results in other studies segmenting fuzzy objects (Caesar et al., 2016). The combination of deep learning and street-level imagery appears promising for the detection of trees in different urban ecosystems. Further, the model is not limited to the use of the same sensor or dataset. Both Mapillary and GSV panoramas showed suitable for urban tree mapping.

We accurately geolocated trees using a monocular depth estimation algorithm and triangulation that requires no additional contextual information. The geolocation of street trees with a mean accuracy of around 4 m was approximately 2 m more accurate than the mean accuracy of 6 m for private trees located in front yards seen from the street. This suggests that the distance from the camera position to trees of interest at time of image capture should be a consideration for future application when selecting or generating a dataset for urban tree mapping. Detection errors influenced our tree geolocation module and we recommend collecting images in a range of 7–14 m away from trees of interest for best positioning results.

Street-level imagery in combination with deep learning brings a new approach to gaining a baseline understanding of urban forests. Accurate masking and geolocation of trees can provide the basis for a variety of quantitative urban forest assessments. Future research directions aimed at optimizing the use of street-level imagery for urban tree mapping include: assessing the influence of spacing between camera positions on triangulation and positional accuracy of tree predictions; evaluating the influence of slope and vertical terrain variability on geolocation performance; improving geolocation performance for areas with high terrain variability; and assessing the influence of seasonality and leaf-off condition on instance segmentation performance.

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.

Acknowledgements

Funding: This work was financed by The University of British Columbia as well as by Genome Canada, Genome British Columbia, and Genome Quebec under the research project Biosurveillance of Alien Forest Enemies (bioSAFE, project number #10106), as part of the 2015 Large-Scale Applied Research Project competition in Natural Resources and the Environment: Sector Challenges – Genomic Solutions. TD is supported by the Banting Postdoctoral Fellowship program (201709BPF-393653-294704) in partnership with the Social Sciences and Humanities Research Council (SSHRC) of Canada. The authors are also very thankful for the constructive comments provided by three anonymous reviews, which have greatly improved the manuscript.

Appendix

Mask Regional Convolutional Neural Network (Mask R-CNN). Mask R-CNN extends Faster R-CNN used in Branson et al. (2018) by adding a segmentation masks prediction for each detected instance or object. It therefore allows us to classify and detect single urban tree instances, create bounding boxes and segmentation masks surrounding the individual trees. Mask R-CNN can be conceptualizes as a two stage algorithm: (1) The first part is also refered to as a Region Proposal Network (RPN), predicting multiple Regions of Interest (RoI). A convolutional backbone architecture, Residual Learning Network coupled with a Feature Pyramind Network (ResNet101-FPN), allows to generate multiple anchor boxes at different scales (Lin et al., 2017). (2) In the so called head of the model, features are then extracted from each RoI which can be associated with the relevant class. All in parallel, the head extracts class and bounding box values, leveraging a RoIPool operation, and creates a binary mask for each detected object using a RoIAlign operation (Zhang et al., 2018). For more detailed information about the architecture of Mask R-CNN please see He et al. (2017).

Hyper-parameters used for training Mask R-CNN. In this study, we chose transfer learning, followed by a layered training approach, and by fine-tuning as a strategy to train Mask R-CNN. After initializing Mask R-CNN with COCO weights, We trained the last 5+ top layers of Mask R-CNN with 50 epochs on COCO Stuff using a learning rate of 1e-4. An epoch refers to all images in the training dataset being run through the entire model and the internal model parameters being updated at least once (Goodfellow et al., 2016). Then, we fine-tuned the model heads (the most shallow or last layers of the model) with the labeled data from Vancouver and Surrey for 30 epochs, followed by another iteration training +5 layers of the Residual Learning Network (ResNet101). After a sparse grid search we found 1e−4/10 to be the most successful learning rate to use for the fine-tuning step. We set all other hyperparameters following recommendations in existing research that employs Mask R-CNN (He et al., 2017; Abdulla, 2017). The full model was trained on a NVIDIA GTX1080 Ti Graphics Processing Unit (GPU), and limited by 11GB of memory to stochastic gradient descent, or a mini-batch size of 1.

References

Abdulla, W., 2017. Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow: matterport/Mask_RCNN, https://github.com/matterport/Mask_RCNN, original-date: 2017-10-19T20:28:34Z, 2017.

Agarwal, S., Furukawa, Y., Snavely, N., Curless, B., Seitz, S.M., Szeliski, R., 2010. Reconstructing Rome. Computer 43 (6), 40–47. https://doi.org/10.1109/MC.2010.175. ISSN 0018-9162.

Alberti, M., 2008. Advances in Urban Ecology: Integrating Humans and Ecological Processes in Urban Ecosystems, Springer, US, ISBN 978-0-387-75509-0, 2008, doi: 10.1007/978-0-387-75510-6, https://www.springer.com/gp/book/9780387755090.

Alberti, M., Marzluff, J.M., Shulenberger, E., Bradley, G., Ryan, C., Zumbrunnen, C., 2003. Integrating Humans into Ecology: Opportunities and Challenges for Studying Urban Ecosystems. Bioscience 53 (12), 1169–1179. https://doi.org/10.1641/0006-3568(2003)053[1169:IHIEOA]2.0.CO;2. ISSN 0006-3568.

Alonzo, M., Bookhagen, B., Roberts, D.A., 2014. Urban tree species mapping using hyperspectral and lidar data fusion. Remote Sens. Environ. 148, 70–83. https://doi.org/10.1016/j.rse.2014.03.018. ISSN 0034-4257, http://www.sciencedirect.com/science/article/pii/S0034425714001047.

Aval, J., Demuynck, J., Zenou, E., Fabre, S., Sheeren, D., Fauvel, M., Adeline, K., Briottet, X., 2018. Detection of individual trees in urban alignment from airborne data and contextual information: A marked point process approach. ISPRS J. Photogram. Remote Sens. 146, 197–210. https://doi.org/10.1016/j.isprsjprs.2018.09.016. ISSN 0924-2716, http://www.sciencedirect.com/science/article/pii/S0924271618302594.

Berland, A., Lange, D.A., 2017. Google Street View shows promise for virtual street tree surveys. Urban Forest. Urban Green. 21, 11–15. https://doi.org/10.1016/j.ufug.2016.11.006. ISSN 1618-8667, http://www.sciencedirect.com/science/article/pii/S1618866716303181.

Bolei, Z., 2017. COCO + Places 2017 Challenge, https://places-coco2017.github.io, 2017.

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. Photogram. Remote Sens. 135, 13–30. https://doi.org/10.1016/j.isprsjprs.2017.11.008. ISSN 0924-2716, http://www.sciencedirect.com/science/article/pii/S0924271617303453.

Caesar, H., Uijlings, J., Ferrari, V., 2016. COCO-Stuff: Thing and Stuff Classes in Context, arXiv:1612.03716 [cs] http://arxiv.org/abs/1612.03716, arXiv: 1612.03716.

Cai, B.Y., Li, X., Seiferling, I., Ratti, C., 2018. Treepedia 2.0: Applying Deep Learning for Large-scale Quantification of Urban Tree Cover. https://arxiv.org/abs/1808.04754.

Cheng, L., Yuan, Y., Xia, N., Chen, S., Chen, Y., Yang, K., Ma, L., Li, M., 2018. Crowd-sourced pictures geo-localization method based on street view images and 3D

reconstruction. ISPRS J. Photogram. Remote Sens. 141, 72–85. https://doi.org/10.1016/j.isprsjprs.2018.04.006. ISSN 0924-2716, http://www.sciencedirect.com/science/article/pii/S0924271618301102.
Chollet, F., 2017. Deep Learning with Python, Manning Publications Co., Greenwich, CT, USA, 1st edn., ISBN 978-1-61729-443-3, 2017.
Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. In: IEEE Conference on Computer Vision and Pattern Recognition, 3213–3223, URL http://openaccess.thecvf.com/content_cvpr_2016/html/Cordts_The_Cityscapes_Dataset_CVPR_2016_paper.html, 2016.
Davis, J., Goadrich, M., 2006. The Relationship Between Precision-Recall and ROC Curves. In: Proceedings of the 23rd International Conference on Machine Learning, ICML '06, ACM, New York, NY, USA, 233–240, ISBN 978-1-59593-383-6, 2006, doi:10.1145/1143844.1143874, http://doi.acm.org/10.1145/1143844.1143874, event-place: Pittsburgh, Pennsylvania, USA.
Duarte, F., Ratti, C., 2017. What Big Data Tell Us About Trees and the Sky in the Cities. In: De Rycke, K., Gengnagel, C., Baverel, O., Burry, J., Mueller, C., Nguyen, M.M., Rahm, P., Thomsen, M.R. (Eds.), Humanizing Digital Reality: Design Modelling Symposium Paris 2017, Springer Singapore, Singapore, 59–62, ISBN 978-981-10-6611-5, 2018, doi:10.1007/978-981-10-6611-5_6, doi: 10.1007/978-981-10-6611-5_6.
Falco, G., Pini, M., Marucco, G., 2017. Loose and Tight GNSS/INS Integrations: Comparison of Performance Assessed in Real Urban Scenarios. Sensors 17 (2), 255. https://doi.org/10.3390/s17020255 https://www.mdpi.com/1424-8220/17/2/255.
Geiger, A., Lenz, P., Stiller, C., Urtasun, R., 2013. Vision meets robotics: The KITTI dataset. Int. J. Robot. Res. 32 (11), 1231–1237. https://doi.org/10.1177/0278364913491297. ISSN 0278-3649.
Godard, C., Mac Aodha, O., Brostow, G.J., 2016. Unsupervised Monocular Depth Estimation with Left-Right Consistency, arXiv:1609.03677 [cs, stat] http://arxiv.org/abs/1609.03677, arXiv: 1609.03677.
Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning, MIT Press, ISBN 978-0-262-03561-3, google-Books-ID: Np9SDQAAQBAJ, 2016.
He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask R-CNN. In: 2017 IEEE International Conference on Computer Vision (ICCV), 2980–2988, 2017, doi:10.1109/ICCV.2017.322.
Hirschmuller, H., 2008. Stereo Processing by Semiglobal Matching and Mutual Information. IEEE Trans. Pattern Anal. Mach. Intell. 30 (2), 328–341. https://doi.org/10.1109/TPAMI.2007.1166. ISSN 0162-8828.
Isaac, K., Beaulieu, L., Owen, C., Danyluk, A., Mulhall, B., 2018. Urban Forest Strategy.
Jang, K.M., Kim, J., Lee, H.-Y., Cho, H., Kim, Y., 2020. Urban Green Accessibility Index: A Measure of Pedestrian-Centered Accessibility to Every Green Point in an Urban Area. ISPRS Int. J. Geo-Inform. 9 (10), 586. https://doi.org/10.3390/ijgi9100586 https://www.mdpi.com/2220-9964/9/10/586, number: 10 Publisher: Multidisciplinary Digital Publishing Institute.
Kang, Y., Zhang, F., Gao, S., Lin, H., Liu, Y., 2020. A review of urban physical environment sensing using street view imagery in public health studies. Annals of GIS 26 (3), 261–275. https://doi.org/10.1080/19475683.2020.1791954. ISSN 1947-5683, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/19475683.2020.1791954.
Ke, Y., Quackenbush, L.J., 2011. A review of methods for automatic individual tree-crown detection and delineation from passive remote sensing. Int. J. Remote Sens. 32 (17), 4725–4747. https://doi.org/10.1080/01431161.2010.494184. ISSN 0143-1161.
Kelly, M., Guo, Q., Liu, D., Shaari, D., 2007. Modeling the risk for a new invasive forest disease in the United States: An evaluation of five environmental niche models. Comput. Environ. Urban Syst. 31 (6), 689–710. https://doi.org/10.1016/j.compenvurbsys.2006.10.002. ISSN 0198-9715, http://www.sciencedirect.com/science/article/pii/S0198971506000913.
Kisantal, M., Wojna, Z., Murawski, J., Naruniec, J., Cho, K., 2019. Augmentation for small object detection, arXiv:1902.07296 [cs] http://arxiv.org/abs/1902.07296, arXiv: 1902.07296.
Krylov, V.A., Kenny, E., Dahyot, R., 2018. Automatic Discovery and Geotagging of Objects from Street View Imagery. Remote Sens. 10 (5), 661. https://doi.org/10.3390/rs10050661 https://www.mdpi.com/2072-4292/10/5/661.
Labelbox Inc., Labelbox: The best way to create and manage training data, https://labelbox.com/, ????.
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. Photogram. Remote Sens., 162 125–136. https://doi.org/10.1016/j.isprsjprs.2020.02.001. ISSN 0924-2716, URL http://www.sciencedirect.com/science/article/pii/S0924271620300356.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553) (2015) 436–444. ISSN 0028-0836, 1476-4687, doi:10.1038/nature14539, http://www.nature.com/articles/nature14539.
Lefèvre, S., Tuia, D., Wegner, J.D., Produit, T., Nassar, A.S., 2017. Toward Seamless Multiview Scene Analysis From Satellite to Street Level. Proc. IEEE 105 (10), 1884–1899. https://doi.org/10.1109/JPROC.2017.2684300. ISSN 0018-9219.
Li, X., Zhang, C., Li, W., Ricard, R., Meng, Q., Zhang, W., 2015. Assessing street-level urban greenery using Google Street View and a modified green view index. Urban Forest. Urban Green. 14 (3), 675–685. https://doi.org/10.1016/j.ufug.2015.06.006. ISSN 1618-8667, http://www.sciencedirect.com/science/article/pii/S1618866715000874.
Li, S., Dragicevic, S., Castro, F.A., Sester, M., Winter, S., Coltekin, A., Pettit, C., Jiang, B., Haworth, J., Stein, A., Cheng, T., 2016. Geospatial big data handling theory and methods: A review and research challenges. ISPRS J. Photogram. Remote Sens. 115, 119–133. https://doi.org/10.1016/j.isprsjprs.2015.10.012. ISSN 0924-2716, http://www.sciencedirect.com/science/article/pii/S0924271615002439.
Li, X., Ratti, C., Seiferling, I., 2017. Mapping Urban Landscapes Along Streets Using Google Street View. In: M.P. Peterson (Ed.), Advances in Cartography and GIScience, Lecture Notes in Geoinformation and Cartography, Springer International Publishing, 341–356, ISBN 978-3-319-57336-6, 2017.
Li, X., Ratti, C., Seiferling, I., 2018. Quantifying the shade provision of street trees in urban landscape: A case study in Boston. USA, using Google Street View, Landscape and Urban Planning 169, 81–91. https://doi.org/10.1016/j.landurbplan.2017.08.011. ISSN 0169-2046, http://www.sciencedirect.com/science/article/pii/S0169204617301950.
Li, X., Chen, W.Y., Sanesi, G., Lafortezza, R., 2019. Remote Sensing in Urban Forestry: Recent Applications and Future Directions. Remote Sens. 11 (10), 1144. https://doi.org/10.3390/rs11101144 https://www.mdpi.com/2072-4292/11/10/1144.
Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature Pyramid Networks for Object Detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Honolulu, HI, 936–944, ISBN 978-1-5386-0457-1, 2017, doi:10.1109/CVPR.2017.106, http://ieeexplore.ieee.org/document/8099589/.
Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., Johnson, B.A., 2019. Deep learning in remote sensing applications: A meta-analysis and review. ISPRS J. Photogram. Remote Sens. 152, 166–177. https://doi.org/10.1016/j.isprsjprs.2019.04.015. ISSN 0924-2716, http://www.sciencedirect.com/science/article/pii/S0924271619301108.
Michels, J., Saxena, A., Ng, A.Y., 2005. High Speed Obstacle Avoidance Using Monocular Vision and Reinforcement Learning. In: Proceedings of the 22Nd International Conference on Machine Learning, ICML '05, ACM, New York, NY, USA, 593–600, ISBN 978-1-59593-180-1, 2005, doi:10.1145/1102351.1102426, http://doi.acm.org/10.1145/1102351.1102426, event-place: Bonn, Germany.
Nielsen, A.B., Östberg, J., Delshammar, T., 2014. Review of Urban Tree Inventory Methods Used to Collect Data at Single-Tree Level (2014) 17.
Nitoslawski, S., Duinker, P., 2016. Managing Tree Diversity: A Comparison of Suburban Development in Two Canadian Cities, Forests 7, doi:10.3390/f7060119.
Nowak, D.J., Hirabayashi, S., Bodine, A., Greenfield, E., 2014. Tree and forest effects on air quality and human health in the United States. Environ. Pollut. 193, 119–129. https://doi.org/10.1016/j.envpol.2014.05.028. ISSN 0269-7491, URL http://www.sciencedirect.com/science/article/pii/S0269749114002395.
Padayachee, A.L., Irlich, U.M., Faulkner, K.T., Gaertner, M., Proches, E., Wilson, J.R.U., Rouget, M., 2017. How do invasive species travel to and through urban environments? Biol. Invasions 19 (12), 3557–3570. https://doi.org/10.1007/s10530-017-1596-9. ISSN 1387-3547, 1573-1464, https://link.springer.com/article/10.1007/s10530-017-1596-9.
Pan, S.J., Yang, Q., 2010. A Survey on Transfer Learning. IEEE Trans. Knowl. Data Eng. 22 (10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191. ISSN 1041-4347, http://ieeexplore.ieee.org/document/5288526/.
Plowright, A.A., Coops, N.C., Eskelson, B.N.I., Sheppard, S.R.J., Aven, N.W., 2016. Assessing urban tree condition using airborne light detection and ranging. Urban Forest. Urban Green. 19, 140–150. https://doi.org/10.1016/j.ufug.2016.06.026. ISSN 1618-8667, http://www.sciencedirect.com/science/article/pii/S1618866715300649.
Plowright, A.A., Coops, N.C., Chance, C.M., Sheppard, S.R.J., Aven, N.W., 2017. Multi-scale analysis of relationship between imperviousness and urban tree height using airborne remote sensing. Remote Sens. Environ. 194, 391–400. https://doi.org/10.1016/j.rse.2017.03.045. ISSN 0034-4257, http://www.sciencedirect.com/science/article/pii/S0034425717301487.
Schmidhuber, J., 2015. Deep Learning in Neural Networks: An Overview. Neural Networks 61, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003. ISSN 08936080, http://arxiv.org/abs/1404.7828, arXiv: 1404.7828.
Seiferling, I., Naik, N., Ratti, C., Proulx, R., 2017. Green streets - Quantifying and mapping urban trees with street-level imagery and computer vision. Landscape Urban Plan. 165, 93–101. https://doi.org/10.1016/j.landurbplan.2017.05.010. ISSN 0169-2046, http://www.sciencedirect.com/science/article/pii/S0169204617301147.
Small, C., 2001. Estimation of urban vegetation abundance by spectral mixture analysis. Int. J. Remote Sens. 22 (7), 1305–1334. https://doi.org/10.1080/01431160151144369. ISSN 0143-1161.
Steele, F., 2016. Urban Forest Climate Adaptation Framework for Metro Vancouver, Tech. Rep., Metro Vancouver, 2016.
Stewart, I., Oke, T., 2010. Thermal differentiation of local climate zones using temperature observations from urban and rural field sites. In: Urban Climate, Keystone, Colorado, 7, 2010.
Stubbings, P., Peskett, J., Rowe, F., Arribas-Bel, D., 2019. A Hierarchical Urban Forest Index Using Street-Level Imagery and Deep Learning. Remote Sens. 11 (12), 1395. https://doi.org/10.3390/rs11121395 https://www.mdpi.com/2072-4292/11/12/1395.
Sweden, M.A. Map data at scale from street-level imagery, https://www.mapillary.com/, ????.
Tippetts, B., Lee, D.J., Lillywhite, K., Archibald, J., 2016. Review of stereo vision algorithms and their suitability for resource-limited systems. J. Real-Time Image Proc. 11 (1), 5–25. https://doi.org/10.1007/s11554-012-0313-2. ISSN 1861-8219.
van den Bosch, M., 2017. Ode Sang, Urban natural environments as nature-based solutions for improved public health – A systematic review of reviews. Environ. Res. 158, 373–384. https://doi.org/10.1016/j.envres.2017.05.040. ISSN 0013-9351, http://www.sciencedirect.com/science/article/pii/S0013935117310241.
Wegner, J.D., Branson, S., Hall, D., Schindler, K., Perona, P., 2016. Cataloging Public Objects Using Aerial and Street-Level Images - Urban Trees, in. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6014–6023

https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Wegner_Cataloging_Public_Objects_CVPR_2016_paper.html.

Yin, D., Wang, L., 2016. How to assess the accuracy of the individual tree-based forest inventory derived from remotely sensed data: a review. Int. J. Remote Sens. 37 (19), 4521–4553. https://doi.org/10.1080/01431161.2016.1214302. ISSN 0143-1161.

Zhang, X., Xia, G.-S., Lu, Q., Shen, W., Zhang, L., 2018. Visual object tracking by correlation filters and online learning. ISPRS J. Photogram. Remote Sens. 140, 77–89. https://doi.org/10.1016/j.isprsjprs.2017.07.009. ISSN 0924-2716, http://www.sciencedirect.com/science/article/pii/S0924271617301715.

Zhen, Z., Quackenbush, L.J., Zhang, L., 2016. Trends in Automatic Individual Tree Crown Detection and Delineation-Evolution of LiDAR Data. Remote Sens. 8 (4), 333. https://doi.org/10.3390/rs8040333 https://www.mdpi.com/2072-4292/8/4/333.