TRACK 6: METHODS

MAPPING THE WALK : A SCALABLE COMPUTER VISION APPROACH FOR GENERATING SIDEWALK NETWORK DATASETS FROM AERIAL IMAGERY

Andres Sevtsuk [1]

[1] Department of Urban Studies and Planning, MIT

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 a strive for economic competitiveness. 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 et al., 2013; Zapata-Diomedi et al., 2019). Economically, walkable and transit-served city environments have also become an important draw for a competitive workforce (Moretti, 2012; Glaeser, 2010) 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 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 Prime), urban transportation models and policies (e.g., metropolitan and urban Travel Demand Models, congestion charging systems in various of cities), as well as mobility 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 or updated, 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 pedestrian infrastructure– 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 (Qin et al., 2018; Cambra et al., 2019; Sun et al., 2019a), which can lead to inaccuracies (e.g., streets with no sidewalks), simplifications (e.g., assumptions that buildings can be directly accessed on both side 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) (Chin et al., 2008; Ellis et al., 2016). 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).

To address these challenges, we introduce Tile2Net –an end-to-end framework for automated mapping of pedestrian infrastruc- ture using aerial imagery. Tile2Net enables users to download orthorectified sub-meter resolution image tiles for a given region from public sources and generate topologically interconnected, georeferenced sidewalk and crosswalk centerlines as well as side- walk, road, and crosswalk polygons. Our goal is to map pedestrian networks "as they are" rather than trying to improve the network connectivity artificially. To achieve this, we use a semantic segmentation model that can detect sidewalk, footpath, and crosswalk polygons from orthorectified tiles. We then use the resulting polygons to create an interconnected network. We pilot tested the approach in Manhattan, NY, Washington, DC, Boston, and Cambridge, MA, and achieved high accuracy in each of these cities. The model can be finetuned based on the topological characteristics of different datasets and cities. Our key contributions are as follows:

1. We provide an end-to-end, open-source framework to create large-scale pedestrian networks from orthorectified imagery(link omitted to satisfy double-blind review requirements).
2. The framework also generates georeferenced polygons of roads, sidewalks (including footpaths) and crosswalks.
3. We offer techniques for the automated creation of annotation masks, using publicly available or user input

datasets to train the semantic segmentation models.

4. Our generalized pedestrian feature detection model–made publicly available–is trained on a selected number of cities with varying street network geometries, building shadow densities, and tree covers (Cambridge, Washington, DC, and New York City parks), making it applicable for other cities with similar environments without any need for additional training.
5. Our solution is adjustable to different city environments, offering various settings to finetune the model on the new dataset, based on the local characteristics of the data.

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 implementation. In Section 4 we present our results. Section 5 presents a discussion of the challenges of automated sidewalk network detection and suggests directions for expanding the work in the future. Section 6 concludes the paper.

2. Literature Review

2.1. *Map generation*

At least five different frameworks for mapping sidewalk infrastructure can be disguised in existing literature and practice, with additional combinations thereof. The main differentiating point between these five categories lies in the method used to detect pedestrian infrastructures such as sidewalks, footpaths, and crosswalks. Figure 1 offers an illustrative summary of these methods.

First, physical site surveys and manual aerial imagery surveys have been used in a number of cities to develop datasets on pedestrian facilities (e.g., in Melbourne, Singapore, and Boston). This involves tracing observable sidewalks and crosswalks from georeferenced aerial imagery, combined with on-the-ground observation and validation (Proulx et al., 2015). Such mapping efforts can produce accurate and high-quality results, but it can also be prohibitively labor intensive and difficult to scale across large regions. In a recent study, 6,400 intersections in San Francisco were manually reviewed and classified based on the crosswalk presence and condition, which took 90 hours 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. How these open-sourced data is generated can vary, but can also include the methods described in this section.

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. In this method, which is a widely used and a common approach in geo-information processing, the boundaries of the resulting polygons are considered as the approximate location of the sidewalks 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. After sidewalk segment geometries are generated, crosswalks can be added by linking the endpoints (i.e., intersections) of the assumed sidewalk intersections perpendicularly across road center- lines (Karimi & Kasemsuppakorn, 2013; Brezina et al., 2017). This approach has several shortcomings, first is the limited extent of the locations such a network can cover. A network constructed based on streets and roads does not include off-road footpaths, pedestrian bridges, skywalks, or underground tunnels. In other words, it is limited to only where roads can go and can generate ar- bitrary sidewalks and crosswalks, which can lead to inaccuracies (e.g., all streets will have sidewalks on both sides), simplifications (e.g., assumptions that buildings can be directly accessed on both sides of a street centerline, while in reality crossing a street is

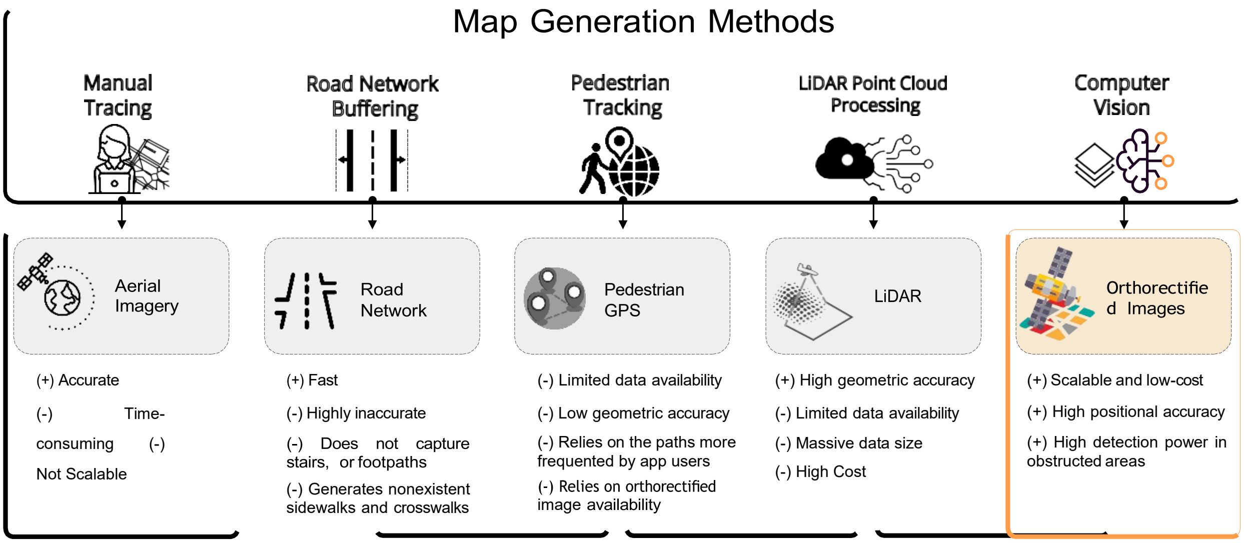
Map Generation Methods
Manual Tracing
Road Network Buffering
Pedestrian Tracking
LiDAR Point Cloud Processing
Computer Vision
Aerial Imagery
Road Network
Pedestrian GPS
LiDAR
Orthorectified Images
(+) Accurate
(-) Time-consuming
(-) Not Scalable
(+) Fast
(-) Highly inaccurate
(-) Does not capture stairs, or footpaths
(-) Generates nonexistent sidewalks and crosswalks
(-) Limited data availability
(-) Low geometric accuracy
(-) Relies on the paths more frequented by app users
(-) Relies on orthorectified image availability
(+) High geometric accuracy
(-) Limited data availability
(-) Massive data size
(-) High Cost
(+) Scalable and low-cost
(+) High positional accuracy
(+) High detection power in obstructed areas

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.

only allowed at specific locations), and misrepresentation (e.g., assuming pedestrian connections based on vehicular routes, where there are none) (Chin et al., 2008; Ellis et al., 2016), each of which can lead to potentially hazardous situations for pedestrians, specifically the more vulnerable population (Saha et al., 2019).

Third, pedestrian pathways have also been identified from Global Positioning System (GPS) trajectories of pedestrian move- ment. This can include data from designated 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 Apps that track their users' location history. Once collected, GPS traces can be merged, sim- plified, 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 locations such as the vicinity of high-rise buildings (Karimi & Kasemsuppakorn, 2013).

The fourth category is LiDAR point cloud processing, which utilizes airborne Light Detection and Ranging (LiDAR) point clouds data. LiDAR devices use active sensing and can be fixed or mounted on mobile objects such as planes and drones (Cura et al., 2018). 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 uses prior knowledge of the unique geometrical shapes and measurements of urban ground elements. 2) Reflectance-based methods utilize the reflectance intensity of different classes of objects to classify the data. The classified points are then normalized based on the laser scanning model, and distance projection is used to create a saliency map. These two methods are often combined to more accurately extract the streetscape features. 3) Scan-based methods take advantage of the scanning pattern to connect the results from consecutive scans into a continuous boundary and refine the segmentation (Ai & Tsai, 2016; Baker & Hou, 2019; Balado et al., 2018). In clustering feature classes, pedestrian path segments are typically assumed to be made of concrete, and parking lots of asphalt (Karimi & Kasemsuppakorn, 2013; Hou & Ai, 2020; Kasemsuppakorn & Karimi, 2013), which as shown by Hosseini et al. (2022), is not the case in many cities. The resulting data represents sidewalks as vector lines or polygons that can be both accurate and scalable (Horva´th et al., 2022; Treccani et al., 2021). Unlike aerial imagery, LiDAR data can be acquired during different hours (day and night), and the data is already georeferenced. However, the lack of spatially dense, universal LiDAR data has limited this approach to relatively few cities overall.

Fifth, and in line with our work, different computer vision techniques have more recently been deployed in a limited number of studies to detect pedestrian infrastructure from aerial images (Ning et al., 2022). The detected features are then converted into georeferenced lines or polygons and go through topological corrections to produce the final network. Among computer vision techniques, semantic segmentation can result in highly 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 et al., 2009; Geiger et al., 2012). To construct a pedestrian network, a segmentation model is first trained to detect different features of the streetscape, such as roads, sidewalks, and crosswalks, from aerial images. Although semantic segmentation has been broadly used to detect roads and building footprints from aerial images (Balali et al., 2015; Iglovikov et al., 2017; Li et al., 2019) and to create road networks (Bastani et al., 2018; Wei et al., 2019; Etten, 2020), it has not been widely implemented for sidewalk mapping so far, possibly due to several technical challenges. First, in order to achieve satisfactory results, semantic segmentation algorithms need to be trained on densely annotated labels, which can be labor-intensive and costly to prepare. Consequently, in applying semantic segmentation models to urban context (Zhang et al., 2018; Wang et al., 2019; Zhou et al., 2021; Kim et al., 2021), researchers often forego retraining or fine-tuning their models on their target datasets and rather rely only on publicly-available models pre-trained on datasets such as CityScapes (Cordts et al., 2016), Mapillary (Neuhold et al., 2017), and ADE20K (Zhou et al., 2017). This reliance on pre-trained models, not specific to the desired task, limits analysis to the pre-defined 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). Second, compared to roads and buildings, detecting sidewalks, footpaths and crosswalks is more challenging since they constitute a small portion of the visual information of aerial images, and their detection can be further inhibited by occlusion from shadow, vegetation, and structures such as bridges or tall buildings (Hosseini et al., 2021). Hence, choosing the right network architecture that can preserve the fine local details while taking the global image context into account is crucial.

2.2. *Semantic segmentation*

The rise of autonomous vehicles and self-driving cars 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 architectures to improve performance over a static resolution approach (Zhao et al., 2017). This has been followed by rapid developments in multi-scale pyramid-style networks (He et al., 2019a; Ding et al., 2018; He et al., 2019b). In particular, HRNet (Sun et al., 2019c; Wang et al., 2020) connects high-to-low resolution convolutions via parallel and repeated multi-scale fusions to better preserve low-resolution representations alongside the high-resolution ones in comparison to previous works (Newell et al., 2016; Chen et al., 2018; Yu et al., 2018). A variant of HRNet, HRNet-W48, has shown superior performance across segmentation benchmarks such as Cityscapes (Cordts et al., 2016) and Mapillary Vista (Sun et al., 2019b), is used as a key component of this proposal's segmentation framework.

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	Image tiles	2018	(MassGIS, 2018)
	Sidewalk Inventory	Off-road footpaths inside parks	2018	(NYC DoITT, 2018)

Manhattan and Brooklyn	Roads	Road polygons	2018	(NYC DoITT, 2018)
	Ortho-imagery	Image tiles	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	Orthophoto SID	2019	(DC GIS, 2020)

Attention-based mechanisms have been adopted in multiple semantic segmentation architectures (Huang et al., 2017; Fu et al., 2019; Chen et al., 2016; Li et al., 2018). Instead of feeding multiple resized images into a shared network and merging the features to make the prediction, which can lead to sub-optimal 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 RAN, a reversed attention mechanism that trains the model on the features which 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 a network architecture that learns to assign a relative weighting between adjacent scales (Tao et al., 2020). This method has shown to be four times more memory efficient and allows for larger crop sizes that can lead to more accurate results. We adopted this architecture in our network generation pipeline due to its superior performance in detecting both high and low-level features while benefiting from its memory-efficient design.

3. Materials and Methods

In this section, we detail the datasets used for training the model, describe our methodology, and discuss how we have ad- dressed the previous challenges of preparing labor-intensive annotation labels for training the algorithm and generalized it to detect pedestrian infrastructure in different urban environments. We also illustrate how initially detected polygon geometries can be con- verted into sidewalk centerlines, bringing the outputs closer to a topologically interconnected network dataset that can be used for pedestrian routing and other network analysis procedures.

3.1. *Data description*

The semantic segmentation model requires pairs of aerial images and their corresponding annotation labels to be trained. We designed an automated method to create annotation labels from publicly available datasets to overcome the annotation bottleneck. Our goal has been to work with scalable, non-proprietary input data and methods that would allow sidewalk mapping to be extended to heterogeneous cities in the U.S. and potentially globally. 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 data that is created from orthorectified images. Next, we provide more details about each one and describe

how they were used in creating the training data. Table 1 shows the datasets used to train the model and their delivery dates.

3.1.1. *High-resolution orthorectified imagery*

Raw aerial images inherently contain distortion caused by sensor orientation, systematic sensor and platform-related geometry errors, terrain relief, and curvature of the earth. Such distortions cause feature displacement and scaling errors, which can result in inaccurate direct 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 et al., 2004; Zhou et al., 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.

Aside from orthoimages provided by U.S. Geological Survey (USGS) (US Geological Survey, 2018), 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. Tile2Net is designed with the capability of automating the data preparation process. It can take as input, the textual name or geographic coordinates of the bounding box of a given region and download the tiles that fall within the bounding box, for the cities where orthoimagery is available.

To create the training data, using Tile2Net, we obtained 11,000 tiles from Washington, DC, 28,000 tiles from Cambridge, and 8,000 tiles from inside NYC parks. Except for Washington, DC, where the tiles are 512x512 pixels, the rest of the tiles come in 256x256 pixels. We choose zoom level 20 for the 256x256 pixel tiles, which corresponds to the zoom level 19 for 512x512 pixels tiles, where each pixel of the image represents 0.19 meters on the surface of the earth. Our experiments training the model with both sizes showed that the model would perform better using 512x512 pixel input images (an increase of roughly 12% in mIoU). Hence, we used the tool to stitch every four neighboring 256x265 pixel tiles to get 512x512 pixel images, creating a total of 20,000 tiles.

3.1.2. *Planimetric GIS data*

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 suitable for creating annotation masks–a priori known and accurate raster polygons that describe the features we seek to automatically detect using semantic segmentation models. An annotation label is like a reference map that corresponds to a given tile, where each pixel color represents the class to which the corresponding pixel in the image belongs (Figure 2(b,c,e,d)).

To prepare the annotation 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. Reliance on existing GIS datasets allows us to prepare large-scale annotation labels using available 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 the available planimetric GIS data, rasterized 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.

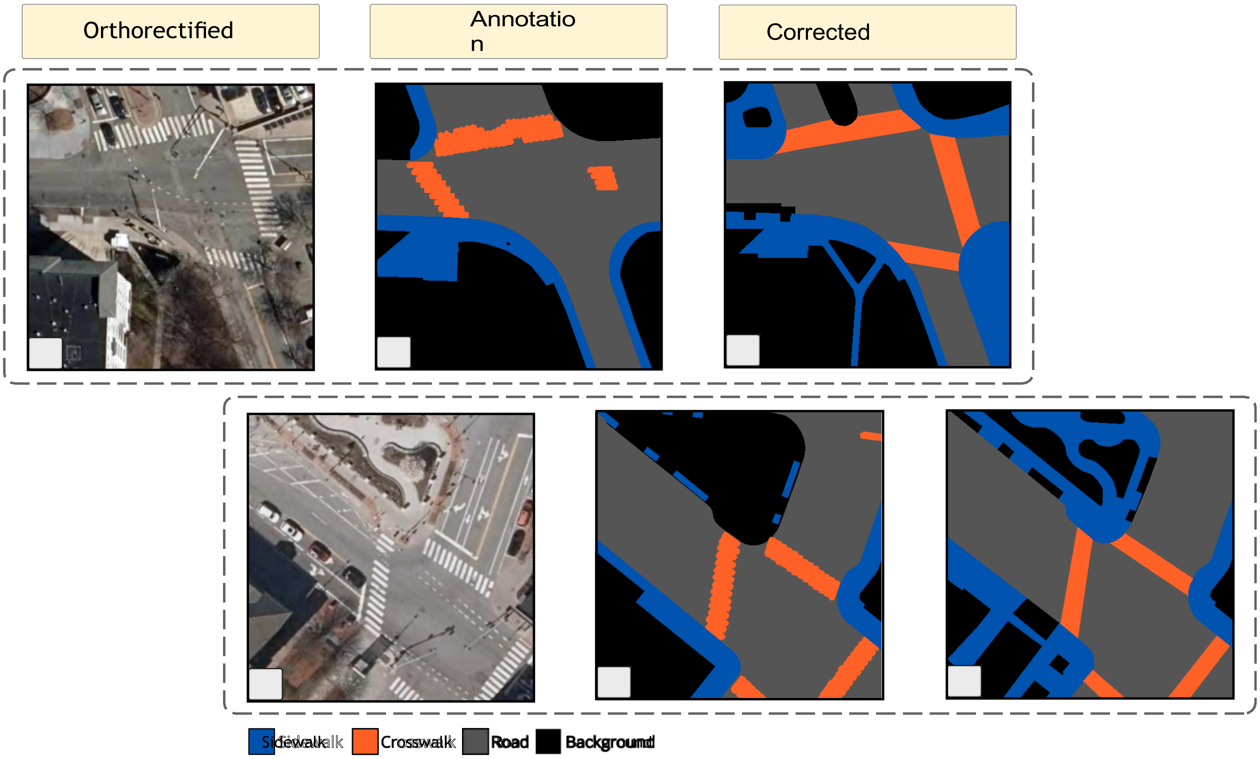

Fig. 2. Examples of the mismatches between the aerial image and the annotation label created from the official data. The manually corrected annotation labels are shown in the last column.

However, challenges remain in creating accurate and consistent training data. The first challenge arises from the lack of consis- tency 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 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 Figure 2, official GIS data can contain numerous errors. Human adjustment and correction may be necessary to bring ground truth annotation labels into alignment with the image data. To achieve that, our research team manually corrected 2,500 tiles of the 12,000 training set, 1,620 image tiles out of 4,000 tiles that were used as our validation set, and 1,500 tiles out of 4,000 test set tiles.

3.2. *Methods*

Tile2Net adopts a multi-scale attention model for detecting pedestrian infrastructure from aerial imagery: sidewalks, cross- walks, stairs, and footpaths that may be separated from streets and roadways (e.g., in parks and open spaces). We combine a semantic segmentation approach with a raster-to-polygon conversion process to generate vector shapefiles of pedestrian infrastruc- ture elements and, separately, a polygon-to-centerline conversion process to produce a topologically interconnected network of pedestrian centerlines. The pipeline has two main parts: 1) Detecting street elements from aerial imagery (Figure 3 (a,b)), and 2) Network construction (Figure 3 (c,d)). In the following, we describe our methods in detail.

3.2.1. *Detecting street elements from aerial imagery*

To detect street elements from aerial imagery, Tile2Net allows users to train a pedestrian feature recognition model on custom, locally-specific data. The trained model can then be used to make inference on unlabeled data. For our semantic segmentation task,

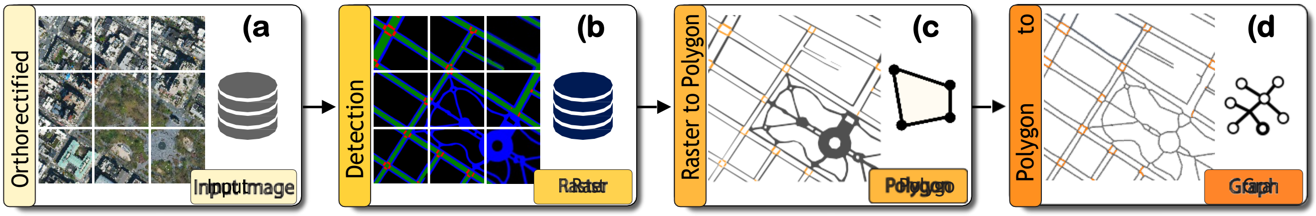

Fig. 3. The proposed network generation pipeline. a) Unlabeled orthorectified tiles are passed through the semantic segmentation model for prediction, b)The model detected sidewalks (blue), crosswalks (red), and roads (green) in the input tiles, c) The sidewalks and crosswalks of the prediction results (raster format) are converted into georeferenced polygons, d) The line representation of the pedestrian network generated from polygons.

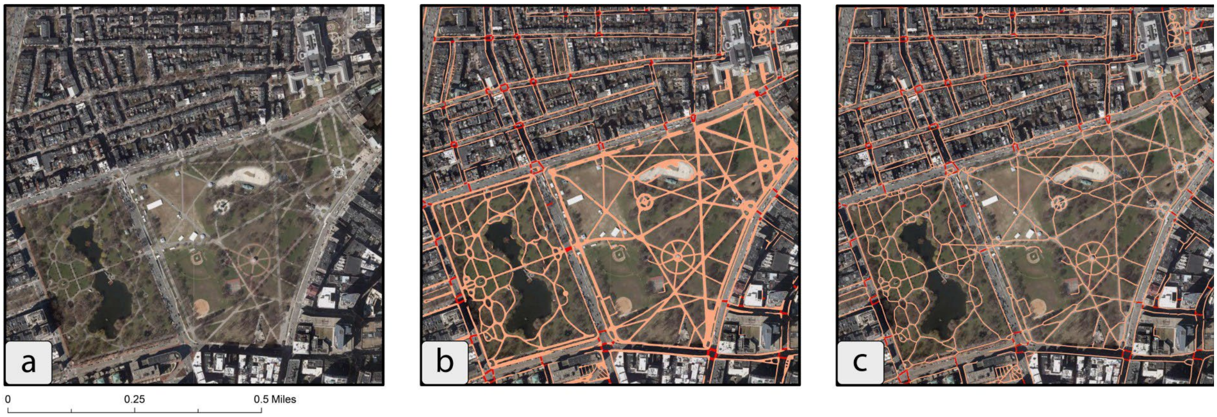

Fig. 4. Boston Commons: a) Aerial image, b) Detected sidewalk and footpath polygons (in orange) and detected crosswalks (in red), c) Fitted sidewalk, crosswalk, and footpath centerlines superimposed on the aerial image.

we adopted the Hierarchical Multi-Scale Attention model (Tao et al., 2020), and used HRNet-W48 Sun et al. (2019c); Wang et al. (2020) with Object-Contextual Representations (Yuan et al., 2019) as the backbone. The computed representation from HRNet- W48 is fed 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 et al., 2019), which accounts for the relationship between pixels instead of only relying on single pixels to calculate the loss.

The 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. After the training phase is completed, the unlabeled orthorectified tiles are passed through the trained model, as shown in Figure 3 (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 (Figure 3 (b)).

3.2.2. *Network creation*

After the pedestrian features were 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; and 2) polygon to centerline conversion to create the final pedestrian network representation. Figure 4 shows the results of these two steps for Boston Commons, which was not part of the training data. Next, we will detail each of these steps.

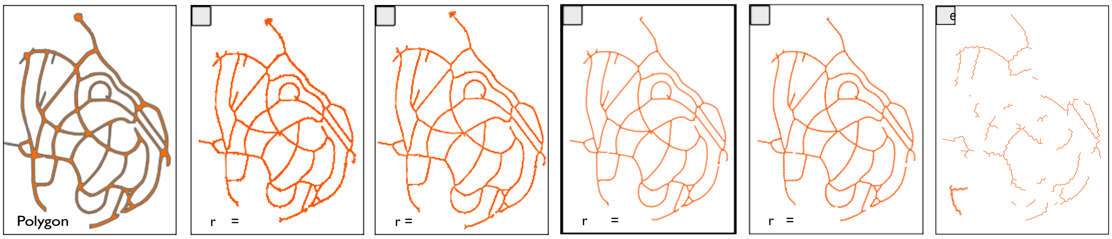

Fig. 5. Impact of different interpolation distances on the resulting centerline created from the input polygon. Small values create extra branches (r=0.5 and r=1) and large values create zigzaggy (r=10) or disjointed lines (r=20). The middle centerline, highlighted with a thicker border, is computed using the interpolation distance computed using our heuristic approach.

3.2.3. *Raster to polygon conversion*

To obtain the vectorized, georeferenced sidewalks, crosswalks, and roads, the detected regions should be converted into poly- gons. To achieve that, we employed connected-component mapping algorithm (Rosenfeld & Pfaltz, 1966; He et al., 2009), 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.

3.2.4. *Polygon to centerline conversion*

In the third and 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 diagram edges not intersecting with the boundary of the object. 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 et al., 2021) as illustrated in Figure 5.

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 Figure 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. To clean and simplify the centerline, we trim branches shorter than an adjustable threshold, which is generally set to half of the average width of the polygon. Crosswalk centerlines were 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.

Table 2. Availability of the official data across different cities. Training: ◦, Evaluation: •

City	**Data type**	**Sidewalk**	**Crosswalk**	**Footpath**
Boston	Polygon	○	-	-
	Centerline	●	-	●
Cambridge	Polygon	○	○	○
	Centerline	●	●	●
Washington DC	Polygon	○	○	○
	Centerline	-	-	-
Manhattan	Polygon	●	-	○
	Centerline	-	-	

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 used R-Tree (Guttman, 1984; Kamel & Faloutsos, 1993) and queried 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. Implementation and Evaluation of Results

This section presents the implementation details and results of using Tile2Net to create city-scale pedestrian networks. We evaluate the performance of our proposed method in two parts. First, we evaluate the results of our semantic segmentation model based on ground truth masks (Section 4.2). Next, we evaluate the accuracy of the constructed maps, both polygons, and centerlines, using the available official data (Section 4.3).Table 2 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 a plain circle, as explained in Section 3.1.

4.1. *Implementation*

The model was trained with a batch size of 16, SGD for the optimizer with polynomial learning rate (Liu et al., 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 downsampling by a factor of two, and a scale of 2 denotes upsampling by a factor of 2 (Tao et al., 2020). We used color augmentation, random horizontal flip, random scaling (0.5x–2.0x), and Gaussian blur on the input tiles to augment the training data and improve the generalizability of the model. The crop size was set to 512x512. The image and annotation 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.

The trained model is then used to make inference to create the city-scale networks; we obtained the tiles corresponding to the bounding box of Boston, Cambridge, Manhattan, and Washington, DC, on zoom level 20. Since smaller tiles result in more disjointed final shapes, we used 1024x1024 pixel tiles stitched using Tile2Net for the inference part. The hierarchical architecture

Table 3. Evaluation metrics on the test set.

Label	IoU	Precision	Recall
Sidewalk	82.67	0.9	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		

of our semantic segmentation network made it possible to choose different scales during the inference. In our experiments using 512x512, 1024x1024, and 2048x2048 pixel tiles during inference, the best results were achieved using 1024x1024 pixel tiles, where the model had enough context to distinguish between different classes.

Tile2Net uses the Geopandas (Jordahl, 2014) and PyGEOS(Wel, Casper van der, 2019) libraries for performing different spatial operations. The raster to polygon conversion was done using the Rasterio library (Gillies et al., 2013). To create the centerlines, we used the Centerline library (Todic, 2016). Momepy (Fleischmann, 2019) was used to handle network cleanups, such as removing the false nodes.

4.2. *Evaluation of the semantic segmentation results*

The trained model outputs four classes in total, two of which were directly used to create the pedestrian networks, i.e., sidewalks and crosswalks, one 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) approach, which is a scale-invariant standard evaluation metric for semantic seg- mentation tasks. 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 since sidewalks and crosswalks comprise a small portion of each image, resulting in a significant class imbalance and an arbitrary high pixel-level accuracy. Table 3 presents the average IoU (mIoU), as well as the class-wise IoU, precision, and recall. The model achieved 84.5% mIoU over all four classes, with sidewalks having 82.7% 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.

4.3. *Evaluation of the constructed maps*

Figure 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. To evaluate the quality of the output vis-a-vis existing official GIS datasets available in each city. We compared both the detected polygons to corresponding city GIS polygons and the detected network segments to a priori known GIS sidewalk networks in each city. Table 2 summarizes the availability of official data across the four cities, and how they were used for both training and evaluation.

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 metrics for detected polygons, showing the total count and the percentage of ground-truth polygons (from the cities' GIS data) that had a matching "detected" polygons spatially intersecting each element. In

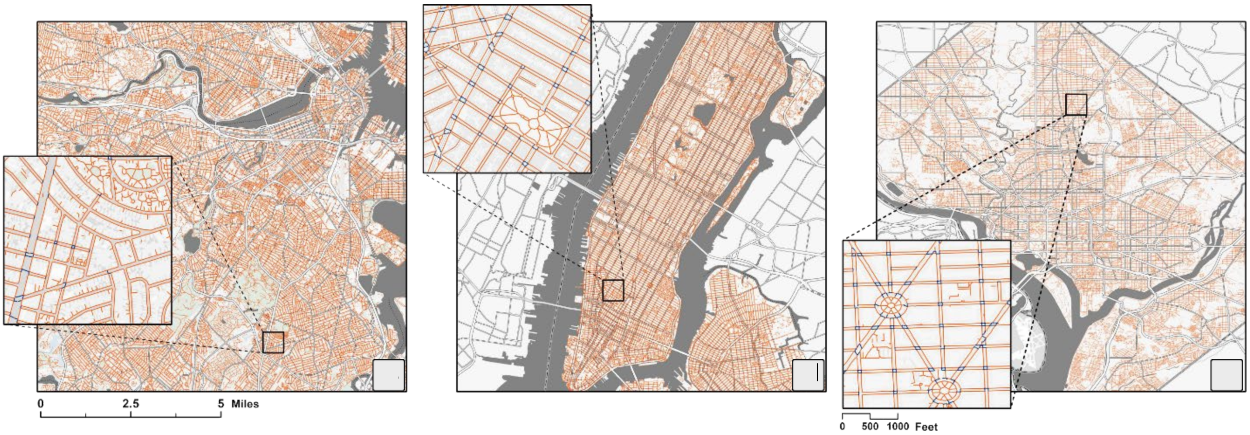

Fig. 6. Model results showing detected sidewalk, crosswalk and footpath centerlines in a) Boston and Cambridge, b) Manhattan and parts of Brooklyn, c) Washington, DC. The maps are shown at the same scale for comparison.

Cambridge, 98.9% of all polygons in official GIS data had overlapped with polygons detected by Tile2Net. In Boston, that number was 98.7%, in Washington, DC, 84.4%, and in Manhattan, 98.2%. 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). This illustrates what percent of the area featured in the official pedestrian polygons overlaps with detected polygons. In Cambridge, 85.9% of the area of official GIS polygons was also covered by detected polygons, 77.9% in Boston, 73.8% in Washington, DC, and 87.5% in Manhattan. Figure 4 illustrates an overlay of detected polygons and network segments in a part of Boston covering the Boston Commons and some blocks around it.

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 2). 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 compute the percentage change of the sidewalk and crosswalk cen- terlines by intersecting the centerlines of each class with the more recent polygon data of that class. We manually examined all the mismatch cases and removed the false positives. Our analysis showed a 23% change from 2011 to 2018 in crosswalks, while sidewalks change was 9.2%, which shows the relative stability of the fixed features such as sidewalks over time. To perform the evaluation, we marked the centroid of each network segment from corresponding city datasets and buffered the centroid by four meters (corresponding to 95th percentile sidewalk width in Boston) to check how many ground-truth network segments have a detected segment within a 4-meter distance of their centroid. 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.

Table 4. Comparison of polygon accuracy results in Cambridge, MA, Boston, MA, New York City, NY, and Washington, DC. The % 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 % detected weighted by area. The mean area overlap % row reports how close in area (from 0-100%) 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	4,684
Match (overlaps with detected)	17,327	24,288	43,963	4,602
% Detected	98.92%	98.72%	84.40%	98.25%
% Detected (weighted by area)	99.62%	99.39%	97.48%	99.91%
Mean area overlap % (weighted by area)	85.9%	77.9%	73.8%	87.5%

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

City	Measures	All	Sidewalk	Crosswalk	Footpath
Cambridge	Official element count	12,792	5,007	2,414	5,371
	Match (within 4m of centroid)	10,631	4,735	2,197	3,699
	% Match	83.1%	94.6%	91.0%	68.9%
Boston	Official element count	110,031	54,864	11,223	37,023
	Match (within 4m of centroid)	86,372	49,806	10,051	23,978
	% Match	78.5%	90.8%	89.6%	64.8%
Manhattan	Official element count	-	-	-	6,239
	Match (within 4m of centroid)	-	-	-	5,309
	% Match	-	-	-	85.1%

Table 6. Network accuracy evaluation in Washington, DC.

City	Measure	All
Washington, DC	OSM swlk element count	11,317
	Match (within 4m of centroid)	8,703
	% Match	76.9%

In Cambridge, our model matched 83.1% of all segments, with notable heterogeneity among different types of elements. Among sidewalks, 94.6% of centerlines had a corresponding detected segment, among crosswalks, 91.0%, and among footpaths, 68.9%. 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). 90.8% of all sidewalk segments in city GIS data and 89.6% of all crosswalks were matched by our results. Footpath matching was again notably lower at 64.8%. In Manhattan, NY, we only had official footpath networks (in parks) available from the city's open data repository. Here, 85.1% of official footpath segments had a corresponding detected segment within a four-meter buffer of their centroid. In Washington, we did not find any official sidewalk centerlines.

For Washington, DC, the comparison could only be performed on more limited data. In Washington, DC, we did not find any official sidewalk centerlines and instead performed the comparison with the available OpenStreetMap sidewalk segments. The results are shown in Table 6. A somewhat lower matching rate with OSM networks was expected and confirmed by the 76.9% 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 notably lower share of footpaths in parks, green areas, and other public spaces, a few caveats need to 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 concurrent with the aerial 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 aerial 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 aerial 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. As also explained in 4.3, the percentage change between the data created based on the 2010 flyovers and the 2018 polygon data was 9.2% for sidewalks and 23% for crosswalks.

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 con- ditions 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. The relatively lower detection accuracy of footpaths is attributable to several factors. On the one hand, feature detection from aerial imagery is hampered by significantly higher levels of tree cover and other vegetation obstructions over footpaths found in parks, courtyards, and campuses. Second, 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 also benefit from further improvement. The network geometry improvements can be categorized into three separate areas. First, as also mentioned in Section 3.2.4, 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. 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. We were partly able to adjust this with automated post-processing routines, but further refinements would be desirable to output continuous centerline networks. There is an extensive body of literature on various skeletonization algorithms (Saha et al., 2016), with some focusing solely on creating the centerlines of the elongated polygons (Lewandowicz & Flisek, 2020; Haunert & Sester, 2008). However, finding the optimal interpolation distance value is beyond the scope of the current research, but as a future direction, we are planning to work on developing algorithms tailored for creating the centerlines of the pedestrian infrastructure.

Second, the resulting network segments are currently not optimized to form singular nodes or endpoints at intersections. Some detected line segments often converge near street corners, forming redundant intersections. This can be addressed in future work

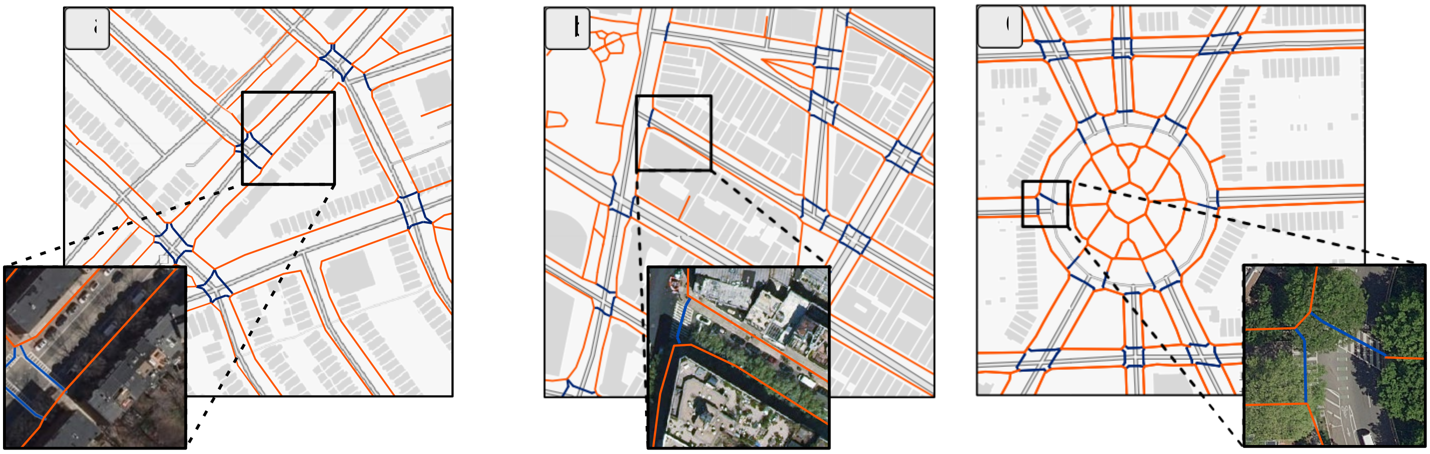

Fig. 7. Mapping obstructed pedestrian facilities in different cities: a) Cambridge, MA. - sidewalks are mapped as continuous despite the heavy shadow, b) Manhattan - sidewalks and crosswalks obstructed by tree foliage and shadow are detected and mapped, c) Washington, DC. - crosswalks covered by vegetation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

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. This threshold distance would ideally be determined contextually, depending on the street widths in each area.

Third, though most computer vision solutions are fundamentally unable to detect sidewalk spatial elements where visual ob- structions 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 sur- prisingly well in occluded areas. Figure 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 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, etc.), 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.

Moreover, in the future, we plan to add additional classes such as driveways, curbs, stairs, and separating public and private footpaths to our detection model. 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 aerial 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).

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 treated as parts of sidewalks. The same problem exists for curb extensions, medians, driveways, and curb-cuts. Moreover, the edges of the road and sidewalk polygons overlap and, in multiple instances, in GIS ground truth data. 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 widely 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, it could have resulted in systemic underdetection in Boston and Cambridge. Such differences are bound to be much 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., it is crucial to train the model specifically for each region.

6. Conclusion

In this paper, we presented Tile2Net, a solution that is able to create accurate pedestrian networks from aerial imagery in an end-to-end fashion. We pilot tested the approach in New York City, Washington, DC, Boston, and Cambridge, with varying street network geometries, building shadow densities, and tree covers and reported on the quality and accuracy of the approach. The resulting networks are created using the most recent orthorectified images, hence, more closely reflect the current urban form and pedestrian infrastructure. 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 et al. (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 (Erath et al., 2015; Sevtsuk et al., 2021; Basu et al., 2022).

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.

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

Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T. et al. (2021). Big self-supervised models advance medical image classification. *arXiv preprint arXiv:2101.05224*, .

Baker, C. D., & Hou, Q. (2019). *Improving Pedestrian Infrastructure Inventory in Massachusetts Using Mobile LiDAR*. Technical Report Massachusetts Department of Transportation (Mass DOT).

Balado, J., D´ıaz-Vilarin˜o, L., Arias, P., & Gonza´lez-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., Madden, 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.

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

Brezina, T., Graser, A., & Leth, U. (2017). Geometric methods for estimating representative sidewalk widths applied to vienna’s streetscape surfaces database.

Journal of Geographical Systems, *19*, 157–174.

Cambra, P. J., Gonc¸alves, 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:

ttps://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.

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.

Cura, R., Perret, J., & Paparoditis, N. (2018). A state of the art of urban reconstruction: street, street network, vegetation, urban feature. *arXiv preprint arXiv:1803.04332*, .

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.

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 and Design*, *43*, 130–151.

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., Ordo´n˜ez 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. In *BMVC* (p. 2). Citeseer volume 1.

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).

Fleischmann, M. (2019). Momepy: Urban morphology measuring toolkit. *Journal of Open Source Software*, *4*, 1807.

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.

Gillies, S., Ward, B., Petersen, A. et al. (2013). Rasterio: Geospatial raster i/o for python programmers. Available online: https://github.com/rasterio/rasterio.

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).

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. (2019a). 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. (2019b). 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.

Horva´th, 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 Jr, R. M. (2021). Sidewalk measurements from satellite images: Preliminary findings. In *Spatial Data Science Symposium 2021 Short Paper Proceedings*. doi:https://doi.org/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*, (p. 103630).

Hou, Q., & Ai, C. (2020). A network-level sidewalk inventory method using mobile lidar and deep learning. *Transportation Research Part C: Emerging Technologies*, *119*, 102772.

Huang, Q., Xia, C., Wu, C., Li, S., Wang, Y., Song, Y., & Kuo, C.-C. J. (2017). Semantic segmentation with reverse attention. *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*, 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. *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., Cerin, E., Adlakha, D., Lowe, M., & Giles-Corti, B. (2021). A generalized framework for measuring pedestrian accessibility around the world using open data. *Geographical Analysis*, .

Liu, W., Rabinovich, A., & Berg, A. C. (2015). Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, .

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*, (p. 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. Qin, H., Curtin, K. M., & Rice, M. T. (2018). Pedestrian network repair with spatial optimization models and geocrowdsourced data. *GeoJournal*, *83*, 347–364. 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., Dash, A., Chen, S., Li, A., Hara, K., & 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., Mu, Y., Wang, X., Liu, W., & 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/.

Todic, F. (2016). Centerline: Calculate the polygon's centerline. https://github.com/fitodic/centerline.

Treccani, D., D´ıaz-Vilarin˜o, L., & Adami, A. (2021). Sidewalk detection and pavement characterisation in historic urban environments from point clouds: Prelimi- nary 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., Liu, D., Mu, Y., Tan, M., Wang, X. et al. (2020). Deep high-resolution representation learning for visual recognition. *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*, 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.

Wel, Casper van der (2019). PyGEOS. https://github.com/pygeos/pygeos.

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. *arXiv preprint arXiv:1909.11065*, .

Zapata-Diomedi, B., Boulange´, 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. *arXiv preprint arXiv:1910.12037*, .

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*, 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).