C2SMART
CONNECTED CITIES WITH SMART TRANSPORTATION

A USDOT University Transportation Center

New York University

Rutgers University

University of Washington

The University of Texas at El Paso

City College of New York

EXPLORING AI-BASED VIDEO SEGMENTATION AND SALIENCY COMPUTATION TO OPTIMIZE IMAGERY-ACQUISITION FROM MOVING VEHICLES

October 2023

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
4. Exploring AI-based Video Segmentation and Saliency Computation to Optimize Imagery-acquisition from Moving Vehicles		5. Report Date October 2023
		6. Performing Organization Code:
7. Author(s) Claudio Silva, Kaan Ozbay, Joao Rulff, Jianzhe Lin, Maryam Hosseini, Eric Tokuda		8. Performing Organization Report No.
9. Performing Organization Name and Address Connected Cities for Smart Mobility towards Accessible and Resilient Transportation Center (C2SMART), 6 Metrotech Center, 4th Floor, NYU Tandon School of Engineering, Brooklyn, NY, 11201, United States		10. Work Unit No. 11. Contract or Grant No. 69A3551747119
12. Sponsoring Agency Name and Address Office of Research, Development, and Technology Federal Highway Administration 6300 Georgetown Pike McLean, VA 22101-2296		13. Type of Report and Period Final report, 3/1/21-10/31/23 14. Sponsoring Agency Code
15. Supplementary Notes		

16. Abstract

In this study, a new dataset and tool were generated describing and mapping street-level infrastructure. The presented dataset, StreetAware, is generated from more than 7 hours of synchronized data collected at urban intersections by specialized Reconfigurable Environmental Intelligence Platform (REIP) sensors developed by the Visualization and Data Analytics (VIDA) Research Center at NYU. To demonstrate these key features of the data, we present four uses for the data that are not possible on many existing datasets. (1) to track objects using the multiple perspectives of multiple cameras from both audio (sound-based localization) and visual modes, (2) to associate audio events with their respective visual representations using audio and video, (3) to track the amount of each type of object in a scene over time, i.e., occupancy, and (4) to measure the speed of a pedestrian while crossing a street using multiple synchronized views and the high-resolution capability of the cameras. Next, 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. This work is 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 datasets we enable less resourceful cities to create datasets describing pedestrian environments which otherwise would not be possible at a comparable cost and time. The results of this work provide a new rich dataset of street level information as well as a tool for mapping pedestrian infrastructure. Both provide a significant contribution for researchers and policymakers working on making the street more accessible and usable.

17. Key Words	18. Distribution Statement No restrictions. This document is available to the public through the National Technical Information Service, Springfield, VA 22161. http://www.ntis.gov		
19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 32	22. Price

Form DOT F 1700.7 (8-72) Reproduction of completed page author

Exploring AI-based Video Segmentation and Saliency Computation to Optimize Imagery-acquisition from Moving Vehicles

PI: Claudio Silva
New York University
0000-0003-2452-2295

Maryam Hosseini
New York University
0000-0001-8329-4638

Co PI: Kaan Ozbay
New York University
0000-0001-7909-6532

Jianzhe Lin
New York University
0000-0001-8456-0475

Joao Rulff de Costa
New York University
0000-0003-3341-7059

Eric K. Tokuda
University of Sao Paolo
0000-0002-6159-2500

C2SMART Center is a USDOT Tier 1 University Transportation Center taking on some of today's most pressing urban mobility challenges. Some of the areas C2SMART focuses on include:

Urban Mobility and Connected Citizens

Urban Analytics for Smart Cities

Resilient, Smart, & Secure Infrastructure

Disruptive Technologies and their impacts on transportation systems. Our aim is to develop innovative solutions to accelerate technology transfer from the research phase to the real world.

Unconventional Big Data Applications from field tests and non-traditional sensing technologies for decision-makers to address a wide range of urban mobility problems with the best information available.

Impactful Engagement overcoming institutional barriers to innovation to hear and meet the needs of city and state stakeholders, including government agencies, policy makers, the private sector, non-profit organizations, and entrepreneurs.

Forward-thinking Training and Development dedicated to training the workforce of tomorrow to deal with new mobility problems in ways that are not covered in existing transportation curricula.

Led by New York University's Tandon School of Engineering, **C2SMART** is a consortium of leading research universities, including Rutgers University, University of Washington, the University of Texas at El Paso, and The City College of NY.

Visit c2smart.engineering.nyu.edu to learn more.

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, by a grant from the U.S. Department of Transportation's University Transportation Centers Program. However, the U.S. Government assumes no liability for the contents or use thereof.

Executive Summary

In this study, a new dataset and tool were generated describing and mapping street-level infrastructure. The presented dataset, StreetAware, is generated from more than 7 hours of synchronized data collected at urban intersections by specialized Reconfigurable Environmental Intelligence Platform (REIP) sensors developed by the Visualization and Data Analytics (VIDA) Research Center at NYU.

To demonstrate these key features of the data, we present four uses for the data that are not possible on many existing datasets. (1) to track objects using the multiple perspectives of multiple cameras from both audio (sound-based localization) and visual modes, (2) to associate audio events with their respective visual representations using audio and video, (3) to track the amount of each type of object in a scene over time, i.e., occupancy, and (4) to measure the speed of a pedestrian while crossing a street using multiple synchronized views and the high-resolution capability of the cameras.

Next, 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. This work is 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 datasets we enable less resourceful cities to create datasets describing pedestrian environments which otherwise would not be possible at a comparable cost and time.

The results of this work provide a new rich dataset of street level information as well as a tool for mapping pedestrian infrastructure. Both provide a significant contribution for researchers and policymakers working on making the street more accessible and usable.

Table of Contents

List of Figures

List of Tables

Section 1. StreetAware: A High-Resolution Synchronized Multimodal Urban Scene Dataset

Introduction

Driven by continuous improvements in computational resources, bandwidth optimization, and latency, activity-rich traffic intersections have been implicated as excellent locations for smart city intelligence nodes. Audio and video sensors located at intersections are, thus, capable of generating large amounts of data. Concomitantly, deep learning and edge computation of these data allow, in real-time, the geospatial mapping and analysis of urban intersection environments, including moving entities, such as pedestrians and vehicles. Intersections are some of the most critical areas for both drivers and pedestrians. They are where vehicles and pedestrian paths most frequently cross. Globally, pedestrians represent 23% of the 1.35 million worldwide road traffic deaths every year with most events occurring at pedestrian crossings. Thus, predicting pedestrian trajectories at intersections and communicating this information to drivers or assisted/autonomous vehicles could help mitigate such accidents. Understanding an intersection scene has significant implications for self-driving vehicles in particular. Figure 1 outlines the concept of enhancing the safety of traffic participants by providing real-time insights into out-of-sight events at intersections using a combination of multimodal sensing and edge and in-vehicle computing. In this example, pedestrians and (semi)autonomous cars are sensed by sight (cameras) and sound (microphones) at intersections, and information is relayed to each car's self-driving system. In this process, edge computing and via the cloud helps extract, in real-time, useful information from the data.

Within the navigation system of an autonomous vehicle, it is important that its control system has detailed, accurate, and reliable information as it approaches such a scene to determine, for instance, the number of road entries into an upcoming crossing or to predict the pedestrian or vehicle trajectories it is on a collision course with. For such purposes, urban analytical data should have high precision, granularity, and variation (such as multiple perspectives of the same area) to be effectively useful.

In this study, we present more than 7 hours of synchronized data collected at urban intersections by specialized Reconfigurable Environmental Intelligence Platform (REIP) sensors developed by the Visualization and Data Analytics (VIDA) Research Center at NYU REIP sensors are capable of dual 5 MP video recording at 15 fps as well as the 12-channel audio at 48 kHz for the recording of pedestrian and vehicle traffic at various locations. We selected three intersections in Brooklyn, New York, with diverse demographic, urban fabric, and built environment profiles and equipped each with four REIP sensors. The sensors were placed at each corner of the intersection and recorded the dynamic of pedestrian and vehicle interaction for several 40 min sessions, resulting in a total of approximated 2 TB of raw audiovisual data. The data were synchronized across all sensors with high accuracy for both modalities (one video frame and one audio sample, respectively) using a custom time synchronization solution

detailed later. High-synchronization is important so that events that happen across cameras and between video and audio can be viewed and analyzed together with reduced effort, and with confidence, those events actually occurred at the time inscribed in the data.

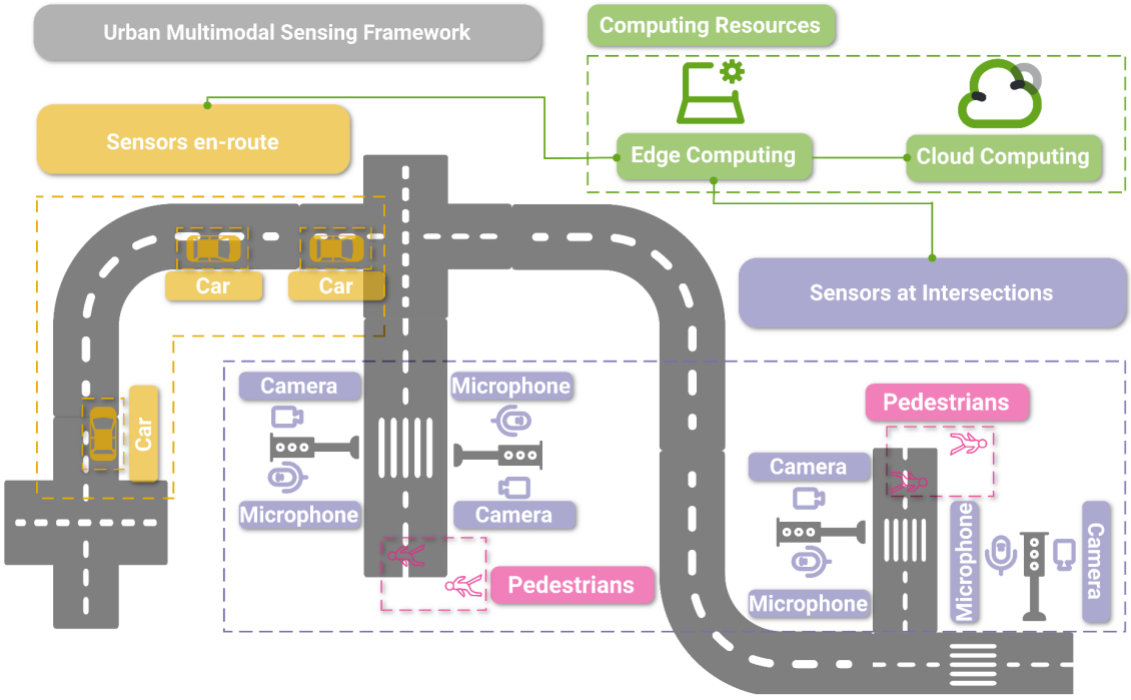

Figure 1. Illustration of the basic concept of combining multimodal sensors at critical nodes (e.g., intersections) with on-device and in-vehicle computing capabilities to provide greater awareness to urban traffic participants.

The presented dataset, which we call StreetAware, is unique to other street-level datasets such as Google Street View because of the following combination of characteristics:

- Multimodal: video, audio, LiDAR;
- Multi-angular: four perspectives;
- High-resolution video: 2592 x 1944 pixels;
- Synchronization across videos and audio streams;
- Fully anonymized: human faces blurred.

To demonstrate these key features of the data, we present four uses for the data that are not possible on many existing datasets. (1) to track objects using the multiple perspectives of multiple cameras from both audio (sound-based localization) and visual modes, (2) to associate audio events with their respective visual representations using audio and video, (3) to track the amount of each type of object in a scene over time, i.e., occupancy, and (4) to measure the speed of a pedestrian while crossing a street using multiple synchronized views and the high-resolution capability of the cameras.

The contributions listed in this chapter include:

- The StreetAware dataset, which contains multiple data modalities and multiple synchronized high-resolution video viewpoints in a single dataset;
- A new method to synchronize high-sample rate audio streams;
- A demonstration of use cases that would not be possible without the combination of features contained in the dataset;

The data presented here will allow other researchers to carry out unique applications of machine learning to urban street-level data, such as pedestrian-vehicle interaction modeling and pedestrian attribute recognition. Such analysis can subsequently help inform policy and design decisions made in the context of urban sensing and smart cities, including accessibility-aware design and Vision Zero initiatives. Among the other possibilities we discuss later, further analysis of our data can also shed light on the optimal configuration needed to record and analyze street-level urban data.

Related Work

Datasets

A handful of related datasets exist. The first is the popular Google Street View, combining street-level photography with navigation technology. Publicly available but not entirely free, Google Maps Street View includes an API and extensive street-level image coverage throughout much of the World's roadways. Unlike StreetAware, Google Street View is a collection of disparate images instead of stationary video recordings of specific places. Moreover, Google Street View often has multiple viewpoints that are in close proximity to one another, but they do not overlap in time. Therefore, synchronization across multiple views is not possible. Another dataset is Mapillary. Mapillary street-level sequences contain more than 1.6 million vehicle-mounted camera images from 30 major cities, distinct cameras, and different viewpoints and capture times, spanning all seasons over a nine-year period. All images are geolocated with GPS and compass, and feature high-level attributes such as road type. Again, these data are not video or synchronized and do not include audio. The next dataset is Urban Mosaic, which is a tool for exploring the urban environment through a spatially and temporally dense data set of 7.7 million street-level images of New York City captured over the period of one year. Similarly, these data are image-only and unsynchronized across views. Another street-level urban data set is SONYC. SONYC consists of 150 million audio recording samples from an acoustic sensor network and is aimed at the development and evaluation of machine listening systems for spatiotemporal urban noise monitoring. However, SONYC does not contain visual information. Finally, there is Urban Sound & Sight (Urbansas), which consists of 12 h of unlabeled audio and video data along with 3 h of manually annotated data, but does not contain multiple views. StreetAware is unique in that it combines stationary, multi-perspective, high-resolution video and audio in a synchronized fashion.

Deep Learning Applications

A number of recent studies have explored the use of deep learning for detecting and analyzing objects in street-level audio and video data. A study by Zhang et al. developed an approach to automatically detect road objects and place them in their correct geolocations from street-level images, relying on two convolutional neural networks to segment and classify. Doiron et al. showed the potential for computer vision and street-level imagery to help researchers study patterns of active transportation and other health-related behaviors and exposures. Using 1.15 million Google Street View (GSV) images in seven Canadian cities, the authors applied PSPnet, and YOLOv3 to extract data on people, bicycles, buildings, sidewalks, open sky, and vegetation to create associations between urban features and walk-to-work rates. Charitidis et al. released a paper in 2023 in which they utilized several state-of-the-art computer vision approaches for object detection. Object detection systems have also been specifically developed for the collection and analysis of street-level imagery in real-time. In "Smart City Intersections: Intelligence Nodes for Future Metropolises", Kostec et al. detail intersections as intelligence nodes using high-bandwidth, low-latency services for monitoring pedestrians and cloud-connected vehicles in real-time.

Pedestrian speed and trajectory prediction are some of the primary computer vision goals in the urban data analytical community, especially in the field of advanced driver assistance systems. The performance of state-of-the-art pedestrian behavior modeling benefits from recent advancements in sensors and the growing availability of large amounts of data. A study by Kuo et al. compared estimations of pedestrian speed from a classical model and a neural network in corridor and bottleneck experiments, with results showing that the neural network can better differentiate the two geometries and more accurately estimate pedestrian speed. Ahmed et al. sought to use a fast region-convolutional neural network (Fast R-CNN), and a Single Shot Detector (SSD) for pedestrian and cyclist detection based on the idea that automated tracking, motion modeling, and pose estimation of pedestrians can allow for a successful and accurate method of intent estimation for autonomous vehicles. Other related studies in the literature include applying deep learning techniques for the prediction of pedestrian behavior on crossings with countdown signal timers, mapping road safety from street view imagery using an R-CNN, and identifying hazard scenarios of non-motorized transportation users through deep learning and street view images in Nanjing, China.

The StreetAware Dataset

REIP Sensors

A multi-view requirement of our data collection could easily be satisfied with off-the-shelf video surveillance systems that often include a set of wireless IP cameras. These cameras transmit their video feeds to a central data storage location (in the form of a local hard drive) which can sometimes be

synchronized with a cloud but is not required for the system's operation. The cameras also include a night mode which can prove beneficial during low-light conditions. However, these cameras rarely provide audio because of privacy concerns and rely on manually configured timing information or NTP (network time protocol) for time-stamping of the video. The latter is a significant barrier to a multi-view analysis of fast-moving objects such as cars. A car traveling at 40 mph covers more than a meter of ground per frame when recorded at 15 fps. Therefore, frame-accurate video synchronization is also a requirement for our dataset and, unfortunately, cannot be met with off-the-shelf security cameras, which are also often operating at reduced frame rates due to limited storage.

There exist commercial motion tracking systems that use high-speed cameras synchronized by NTP. Although these cameras provide high temporal resolution and accuracy for video, they would be insufficient for the synchronization of audio data (sub-millisecond timing accuracy required). Furthermore, such cameras are typically designed for indoor infrared light imaging, are costly, and rely on an Ethernet interface for synchronization and data transfer which makes them impractical for larger-scale urban deployments. Ultimately, the sensors used in this study are custom-built in our lab. Our sensors (REIP sensors) provide high-resolution video and audio recording with an in-built synchronization solution.

Data Collection

Three intersection locations were selected to acquire the dataset with different road configurations and pedestrian demographics as described below:

- **Commodore Barry Park:** This intersection is adjacent to a public school. It has a low-to-medium frequency of traffic making it an uncrowded intersection.
- **Chase Center:** This intersection is adjacent to the Chase Bank office building within Brooklyn's MetroTech Center. It is also an active pedestrian intersection.
- **DUMBO:** The intersection of Old Fulton Street and Front Street is under the Brooklyn Bridge. Being a tourist destination, this intersection is the busiest of the three. Because of smaller crosswalks and heavy traffic, it provides challenges such as occlusion and a diverse range of pedestrian types.

Overhead map locations and the sensors' positions for the recording sessions at Commodore Barry Park as an example are shown in Figure 2. We used four REIP sensors at each intersection, one placed at each corner of the intersection. We recorded several 30-45 min long sessions at each intersection. Four at Commodore Barry Park, three at Chase Center, and four at DUMBO. This results in about 200 GB of raw audiovisual data recorded by each sensor per location (limited by the sensor's max storage capacity of 250 GB). In total, we collected approximately 2 TB of raw data.

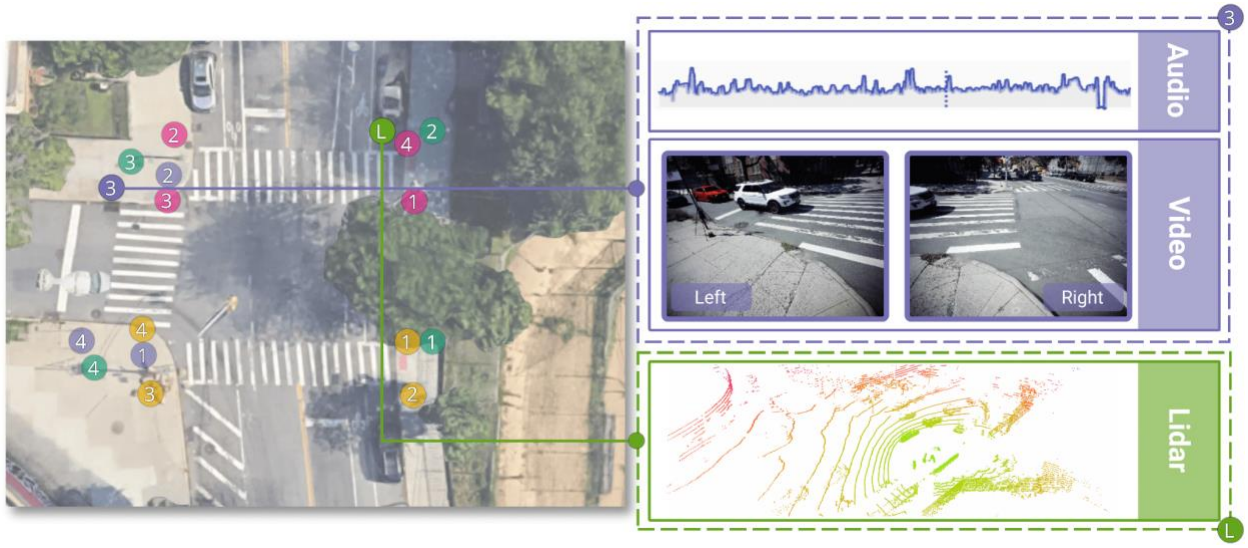

Figure 2. Illustration of the sensor positions and data types at the Commodore Barry Park intersection.

The data acquisition pipeline of the sensors is shown in Figure 3. The output of the sensor's data acquisition pipeline contains three types of data: 500 MB chunks of video data (approximately one minute of recording, depending on the intersection), JSON files containing batches of timestamps for each frame in the video data chunks, and 5-second-long chunks of audio data with its timing information embedded as extra audio channels. We spare the users from working with the sensor's raw data by preprocessing it, including anonymization and synchronization. We also use a space-efficient video codec H.264 instead of the camera's original MJPEG data stream. Table 1 summarizes the specifications of the processed dataset that we are releasing.

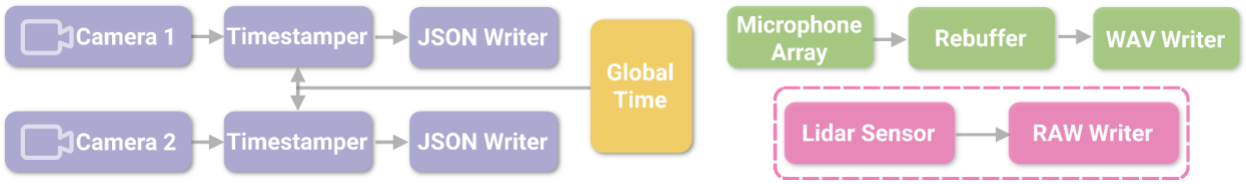

Figure 3. The sensor's data acquisition pipeline built using software blocks available in the REIP SDK. It contains separate tasks for each camera and the microphone array. The LiDAR data acquisition is performed on a separate machine (orchestrator laptop).

Table 1: Dataset specifications after processing, featuring 3 data modalities (audio, video, and LiDAR) with synchronized footage.

Feature	Specification
Number of geographic locations	3
Number of recording sessions	11
Typical recording length	30-45 min
Total unique footage time	465 min (7.75 h)
Total number of image frames	~403,000
Video resolution	2592 X 1944 pixels
Number of data modalities	3
Synchronized and anonymized	True
Video synchronization tolerance	2 frames
Audio synchronization tolerance	1 sample
Total audio & video size	236 GB
Total LiDAR size	291 GB
Total size	527 GB

Data Synchronization

In this section, we detail our synchronization techniques, first for audio, then for video data modality. The synchronization techniques are independent for each modality.

Audio: Audio synchronization is a challenging task because audio data is being sampled at a very high rate, 48 kHz in the case of our sensors. Furthermore, the speed of sound wave propagation in the air is 343 m/s which translates into a synchronization accuracy requirement of less than one millisecond, across all sensors, for any meaningful audio-based sound source locations to work. Such accuracy cannot be achieved by simply attaching a timestamp to the chunks of audio provided by the driver because of the large 'jitter' of such timestamps caused by the random operating system (OS) interrupts on the computing platform. Therefore, the synchronization information must be embedded into the audio data itself before it even makes it to the audio driver of the OS. In this subsection, we introduce a novel method for high-accuracy audio synchronization by means of embedding a special signal into a dedicated audio channel of the audio interface.

The radio module of each sensor is receiving a global timestamp from a master radio transmitting it at a rate of 1200 Hz. Unlike the operating system of the computing platform, the microcontroller operating the radio module via Serial Peripheral Interface (SPI) can be programmed to process the incoming data packets from the master radio in a very deterministic way. Specifically, the packet arrival interrupt request (IRQ) signal from the radio module causes the MCU to interrupt its current routine and execute a function that decodes the latest timestamp from the data payload of the packet and phase-adjusts the

MCU's internal timer to match the master radio's clock. The jitter of the continuously adjusted slave clock is less than 1 µs with the nRF24L01+ 2.4 GHz radio module. The timer in turn generates a special synchronization signal connected to one of the inputs of the MCHStreamer device that we use as an audio interface.

We are using a simple UART-like serial protocol with one start bit, a 32-bit payload, and a more than 200 audio sample-long stop bit to generate the audio synchronization signal. The start bit and payload bits are five audio samples wide for more reliable encoding. Such a signal is easy to decode during audio processing, and a single audio sample synchronization accuracy is achieved because the start bit of the sequence is aligned with the time of arrival of the timestamp from the master radio, and the micro-controller has a deterministic delay when processing this information.

Video: Video recording is inherently occurring at a much lower sampling rate than audio. For instance, the cameras in REIP sensors are configured to record at 15 fps. That corresponds to an approximately 67 ms period between consecutive frames. The radio module receives a new global timestamp every 0.83 ms, which is almost two orders of magnitude more frequent. Therefore, it makes sense for video recording to timestamp each frame as it is being received by the driver and calibrates the latency between the moment of assignment of this timestamp and when the frame is actually exposed instead of inventing a way of embedding the timing information directly into the image data during exposure as we did for audio. However, this approach comes with new challenges, such as timestamp jitter and lost frames.

There are three timestamps assigned to each video frame: (1) the GStreamer timestamp, which starts from zero and is defined by the camera driver upon arrival of the frame into the queue from the USB, (2) the Python timestamp representing the current system time which is added using the time.time() function when the frame is released by GStreamer into the data acquisition and processing pipeline powered by REIP SDK, and (3) the Global timestamp added to the frame metadata at the same time as the Python timestamp which is the latest global timestamp communicated to the global time block from the MCU via USB 1.1 interface, introducing extra jitter.

We developed a method for reducing the jitter of global timestamps to virtually zero before rendering the synchronized video streams (Figure 4). The main source of timestamp jitter is operating system interrupts that happen when the computing platform, for example, needs to process various I/O events or perform memory management. That is why GStreamer timestamps have the least amount of jitter because they are defined when the OS handles USB 2.0 data transfers from the camera. That is also why we are starting with GStreamer timestamps to reliably detect if and when there are any frames lost by looking for gaps larger than the expected period of the camera's frame rate. After correcting for lost frames, we then convert these timestamps into a global timeline through a couple of regressions incorporating the information from other types of timestamps without adding jitter.

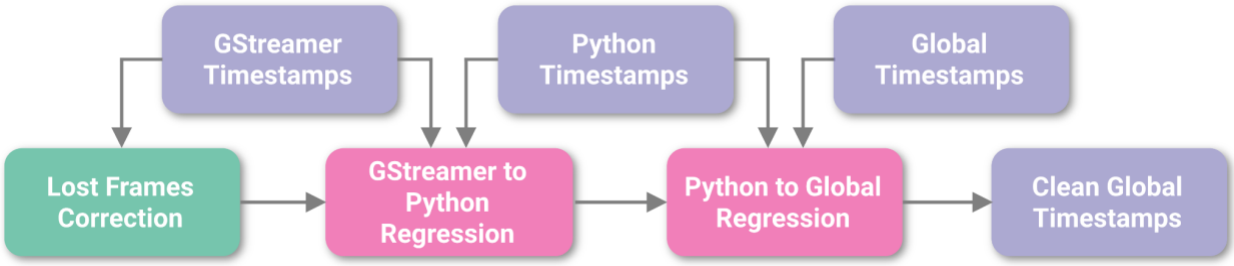

Figure 4. Diagram illustrating timestamp processing.

We start with the least jittery GStreamer timestamps and identify any lost frames so that we can reconstruct the original timeline and average period for the saved frames. We then convert these reference timestamps into a global timeline through a series of regression steps that incorporate the information from other kinds of timestamps without adding jitter.

In addition to correcting for lost frames and eliminating jitter, our method is also fixing any queue overflow issues that often result in the jamming of multiple frames one after another with very similar Python and Global timestamps. This happens when the queue is emptied out quickly after a prolonged operating system interrupt. Another less common issue is when the frames saved to the disk get corrupted due to high data flow or during the copying of the data from the sensors to a server. The solution requires the corresponding timestamps to be deleted from the metadata, and the associated non-decodable frames are considered lost.

To further validate the video synchronization, we render a surveillance-style mosaic video using processed frames from all eight cameras at a given intersection and a global timeline produced by the synchronization of timestamps. Figure 5 shows a mosaic of the frames at the moment at the Chase Center intersection. Essential for many analysis applications, at any given moment, the recording of all traffic remains in sync from multiple viewpoints. Frames for which a camera did not successfully record data are temporarily made black in the camera's associated block in the mosaic.

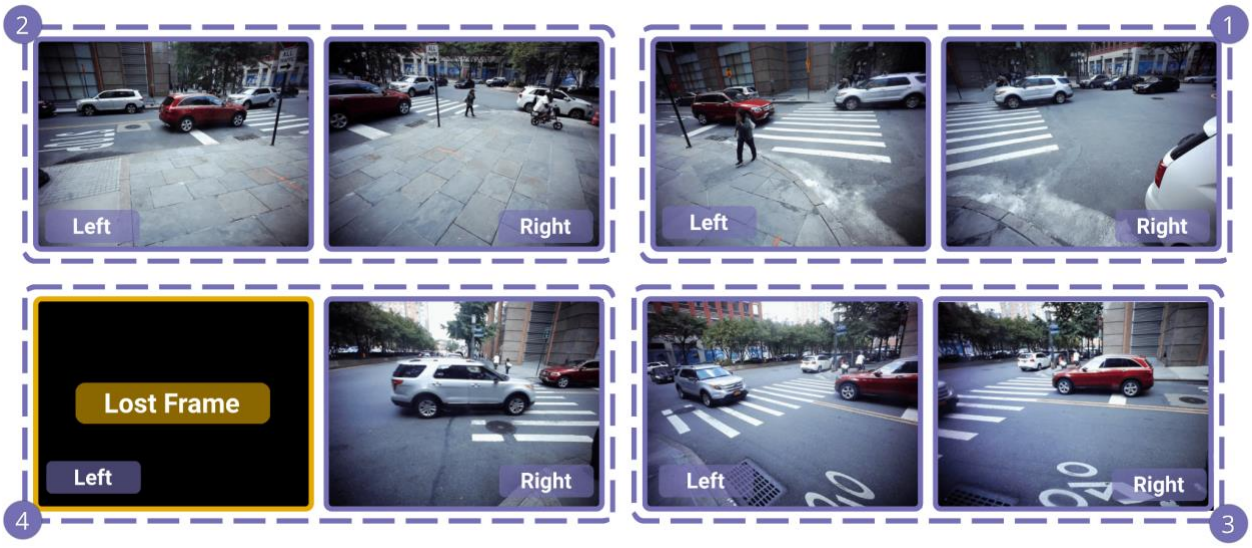

Figure 5. Mosaic rendering of the synchronized frames from recording session one at the Chase Center intersection that can be played as a video.

Four sensors with two cameras each (numbered in the corners) provide eight different views for comprehensive analysis of the intersection. If a camera did not successfully record during a particular frame, its block is turned black, such as the left camera of sensor 4 in this example.

Use Cases

In this section, we will demonstrate four use cases highlighting the potential applications of StreetAware. First, we present two examples of how such data can enhance pedestrians' safety in large urban areas by (1) informing pedestrians and incoming traffic of occluded events using multiple sensors and sound-based localization, and (2) associating audio events (such as the presence of loud engines) with their respective visuals. Second, we present easily quantifiable metrics that can be extracted from the data using the \name infrastructure framework: (3) calculating object counts (occupancy) over time, and (4) measuring pedestrian speed during crosswalk traversal.

Audio Source Localization

As the number of sensors deployed in urban environments increases, cities have the potential to become more human-centered by prioritizing pedestrians over cars. Adaptive traffic and pedestrian signal timing is one example of how an intelligent sensing platform can be used to provide a safer environment for pedestrians. By making the signal timing adjustable to the volume of foot traffic as well as the needs of different groups of people, we can allocate longer signal timing to, for example, crowded intersections or pedestrians with special needs such as the elderly, pregnant, or those with vision

impairments. Most traffic monitoring systems use one or two fixed cameras for each intersection. However, the complex configuration of intersections in large cities makes it challenging for one or two cameras to count and detect every traffic participant at a busy intersection. They have inherent limitations of fixed field of view and susceptibility to occlusions.

In this first use case, we demonstrate how a synchronized multisensor setting can leverage a data modality, such as audio, to localize sound-emitting traffic participants and reduce the chance an object is completely obstructed by another. The ability of the sensors to “listen” as well as “see” allows the sensor network to remain resilient against occlusions and dead zones. Figure 6 shows an example of detecting the position of a bicyclist using sound, regardless of whether or not the bicyclist is in any of the cameras' field of view thanks to the diffraction property of the sound waves. In order to reconstruct the position of the bicyclist ringing the bell, we first annotate the high amplitude peaks, in the audio data, synchronized using the common time scale as reconstructed from the dedicated audio channel with the serialized timestamps. With the known sensor positions, one can find the sound source position at given time by minimizing the errors:

$$\boldsymbol{p}, \boldsymbol{t} = \underset{\boldsymbol{p},\boldsymbol{t}}{\arg\min} \sum_{i=1}^{4} \left(\| p - p_i \| - c \cdot |t - t_i| \right)^2 ,$$

All four sensors must hear the sound for this to be a well-posed problem. The results are shown in Figure 6 and are in good agreement with the video footage from the same sensors. There are examples of when audio-based localization was not possible because of noise pollution by a bus and vice versa when the object was out of the field of view of the cameras but could still be heard which illustrates the benefits of such a complementary multimodal approach. This audio-based sound source localization would not be possible without the synchronization technique presented in this paper.

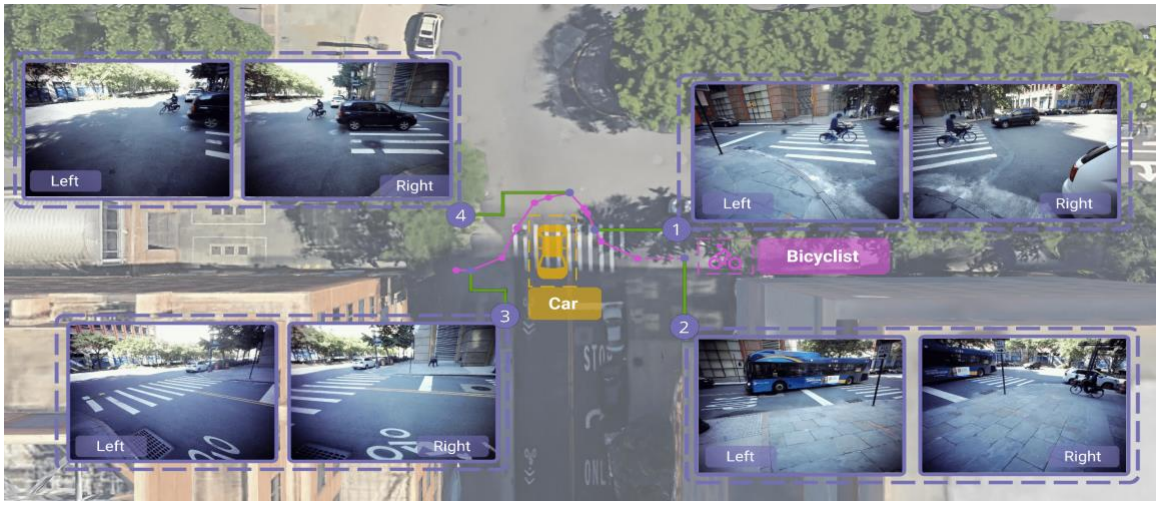

Figure 6. Audio-based localization of a bicyclist crossing the street at Chase Center and ringing the bell repeatedly (magenta points).

In chronological order: Sensor 2 can see the bicyclist approaching the intersection, but localization of the bell ring is not possible because two sensors were occluded by a noisy bus; Sensor 1 view confirms the position of the bicyclist taking a right turn; Sensor 4 footage reveals the reason for the bicyclist's curved trajectory---the black car did not stop to yield the right of way; Eventually, the bicyclist is no longer in the field of view of Sensor 3, but can still be localized thanks to the diffraction of the bell's sound waves.

Occupancy Tracking and Pedestrian Speed

Stakeholders interested in monitoring the level of activity and quantity of pedestrians and traffic in an area could make use of StreetAware. Here, we present an example in which we evaluate the occupancy of one of the intersections during a recording session. First, the dataset is evaluated with HRNet, an object detection, human pose estimation, and segmentation algorithm. Adapted from the Faster R-CNN network, HRNet is capable of performing state-of-the-art bottom-up segmentation via high-resolution feature pyramids. The network is trained on the COCO dataset. We detect six classes: person, car, bicycle, truck, motorcycle, and bus.

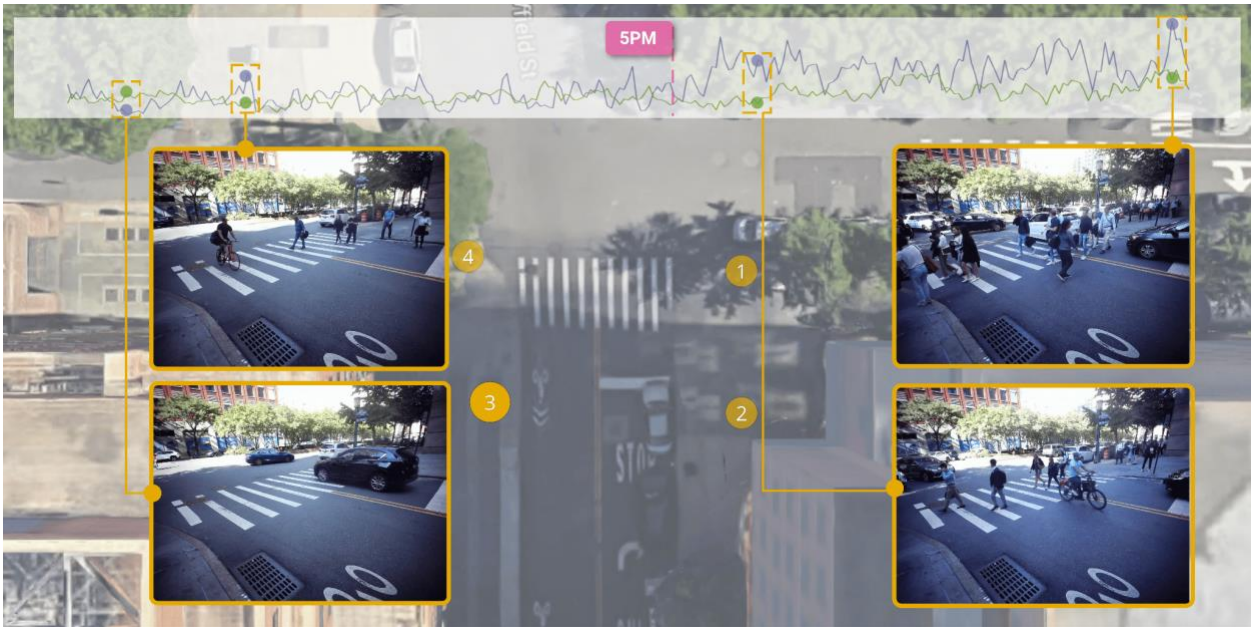

Figure 7. Chase Center intersection occupancy by object type during a recording session in the afternoon. Purple lines representing people and green lines representing cars on the top chart.

In the figure, the four sensors collecting data during this session are represented by circles. There is a significant increase in the pedestrian count (blue) around 5 p.m. as people leave work. Moreover, it is possible to detect traffic light cycles based on the ratio of the number of pedestrians versus vehicle counts.

Using this detection framework, Figure 7 shows the total count of the various urban scene entities throughout an entire recording session at the Chase Center intersection. We intentionally choose this particular recording session because it was conducted around 5 p.m. when people are finishing their workday and traveling home. This activity results in a spike in pedestrian and car traffic. There are roughly three times as many pedestrians counted (most crossing a street) toward the end of the recording than at the beginning. This trend also inversely correlates with car count, presumably due to cars yielding the right-of-way to pedestrians. We do not observe as much change in the number of cars or other motorized vehicles because this intersection is typically more consistently busy throughout the day and there is limited space along the street curbs to park cars compared to pedestrians on sidewalks. Parked cars present a certain level of static background count for the car object class. Measuring and predicting pedestrian behavior such as their speed and trajectory are of interest to the computer vision and urban design research communities.

Discussion

In this study, we collected unique data about traffic and pedestrians from three urban intersections using customized high-resolution audio-video sensors. The novel data includes multiple modalities (audio, video, and LiDAR) with highly accurate temporal information and synchronization. Since the data were recorded in New York City, many demographics are captured. This is particularly important since some of these groups, such as wheelchair users and people with varying levels and types of disabilities, are absent from large-scale datasets in computer vision and robotics, creating a steep barrier to developing accessibility-aware autonomous systems. Identifying pedestrians with disabilities, the qualities of their behavior, and ease at traversing the sensed urban environment is an area of possible exploration with datasets such as this one.

Overall, the REIP sensors have demonstrated great versatility in data acquisition pipelines and operating conditions. They even withstood, without damage, a sudden rain incident during one of the recording sessions at Commodore Barry Park. The majority of the sensor's hardware is enclosed within an aluminum weatherproof housing with heat sinks, however, we still experienced occasional periods of lost frames, even during operation in shadows, due to the random operating system interrupts and throttling of the CPUs. Therefore, any long-term deployments would need to account for these issues in a comprehensive way.

The data presented in this study are limited in a few ways. First, the geographic coverage is narrow. Though the activity at each site is somewhat varied, ultimately, data were only collected at three intersections in a single borough in a single highly-developed city in the United States. Second, compared to some other available datasets, StreetAware lacks diverse environmental conditions such as nighttime, precipitation, and fog. However, we did preserve some of the more challenging recording sessions where select sensors experienced an increased amount of occlusion from vegetation during windy conditions. Moreover, third, the data are quite raw---the audio and video recordings are not labeled (e.g., objects, actions, sound sources, etc.) and the LiDAR files provided are unprocessed. In its current form, a user would not be able to query the data for information or have an idea of what is happening over time in a scene without manually inspecting the data or performing further processing and analysis.

Conclusion

In this chapter, we presented the StreetAware dataset, which contains synchronized multi-perspective street-level audio and video in a single dataset. We also presented a new method to synchronize high-sample rate audio streams and demonstrated unique use cases for the data; in the process, we describe the limitations and real-world implementation of REIP sensors.

Section 2. Mapping the Walk: A Scalable Computer Vision Approach for Generating Sidewalk Network Datasets from Aerial Imagery

Introduction

After a century of car-oriented urban growth, 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 CO2 emissions attributable to the transport sector, 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. 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. Economically, walkable and transit-served city environments have also become an important draw for a competitive workforce and now command some of the highest-priced real estates in American cities.

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), 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.

A number of studies have highlighted the inadequacy of using street-centerline networks for pedestrian routing, 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). For instance, Chin et al., 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. Lack of accurate sidewalk routing data threatens their independence and decreases their quality of life.

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 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 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 datasets we enable less resourceful cities to create datasets describing pedestrian environments which otherwise would not be possible at a comparable cost and time.

In this chapter, we make the following contributions:

1. We present the design and implementation of Tile2Net as an end-to-end, open-source tool for creating large-scale pedestrian networks from orthorectified aerial imagery.
2. We show the calibration procedure of a high-performing scene classification model for detecting sidewalks, crosswalks, and footpaths. We have custom trained Tile2Net on around 20,000 detailed images in Cambridge, New York City, and Washington, where detailed GIS data of pedestrian infrastructure was available. 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. We discuss how adjustable TIle2Net is to different city environments, offering various settings to finetune the model on new data based on local environmental characteristics.

Related Work

Map Generation

At least five different methods for mapping sidewalk infrastructure can be distinguished in existing literature and practice, with additional combinations, as depicted in Figure 8.

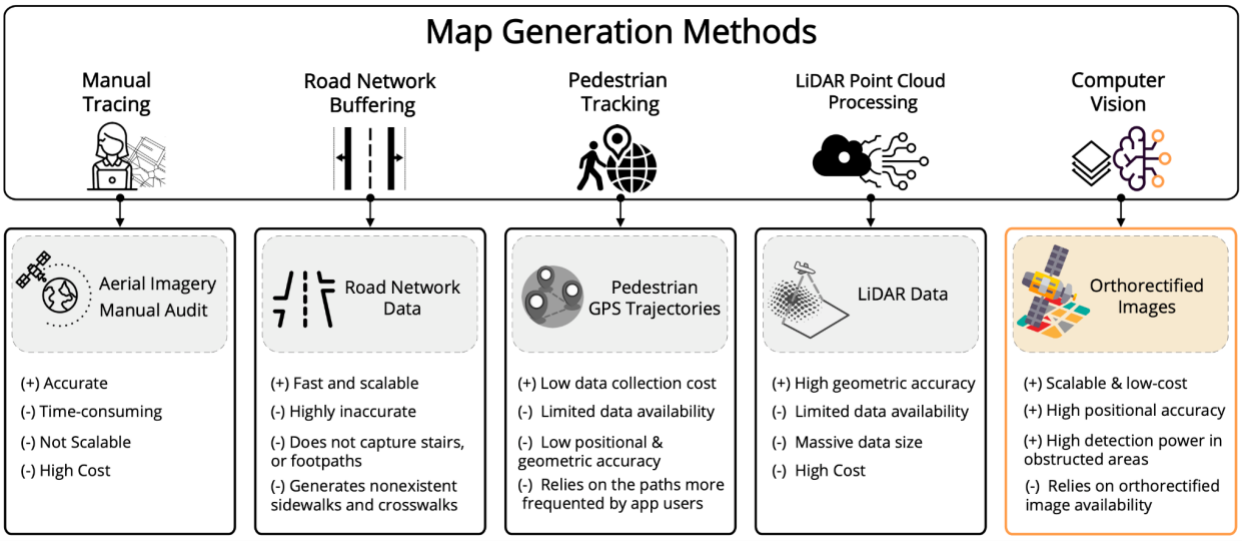

Figure 8. Different methods of map generation.

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.

First, physical site surveys and manual aerial image surveys have been used in a number of cities to develop datasets on pedestrian facilities. This involves human tracing of observable sidewalks and crosswalks from georeferenced aerial imagery, combined with on-the-ground observation and validation. 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, 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.

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.

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 App. 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, simplified, and joined into contiguous network datasets. 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.

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. The resulting data represent sidewalks as vector lines or polygons that can be both accurate and scalable. 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. 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. Although semantic segmentation has been broadly used to detect roads and building footprints from aerial or satellite images and to create road networks 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. 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, researchers often forego retraining or fine-tuning their models on target datasets and rather rely on publicly-available pre-trained datasets such as CityScapes, Mapillary, and ADE20K. Relying on pre-trained models, not specific to the task, limits analysis to the classes included in those datasets. Further, pre-trained models not fine-tuned on domain-specific data can yield suboptimal performance. 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. Hence, choosing the right network architecture that can preserve local details while accounting for global image context is crucial.

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 architectures to improve performance over a static resolution approach. This has been followed by rapid developments in multi-scale pyramid-style networks. In particular, HRNet 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. A variant of HRNet, HRNet-W48, which has shown superior performance across segmentation benchmarks such as Cityscapes and Mapillary Vista, is used as a key component of our segmentation framework described below.

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, NetworkX, and Geopandas. 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 256x256-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 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.

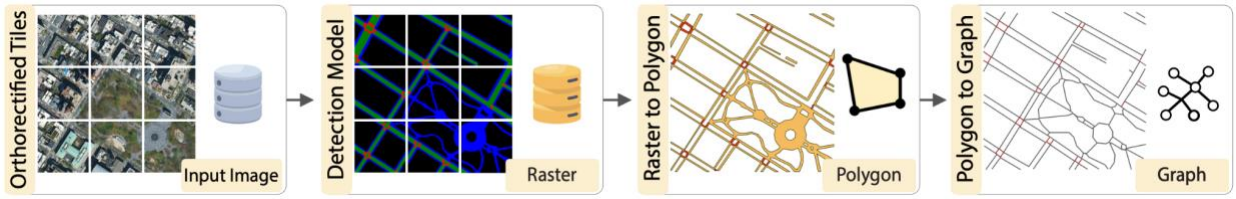

Figure 9. The proposed network generation pipeline:
a) Unlabeled orthorectified tiles are passed through the semantic segmentation model for prediction, b) The model detected sidewalks (blue), crosswalks (red), and roads (green) in the input tiles, c) The sidewalks and crosswalks of the prediction results (raster format) are converted into georeferenced polygons, d) The line representation of the pedestrian network generated from polygons.

Figure 9 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.

Detecting sidewalks, footpaths, and crosswalks from aerial imagery

Our semantic segmentation model takes an input image, 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. 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, and large objects with less details (e.g., roads) are best detected at a lower zoom level. 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 512x512, 1024x1024, and 2048x2048 pixel tiles, the best results were achieved using 1024x1024 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-meter/pixel images.

We used HRNet-W48 with Object-Contextual Representations as the backbone, since HRNet maintains twice as high a resolution representation as other popular backbones such as WiderResnet38. 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, which accounts for the relationship between pixels instead of only relying on single pixels to calculate the loss.

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. and international cities, and 2) planimetric GIS data representing the same elements as seen in orthorectified images.

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-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. 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) 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. For the purposes of this study, we used orthorectified images provided by Massachusetts, Washington, DC, and New York to train the model and pilot test the approach. We obtained 11,000 tiles from Washington, DC, 28,000 tiles from Cambridge, MA 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, 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.

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. Since planimetric data are created using orthorectified images, they are also suitable for creating labels for semantic segmentation models. An annotated image is a reference image where each pixel value describes the label to which the pixel in the aerial image belongs as shown in Figure 10-b,c,e,d.

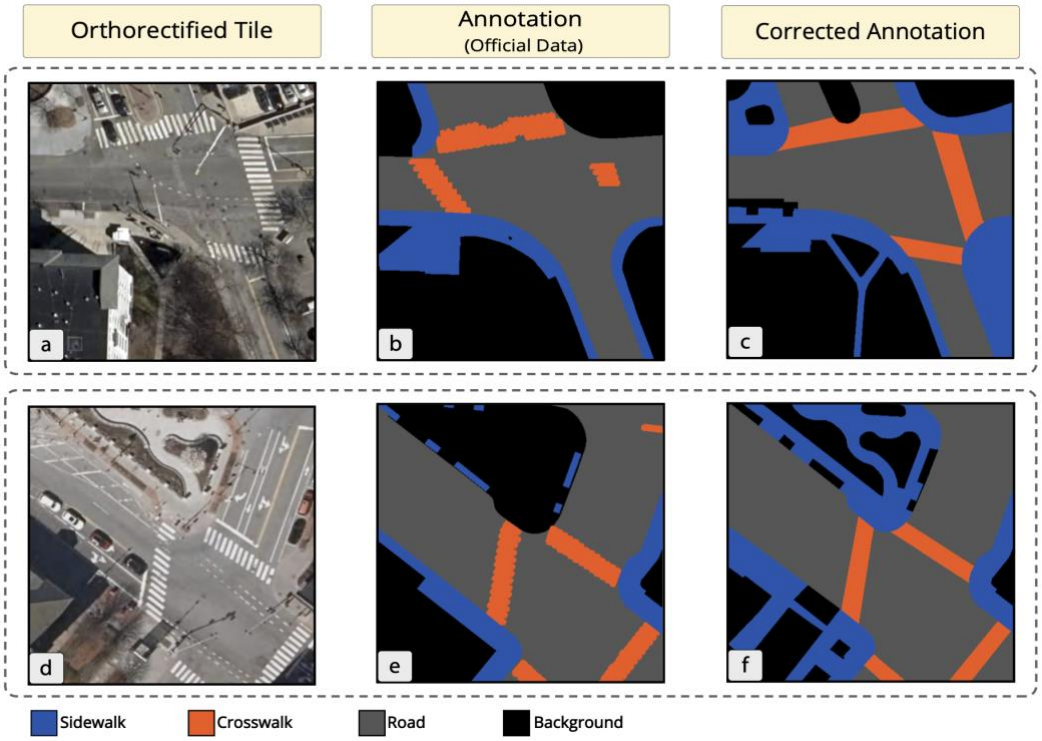

Figure 10. Examples of the mismatches between the aerial image and the label created from the official data. The manually corrected labels are shown in the last column.

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. 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 Figure 11, 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 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.

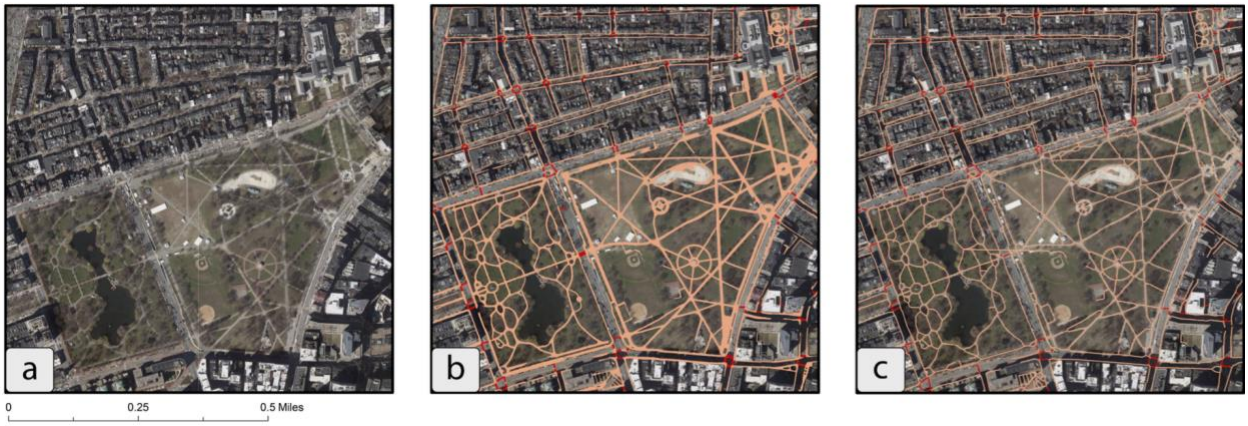

Figure 11. Boston Commons: a) Aerial image, b) Detected sidewalk and footpath polygons (in orange) and detected crosswalks (in red), c) Fitted sidewalk, crosswalk, and footpath centerlines superimposed on the aerial image.

Implementation of the detection model

The model was trained with a batch size of 16, SGD for the optimizer with polynomial learning rate, momentum 0.9, weight decay $5e^{-}$, 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. We used color augmentation, random horizontal flip, random scaling, 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 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 (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.

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 fine tuning 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. 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: Evaluation metrics on the test set.

Label	IoU (%)	Precisions	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		

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.

Using the trained model

After the training phase is completed, the unlabeled orthorectified tiles are passed through the trained model, as shown in Figure 10-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 10-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 (Figure 10-c); and 2) polygon to centerline conversion to create the final pedestrian network representation (Figure 10-d). Figure 11 shows the results of these last two steps for Boston Commons, which was not part of the training data.

Raster to polygon conversion: To obtain vectorized and georeferenced polygons from the detected sidewalk, crosswalk, and road highlight {raster regions}, we employed a connected-component mapping algorithm,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.

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. Next, a dense Voronoi diagram is computed to extract the centerlines of the sidewalk polygons. The centerline is constructed by linking the internal Voronoi diagram edges not intersecting with the boundary of the object as shown in Figure 12.

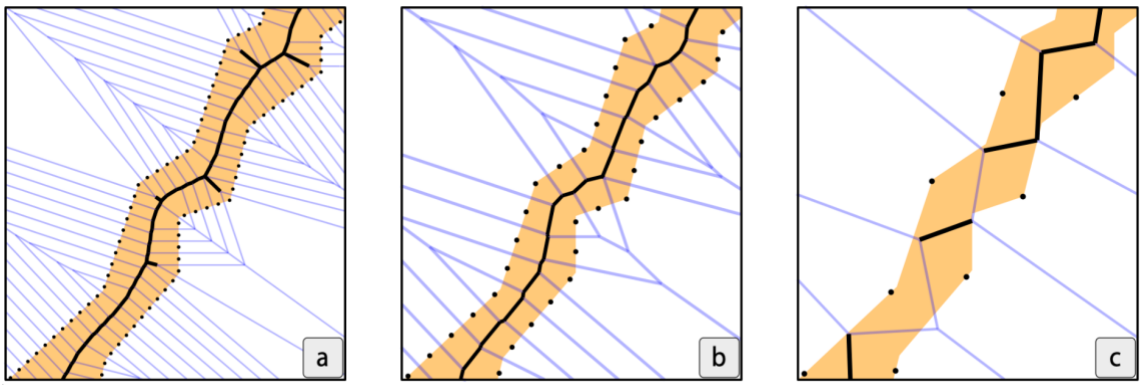

Figure 12. Construction of centerline using Dense Voronoi method (DV) with different interpolation distances (d), which is the maximum distance between the sampled points (black points) on the polygon's boundary.

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.

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 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.

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).

Figure 13 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.

Figure 13. Model results showing detected sidewalk, crosswalk and footpath centerlines: a) Boston and Cambridge, b) Manhattan and parts of Brooklyn, c) Washington, DC. The maps are shown at the same scale for comparison.

Table 3 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. 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 3.

Table 3: Comparison of polygon accuracy results in Cambridge, MA, Boston, MA, New York City, NY, and Washington, DC.

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
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%

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 0-100\%) the detected polygons are to the city dataset, on average (including those city polygons that remained undetected).

The last row of Table 3 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.

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 in Table 4. 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.

Table 4: 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	83.10%	94.56%	91.01%	68.86%
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	78.49%	90.78%	89.56%	64.76%
Manhattan	Official element count	-	-	-	6,239
	Match (within 4m of centroid)	-	-	-	5,309
	Match	-	-	-	85.09%

To evaluate the accuracy of the generated network centerlines, we 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 4m 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 4.

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. 90.78% of all sidewalk segments in city GIS data andc89.56% 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.

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 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 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 differences 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 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.

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

1. Rosenfeld and Pfaltz, “Sequential Operations in Digital Picture Processing.”
2. Douglas and Peucker, “Algorithms for the Reduction of the Number of Points Required to Represent a Digitized Line or Its Caricature.”
3. Brandt and Algazi, “Continuous Skeleton Computation by Voronoi Diagram.”
4. Guttman, “R-Trees: A Dynamic Index Structure for Spatial Searching.”
5. Liu, Rabinovich, and Berg, “Parsenet: Looking Wider to See Better.”
6. Tao, Sapra, and Catanzaro, “Hierarchical Multi-Scale Attention for Semantic Segmentation.”
7. Zhu et al., “Improving Semantic Segmentation via Video Propagation and Label Relaxation.”
8. MassGIS, “MassGIS Data: 2018 Aerial Imagery.”
9. DC GIS, “Aerial Photography (Orthophoto SID) - 2019.”
10. NYC GIS, “NYS Statewide Digital Orthoimagery Program.”
11. Quackenbush, “A Review of Techniques for Extracting Linear Features from Imagery.”
12. Sun et al., “Deep High-Resolution Representation Learning for Human Pose Estimation”; Wang et al., “Deep High-Resolution Representation Learning for Visual Recognition.”
13. Yuan, Chen, and Wang, “Object-Contextual Representations for Semantic Segmentation.”
14. Wu, Shen, and Van Den Hengel, “Wider or Deeper: Revisiting the Resnet Model for Visual Recognition.”
15. Zhao et al., “Region Mutual Information Loss for Semantic Segmentation.”
16. U.S. Geological Survey, “USGS EROS Archive - Aerial Photography - High Resolution Orthoimagery (HRO).”
17. Tucker, Grant, and Dykstra, “NASA’s Global Orthorectified Landsat Data Set”; Zhou et al., “A Comprehensive Study on Urban True Orthorectification.”
18. Tao, Sapra, and Catanzaro, “Hierarchical Multi-Scale Attention for Semantic Segmentation.”
19. Boeing, “OSMnx: New Methods for Acquiring, Constructing, Analyzing, and Visualizing Complex Street Networks.”
20. Hagberg and Conway, “NetworkX: Network Analysis with Python.”
21. Jordahl, “GeoPandas: Python Tools for Geographic Data.”
22. Zhang et al., “Representing Place Locales Using Scene Elements”; Wang et al., “Impacts of the Water Absorption Capability on the Evaporative Cooling Effect of Pervious Paving Materials”; Zhou et al., “Using Google Street View Imagery to Capture Micro Built Environment Characteristics in Drug Places, Compared with Street Robbery”; Kim et al., “Decoding Urban Landscapes: Google Street View and Measurement Sensitivity.”
23. Cordts et al., “The Cityscapes Dataset for Semantic Urban Scene Understanding.”
24. Neuhold et al., “The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes.”
25. Zhou et al., “Scene Parsing through Ade20k Dataset.”
26. Ahn and Kwak, “Learning Pixel-Level Semantic Affinity with Image-Level Supervision for Weakly Supervised Semantic Segmentation.”
27. Azizi et al., “Big Self-Supervised Models Advance Medical Image Classification.”
28. Hosseini et al., “Sidewalk Measurements from Satellite Images: Preliminary Findings.”
29. Zhao et al., “Pyramid Scene Parsing Network.”
30. He, Deng, and Qiao, “Dynamic Multi-Scale Filters for Semantic Segmentation.”
31. Sun et al., “Deep High-Resolution Representation Learning for Human Pose Estimation”; Wang et al., “Deep High-Resolution Representation Learning for Visual Recognition.”

32. Newell, Yang, and Deng, “Stacked Hourglass Networks for Human Pose Estimation”; Chen et al., “Cascaded Pyramid Network for Multi-Person Pose Estimation”; Yu et al., “Deep Layer Aggregation.”
33. Kasemsuppakorn and Karimi, “Pedestrian Network Extraction from Fused Aerial Imagery (Orthoimages) and Laser Imagery (Lidar).”
34. Yang et al., “Pedestrian Network Generation Based on Crowdsourced Tracking Data.”
35. Cura, Perret, and Paparoditis, “A State of the Art of Urban Reconstruction: Street, Street Network, Vegetation, Urban Feature.”
36. Ai and Tsai, “Automated Sidewalk Assessment Method for Americans with Disabilities Act Compliance Using Three-Dimensional Mobile Lidar.”
37. Horváth, Pozna, and Unger, “Real-Time LIDAR-Based Urban Road and Sidewalk Detection for Autonomous Vehicles.”
38. Luo et al., “Developing an Aerial-Image-Based Approach for Creating Digital Sidewalk Inventories.”
39. Ess et al., “Segmentation-Based Urban Traffic Scene Understanding.”
40. Balali, Rad, and Golparvar-Fard, “Detection, Classification, and Mapping of US Traffic Signs Using Google Street View Images for Roadway Inventory Management.”
41. Bastani et al., “Roadtracer: Automatic Extraction of Road Networks from Aerial Images.”
42. Proulx, Zhang, and Grembek, “Database for Active Transportation Infrastructure and Volume.”
43. Cottrill et al., “Future Mobility Survey: Experience in Developing a Smartphone-Based Travel Survey in Singapore.”
44. Liu, “Fuzzy Modularity and Fuzzy Community Structure in Networks.”
45. Cambra, Gonçalves, and Moura, “The Digital Pedestrian Network in Complex Urban Contexts: A Primer Discussion on Typological Specifications”; Sun et al., “Deep High-Resolution Representation Learning for Human Pose Estimation.”
46. Ellis et al., “Connectivity and Physical Activity: Using Footpath Networks to Measure the Walkability of Built Environments.”
47. Chin et al., “Accessibility and Connectivity in Physical Activity Studies: The Impact of Missing Pedestrian Data.”
48. Saha et al., “Project Sidewalk: A Web-Based Crowdsourcing Tool for Collecting Sidewalk Accessibility Data At Scale”; Zhang and Zhang, “Pedestrian Network Analysis Using a Network Consisting of Formal Pedestrian Facilities: Sidewalks and Crosswalks.”
49. Cohen and Dalyot, “Route Planning for Blind Pedestrians Using OpenStreetMap”; El-Taher et al., “A Systematic Review of Urban Navigation Systems for Visually Impaired People”; Delboni Lomba and Godoy da Silva, “Informed Search Algorithm for Route Optimization for Visually Impaired People: Possibility of Intelligent Assistive Technology.”
50. Walker and Johnson, “Peak Car Ownership: The Market Opportunity of Electric Automated Mobility Services.”
51. EPA, “Sources of Greenhouse Gas Emissions.”
52. Cervero, *The Transit Metropolis: A Global Inquiry*; Speck, *Walkable City: How Downtown Can Save America, One Step at a Time*.
53. Frank and Engelke, “The Built Environment and Human Activity Patterns: Exploring the Impacts of Urban Form on Public Health”; Grasser et al., “Objectively Measured Walkability and Active Transport and Weight-Related Outcomes in Adults: A Systematic Review”; Zapata-Diomedi et al., “Physical Activity-Related Health and Economic Benefits of Building Walkable Neighbourhoods: A Modelled Comparison between Brownfield and Greenfield Developments.”

54. Moretti, *The New Geography of Jobs*; Glaeser, *Triumph of the City: How Our Greatest Invention Makes Us Richer, Smarter, Greener, Healthier, and Happier*.
55. Leinberger and Lynch, “Foot Traffic Ahead: Ranking Walkable Urbanism in America’s Largest Metros.”
56. Lin et al., “Microsoft Coco: Common Objects in Context.”
57. Sighencea, Stanciu, and Căleanu, “A Review of Deep Learning-Based Methods for Pedestrian Trajectory Prediction.”
58. Tordeux et al., “Prediction of Pedestrian Speed with Artificial Neural Networks.”
59. Ahmed et al., “Pedestrian and Cyclist Detection and Intent Estimation for Autonomous Vehicles: A Survey.”
60. Girshick, “Fast R-Cnn.”
61. Liu et al., “Ssd: Single Shot Multibox Detector.”
62. Fourkiotis et al., “Applying Deep Learning Techniques for the Prediction of Pedestrian Behaviour on Crossings with Countdown Signal Timers.”
63. Sainju and Jiang, “Mapping Road Safety Features from Streetview Imagery: A Deep Learning Approach.”
64. Wang, Liu, and Luo, “Identification and Improvement of Hazard Scenarios in Non-Motorized Transportation Using Multiple Deep Learning and Street View Images.”
65. Miranda et al., “Urban Mosaic: Visual Exploration of Streetscapes Using Large-Scale Image Data.”
66. Fuentes et al., “Urban Sound & Sight: Dataset and Benchmark for Audio-Visual Urban Scene Understanding.”
67. Zhang, Fan, and Li, “Automated Detecting and Placing Road Objects from Street-Level Images.”
68. Doiron et al., “Predicting Walking-to-Work Using Street-Level Imagery and Deep Learning in Seven Canadian Cities.”
69. Zhao et al., “Pyramid Scene Parsing Network.”
70. Redmon and Farhadi, “Yolov3: An Incremental Improvement.”
71. Charitidis et al., “StreetScouting: A Deep Learning Platform for Automatic Detection and Geotagging of Urban Features from Street-Level Images.”
72. Sukel, Rudinac, and Worring, “Urban Object Detection Kit: A System for Collection and Analysis of Street-Level Imagery.”
73. Google Maps Platform, “Google Street View Static API.”
74. Neuhold et al., “The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes.”
75. Kostić et al., “Smart City Intersections: Intelligence Nodes for Future Metropolises.”
76. Sighencea, Stanciu, and Căleanu, “A Review of Deep Learning-Based Methods for Pedestrian Trajectory Prediction.”
77. Ballardini et al., “Urban Intersection Classification: A Comparative Analysis.”
78. Piadyk et al., “REIP: A Reconfigurable Environmental Intelligence Platform and Software Framework for Fast Sensor Network Prototyping.”