Computers, Environment and Urban Systems 101 (2023) 101950

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

Mapping the walk: A scalable computer vision approach for generating sidewalk network datasets from aerial imagery

Maryam Hosseini [a,*], Andres Sevtsuk [a], Fabio Miranda [d], Roberto M. Cesar Jr [e], Claudio T. Silva [b,c]

[a] Department of Urban Studies and Planning, Massachusetts Institute of Technology (MIT), MA, USA
[b] Department of Computer Science and Engineering, New York University, NY, USA
[c] Center for Data Science, New York University, NY, USA
[d] Department of Computer Science, University of Illinois at Chicago (UIC), IL, USA
[e] University of São Paulo (USP), SP, Brazil

ARTICLE INFO

Keywords:
Pedestrian network extraction
Semantic segmentation
Automated network generation
Pedestrian infrastructure

ABSTRACT

While cities around the world are increasingly promoting streets and public spaces that prioritize pedestrians over vehicles, significant data gaps have made pedestrian mapping, analysis, and modeling challenging to carry out. Most cities, even in industrialized economies, still lack information about the location and connectivity of their sidewalks, making it difficult to implement research on pedestrian infrastructure and holding the technology industry back from developing accurate, location-based Apps for pedestrians, wheelchair users, street vendors, and other sidewalk users. To address this gap, we have designed and implemented an end-to-end open-source tool— Tile2Net —for extracting sidewalk, crosswalk, and footpath polygons from orthorectified aerial imagery using semantic segmentation. The segmentation model, trained on aerial imagery from Cambridge, MA, Washington DC, and New York City, offers the first open-source scene classification model for pedestrian infrastructure from sub-meter resolution aerial tiles, which can be used to generate planimetric sidewalk data in North American cities. Tile2Net also generates pedestrian networks from the resulting polygons, which can be used to prepare datasets for pedestrian routing applications. The work offers a low-cost and scalable data collection methodology for systematically generating sidewalk network datasets, where orthorectified aerial imagery is available, contributing to over-due efforts to equalize data opportunities for pedestrians, particularly in cities that lack the resources necessary to collect such data using more conventional methods.

1. Introduction

After a century of car-oriented urban growth (Walker & Johnson, 2016), cities around the world are implementing policies and plans that aim to make their neighborhoods and streets more walkable and transit oriented. Renewed attention to walkability is driven simultaneously by the impending climate crisis, public health concerns, and inter-city economic competition. With more than a third of all CO_2 emissions attributable to the transport sector (EPA, 2021), it has become clear that climate goals will not be reached unless urban populations start driving less and relying more on walking and public transportation (Cervero, 1998; Speck, 2013). From a health perspective, more walkable cities have been found to have lower obesity and inactivity-related conditions, respiratory diseases, and lower overall public health expenditures (Frank & Engelke, 2001; Grasser, Van Dyck, Titze, & Stronegger, 2013; Zapata-Diomedi et al., 2019). Economically, walkable and transit-served city environments have also become an important draw for a competitive workforce (Glaeser, 2010; Moretti, 2012) and now command some of the highest-priced real estates in American cities (Leinberger & Lynch, 2014).

Despite the growing, multi-pronged importance of pedestrian-oriented city design, the necessary geospatial data for pedestrian infrastructure mapping and modeling remains far behind vehicular infrastructure data. Digital mapping of vehicular road networks expanded rapidly in the 1990s, led by Federal legislation (President Clinton 1994), municipal governments' investments, as well as private companies such

* Corresponding author at: MIT City Form Lab, Department of Urban Sudies and Planning, 77 Massachusetts Ave, Suite 10-402, Cambridge, MA 02139, USA.
E-mail addresses: maryamh@mit.edu (M. Hosseini), asevtsuk@mit.edu (A. Sevtsuk), fabiom@uic.edu (F. Miranda), rmcesar@usp.br (R.M. Cesar), csilva@nyu.edu (C.T. Silva).

https://doi.org/10.1016/j.compenvurbsys.2023.101950
Received 18 April 2022; Received in revised form 20 January 2023; Accepted 10 February 2023
Available online 22 February 2023

as Navteq and TomTom that operationalized roadway mapping in cities across the world. Assembly and wide-scale dissemination of such data has been instrumental to numerous technologies that use road network data as a key input: mapping and routing applications (e.g., Google Maps, TransitApp), transportation service technologies (e.g., Uber, Amazon), urban transportation models and policies (e.g., metropolitan and urban Travel Demand Models, congestion charging systems in cities including London, Singapore, and Stockholm), data specification standards (e.g., Google's General Transit Feed Specification, and the City of Los Angeles' Mobility Data Specification).

Transportation debates are often skewed towards topics rich in data – vehicle throughput, for instance, which is monitored on individual streets in many cities, is a key parameter for new road design and investment. Not only is comparable data describing pedestrian throughput on sidewalks typically unknown, the locations and types of sidewalks are also rarely mapped, contributing to systemic underinvestment in the pedestrian realm. When pedestrian accessibility is analyzed, it is often done using simplified road-centerline data, not the actual sidewalks, footpaths, and road crossings (Liu et al., 2021).

A number of studies have highlighted the inadequacy of using street-centerline networks for pedestrian routing (Cambra, Gonçalves, & Moura, 2019; Sun, Su, Ren, & Guan, 2019), which can lead to inaccuracies (e.g., streets with no sidewalks), simplifications (e.g., assumptions that buildings can be directly accessed on both sides of a street centerline, while in reality crossing a street is only allowed at certain locations), and misrepresentation (e.g., assuming pedestrian connections based on vehicular routes, where there are none (Ellis et al., 2016)). For instance, (Chin, Van Niel, Giles-Corti, & Knuiman, 2008), who examined pedestrian access in Perth, Australia, found up to 120% difference in pedestrian connectivity using road centerlines as opposed to sidewalk centerlines. Not only can road-network data be imprecise for pedestrian needs, it can also be hazardous for the more vulnerable street users, such as vision-, hearing- or mobility-challenged travelers, wheelchair-bound travelers, the elderly, and the young (Saha et al., 2019; Zhang & Zhang, 2019). Lack of accurate sidewalk routing data threatens their independence and decreases their quality of life (Cohen & Dalyot, 2021; Delboni Lomba & Godoy da Silva, 2022; El-Taher, Taha, Courtney, & Mckeever, 2021).

To address these challenges, we introduce Tile2Net–a new open-source tool for automated mapping of pedestrian infrastructure using aerial imagery. Tile2Net enables users to download orthorectified sub-meter resolution image tiles for a given region from public sources, which are used to generate topologically georeferenced sidewalk, crosswalk, and footpath polygons as well as their interconnected centerlines. By using available official network and polygon data as a ground truth, we would like to investigate to what extent the automatically generated networks using computer vision models can produce accurate results. Our goal is to map pedestrian networks "as they are" rather than trying to improve the network connectivity artificially. To achieve this, we train and implement a semantic segmentation model that can detect these pedestrian infrastructure elements from orthorectified aerial tiles. We pilot test the approach in Manhattan, NY, Washington, DC, Boston, and Cambridge, MA, and report the accuracy measures in each of these cities. This work is as an important step towards a robust and open-source framework that enables comprehensive digitization of pedestrian infrastructure, which we argue to be a key missing link to more accurate and reliable pedestrian modeling and analyses. By offering low-cost solutions to create planimetric dataset describing pedestrian environment we enable less resourceful cities to create datasets describing pedestrian environment which otherwise would not be possible at a comparable cost and time.

Our key contributions are:

1. We designed and implemented Tile2Net as an end-to-end, open-source tool for creating large-scale pedestrian networks from orthorectified aerial imagery. https://github.com/VIDA-NYU/tile2net.
2. We calibrated a high-performing scene classification model for detecting sidewalks, crosswalks, and footpaths. We have custom trained Tile2Net on around 20,000 detailed images (where each 515 $\times$ 512 pixel image covers a roughly 9500 square-meter or 2.35 acre area) in Cambridge, MA, New York City, NY, and Washington, DC, where detailed GIS data of pedestrian infrastructure was available. The results of this training process are available through Tile2Net, allowing the tool to be applied to other cities (where no prior training was performed) to automatically detect and map pedestrian infrastructure. Our GitHub repository includes, to the best of our knowledge, the first publicly available scene classification model for detecting sidewalks, crosswalks, and footpaths from orthorectified aerial tiles.
3. Our solution is adjustable to additional training in different city environments, offering various settings to finetune the model on new data based on local environmental characteristics. Tile2Net automates the creation of labels using GIS datasets that are needed for retraining the model for different urban conditions.

The paper is organized as follows: In Section 2, we review existing literature on sidewalk mapping. In Section 3, we describe our methodology, data sources, and the Tile2Net functionalities. Section 4 presents our results of applying the model in three East Coast cities. Section 5 discusses the challenges of automated sidewalk network detection and suggests directions for expanding the work in the future.

2. Literature review

2.1. Map generation

At least five different methods for mapping sidewalk infrastructure can be distinguished in existing literature and practice, with additional combinations thereof (Fig. 1).

First, physical site surveys and manual aerial image surveys have been used in a number of cities to develop datasets on pedestrian facilities (e.g., in Melbourne, Singapore, and Boston). This involves human tracing of observable sidewalks and crosswalks from georeferenced aerial imagery, combined with on-the-ground observation and validation (Proulx, Zhang, & Grembek, 2015). Such mapping efforts can produce accurate and high-quality results, but they can also be prohibitively labor-intensive and difficult to scale across large regions. In a recent study, 6400 intersections in San Francisco were manually reviewed and classified based on the crosswalk presence and condition, which took 90 h for a researcher to complete (Moran, 2022). Some cities have relied on crowd-sourcing sidewalk mapping to a community of online users (Sachs, 2016). Custom-built mapping platforms, such as OpenSidewalks (TCAT, 2016), WalkScope (Placematters and WalkDenver, 2014), or global open-access platforms like OpenStreetMap, enable users to view and edit available datasets collectively.

Second, network buffering uses a geospatial road centerline network as a reference, which is offset on both sides to generate polygons whose boundaries approximate the right-of-way of the roadway. Boundaries of the resulting polygons are considered as approximate locations of sidewalk segments, assuming that (1) pedestrian path segments only exist along roads, (2) sidewalks exist along both sides of selected roads, and (3) crosswalks are located at every intersection. Buffer distances can include road right-of-way or road-width dimensions from the vehicular road centerline network dataset.

Third, pedestrian pathways have also been identified from Global Positioning System (GPS) trajectories of pedestrian movement. This can include data from GPS tracking devices that are handed out to consenting participants or collected from their smartphone tracking Apps (Cottrill et al., 2013). Third-party data aggregators, such as StreetLightData and Cuebiq collect GPS trace data from hundreds of different

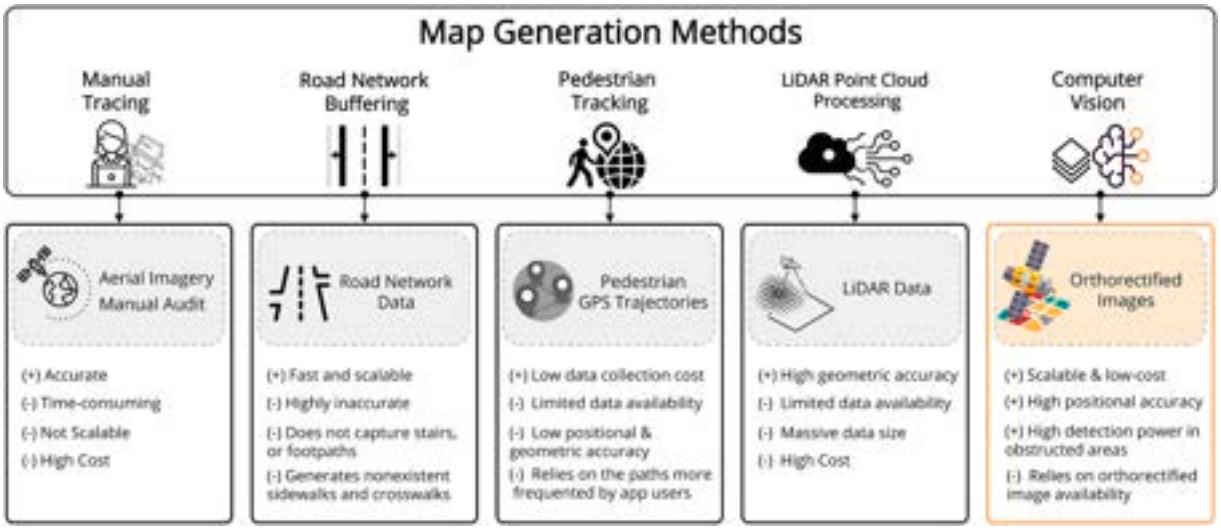

Fig. 1. Different methods of map generation. Each box presents the main data sources (shaded parts), as well as the strengths (+) and weaknesses (−) of each method. The last box highlighted in orange denotes the method used in this paper.

Apps that track their users' location history. Once collected, GPS traces can be merged, simplified, and joined into contiguous network datasets (Kasemsuppakorn & Karimi, 2013). The results can effectively illustrate where people (or at least App users) actually walked, but they may ignore segments not frequented by smartphone or App users (Yang et al., 2020). Moreover, the accuracy of the final network relies heavily on the positional accuracy of the GPS trajectories, which can be noisy, specifically in the vicinity of high-rise buildings (Karimi & Kasemsuppakorn, 2013).

The fourth category uses airborne Light Detection and Ranging (LiDAR) point cloud data. LiDAR devices use active sensing and can be mounted on mobile objects such as planes and drones (). In general, three main methods have been used for processing LiDAR point cloud data to extract road and sidewalk features: 1) geometry-based methods, which use prior knowledge of unique geometrical shapes and measurements of urban ground elements, 2) reflectance-based methods, which utilize the reflectance intensity of different object classes to classify them, 3) scan-based methods, which take advantage of the scanning pattern to connect results from consecutive scans into a continuous boundary to refine object segmentation (Ai & Tsai, 2016; Ai & Hou, 2019; Balado, Daz-Vilariño, Arias, & González-Jorge, 2018). The resulting data represent sidewalks as vector lines or polygons that can be both accurate and scalable (Horváth, Pozna, & Unger, 2022; Treccani, Díaz-Vilariño, & Adami, 2021). However, the limited availability of spatially dense and open-access LiDAR data has constrained this approach to relatively few cities overall.

Fifth, and in line with our work, computer vision techniques have recently been deployed in a limited number of studies to detect pedestrian infrastructure from aerial or satellite images (Luo, Wu, Wei, Boriboonsomsin, & Barth, 2019; Ning, Ye, Chen, Liu, & Cao, 2022). Among computer vision techniques, semantic segmentation has been shown to result in reasonably accurate detection and localization of infrastructure elements. This method makes dense predictions, inferring labels for each pixel of an image, hence, giving each one a semantic meaning (Ess, Mueller, Grabner, & Van Gool, 2009; Geiger, Lenz, & Urtasun, 2012). Although semantic segmentation has been broadly used to detect roads and building footprints from aerial or satellite images (Balali, Rad, & Golparvar-Fard, 2015; Iglovikov, Mushinskiy, & Osin, 2017; Li et al., 2019) and to create road networks (Bastani et al., 2018; Etten, 2020; Wei, Zhang, & Ji, 2019), it has not been widely implemented for sidewalk and crosswalk mapping so far, possibly due to technical challenges and costs involved in training robust models. Existing examples of sidewalk and crosswalk detection models using aerial or satellite images suffer from relatively low prediction accuracy. For instance, Luttrell IV (2022) experimented with an aerial image-only semantic segmentation model to detect crosswalks and achieved a 55.59% accuracy and a 15.41% recall rate[1] The results substantially improved when street-level images were incorporated. In another study, Ning et al. (2022) implemented a segmentation model, which predicted sidewalks with a 55% precision and a 74% recall rate. Here again, the results improved by using street-level images, limiting its applicability to only regions of the world where such data is available.

To train semantic segmentation models, densely annotated labels are needed, which are often labor-intensive and costly to prepare. Consequently, in applying semantic segmentation models to urban context (Kim, Lee, Hipp, & Ki, 2021; Wang et al., 2019; Zhang, Zhang, Liu, & Lin, 2018; Zhou et al., 2021), researchers often forego retraining or fine-tuning their models on target datasets and rather rely on publicly-available pre-trained datasets such as CityScapes (Cordts et al., 2016), Mapillary (Neuhold, Ollmann, Rota Bulo, & Kontschieder, 2017), and ADE20K (Zhou et al., 2017). Relying on pre-trained models, not specific to the task, limits analysis to the classes included in those datasets (Ahn & Kwak, 2018). Further, pre-trained models not fine-tuned on domain-specific data can yield sub-optimal performance (Azizi et al., 2021). Compared to roads and buildings, detecting sidewalks, footpaths and crosswalks is more challenging since they constitute a relatively small portion of the visual field, and their detection can be further inhibited by occlusion from shadow, vegetation, and structures (Hosseini et al., 2021). Hence, choosing the right network architecture that can preserve local details while accounting for global image context is crucial.

2.2. Semantic segmentation

The feature detection mechanism we use in Tile2Net relies on semantic segmentation. Research on automated vehicles has created significant demand for fast and efficient algorithms that can extract both high and low-level information from urban scenes, leading to notable improvements in the field of scene parsing, specifically pixel-wise classification, commonly referred to as semantic segmentation. Early work incorporated multi-resolution processing into segmentation

[1] Precision measures the proportion of positive identifications that was actually correct (True Positive/True Positive+False Positive), recall refers to the percentage of total relevant results correctly classified by the algorithm (True Positive/True Positive +False Negative).

architectures to improve performance over a static resolution approach (Zhao, Shi, Qi, Wang, & Jia, 2017). This has been followed by rapid developments in multi-scale pyramid-style networks (He, Deng, & Qiao, 2019; Ding, Jiang, Shuai, Liu, & Wang, 2018; He, Deng, Zhou, Wang, & Qiao, 2019). In particular, HRNet (Sun et al., 2019; Wang et al., 2020) connects high-to-low resolution convolutions via parallel and repeated multi-scale fusion to better preserve low-resolution representations alongside high-resolution ones in comparison to previous work (Chen et al., 2018; Newell, Yang, & Deng, 2016; Yu, Wang, Shelhamer, & Darrell, 2018). A variant of HRNet, HRNet-W48, which has shown superior performance across segmentation benchmarks such as Cityscapes (Cordts et al., 2016) and Mapillary Vista (Sun, Xiao, Liu, & Wang, 2019), is used as a key component of our segmentation framework below.

Attention-based mechanisms have been adopted in multiple semantic segmentation architectures (Chen, Yang, Wang, Xu, & Yuille, 2016; Fu et al., 2019; Huang et al., 2017; Li, Xiong, An, & Wang, 2018). Instead of feeding multiple resized images into a shared network and merging the features to make predictions, which can lead to suboptimal results, the attention mechanism learns to assign different weights to multi-scale features at a pixel-level and uses the weighted sum of score-maps across all scales for the final prediction (Chen et al., 2016). Huang et al. (2017) proposed a reversed attention mechanism that trains the model on features that are not associated with the target class. The network has three branches that simultaneously perform direct, reverse, and reversed-attention learning. Hierarchical multi-scale attention is thus a network architecture that learns to assign relative weights between adjacent scales (Tao, Sapra, & Catanzaro, 2020). This method has shown to be more memory efficient and can lead to more accurate results; we have therefore integrated it as part of our network generation pipeline.

3. Tile2Net

Tile2Net is an end-to-end open-source Python tool that downloads and combines orthorectified tiles from publicly available data sources, detects street elements from these tiles, creates sidewalk, crosswalk, and footpath polygons, and ultimately generates pedestrian networks. We chose Python because of its popularity among data analysts and urban scientists, with a myriad of popular packages that can be used in conjunction with Tile2Net for richer network analytics, including OSMnx (Boeing, 2017), NetworkX (Hagberg & Conway, 2020), and Geopandas (Jordahl, 2014). Tile2Net's functionalities are exposed through an easy-to-use API that can be used in interactive environments, such as Jupyter Notebooks.

Tile2Net is designed to work with slippy map tiles, a system that uses Web Mercator coordinates and constructs a map from 256*x*256–pixel square tiles, referenced by the tile coordinates and a zoom level. At successively higher zoom levels, the number of tiles increases by a factor of four. The tool then follows this system and works at grid and tile levels–i.e., for a region of interest, it defines a slippy map-based grid of tiles. The user can initialize this process in two ways: specifying an address (e.g., Washington Square park, Manhattan, NYC, USA) that then is geocoded using the Nominatim API, or passing the top-left and bottom-right coordinates of the bounding box of the region. Tile2Net will create the tile grid and provide a number of functionalities for users, such as downloading the orthoimagery tiles that fall within its bounding box or merging tiles to create larger ones. Then, Tile2Net will use the trained model (detailed in 3.1) to detect roads, sidewalks, crosswalks, and footpaths in each tile and create geo-referenced vector data (polygons and networks) from segmentation results, which are initially in raster format.

We train the detection model on thousands of orthorectified aerial tiles from Cambridge, MA, New York City, NY, and Washington, DC, which allows the tool to be used for extracting such data in the North American context, or other cities with similar urban fabrics, without needing any further training. However, Tile2Net also allows users to retrain the feature detection model in new contexts, where pedestrian infrastructure may visibly differ from our initial cities. For users interested in modifying the model or further training it, Tile2Net offers the capability to automatically create labels (given authoritative data). This can substantially reduce a common bottleneck–preparing thousands of labels manually. Retraining can be initialized with our trained weights, which can lead to significant time and cost savings.

Fig. 2 illustrates the Tile2Net data processing pipeline. We combine a semantic segmentation approach with a raster-to-polygon conversion process to generate polygon shapefiles of pedestrian infrastructure elements and, separately, a polygon-to-centerline conversion process to produce a topologically interconnected network of pedestrian centerlines. In the following, we start by describing the data we used and the procedures we chose to detect the features of interest in our training procedures (Section 3.1). Section 3.2 describes implementation and training details, and Section 3.3 presents the training results. Section 3.4 describes how the trained model can be used.

3.1. Detecting sidewalks, footpaths, and crosswalks from aerial imagery

Our semantic segmentation model takes an input image, makes dense predictions inferring labels for each pixel, and outputs a feature map showing whether and where the objects of interest are recognized in the image tile. For this task, we adopted the Hierarchical Multi-Scale Attention model (Tao et al., 2020). The idea behind multi-scale architecture is to combine the predictions from multiple scales of the input image. Fine details (e.g., narrow footpaths, poles in the background, etc.) can be best detected in higher zoom levels or larger images (2× scale for instance), and large objects with less details (e.g., roads) are best detected at a lower zoom level (0.5 scale for instance). The model learns which image scale works best for different objects and uses that scale to make the prediction. The hierarchical architecture of our semantic segmentation network makes it possible to choose different scales during the inference. In our experiments, using 512 × 512, 1024 × 1024, and 2048 × 2048 pixel tiles, the best results were achieved using 1024 × 1024 pixel tiles, where the model had enough context to distinguish between different classes. Images should be in zoom levels where sidewalks can be visible; for instance, sidewalks are not visible from 3-m/pixel images. In general, the model can work with resolutions 17 to 23.

We used HRNet-W48 (Sun, Zhao, et al., 2019; Wang et al., 2020) with Object-Contextual Representations (Yuan, Chen, & Wang, 2019) as the backbone, since HRNet maintains twice as high a resolution representation as other popular backbones such as WiderResnet38 (Wu, Shen, & Van Den Hengel, 2019). The computed representation from HRNet-W48 is fed into the OCR module, which computes the weighted aggregation of all the object region representations to augment the representation of each pixel. The augmented representations are the input for the attention model. For the primary loss function, we used Region Mutual Information (RMI) loss (Zhao, Wang, Yang, & Cai, 2019), which accounts for the relationship between pixels instead of only relying on single pixels to calculate the loss.

3.1.1. Training data description

The semantic segmentation model requires a set of aerial images and their corresponding labels to be trained. Two main data sources were used to create our training set: 1) high-resolution orthorectified imagery that is available across numerous U.S. (US Geological Survey, 2018) and international cities, and 2) planimetric GIS data representing the same elements as seen in orthorectified images. Table 1 shows the datasets obtained from Cambridge, MA, Washington, DC, and New York City, NY we used to train the model, including which class of feature was used (e.g. road polygons, sidewalk polygons, etc.) and their dates.

High-resolution orthorectified imagery. A key input to detecting pedestrian infrastructure elements in our pipeline is sub-meter resolution orthorectified imagery. Raw aerial images inherently contain distortions caused by sensor orientation, systematic sensor and platform-

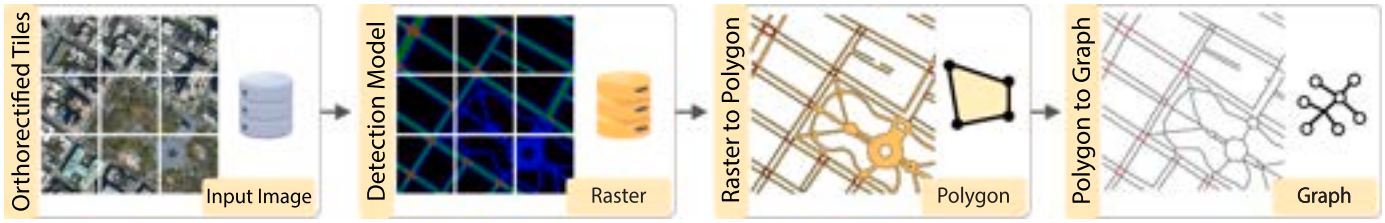

Fig. 2. The proposed network generation pipeline. a) Unlabeled orthorectified tiles are passed through the semantic segmentation model for prediction, b) The model detected sidewalks (blue), crosswalks (red), and roads (green) in the input tiles, c) The sidewalks and crosswalks of the prediction results (raster format) are converted into georeferenced polygons, d) The line representation of the pedestrian network generated from polygons. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

Table 1
Datasets used for training the model and their sources.

City	Dataset	Features	Date	Source
Cambridge, MA	Sidewalks	Sidewalk polygons	2018	(Cambridge GIS, 2018a)
	Roads	Roads polygons	2018	(Cambridge GIS, 2018d)
	Pavement Markings	Crosswalk polygons	2018	(Cambridge GIS, 2018b)
	Public Footpaths	Paved & unpaved	2018	(Cambridge GIS, 2018c)
	Ortho-imagery	Orthorectified imagery	2018	(MassGIS, 2018)
Manhattan and Brooklyn, NY	Sidewalk Inventory	Off-road footpaths inside parks	2018	(NYC DoITT, 2018)
	Roads	Road polygons	2018	(NYC DoITT, 2018)
	Ortho-imagery	Orthorectified imagery	2018	(NYC GIS, 2018)
Washington, DC	Sidewalk Inventory	Sidewalk and crosswalk polygons	2019	(DC GIS, 2019b)
	Road	Road polygons	2019	(DC GIS, 2019a)
	Ortho-imagery	Orthorectified imagery	2020	(DC GIS, 2020)

related geometry errors, terrain relief, and curvature of the earth. Such distortions cause feature displacement and scaling errors, which can result in inaccurate measurement of distance, angles, areas, and positions, making raw images unsuitable for feature extraction and mapping purposes. Orthorectification removes these distortions and creates accurately georeferenced images with a uniform scale and consistent geometry (Tucker, Grant, & Dykstra, 2004; Zhou, Chen, Kelmelis, & Zhang, 2005). The orthoimagery tile system also makes it possible to convert between positional coordinates of tiles in x/y/z (where z represents the zoom level) and geographical coordinates.

High resolution orthorectified images are becoming increasingly accessible. In the United States, U.S. Geological Survey (USGS) (US Geological Survey, 2018) provides high-resolution orthorectified across almost the whole country. Many other countries across Europe, Asia, and the Global South also acquire high resolution orthorectified images and make them publicly available. Moreover, various commercial companies such as MAXAR and Planet Scope sell orthorectified image data. Additionally, there are some state-wide programs dedicated to producing digital ortho-imagery on different zoom levels, which may offer more recent data. For the purposes of this study, we used orthorectified images provided by Massachusetts (MassGIS, 2018), Washington, DC (DC GIS, 2020), and New York (NYC GIS, 2018) to train the model and pilot test the approach. We obtained 11,000 tiles from Washington, DC, 28,000 tiles from Cambridge, MA and 8000 tiles from inside NYC parks. Except for Washington, DC, where the tiles are 512*x*512-pixels, the rest of the tiles come in 256*x*256-pixels. We choose zoom level 20 for the 256*x*256-pixel tiles, where each pixel of the image represents 0.19 m on the surface of the earth. Our experiments training the model with both sizes showed that the model would perform better using 512*x*512-pixel input images (an increase of roughly 12% in mIoU). Hence, we used the tool to stitch every four neighboring 256*x*265-pixel tiles to get 512*x*512-pixel images, creating a total of 20,000 tiles.

Planimetric GIS data. Many GIS datasets have been created using planimetric mapping. Planimetric mapping involves extracting features from orthoimagery to create maps that only capture the horizontal distance between the features irrespective of elevation (Quackenbush, 2004). Since planimetric data are created using orthorectified images, they are also suitable for creating labels for semantic segmentation models –a priori known and accurate raster polygons that describe the features we seek to detect automatically. An annotated image is a reference image where each pixel value describes the label to which the pixel in the aerial image belongs (Fig. 3(b,c,e,d)).

To prepare labels, Tile2Net primarily relies on available GIS data on sidewalk, crosswalk, and footpath locations in select city environments. In this study, we used the publicly available planimetric data on sidewalks, footpaths, and crosswalks in parts of Cambridge, Washington, DC, and selected sites from inside the parks of New York City (Table 1). Reliance on existing GIS datasets allows us to prepare large-scale labels using GIS data rather than manually annotating a huge number of images. Tile2Net takes the bounding box of each tile, finds the corresponding sidewalk, footpath, crosswalk, and road polygons from given planimetric GIS data, rasterizes the GIS polygons into pixel regions, and outputs annotated image tiles with four total classes: sidewalks (including footpaths), crosswalks, roads, and background, representing each class with a distinct color. These annotations are used as ground truth data for training the model.

However, challenges remain in creating accurate and consistent training data. The first challenge arises from the lack of consistency between the mapping standards used by different municipalities. Moreover, since GIS data on pedestrian infrastructure does not necessarily reflect the exact conditions that are represented in our aerial images, there can be a temporal difference between tiles and GIS data as the creation of GIS data may have relied on a different underlying data source. As illustrated in Fig. 4, official GIS data can contain numerous errors. Human adjustment and correction may be necessary to bring ground truth labels into alignment with the image data. To achieve that, our research team manually corrected 2500 tiles of the 12,000 training set, 1620 image tiles out of 4000 tiles that were used as our validation set, and 1500 tiles out of 4000 test set tiles.

3.2. Implementation of the detection model

The model was trained with a batch size of 16, SGD for the optimizer with polynomial learning rate (Liu, Rabinovich, & Berg, 2015), momentum 0.9, weight decay $5e^{-4}$, and an initial learning rate of 0.002. The multi-scale setting used 0.5, 1, 1.5, and 2, where a 0.5 scale denotes scaling the image down by a factor of two, and a scale of 2 denotes scaling the image up by a factor of 2 (Tao et al., 2020). We used color augmentation, random horizontal flip, random scaling (0.5×–2.0×), and Gaussian blur on the input tiles to augment the training data and

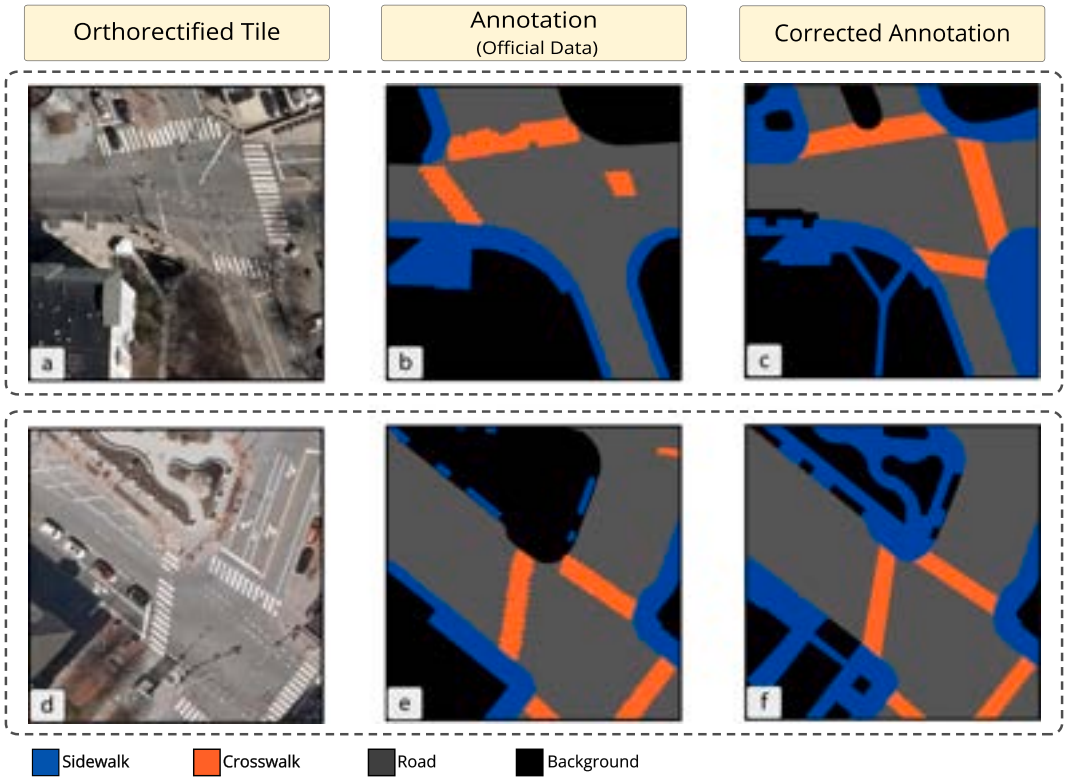

Fig. 3. Examples of the mismatches between the aerial image and the label created from the official data. The manually corrected labels are shown in the last column.

Fig. 4. Boston Commons: a) Aerial image, b) Detected sidewalk and footpath polygons (in orange) and detected crosswalks (in red), c) Fitted sidewalk, crosswalk, and footpath centerlines superimposed on the aerial image. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

improve the generalizability of the model. The crop size was set to 512*x*512. The aerial image and annotated image pairs were split into three parts: 60% of the tiles were used to train the model, 20% of the tiles to validate, and 20% were held-out to test the model in the final stage. To handle the class imbalance, we employed class uniform sampling in the data loader, which chooses equal samples for each class (Zhu et al., 2019) (classes like road and background are present in almost all images, whereas crosswalks can appear less frequently), and the class uniform percentage was set to 0.5. The segmentation model was trained for 310 epochs using 4 NVIDIA RTX8000 GPUs with 48 GB of RAM each.

3.3. Training results

The trained model outputs four classes in total, two of which were directly used to create the pedestrian networks (sidewalk including footpaths, and crosswalks), one—roads—was used to draw local attributes for finetuning the network creation parameters, and the background, which contains all other elements not used in this study. To evaluate the performance of the model, we used the Jaccard index, commonly referred to as the Intersection over Union (IoU), which is a scale-invariant standard evaluation metric for semantic segmentation tasks. IoU measures the overlap area between the prediction and the ground truth divided by the area of union between the two. It ranges between 0 and 1, with one showing the perfect overlap. Class-specific

accuracy measures are also calculated to assess the model's performance in classifying objects of different classes. We did not rely on the more biased pixel-level accuracy metrics since sidewalks and crosswalks comprise a small portion of each image, which would result in a significant class imbalance and an arbitrary high pixel-level accuracy.

Table 2 presents the average IoU (mIoU) across all classes, as well as the class-wise IoU, precision, and recall. The model achieved 84.51% mIoU over all four classes, with sidewalks having 82.67% IoU and crosswalks having 75.42% IoU. The lower accuracy of the crosswalks can be attributed to the more temporal nature of the crosswalks and the fact that they can get faded and, in some cases, not even visible to human eyes.

3.4. Using the trained model

After the training phase is completed, the unlabeled orthorectified tiles are passed through the trained model, as shown in Fig. 2(a) the prediction model outputs a raster image where each pixel has a value corresponding to one of our four classes: sidewalk, crosswalk, road, and background (Fig. 2(b)). After the pedestrian features are detected from the input images, Tile2Net takes the model's prediction in raster format and performs 1) raster to polygon conversion, which can save the output polygons in different formats such as GeoJSON and shapefiles, usable across multiple GIS tools (Fig. 2(c)); and 2) polygon to centerline conversion to create the final pedestrian network representation (Fig. 2(d)). Fig. 4 shows the results of these last two steps for Boston Commons, which was not part of the training data.

3.4.1. Raster to polygon conversion

To obtain vectorized and georeferenced polygons from the detected sidewalk, crosswalk, and road raster regions, we employed a connected-component mapping algorithm (He, Chao, Suzuki, & Wu, 2009; Rosenfeld & Pfaltz, 1966),in which the connected cells of the same category in the raster image form regions or raster polygons. These regions are then georeferenced, using an affine transformation, which preserves lines and parallelism and maps the raster pixels into the geographic coordinates.

3.4.2. Polygon to centerline conversion

In the final step Tile2Net calculates the centerlines for each polygon. Given that the initially detected regions are pixel-precise, we first simplify the polygons using the Douglas-Peucker algorithm (Douglas & Peucker, 1973). Next, a dense Voronoi diagram is computed to extract the centerlines of the sidewalk polygons (Brandt & Algazi, 1992). The centerline is constructed by linking the internal Voronoi diagram edges not intersecting with the boundary of the object as shown in Fig. 5 (see Appendix A for more details).

To clean and simplify the centerline, we trim branches shorter than an adjustable threshold. Crosswalk centerlines are created by joining the centroids of the smaller edges of the minimum rotated rectangles for each polygon. The crosswalk centerlines are then connected to their nearest sidewalk lines. The resulting vector lines form the basis of our pedestrian network.

Following this step, the network goes through algorithmic post-processing operations to correct its topology: removing false nodes and removing the isolated lines. To close the small gaps, we use R-Tree (Guttman, 1984; Kamel & Faloutsos, 1993) and query for gaps smaller than certain thresholds. Then we extrapolate both lines to meet in the center of the gap. These operations help refine the detected pedestrian centerlines into a topologically continuous network while avoiding undue corrections and additions where connections between sidewalk segments are lacking 4.

Table 2
Evaluation metrics on the test set.

Label	IoU (%)	Precision	Recall
Sidewalk	82.67	0.90	0.92
Road	86.04	0.91	0.94
Crosswalk	75.42	0.86	0.86
Background	93.94	0.97	0.96
mIoU (%)	84.51		

4. Evaluation of results

This section presents the implementation details and results of using Tile2Net to create city-scale pedestrian networks in Cambridge, MA, Boston, MA, which was not used for training at all, New York City (where only footpaths in Manhattan parks were used for training) and Washington, DC. We evaluate the accuracy of the constructed maps—both polygons and centerlines—using the available official data of such elements from the respective cities. Table 3 presents an overview of the available ground truth data used in our evaluation. The polygon data was partly used in our training process (denoted by T), as explained in Section 3.1.1. No GIS centerline data was used for training the model in any of the cities.

Fig. 6 presents the model outputs in Boston and Cambridge, Manhattan, parts of Brooklyn, and Washington, DC. All cities are shown at the same scale for comparison. For polygon comparisons, comprehensive and public data for sidewalks, crosswalks, and footpaths, was available in Cambridge, and Washington, DC. In Boston, only sidewalk GIS polygons were available, and Manhattan's sidewalk data includes the footpath polygons.

Table 4 presents class-level evaluation results for detected polygons, showing the total count and the percentage of ground-truth polygons (from the cities' GIS data) that had a matching "detected" polygon spatially intersecting each element using GIS. In Cambridge, 98.92% of all polygons in official GIS data had overlapped with polygons detected by Tile2Net. In Boston, the result was 98.72%, in Washington, DC, 84.40%, and in Manhattan, 98.25%. Since most of the unmatched polygons were small in size, we also report the area-weighted overlap percentages in Table 4.

The last row of Table 4 reports the mean aerial overlap percent between official GIS pedestrian infrastructure polygons and polygons detected by Tile2Net (also weighted by size). Analogous to IoU, this measure illustrates what percent of the area featured in the official pedestrian polygons overlaps with detected polygons. In Cambridge, 85.90% of the area of official GIS polygons was also covered by detected polygons, 77.90% in Boston, 73.80% in Washington, DC, and 87.50% in Manhattan. The lower overlap between detected and official city polygons in Boston is likely due to the fact that no tiles from Boston were used to train the model. A lower match in Washington DC primarily results from a mismatch between the city's official sidewalk polygon data layer and the imagery we used, as well as a higher proportion of tree-covered footpaths in many parts of the city.

To evaluate the accuracy of the networks extracted from the imagery, we compared them against the publicly available sidewalk, crosswalk, and footpath centerline shapefiles of each city, where available (Table 3, Table 5). All three types of pedestrian infrastructure centerlines were available in Cambridge. In Boston, the sidewalk centerline dataset includes crosswalks, and in Manhattan, only footpath centerlines were available for comparison. However, in Cambridge and Boston, centerline data dates back to 2011.To investigate the reliability of the centerline data for evaluation, we analyzed the Cambridge data, where more recent polygon data (2018) are available for both sidewalks and crosswalks. We manually examined all the mismatch cases and removed the false positives (i.e., cases where a polygon was falsely selected as being a match). Our analysis showed a 23% change from 2011 to 2018 in crosswalks, while sidewalks change was 9.2%, which illustrates the gradual change of seemingly fixed urban features such as sidewalks over time.

To evaluate the accuracy of the generated network centerlines, we

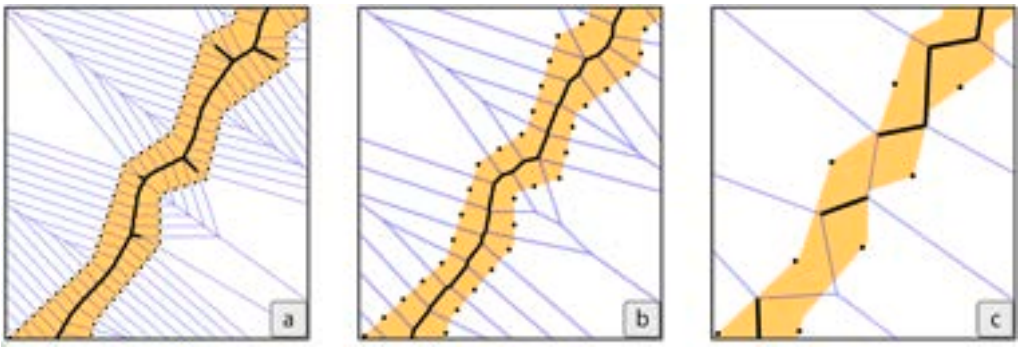

Fig. 5. Construction of centerline using Dense Voronoi method (DV) with different interpolation distances (d), which is the maximum distance between the sampled points (black points) on the polygon's boundary. The black line in the middle is the resulting centerline before any cleanups were applied. a) d = 0.2 resulting in a smoother line but with more extra branches, b) d = 0.5 created less smooth line but no extra branches, c) d = 2 resulted in a broken line.

Table 3
Availability of the official data across different cities. Training: T, Evaluation: E.

City	Data type	Sidewalk	Crosswalk	Footpath
Boston	Polygon	E	–	E
	Centerline	E	E	E
Cambridge	Polygon	T, E	T, E	T, E
	Centerline	E	E	E
Washington DC	Polygon	T, E	T, E	T, E
	Centerline	–	–	–
Manhattan	Polygon	T, E	–	T, E
	Centerline	–	–	E

first marked the centroids of network segments from a corresponding city dataset and buffered the centroids by four meters (corresponding to 95th percentile sidewalk width in Boston). We then spatially joined these centroids with our detected segments using spatial intersection analysis in GIS. The results thus report cases where Tile2Net had generated a network element within 4 m from a segment centroid in a city network dataset. We relied on centroids rather than full segments or endpoints to avoid matching intersecting line segments around network nodes. The results are reported in Table 5.

In Cambridge, our model matched 83.10% of all segments, with notable heterogeneity among different types of elements. Among sidewalks, 94.56% of centerlines had a corresponding detected segment, among crosswalks, 91.01%, and among footpaths, 68.86%. The lower matching rates among footpaths were expected due to more frequent tree cover over footpaths in parks and green spaces. Network matching in Boston was fairly similar across the same network types (Table 5). 90.78% of all sidewalk segments in city GIS data and 89.56% of all crosswalks were matched by our results. Footpath matching was again notably lower at 64.76%. In Manhattan, we only had official footpath networks (in parks) available from the city's open data repository. Here, 85.09% of official footpath segments had a corresponding detected segment within a four-meter buffer of their centroid.

For Washington, DC, the comparison could only be performed on more limited data. We did not find any official sidewalk centerlines and instead performed the comparison with the available OpenStreetMap pedestrian network. The results are shown in Table 6. A somewhat lower matching rate with OSM networks was expected (due to incompleteness of OSM sidewalk data in DC) and confirmed by the 76.90% match across all categories since OSM sidewalk networks are not official data, following different standards than those prepared by city governments. Though our inspection of results confirmed that both sidewalks and crosswalks again matched more closely than footpaths in parks, no type attributes for such comparison were available in the OSM network.

5. Discussion

While the automated pedestrian infrastructure mapping methodology we explored was able to capture a 90% or higher share of sidewalks and crosswalks featured in city GIS datasets, and a lower share of footpaths in parks, green areas, and other public spaces, a few caveats must be highlighted to interpret these results. First, the sidewalk, crosswalk, and footpath data available for validation in Cambridge, Boston, Washington, DC, and New York City are not necessarily temporally

Fig. 6. Model results showing detected sidewalk, crosswalk and footpath centerlines in a) Boston and Cambridge, b) Manhattan and parts of Brooklyn, c) Washington, DC. The maps are shown at the same scale for comparison.

Table 4
Comparison of polygon accuracy results in Cambridge, MA, Boston, MA, New York City, NY, and Washington, DC. Feature detected indicates what proportion of polygons in the city dataset had a corresponding detected polygon that overlaps with it. Since many of the undetected polygons are small in area, we also report the percentage of detected features weighted by area. The mean area overlap area reports how close in area (from 0 to 100%) the detected polygons are to the city dataset, on average (including those city polygons that remained undetected).

Measures	Cambridge, MA	Boston, MA	Washington, DC	New York City, NY
Official data polygon count	17,516	24,604	52,087	4684
Match (overlaps with detected)	17,327	24,288	43,963	4602
Features Detected	98.92%	98.72%	84.40%	98.25%
Features Detected (weighted by area)	99.62%	99.39%	97.48%	99.91%
Mean area overlap (weighted by area)	85.90%	77.90%	73.80%	87.50%

Table 5
Comparison of network accuracy results in Cambridge, Boston, and Manhattan.

City	Measures	All	Sidewalk	Crosswalk	Footpath
Cambridge	Official element count	12,792	5007	2414	5371
	Match (within 4 m of centroid)	10,631	4735	2197	3699
	Match	83.10%	94.56%	91.01%	68.86%
Boston	Official element count	110,031	54,864	11,223	37,023
	Match (within 4 m of centroid)	86,372	49,806	10,051	23,978
	Match	78.49%	90.78%	89.56%	64.76%
Manhattan	Official element count	–	–	–	6239
	Match (within 4 m of centroid)	–	–	–	5309
	Match	–	–	–	85.09%

Table 6
Network accuracy evaluation in Washington, DC.

City	Measure	Features
Washington, DC	OSM swlk element count	11,317
	Match (within 4 m of centroid)	8703
	Match	76.90%

concurrent with the imagery we used for feature detection. This can lead to expected differences between ground truth and detected features. For instance, in Cambridge, the GIS data we used for validation was last updated to reflect the year 2010 flyover conditions according to the city's metadata, but the image tiles we used as input for feature detection were captured in 2018. The Boston sidewalk and crosswalk centerline data were last updated to reflect 2011 conditions, while our Boston image tiles were captured in 2018. Some pedestrian elements in views are therefore not featured in the cities' GIS data and vice versa, possibly because they were altered before or after the images were captured. Our tests in Section 4 showed that the percentage change between data created based on 2010 flyovers and 2018 polygons was 9.2% for sidewalks and 23% for crosswalks. A similar proportion of matching difference is thus expected between the cities' GIS data and our results.

Second, we also noted errors in the cities' GIS datasets, where pedestrian infrastructure elements were missing or different from the Google Street View conditions dated to the same year. Given that the city datasets were likely prepared with a combination of automated feature detection and human correction, some error is expected. While these were the only data available to construct a quasi-official comparison of our results, these caveats are also partially responsible for the differences between detected and official pedestrian network elements.

The model can be improved with training and validation data that are both temporally and geometrically identical to the conditions captured in the image tiles used for feature detection. If city GIS data is versioned by year, the ground truth GIS data used for training the model could be dated back to an antecedent year that matches the image tiles and additionally humanly corrected to eliminate omissions and errors. This can ensure in future work that the detected polygons best match ground-truth polygons.

Future work should also explore the use of different semantic segmentation architectures in the detection model than the attention model used within Tile2Net so far. While prior studies reviewed above have used different segmentation architectures than here with notably lower accuracy results, a meaningful results comparison would have to use identical input datasets. The relatively lower detection accuracy of footpaths is attributable to several factors. On the one hand, feature detection from imagery is hampered by significantly higher levels of tree cover and other vegetation obstructions over footpaths found in parks, courtyards, and campuses. Footpaths also tend to have more complex geometries with winding and non-gridiron layouts, resulting in a much higher and more detailed segment count than on sidewalks and crosswalks. A complex curving footpath in a park made up of several segments may have a matching detected segments on some but not all of its segmented parts.

The polygon to centerline fitting part could benefit from further improvement. The Voronoi skeleton approach (Brandt & Algazi, 1992) we used for converting polygons to centerlines is very sensitive to the interpolation distance parameter and is not optimized for extracting the centerline of elongated polygons (see Appendix A). Moreover, the algorithm fits centerlines into discrete polygons and is not optimized for fitting the centerlines such that the endpoints of one skeleton topologically connect to the skeleton of another polygon, resulting in discontinuities between polygons. The resulting network segments are currently not optimized to form singular nodes or endpoints at intersections. Numerous detected line segments often converge near street corners, forming redundant intersections. We were partly able to adjust this with automated post-processing routines, but further refinements would be desirable to output continuous centerline networks. This can be addressed in future work by improving the algorithmic procedures to join endpoints into a single overlapping endpoint located at the geometric centroid of the multiple nodes found within a given distance. There is an extensive body of literature on various skeletonization algorithms (Saha, Borgefors, & di Baja, 2016), with some focusing solely on creating centerlines of elongated polygons (Haunert & Sester, 2008; Lewandowicz & Flisek, 2020). Finding the optimal interpolation distance value is beyond the scope of the current research, but presents a future work direction.

Though most computer vision solutions are fundamentally unable to detect sidewalk spatial elements where visual obstructions exist, lower detection accuracy in tree-covered regions was expected. Nevertheless, since our model was trained on planimetric GIS data, where pedestrian infrastructure elements were present regardless of obstructions, our model performed surprisingly well in occluded areas. Fig. 7 shows examples of the created network in sample areas of Cambridge, MA, Manhattan, and Washington, DC. In each case, the detection model correctly classified sidewalks and crosswalks, creating a continuous network despite the heavy shadow concentration on sidewalks (a), shadow and vegetation obstructing sidewalks, and crosswalks (b), and vegetation obstructing curbs and crosswalks (c).

Future work could further examine ways to fill in missing gaps in the resulting networks using probabilistic techniques. For instance, if additional detection classes, such as "tree" or "shadow," are added to the

Fig. 7. Mapping obstructed pedestrian facilities in different cities: a) Cambridge, MA. - sidewalks are mapped as continuous despite the heavy shadow, b) Manhattan - sidewalks and crosswalks obstructed by tree foliage and shadow are detected and mapped, c) Washington, DC. - crosswalks covered by vegetation are correctly detected and mapped.

semantic segmentation procedure, then these could be used in the network correction procedures to automatically connect gaps under trees and shadows. Yet, any automated correction for missing network links faces the hazard of erroneously creating pedestrian segments where they are not visible and hence may not exist. When networks are prepared for vulnerable street users (e.g., wheelchair users, mobility-impaired users), for whom network accuracy is critical, automated network correction procedures are likely futile, and improvements can only be made from ground surveys or Google Street View images.

The model is presently limited to detecting only sidewalk and crosswalk elements, which may not be appropriate in cities, where considerable parts of the pedestrian infrastructure are invisible from aerial imagery–overground foot-bridges, under-ground pedestrian crossings, covered pathways, and public pathways inside buildings. Additional efforts will be needed to combine sidewalk and crosswalk detection with invisible indoor elements in the contexts where the latter are significant (e.g., Hong Kong, Singapore, Minneapolis, and Montreal, to name a few). Moreover, additional classes such as driveways, curbs, stairs could be added to our detection model.

The lack of standardized training data across different cities also posed challenges in our work. For instance, different cities have captured and mapped sidewalks with varying levels of detail. In Washington, DC, unpaved planter areas were excluded from sidewalk polygons, whereas in Boston and NYC, they were included as parts of sidewalks. The same problem exists for curb extensions, medians, driveways, and curb-cuts. Crosswalk representation presented another source of variation among different cities. While they were mapped as part of sidewalk inventory data in Washington DC, in Boston, they were only presented in the sidewalk centerline dataset; hence, with no information available about their size and shape. In Cambridge, they were part of both the sidewalk centerline data and a separate dataset on road markings, where pedestrian zebras were represented as polygons.

Beyond heterogeneity in training data, the physical features, materials, and dimensions of sidewalks and crosswalks can also vary between cities. We observed multiple instances of faded crosswalks that made it challenging for semantic segmentation to detect. We also noted differences in both sidewalk materials and crosswalk materials across cities. Whereas very few crosswalks are paved in brick in NYC, they are common in Cambridge and Boston. Had we trained the algorithm on NYC alone, it could have resulted in systemic under-detection in Boston and Cambridge. Such differences are bound to be bigger between international cities, where construction materials, crosswalk marking conventions, and infrastructure dimensions vary more considerably than between the three East Coast cities included in our study. When extending the model to new contexts, especially outside the U.S., the model can be retrained specifically for each region.

While the results are promising, we emphasize the need for expanding the work to additional cities and regions globally, where locally specific training may be needed to achieve high detection accuracy. However, the retraining for new regions can be done at much lower cost since our pre-trained model can be used for transfer-learning and domain adaptations with significantly less data compared to the initial training.

The resulting sidewalk and crosswalk dataset can be further combined with attribute information that may be useful for various pedestrian analytics. For instance, as shown by Hosseini et al. (2021), the captured sidewalk and crosswalk polygons can be used to measure the width of each sidewalk segment. Furthermore, using results by Hosseini, Miranda, Lin, and Silva (2022), who developed a method for detecting sidewalk surface materials from Google Street View imagery, our sidewalk segments can be joined with corresponding geotagged material information, instead of having to aggregate the data from left and right sidewalks into road centerlines. Such measurable attributes can impact the quality and attractiveness of sidewalks, and have been shown to affect pedestrian route choice and perceived route length (Basu, Sevtsuk, & Li, 2022; Erath, van Eggermond, Ordóñez Medina, & Axhausen, 2015; Sevtsuk, Basu, Li, & Kalvo, 2021).

Having pedestrian paths represented as continuous, topologically connected network datasets could open up new (and overdue) efforts for pedestrian routing, flow analysis, and potential location-based or delivery services. Transit-first policies, walkable-streets initiatives, step-free access for public transport, and vision zero goals represent but few planning and policy areas which could benefit from citywide sidewalk and crosswalk datasets.

Author statement

The authors have no financial interests to declare.

Declaration of Competing Interest

None.

Acknowledgement

We would like to thank our colleagues at New York University for their help in this research.

This work was supported in part by: C2SMART, the Moore-Sloan Data Science Environment at NYU; NASA; NSF awards CNS-1229185, CCF-1533564, CNS-1544753, CNS-1730396, CNS-1828576, CNS-1626098; and the NVIDIA NVAIL at NYU. Roberto M.CesarJr is grateful to São Paulo Research Foundation (FAPESP) grants \#2015/22308-2

and \#2019/01077-3, CNPq and MCTI PPI-SOFTEX TIC 13 DOU 01245.010222/2022-44. Claudio T. Silva is partially supported by the DARPA D3M program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

Appendix A. Polygon to line transformation: a closer look

The border density parameter, called interpolation distance, densifies the input geometry's border by placing additional points at that given distance. If the interpolation distance is too small, the output will have many unwanted branches, while large values can lead to zigzaggy and disjointed centerlines (Lewandowicz & Flisek, 2020; Li, Guan, Yu, Chiang, & Knoblock, 2021) as illustrated in Fig. A.8.

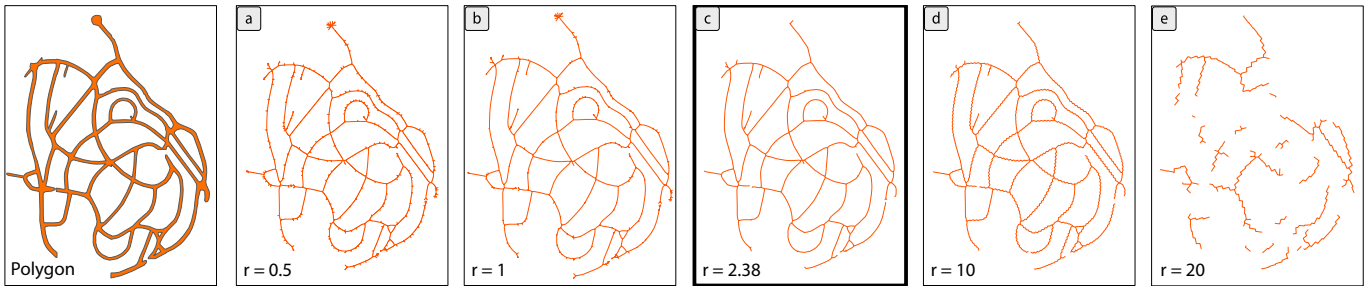

Fig. A.8. Impact of different interpolation distances on the resulting centerline created from the input polygon. Small values create extra branches ($r = 0.5$ and $r = 1$) and large values create zigzaggy ($r = 10$) or disjointed lines ($r = 20$). The middle centerline, highlighted with a thicker border, is computed using the interpolation distance computed using our heuristic approach.

Finding the optimal interpolation distance is beyond the scope of the current work. To approximate a suitable parameter for each polygon, we used a heuristic approach and selected a sample of 400 polygons of varying areas and perimeters. Next, for each polygon, we tested different interpolation distances ranging from 0.5 to 20, using a 0.5 step (i.e., total of 40 different parameters) and chose the line with the highest connectivity and the least number of extra branches which best represents our irregular shapes. For each polygon, we record the interpolation distance that results in the best centerline, as well as the polygon area, perimeter, average width, number of vertices, area to minimum bounding box area ratio, and area to perimeter ratio. We used a polynomial regression model and concluded that the area to perimeter ratio is a significant factor in choosing the interpolation distance. Using the derived coefficient, we compute the interpolation distance of each polygon for centerline creation. In Fig. 5 the centerline highlighted with a thicker border is computed using the interpolation distance derived from our heuristic approach ($r = 2.38$), having smooth lines which follow the form of the input polygons with very few extra branches compared to smaller values. The coefficient can be finetuned on new datasets.

References

Ahn, J., & Kwak, S. (2018). Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4981–4990).

Ai, C., & Hou, Q. (2019). Improving Pedestrian Infrastructure Inventory in Massachusetts Using Mobile LiDAR (No. 19-007). Massachusetts. *Dept. of Transportation. Office of Transportation Planning*.

Ai, C., & Tsai, Y. (2016). Automated sidewalk assessment method for americans with disabilities act compliance using three-dimensional mobile lidar. *Transportation Research Record: Journal of the Transportation Research Board*, 25–32.

Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., & Natarajan, V. (2021). Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 3478–3488).

Balado, J., Daz-Vilariño, L., Arias, P., & González-Jorge, H. (2018). Automatic classification of urban ground elements from mobile laser scanning data. *Automation in Construction, 86*, 226–239.

Balali, V., Rad, A. A., & Golparvar-Fard, M. (2015). Detection, classification, and mapping of us traffic signs using google street view images for roadway inventory management. *Visualization in Engineering, 3*, 15.

Bastani, F., He, S., Abbar, S., Alizadeh, M., Balakrishnan, H., Chawla, S., … DeWitt, D. (2018). Roadtracer: Automatic extraction of road networks from aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4720–4728).

Basu, R., Sevtsuk, A., & Li, X. (2022). *How do street attributes affect willingness-to-walk? City-wide pedestrian route choice analysis using big data from Boston and San Francisco. Transportation Research A,. Upcoming*.

Boeing, G. (2017). Osmnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems, 65*, 126–139.

Brandt, J. W., & Algazi, V. R. (1992). Continuous skeleton computation by voronoi diagram. *CVGIP: Image understanding, 55*, 329–338.

Cambra, P. J., Gonçalves, A., & Moura, F. (2019). The digital pedestrian network in complex urban contexts: A primer discussion on typological specifications. *Finisterra, 54*, 155–170.

Cambridge GIS. (2018a). Cambridge sidewalk. Retrieved from: https://www.cambridgema.gov/GIS/gisdatadictionary/Basemap/BASEMAP_Sidewalks.

Cambridge GIS. (2018b). Pavement markings. Retrieved from: https://www.cambridgema.gov/GIS/gisdatadictionary/Traffic/TRAFFIC_PavementMarkings.

Cambridge GIS. (2018c). Public footpaths. Retrieved from: https://www.cambridgema.gov/GIS/gisdatadictionary/Basemap/BASEMAP_PublicFootpaths.

Cambridge GIS. (2018d). Roads. Retrieved from: https://www.cambridgema.gov/GIS/gisdatadictionary/Basemap/BASEMAP_Roads.

Cervero, R. (1998). *The transit metropolis: A global inquiry*. Island press.

Chen, L.-C., Yang, Y., Wang, J., Xu, W., & Yuille, A. L. (2016). Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3640–3649).

Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., & Sun, J. (2018). Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7103–7112).

Chin, G. K., Van Niel, K. P., Giles-Corti, B., & Knuiman, M. (2008). Accessibility and connectivity in physical activity studies: The impact of missing pedestrian data. *Preventive Medicine, 46*, 41–45.

Cohen, A., & Dalyot, S. (2021). Route planning for blind pedestrians using openstreetmap. *Environment and Planning B: Urban Analytics and City Science, 48*, 1511–1526.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3213–3223).

Cottrill, C. D., Pereira, F. C., Zhao, F., Dias, I. F., Lim, H. B., Ben-Akiva, M. E., & Zegras, P. C. (2013). Future mobility survey: Experience in developing a smartphone-based travel survey in Singapore. *Transportation Research Record, 2354*, 59–67.

DC GIS. (2019a). Roads 2019. Available online https://opendata.dc.gov/datasets/DCGIS::roads-2019/.

DC GIS. (2019b). Sidewalks 2019. Available online https://opendata.dc.gov/datasets/sidewalks-2019/.

DC GIS. (2020). Aerial photography (orthophoto sid) - 2019. Available online https://opendata.dc.gov/documents/DCGIS::aerial-photography-download-orthophoto-sid-2019.

Delboni Lomba, L. F., & Godoy da Silva, J. (2022). Informed search algorithm for route optimization for visually impaired people: Possibility of intelligent assistive technology. *Journal of Location Based Services*, 1–16.

Ding, H., Jiang, X., Shuai, B., Liu, A. Q., & Wang, G. (2018). Context contrasted feature and gated multi-scale aggregation for scene segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2393–2402).
Douglas, D. H., & Peucker, T. K. (1973). Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: the international journal for geographic information and geovisualization, 10*, 112–122.
Ellis, G., Hunter, R., Tully, M. A., Donnelly, M., Kelleher, L., & Kee, F. (2016). Connectivity and physical activity: Using footpath networks to measure the walkability of built environments. *Environment and Planning. B, Planning & Design, 43*, 130–151.
El-Taher, F. E.-Z., Taha, A., Courtney, J., & Mckeever, S. (2021). A systematic review of urban navigation systems for visually impaired people. *Sensors, 21*, 3103.
EPA. (2021). Sources of greenhouse gas emissions. Retrieved from https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions.
Erath, A. L., van Eggermond, M. A., Ordóñez Medina, S. A., & Axhausen, K. W. (2015). Modelling for walkability: Understanding pedestrians' preferences in Singapore. In *14th international conference on travel behavior research (IATBR 2015)*. IVT, ETH Zurich.
Ess, A., Mueller, T., Grabner, H., & Van Gool, L. (2009). *Segmentation-based urban traffic scene understanding. British Machine Vision Association and Society for Pattern Recognition (BMVC 2009) (p. 2)*. In .
Etten, A. V. (2020). City-scale road extraction from satellite imagery v2: Road speeds and travel times. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 1786–1795).
Frank, L. D., & Engelke, P. O. (2001). The built environment and human activity patterns: Exploring the impacts of urban form on public health. *Journal of Planning Literature, 16*, 202–218.
Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., & Lu, H. (2019). Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3146–3154).
Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? The Kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition* (pp. 3354–3361). IEEE.
Glaeser, E. (2010). *Triumph of the City: How our greatest invention makes us richer, smarter, greener, healthier, and happier*. Penguin Press.
Grasser, G., Van Dyck, D., Titze, S., & Stronegger, W. (2013). Objectively measured walkability and active transport and weight-related outcomes in adults: A systematic review. *International Journal of Public Health, 58*, 615–625.
Guttman, A. (1984). R-trees: A dynamic index structure for spatial searching. In *Proceedings of the 1984 ACM SIGMOD international conference on management of data* (pp. 47–57).
Hagberg, A., & Conway, D. (2020). Networkx: Network analysis with python. *URL: https://networkx.github.io*.
Haunert, J.-H., & Sester, M. (2008). Area collapse and road centerlines based on straight skeletons. *GeoInformatica, 12*, 169–191.
He, J., Deng, Z., & Qiao, Y. (2019). Dynamic multi-scale filters for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3562–3572).
He, J., Deng, Z., Zhou, L., Wang, Y., & Qiao, Y. (2019). Adaptive pyramid context network for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7519–7528).
He, L., Chao, Y., Suzuki, K., & Wu, K. (2009). Fast connected-component labeling. *Pattern Recognition, 42*, 1977–1987.
Horváth, E., Pozna, C., & Unger, M. (2022). Real-time lidar-based urban road and sidewalk detection for autonomous vehicles. *Sensors, 22*, 194.
Hosseini, M., Araujo, I. B., Yazdanpanah, H., Tokuda, E. K., Miranda, F., Silva, C. T., & Cesar, R. M., Jr. (2021). Sidewalk measurements from satellite images: Preliminary findings. In *Spatial data science symposium 2021 short paper proceedings*. 10.25436/E2QG6F.
Hosseini, M., Miranda, F., Lin, J., & Silva, C. T. (2022). Citysurfaces: City-scale semantic segmentation of sidewalk materials. *Sustainable Cities and Society*, 103630.
Huang, Q., Xia, C., Wu, C., Li, S., Wang, Y., Song, Y., & Kuo, C.-C. J. (2017). Semantic segmentation with reverse attention. In *British Machine Vision Association and Society for Pattern Recognition (BMVC 2017), (Oral Presentation)*. arXiv preprint. arXiv:1707.06426.
Iglovikov, V., Mushinskiy, S., & Osin, V. (2017). Satellite imagery feature detection using deep convolutional neural network: A kaggle competition. *arXiv preprint. arXiv:1706.06169*.
Jordahl, K. (2014). Geopandas: Python tools for geographic data. https://github.com/geopandas/geopandas.
Kamel, I., & Faloutsos, C. (1993). *Hilbert R-tree: An improved R-tree using fractals* (Technical Report).
Karimi, H. A., & Kasemsuppakorn, P. (2013). Pedestrian network map generation approaches and recommendation. *International Journal of Geographical Information Science, 27*, 947–962.
Kasemsuppakorn, P., & Karimi, H. A. (2013). Pedestrian network extraction from fused aerial imagery (orthoimages) and laser imagery (lidar). *Photogrammetric Engineering & Remote Sensing, 79*, 369–379.
Kim, J. H., Lee, S., Hipp, J. R., & Ki, D. (2021). Decoding urban landscapes: Google street view and measurement sensitivity. *Computers, Environment and Urban Systems, 88*, Article 101626.
Leinberger, C. B., & Lynch, P. (2014). *Foot traffic ahead: Ranking walkable urbanism in america's largest metros*. Transportation Research Board.
Lewandowicz, E., & Flisek, P. (2020). A method for generating the centerline of an elongated polygon on the example of a watercourse. *ISPRS International Journal of Geo-Information, 9*, 304.
Li, H., Xiong, P., An, J., & Wang, L. (2018). Pyramid attention network for semantic segmentation. In *British Machine Vision Association and Society for Pattern Recognition (BMVC 2018)*. arXiv preprint. arXiv:1805.10180.
Li, W., He, C., Fang, J., Zheng, J., Fu, H., & Yu, L. (2019). Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source gis data. *Remote Sensing, 11*, 403.
Li, Z., Guan, R., Yu, Q., Chiang, Y.-Y., & Knoblock, C. A. (2021). Synthetic map generation to provide unlimited training data for historical map text detection. In *Proceedings of the 4th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery* (pp. 17–26).
Liu, S., Higgs, C., Arundel, J., Boeing, G., Cerdera, N., Moctezuma, D., … Giles-Corti, B. (2021). A generalized framework for measuring pedestrian accessibility around the world using open data. *Geographical Analysis, 54*(3), 559–582.
Liu, W., Rabinovich, A., & Berg, A. C. (2015). Parsenet: Looking wider to see better. *arXiv preprint*. arXiv:1506.04579.
Luo, J., Wu, G., Wei, Z., Boriboonsomsin, K., & Barth, M. (2019). Developing an aerial-image-based approach for creating digital sidewalk inventories. *Transportation Research Record, 2673*, 499–507.
Luttrell, J. B., IV (2022). *Data collection and machine learning methods for automated pedestrian facility detection and mensuration*.
MassGIS. (2018). MassGIS Data: 2018 Aerial Imagery. https://www.mass.gov/info-details/massgis-data-2018-aerial-imagery.
Moran, M. E. (2022). *Where the crosswalk ends: Mapping crosswalk coverage via satellite imagery in San Francisco*. Environment and Planning B: Urban Analytics and City Science, 23998083221081530.
Moretti, E. (2012). *The new geography of jobs*. Houghton Mifflin Harcourt.
Neuhold, G., Ollmann, T., Rota Bulo, S., & Kontschieder, P. (2017). The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision* (pp. 4990–4999).
Newell, A., Yang, K., & Deng, J. (2016). Stacked hourglass networks for human pose estimation. In *European conference on computer vision* (pp. 483–499). Springer.
Ning, H., Ye, X., Chen, Z., Liu, T., & Cao, T. (2022). Sidewalk extraction using aerial and street view images. *Environment and Planning B: Urban Analytics and City Science, 49*, 7–22.
NYC DoITT. (2018). New york city planimetrics data. Retrieved from: https://github.com/CityOfNewYork/nyc-planimetrics.
NYC GIS. (2018). NYS Statewide Digital Orthoimagery Program. Available: https://gis.ny.gov/gateway/orthoprogram/index.cfm.
Placematters and WalkDenver. (2014). Walkscope. http://www.walkscope.org/.
Proulx, F. R., Zhang, Y., & Grembek, O. (2015). Database for active transportation infrastructure and volume. *Transportation Research Record, 2527*, 99–106.
Quackenbush, L. J. (2004). A review of techniques for extracting linear features from imagery. *Photogrammetric Engineering & Remote Sensing, 70*, 1383–1392.
Rosenfeld, A., & Pfaltz, J. L. (1966). Sequential operations in digital picture processing. *Journal of the ACM (JACM), 13*, 471–494.
Sachs, D. (2016). A complete map of Denver's walking network is now within reach. https://denver.streetsblog.org/2016/06/29/a-complete-map-of-denvers-walking-network-is-now-within-reach/.
Saha, M., Saugstad, M., Maddali, H. T., Zeng, A., Holland, R., Bower, S., … Froehlich, J. (2019). Project sidewalk: A web-based crowdsourcing tool for collecting sidewalk accessibility data at scale. In *Proceedings of the 2019 CHI conference on human factors in computing systems CHI '19*. Association for Computing Machinery.
Saha, P. K., Borgefors, G., & di Baja, G. S. (2016). A survey on skeletonization algorithms and their applications. *Pattern Recognition Letters, 76*, 3–12.
Sevtsuk, A., Basu, R., Li, X., & Kalvo, R. (2021). A big data approach to understanding pedestrian route choice preferences: Evidence from San Francisco. *Travel Behaviour and Society, 25*, 41–51.
Speck, J. (2013). *Walkable city: How downtown can save America, one step at a time*. Macmillan.
Sun, C., Su, J., Ren, W., & Guan, Y. (2019a). Wide-view sidewalk dataset based pedestrian safety application. *IEEE Access, 7*, 151399–151408.
Sun, K., Xiao, B., Liu, D., & Wang, J. (2019b). Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5693–5703).
Sun, K., Zhao, Y., Jiang, B., Cheng, T., Xiao, B., Liu, D., … Wang, J. (2019c). High-resolution representations for labeling pixels and regions. *arXiv preprint*. arXiv:1904.04514.
Tao, A., Sapra, K., & Catanzaro, B. (2020). Hierarchical multi-scale attention for semantic segmentation. *arXiv preprint. arXiv:2005.10821*.
TCAT. (2016). *OpenSidewalks, Openly mapping for the pedestrian experience, Taskar Center for Accessible Technology (TCAT)*. University of Washington. https://www.opensidewalks.com/.
Treccani, D., Díaz-Vilariño, L., & Adami, A. (2021). Sidewalk detection and pavement characterisation in historic urban environments from point clouds: Preliminary results. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 43*, 243–249.
Tucker, C. J., Grant, D. M., & Dykstra, J. D. (2004). Nasa's global orthorectified landsat data set. *Photogrammetric Engineering & Remote Sensing, 70*, 313–322.
US Geological Survey. (2018). *USGS EROS Archive - Aerial Photography - High Resolution Orthoimagery (HRO)*. https://doi.org/10.5066/F73X84W6
Walker, J., & Johnson, C. (2016). Peak car ownership: The market opportunity of electric automated mobility services. Retrieved from https://www.auto-mat.ch/wAssets/docs/170327_CWRRMI_POVdefection_ExecSummary_L12.pdf.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., , … Wang, X., et al. (2020). Deep high-resolution representation learning for visual recognition. In *IEEE transactions on pattern analysis and machine (intelligence)*.

Wang, R., Liu, Y., Lu, Y., Zhang, J., Liu, P., Yao, Y., & Grekousis, G. (2019). Perceptions of built environment and health outcomes for older chinese in Beijing: A big data approach with street view images and deep learning technique. *Computers, Environment and Urban Systems, 78*, Article 101386.

Wei, Y., Zhang, K., & Ji, S. (2019). Road network extraction from satellite images using cnn based segmentation and tracing. In *IGARSS 2019–2019 IEEE international geoscience and remote sensing symposium* (pp. 3923–3926). IEEE.

Wu, Z., Shen, C., & Van Den Hengel, A. (2019). Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition, 90*, 119–133.

Yang, X., Tang, L., Ren, C., Chen, Y., Xie, Z., & Li, Q. (2020). Pedestrian network generation based on crowdsourced tracking data. *International Journal of Geographical Information Science, 34*, 1051–1074.

Yu, F., Wang, D., Shelhamer, E., & Darrell, T. (2018). Deep layer aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2403–2412).

Yuan, Y., Chen, X., & Wang, J. (2019). Object-contextual representations for semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16* (pp. 173–190). Springer International Publishing. arXiv preprint.arXiv:1909.11065.

Zapata-Diomedi, B., Boulangé, C., Giles-Corti, B., Phelan, K., Washington, S., Veerman, J. L., & Gunn, L. D. (2019). Physical activity-related health and economic benefits of building walkable neighbourhoods: A modelled comparison between brownfield and greenfield developments. *International Journal of Behavioral Nutrition and Physical Activity, 16*, 1–12.

Zhang, F., Zhang, D., Liu, Y., & Lin, H. (2018). Representing place locales using scene elements. *Computers, Environment and Urban Systems, 71*, 153–164.

Zhang, H., & Zhang, Y. (2019). Pedestrian network analysis using a network consisting of formal pedestrian facilities: Sidewalks and crosswalks. *Transportation Research Record, 2673*, 294–307.

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2881–2890).

Zhao, S., Wang, Y., Yang, Z., & Cai, D. (2019). Region mutual information loss for semantic segmentation. *Advances in Neural Information Processing Systems, 32* (NeurIPS 2019).

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., & Torralba, A. (2017). Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 633–641).

Zhou, G., Chen, W., Kelmelis, J. A., & Zhang, D. (2005). A comprehensive study on urban true orthorectification. *IEEE Transactions on Geoscience and Remote Sensing, 43*, 2138–2147.

Zhou, H., Liu, L., Lan, M., Zhu, W., Song, G., Jing, F., Zhong, Y., Su, Z., & Gu, X. (2021). Using google street view imagery to capture micro built environment characteristics in drug places, compared with street robbery. *Computers, Environment and Urban Systems, 88*, Article 101631.

Zhu, Y., Sapra, K., Reda, F. A., Shih, K. J., Newsam, S., Tao, A., & Catanzaro, B. (2019). Improving semantic segmentation via video propagation and label relaxation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8856–8865).