Han *et al. Urban Informatics* (2024) 3:22
https://doi.org/10.1007/s44212-024-00053-9

Urban Informatics
The International Society for Urban Informatics

ORIGINAL ARTICLE Open Access

Understanding pedestrian movement using urban sensing technologies: the promise of audio-based sensors

Chaeyeon Han[1], Pavan Seshadri[2], Yiwei Ding[2], Noah Posner[3], Bon Woo Koo[4], Animesh Agrawal[1], Alexander Lerch[2] and Subhrajit Guhathakurta[1*]

Abstract

While various sensors have been deployed to monitor vehicular flows, sensing pedestrian movement is still nascent. Yet walking is a significant mode of travel in many cities, especially those in Europe, Africa, and Asia. Understanding pedestrian volumes and flows is essential for designing safer and more attractive pedestrian infrastructure and for controlling periodic overcrowding. This study discusses a new approach to scale up urban sensing of people with the help of novel audio-based technology. It assesses the benefits and limitations of microphone-based sensors as compared to other forms of pedestrian sensing. A large-scale dataset called ASPED is presented, which includes high-quality audio recordings along with video recordings used for labeling the pedestrian count data. The baseline analyses highlight the promise of using audio sensors for pedestrian tracking, although algorithmic and technological improvements to make the sensors practically usable continue. This study also demonstrates how the data can be leveraged to predict pedestrian trajectories. Finally, it discusses the use cases and scenarios where audio-based pedestrian sensing can support better urban and transportation planning.

Keywords Sensors, Audio-based, Pedestrian, Active mobility

1 Introduction

A significant component of smart city initiatives has been the deployment of sensor technologies to monitor and control various city services and functions. Cities use a variety of sensors to assess how urban services are being delivered and accessed, which helps alleviate bottlenecks and trigger advance warnings about potential service disruptions. Understanding the temporal and spatial variation in the demand for urban services can also lead to better resource use, more equitable service delivery, and greater sustainability and resilience. Various sensors are now deployed in the urban environment, particularly in transportation, but also to monitor environmental conditions, the flow of energy, water, and waste, and in tracking criminal activities (Lee et al., 2014). More recently, with the growing interest in active mobility and walkability, several cities have experimented with various technologies to sense people.

While the movement of vehicles has been an important component of traffic planning, there has been less effort in understanding the movement of pedestrians in cities, particularly in the United States. By understanding pedestrian flows, we can design better and more equitable walkable environments and gain useful insights into the spaces people linger in or avoid, as well as about the activities they pursue in the public domain. Accurate

*Correspondence:
Subhrajit Guhathakurta
subhro.guha@design.gatech.edu
[1] Center for Spatial Planning Analytics and Visualization, Georgia Tech, Atlanta, GA, USA
[2] Music Informatics Group, Georgia Tech, Atlanta, GA, USA
[3] IPaT, Georgia Tech, Atlanta, GA, USA
[4] School of Urban & Regional Planning, Toronto Metropolitan University, Toronto, ON, Canada

Springer

prediction of individual and social behavior in public spaces has powerful implications for urban planning.

The detection of pedestrians has been mainly based on video data analysis (Li et al., 2016; Rahman et al., 2019) or through infrared counters (Mathews & Poigné, 2009; Yang et al., 2010, 2011); both are several times more expensive than audio sensing. More sophisticated alternatives that are sometimes considered for pedestrian sensing, such as radar, radio beams, inductive loops, and piezoelectric strips, are also costly to deploy and maintain (Ozan et al., 2021; Ozbay et al., 2010). In this paper, we explore the potential for microphone-based sensors, combined with methods developed for the analysis of highly complex musical audio signals, to be adapted to sensing pedestrians. As sensors, microphones offer several advantages: they are affordable, have low power requirements, can cover large angles up to 360 degrees, and can capture otherwise unobserved data given that sound travels around objects and thus cannot be easily blocked. Recent research has also demonstrated that soundscapes detected through audio devices provide sufficient visual information of the same places (Zhuang et al., 2024). These advantages are counterbalanced by challenges that require investigation, experimentation, and mitigation, including the requirement to develop more advanced processing algorithms for extracting meaningful information as multiple sound signals are superposed in a poly-timbral mixture with unknown directional information; the challenge of positioning microphone-based sensors in ways that optimize data collection; and challenges beyond technology such as issues of privacy, anonymity, and sanitization of data.

Given that prior research has provided little guidance on how to mitigate many of these challenges, we have addressed this gap by conducting pilot experiments in a campus setting to identify an appropriate and sustainable way of achieving continuous data capture and processing. In this paper, we report on our preliminary tests of using microphones to sense people while also demonstrating how strategically placed sensors can offer valuable information about pedestrian flows. We demonstrate our approach by using audio data to detect pedestrians. However, the flow prediction algorithm uses the information we retrieved from the video footage used to label the audio files. We use these separate approaches because our algorithms to detect pedestrians using audio are not at the level of accuracy that video-based methods have reached. However, the flow prediction algorithms we develop are agnostic to the method of obtaining pedestrian counts. We expect that audio-based pedestrian count data with improved accuracy can easily substitute for the video-based data we use now for pedestrian flow tracking.

The rest of the paper is organized as follows: the next section, Section 2, assesses current pedestrian sensing technologies and provides a brief overview of their technical challenges. Section 3 reviews the pedestrian flow prediction approaches that inform our study. Next, in Section 4, we discuss the collection and curation of the ASPED dataset, which was used in our preliminary analyses. In this section, we also present our methodological approach and the design of experiments. Section 5 presents our results from the experiments in pedestrian detection and for predicting pedestrian flows. Section 6 offers a discussion of the promise and broader impacts of audio-sensing technology for sensing people and concludes with some closing remarks.

2 Pedestrian sensing technologies

While pedestrian sensors have been around for over 30 years, audio-based sensing has — to the best knowledge of the authors — never been attempted. This is probably because data science and machine learning-based technologies were less mature and are only now starting to transform many domains. Several pedestrian sensor systems have been operating in Europe and Australia since the early 1990s, and systematic efforts to deploy such systems in the U.S. began soon after 2000. The Federal Highway Administration commissioned a study in 2001 to evaluate whether automated pedestrian detectors at intersections can reduce pedestrian-vehicle conflicts when compared with standard pedestrian push-button walk signals (Hughes et al., 2001). This study also compared a number of different pedestrian detection technologies, including microwave sensors, infrared detectors, and video. The detection technologies had a significant failure rate, and the authors concluded that “improvements are needed in detection accuracy to reduce the number of false alarms and missed calls at intersections” (Hughes et al., 2001, p.16). Subsequently, the National Bicycle and Pedestrian Documentation Project (NBPDP) began in 2004 to standardize data collection of bicycles and pedestrians, initially using manual short duration counts (Ozan et al., 2021). Soon, multiple federally funded studies were commissioned to evaluate various technologies to detect pedestrians in different locations, followed by similar efforts by many state and metropolitan government agencies (Fields, 2012; Figliozzi et al., 2014; Minge et al., 2017; Ryus et al., 2014). An excellent review of the reports and studies to date for bicycle and pedestrian data collection efforts and an evaluation of various sensor technologies to capture such data can be found in Ozan et al. (2021).

Most past deployments of sensors to detect active mobility cannot distinguish pedestrians and bicyclists (Ozan et al., 2021). The technologies used in these sensors, such as active and passive infrared, laser scanning, radar, and radio beams, are designed to detect interrupted or reflected pulses from human bodies or their heat signatures. Therefore, exposed humans, whether pedestrians, bicyclists, or people using micromobility options, will all be detected without distinction. Other technologies typically used in detecting vehicles, such as pneumatic tubes, inductive loops, and magnetometers are not appropriate for pedestrian tracking but can be used for sensing bicyclists. Other less common sensing technologies, such as ultrasonic sensors, pressure sensing mats, piezoelectric strips, and various hybrid technologies, have also been deployed for pedestrian counts. Each of these sensors offers particular advantages but also comes with several challenges and limitations.

In the commercial realm, several companies have offered their own solutions to communities for tracking active mobility trips. Among the commercial offerings, EcoCounter, MetroCount, Jamar, and StreetLight are some of the more popular vendors of bicycle and pedestrian data. Eco-Counter, a French company with offices in North America and across Europe, utilizes various technologies for its pedestrian and bicycle counters, including pneumatic tubes, passive infrared, inductive loops, and mixed infrared/inductive loops. Metrocount has developed bicycle counters (RidePod®BT) using pneumatic tubes and pedestrian and bicycle counters that utilize piezoelectric technologies (RidepPod®BP). More recently, companies such as Miovision and Numina have been offering video image-based sensors. These image-based sensors use advanced neural networks to detect and isolate different transport and pedestrian traffic modes. Perhaps the most advanced private player in this domain is StreetLight Data, which is pioneering the development of proprietary artificial intelligence (AI) algorithms applied to data from millions of mobile devices, Internet of Things (IoT) sensors, and other geospatial databases to estimate traffic mode and volumes at a fine geographic scale. Regardless of the technology and their providers, all the technologies noted above have a price point that is well above $1,500 per sensor, making them cost-prohibitive for wide-scale adoption across urban areas. Therefore, they are mostly used in a few selected locations and are applied to track active mobility in small areas within the city.

At this time, the most widely used pedestrian sensing technologies are: (i) video image processing, and (ii) active and passive infrared counters.

2.1 Video-based sensing using computer vision

In the absence of advanced machine learning approaches and training data sets, video technology was initially used to count pedestrians manually from recorded video (Ozan et al., 2021). With the rapid progress in computer vision-based algorithms and growing availability of pre-tagged image datasets in the last two decades, detection of people and objects from video or static images has become easier and more efficient (Baker et al., 2011; Barron et al., 1994; Dollar et al., 2009; Fei-Fei et al., 2006; Martin et al., 2004; Scharstein & Szeliski, 2002). This research area has seen rapid development also due to its importance for autonomous, self-driving vehicles. In a review article from 2012, Dollar et al. point out that a performance drop is noticed across various systems if pedestrians are represented by less than 80 pixels (Dollar et al., 2012). Even in the best case, however, up to 20 percent of pedestrians can be missed. The data quality strongly impacts the miss rate, as can be seen through strong variance over different datasets (Brunetti et al., 2018) and weather conditions (Li et al., 2020).

It should be noted, however, that most of the work studying pedestrian detection aims at solving the detection problem under the very challenging circumstances of an autonomous vehicle, where the video camera itself is moving, the pedestrians not only have to be counted but individually tracked, the video feed potentially has limited resolution, and the error of missing a pedestrian can have fatal consequences. Our objective is different and can tolerate a somewhat higher margin of error in prediction.

2.2 Infrared counters

Three types of infrared sensors have been used for pedestrian detection: (i) active, (ii) passive, and (iii) target reflective. An active counter uses an invisible beam, which, when interrupted by a body, registers a pedestrian count. A passive infrared counter senses the heat emitted by human bodies passing through the sensing area to detect pedestrians. Target-reflective devices also use an invisible beam that is bounced back from a reflector mounted on the opposite side of a sensing area. The absence of the reflected beam indicates the presence of pedestrians.

Several studies evaluated the performance of infrared counters and found them to be inaccurate, mostly because they systematically undercounted pedestrians. A study of trail users in Indiana that used infrared counters concluded that the sensor systematically undercounted trail users by 15% (Wolter & Lindsey, 2001). Other studies have also found similar results, such as a study conducted in San Diego County that found the undercount rate to be 15% to 21% from active infrared counters and 12% to 48% from passive infrared counters (Jones et al., 2010). The latest tests conducted in New Jersey found that the infrared undercounting error can be more than 20% at sites with high volumes (Yang et al., 2010, 2011).

These field tests confirm that multiple pedestrians at the same time can confuse the infrared counters, especially when the bodies are lined up along the beam.

3 Advances in pedestrian flow prediction

The data gathered on pedestrian movement has proven invaluable for predicting and modeling human mobility patterns on an urban scale (Yabe et al., 2023). Predicting pedestrian flow plays a significant role in various fields, such as analyzing transportation and activity behaviors (Ai et al., 2019; González et al., 2008; Jiang et al., 2017), managing disaster responses (Oliver et al., 2020; Yabe et al., 2023), assessing public safety, and urban planning (Ratti et al., 2006).

Recent advancements in pedestrian flow prediction show a growing inclination towards using AI (Kitano et al., 2019), particularly for handling the complexities of spatial-temporal prediction. The emergence of deep learning techniques, particularly the integration of Convolutional Neural Networks (CNNs) for spatial analysis and Recurrent Neural Networks (RNNs) for temporal patterns, has significantly improved the accuracy of these predictions (Ai et al., 2019). Also, Deo and Trivedi (2021) introduced a distinctive approach for grid-based prediction, a critical aspect of pedestrian and vehicle trajectory forecasting. They proposed multimodal trajectory forecasts on scenarios sampled from a grid-based policy, which is learned using maximum entropy inverse reinforcement learning (MaxEnt IRL).

For regions with irregular spatial correlations, Graph Convolutional Networks (GCNs) have been heralded for their ability to model complex interactions between nodes effectively (Liu et al., 2021; Xia et al., 2021). These networks, where nodes represent areas and edges signify road links or Origin-Destination trajectories, have effectively processed complex spatial interactions. Liu et al. (2021) utilized camera-detected pedestrian data and employed their unique GCN model, which is capable of directly processing the spatial topology of road networks. Xia et al. (2021) introduced a 3-dimensional GCN (3DGCN) model for dynamic spatial-temporal graph prediction challenges, integrating Point-of-Interest (POI) data for enhanced accuracy. Furthermore, Sun et al. (2022) proposed a hybrid of GCN and fully-connected neural networks, with a multi-view fusion module, to capture and predict spatial correlations and crowd flow dynamics. Studies have shown that these methods adeptly handle non-linear spatial dependencies and time-varying trends often present in pedestrian movements.

The inclusion of external factors, such as inter-region traffic and weather conditions, has also enriched pedestrian flow predictions. For instance, Zhang and Kabuka (2018) and Zhang et al. (2017) considered weather and day-specific influences, while Lin et al. (2019) utilized the DeepSTN+ model, integrating POIs and temporal elements to better incorporate spatial dependencies. Zhang et al. (2018) developed the ST-ResNet model to predict crowd inflow and outflow, incorporating weather and time data.

The studies mentioned previously have employed various data sources to enhance pedestrian flow predictions, including mobile GPS (Global Positioning System), social media (Zhang et al., 2017), as well as data from bicycle sharing systems (Lin et al., 2019) and taxi GPS records (Sun et al., 2022). However, accessing this raw data for effective use in pedestrian mobility prediction poses a significant challenge for researchers. Often, simply having data is not sufficient; there is also a need for developing specialized software and systems, typically in collaboration with cellphone companies (Ratti et al., 2006). In this context, our audio-based pedestrian count dataset offers a potential solution to these challenges, providing a unique and valuable open resource data for pedestrian mobility research.

4 Data and methods

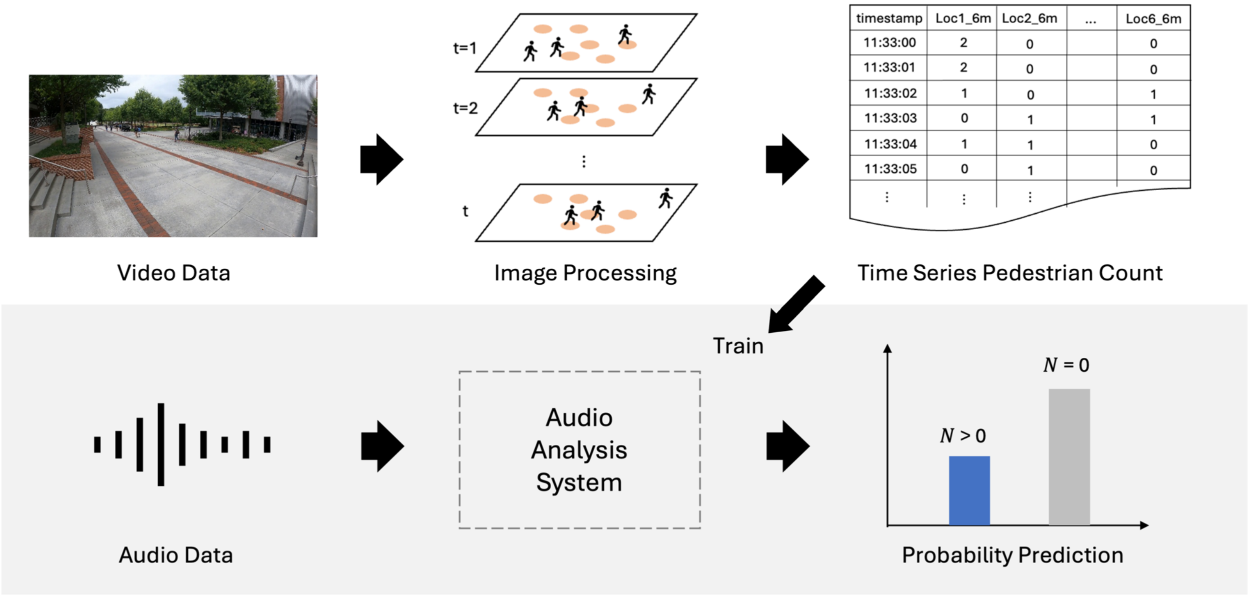
4.1 Data

4.1.1 Data collection and processing

To evaluate the capability of microphone-based sensors in identifying pedestrian presence, we collected and analyzed a unique dataset, named Audio Sensing for PEdestrian Detection (ASPED) (Seshadri et al., 2024).[1] This dataset serves as the foundation for a series of experiments that explore the possibility of audio sensing for pedestrian detection. The overall workflow is illustrated in Fig. 1.

To facilitate data collection, we captured data using off-the-shelf recording devices. The audio collection setup used Tascam DR-05X recorders with power banks for extended recording duration, Saramonic SR-XM1 microphones to avoid RF interference issues of the Tascam's built-in mics, and 5L OverBoard Dry Flat Bags for weatherproofing while maintaining audio permeability. For video data, we used GoPro HERO9 Black cameras with USB pass-through doors connected to Anker PowerCore III Elite 26K power banks for longer recording. The power banks were enclosed in Seahorse 56 OEM Micro Hard Cases, modified with a drilled hole, a Wraparound Plastic Submersible Cord Grip for the cord, and a 90-degree USB connector for better positioning and fit. For synchronizing time across cameras, we displayed the time from www.time.gov on a mobile device to each camera after the recording started, followed by a whistle blow

[1] urbanaudiosensing.github.io/ASPED.html, last access date Jan 31, 2024



timestamp	Loc1_6m	Loc2_6m	...	Loc6_6m
11:33:00	2	0		0
11:33:01	2	0		0
11:33:02	1	0		1
11:33:03	0	1		1
11:33:04	1	1		0
11:33:05	0	1		0
⋮	⋮	⋮		

Fig. 1 The overall workflow of training an audio analysis system to detect pedestrians. Note: *N* refers to the number of pedestrians detected

to mark the exact time, aiding in syncing with the audio recorders. In larger areas, multiple whistle signals were used.

The recording devices were set up in two locations on the Georgia Tech campus: *Cadell Courtyard* and *Tech Walkway*, both situated near dining areas but closed to vehicles. Due to the battery life of the devices, recording sessions were limited to approximately 2 days each.

To extract the pedestrian count per video frame, we use the Mask2Former model (Cheng et al., 2022) to detect people per frame at 1 frame per second. We use the specific implementation by OpenMMLab[2] which was trained on the Microsoft COCO dataset. This algorithm was parametrized with a prediction threshold of 0.7. For each frame of the video, the algorithm identified the 'person' class and generated bounding boxes around them. Subsequently, to analyze the proximity of these detected pedestrians to the audio recorders, circular buffers with various radii $r \in [1\text{ m}, 3\text{ m}, 6\text{ m}, 9\text{ m}]$ were superimposed on the video frames. These buffers were centered around the poles where the audio recorders were mounted. The orientation of the buffers was adjusted to align with the perspective of each specific video recording. Finally, every frame was annotated with the number of pedestrians detected if the bottom-center point of pedestrian bounding boxes intersected with the recorder buffers; otherwise, it was labeled as no-pedestrian present.

Note that the experiment presented in the following only utilizes binary labels (pedestrian present vs. no pedestrian present or pedestrian count $N \geq N_{\text{Threshold}}$ vs. count $N < N_{\text{Threshold}}$), however, the dataset labels reflect the actual pedestrian count.

4.1.2 Data description

Overall, we captured one frame-per-second video recording totaling 3,406,229 video frames, accompanied by nearly 2,600 h of audio, in five recording sessions. All recording days were weekdays with only one exception.

A noteworthy characteristic of the dataset is its imbalance: most of the time, there is no pedestrian close to the microphones. Therefore, the number of frames annotated with a pedestrian count of zero is considerably higher than those with nonzero pedestrians. Across the five collected sessions, the percentage of frames with one or more pedestrians ranges from 4.58% to 10.75% with a mean of 8.79%.

By looking at the distribution over different hours in a day in Fig. 2, we observe that most pedestrian events happen —unsurprisingly— during the daytime and that there is a small peak during lunchtime around 12 PM when 20% to 30% of the frames have pedestrians in proximity to the microphone. Between 1 and 4 AM, on the other hand, there are hardly any frames with pedestrians.

This label imbalance is important to keep in mind as it requires a careful design of machine learning models, training strategies, and evaluation methodology. Otherwise, the model might easily achieve (seemingly) high accuracy

[2] openmmlab.com, last access date Sep 5, 2023

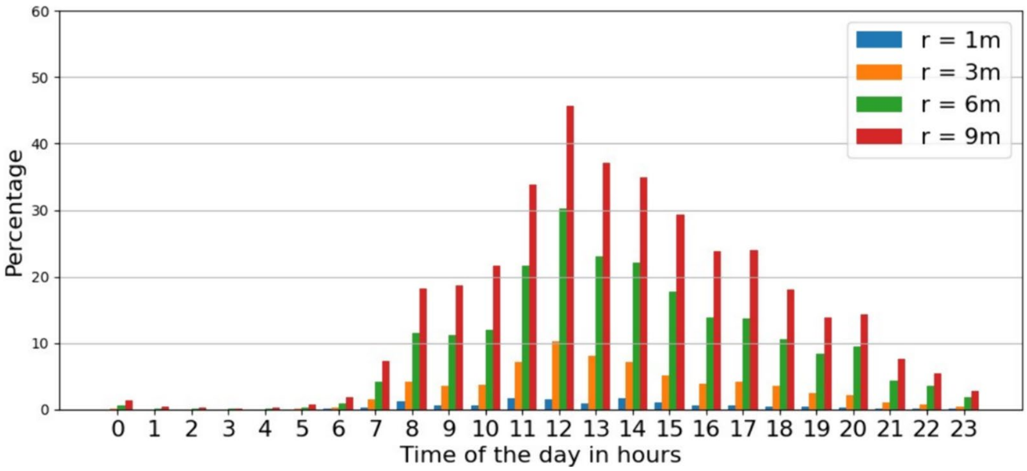


Fig. 2 Percentage of labels with pedestrians in different hours during a day

by simply predicting everything as no-pedestrian without learning anything meaningful.

We conducted experiments using data from audio and video sensors to achieve two objectives: (i) detect the presence of pedestrians near audio sensors using audio data (specifically when $N \neq 0$) that is annotated with the help of video frames from the same place and time, and (ii) predict the count of pedestrians near each sensor based on information from video frames in the ensuing seconds. In the subsequent sections, we explain our methodological approach for addressing each objective.

4.2 Audio sensing for pedestrian detection

4.2.1 Models

As mentioned above, the recorded audio signals are a superposition of many source signals comprising an auditory scene, which may or may not indicate pedestrians. Sounds indicating pedestrians can include, e.g., footsteps or speech, while other sounds can originate from various sources, including traffic, animals, construction, etc. All of these sounds are mixed together in a time-variant superposition with different volumes. This complexity requires sophisticated audio analyses such as the ones used in the field of musical audio analysis (Lerch, 2023). In this field, traditional approaches based on audio descriptors extracted from the time or spectral domain of the signal with a subsequent classifier (e.g., Support Vector Machine) have been nearly completely replaced by deep learning approaches as such approaches have shown superior performance across a wide variety of tasks, particularly challenging tasks with complex mixtures of audio sources.

A more detailed schematic of the "Audio Analysis System" introduced in Fig. 1 is presented in Fig. 3. Essentially, a binary classifier is trained which predicts whether a pedestrian is recorded in a given audio sample. Here, a t-second audio recording is chunked into individual 1 s segments, transformed into a Mel Spectrogram, and fed into an audio encoder network that learns to extract the information relevant for pedestrian detection. The result is a 128-dimensional vector representing each second of audio. This vector is then fed into a temporal aggregation model to add contextual information about the activity across the entire audio sample. The output of the temporal aggregation is then the input to a classifier estimating the likelihood of pedestrians for each segment of audio.

We investigated the following models to serve as the "Audio Encoder" shown in Fig. 3. These are methods for audio event classification at varying levels of complexity.

1. VGGish Pre-Trained Features (0 learnable parameters)
2. Convolutional Neural Network (2M learnable parameters)
3. Audio Spectrogram Transformer (80M learnable parameters)

The first model, denoted as VGGISH, uses a pre-trained VGGish network to extract audio information. The VGGish network (Hershey et al., 2017) is originally trained using AudioSet, a dataset of short audio events used for audio classification tasks (Gemmeke et al., 2017). This network is not updated during our training process. The second model, denoted as CONV, uses a six-layer convolutional network, trained from scratch using our data. The third model, denoted as AST, uses the Audio Spectrogram Transformer, which is a current state-of-the-art (SoTA) model for audio classification tasks (Gong et al., 2021). This model does not use the temporal aggregation model but

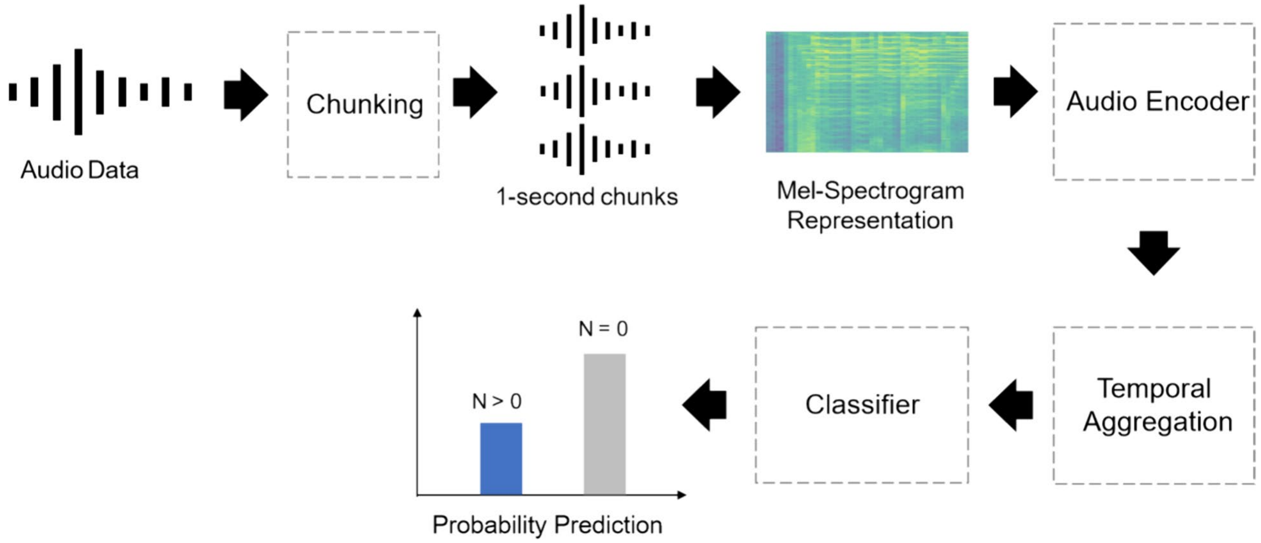


Fig. 3 The overall workflow of our audio analysis system. Note: *N* refers to the number of pedestrians detected

produces per-second probabilities without longer context. For the temporal aggregation for VGGISH and CONV, we use a transformer encoder, a SoTA model used for sequence modeling (Vaswani et al., 2017). All models are trained using a binary cross entropy loss for binary classification. For details regarding hyperparameters and the implementation of our models, please see our ASPED dataset publication (Seshadri et al., 2024).

4.2.2 Experimental setup

By undertaking three experiments, we evaluated several attributes within our workflow, including model type, detection boundary, and strength of pedestrian signals.

> **Experiment 1 — Comparison of performance of each audio encoder model**: We trained each of the aforementioned models in identical routines to determine relative performance.
> **Experiment 2 — Comparison of performance across buffers of different radii**: We investigated the effect of the recording radius on the performance of each model by training and testing each model for different buffer radii $r \in [1\text{ m}, 3\text{ m}, 6\text{ m}, 9\text{ m}]$. We anticipate that smaller radii have less but stronger signals of pedestrians, while larger radii have more but weaker signals of pedestrians.
> **Experiment 3 — Impact of training/testing thresholds on performance**: To investigate the model's response to the strength of pedestrian signals, we set thresholds during training and testing for binary classification (i.e., only considering pedestrian counts above $N_{\text{Threshold}}$-pedestrians as pedestrian events). We expect that pedestrian events under high thresholds correlate to stronger signals, while those under low thresholds correlate to weaker signals. In this study, we investigated setting $N_{\text{Threshold}} \in [1, 2, 3, 4]$ during both model training and testing.

Due to the considerable class imbalance in our dataset, a classifier trained in a standard approach would likely learn to simply ignore the underrepresented class and predict the overrepresented class. To mitigate this, we add two enhancements to our training routine. First, during each training batch, we oversample the underrepresented class with replacement, such that each batch contains roughly half audio samples with pedestrian events. This exposes the model to each class roughly equally per-batch, with complete exposure to the dataset happening over multiple epochs of training, rather than per each epoch. Second, we apply a weighting function to our loss term, which roughly weights pedestrian events and no-pedestrian events equally in the loss term, such that both contribute to model learning equally. The function and overall loss term are shown below:

$$\ell = \lambda \ell_{\text{BCE+}} + (1 - \lambda)\ell_{\text{BCE-}} \tag{1}$$

$$\lambda = \begin{cases} \dfrac{1/_{num^+}}{1/_{num^+} + 1/_{num^-}}, \textit{if } num^+ \neq 0 \\ 0, \textit{if } num^+ = 0 \end{cases} \tag{2}$$

4.3 Pedestrian flow prediction

4.3.1 Experimental setup

We also conducted a pilot study on a street level at a location referred to as *Cadell Courtyard* within our dataset.

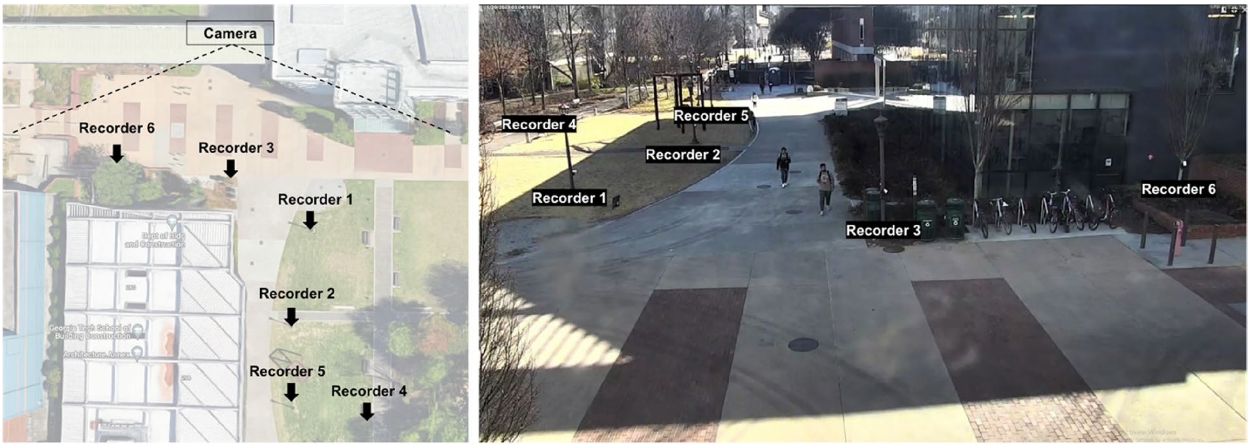


Fig. 4 The aerial (left) and the surveillance camera (right) view of the experiment site

Our ultimate goal is to forecast the inflow and outflow of pedestrians across street networks. However, this paper presents an initial step, where we concentrate on short-term predictions of the distribution of pedestrians across our sensors over time. Additionally, we propose a framework describing how our dataset could be improved and leveraged to conduct inflow-outflow prediction.

To achieve short-term pedestrian flow prediction, we undertake the following tasks: (i) pedestrian detection from video sensors, (ii) training a CNN using this data, and (iii) utilizing the sliding window method for short-term prediction. Earlier sections of this paper have outlined the experimentation of the first step, and we are actively working on the subsequent stages as described in this section.

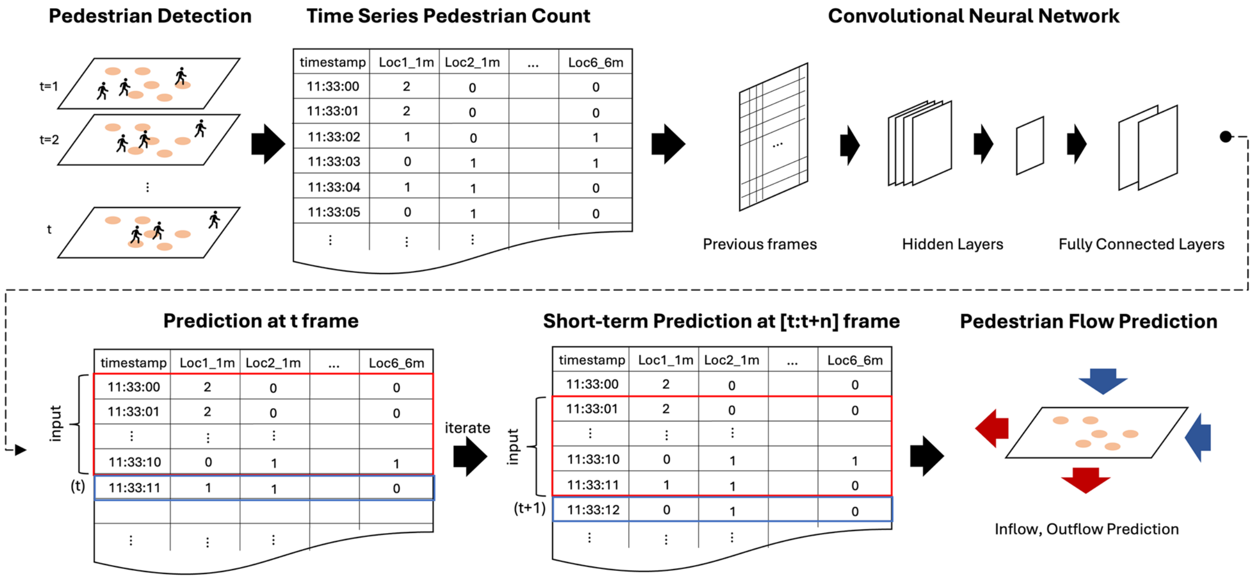
At the experiment site, we set up six audio recorders and employed a surveillance camera overseeing the entire area (Fig. 4). Recognizing that our methods for audio-based pedestrian sensing are in the process of refinement, we relied on the video feeds that are used to annotate the audio data. The intention is to develop pedestrian flow estimation algorithms with data extracted from video feeds and then update the validated models with audio-based sensor data when such data are deemed robust. Since our audio data does not contain information about pedestrian direction, we extracted only the count of pedestrians from video frames despite the capability to predict pedestrian directions.

While simple mathematical methods like linear regression can be effective for such small-scale, densely arranged sensor networks, we encountered several limitations with this approach. For instance, linear regression requires creating separate models for each recorder, a process that can be inefficient for larger-scale predictions. Additionally, this method fails to account for the fluctuations in pedestrian flow that occur at different times of the day or during various seasons.

To address these issues and to better represent the non-linear relationships between recorders, we formulated a predictive framework for estimating pedestrian flow utilizing a CNN (Fig. 5). This method processes the data of pedestrians detected in the past 11 frames (i.e., 11 s) from all recorders, to simultaneously predict pedestrian counts at each recorder location and for every radial distance (1 m, 3 m, 6 m, and 9 m). The input for this CNN includes a comprehensive set of 25 data columns from all recorder locations and radii (6 locations times 4 boundaries), along with the timestamp, which could provide valuable insights into the time of day or date, potentially influencing pedestrian traffic flow patterns.

Before feeding the data into the input layer, timestamps were converted to UNIX timestamps for easier processing. Additionally, we excluded periods where all recorders consistently reported zero pedestrians over 11 consecutive frames, as these provide no useful data for prediction. Including these frames, our test accuracy would always be close to perfect, as over 90% of our initial dataset has no pedestrians. We combined the five sessions in our dataset, which we randomly divided into training and testing sets using an 80/20 split. This approach allows for extensive training data and covers a wide range of times when pedestrian activity occurs. To standardize the data, we centered it around zero by subtracting the mean and scaled it by dividing it by the standard deviation, calculated element-wise for each input feature. These statistics were also applied to pre-process the test set data.

For CNN, we used a basic shallow architecture with 4 2-dimensional convolutional layers, 1 max pool layer, and 2 fully connected layers. We used a stochastic gradient


Pedestrian Detection
t=1
t=2
⋮
t
Time Series Pedestrian Count
Convolutional Neural Network
...
Previous frames
Hidden Layers
Fully Connected Layers
Prediction at t frame
input
(t)
iterate
Short-term Prediction at [t:t+n] frame
input
(t+1)
Pedestrian Flow Prediction
Inflow, Outflow Prediction

timestamp	Loc1_1m	Loc2_1m	...	Loc6_6m
11:33:00	2	0		0
11:33:01	2	0		0
11:33:02	1	0		1
11:33:03	0	1		1
11:33:04	1	1		0
11:33:05	0	1		0
⋮	⋮	⋮		

timestamp	Loc1_1m	Loc2_1m	...	Loc6_6m
11:33:00	2	0		0
11:33:01	2	0		0
⋮	⋮	⋮		
11:33:10	0	1		1
11:33:11	1	1		0
⋮	⋮	⋮		

timestamp	Loc1_1m	Loc2_1m	...	Loc6_6m
11:33:00	2	0		0
11:33:01	2	0		0
⋮	⋮	⋮		
11:33:10	0	1		1
11:33:11	1	1		0
11:33:12	0	1		0
⋮	⋮	⋮		

Fig. 5 Pedestrian flow prediction framework

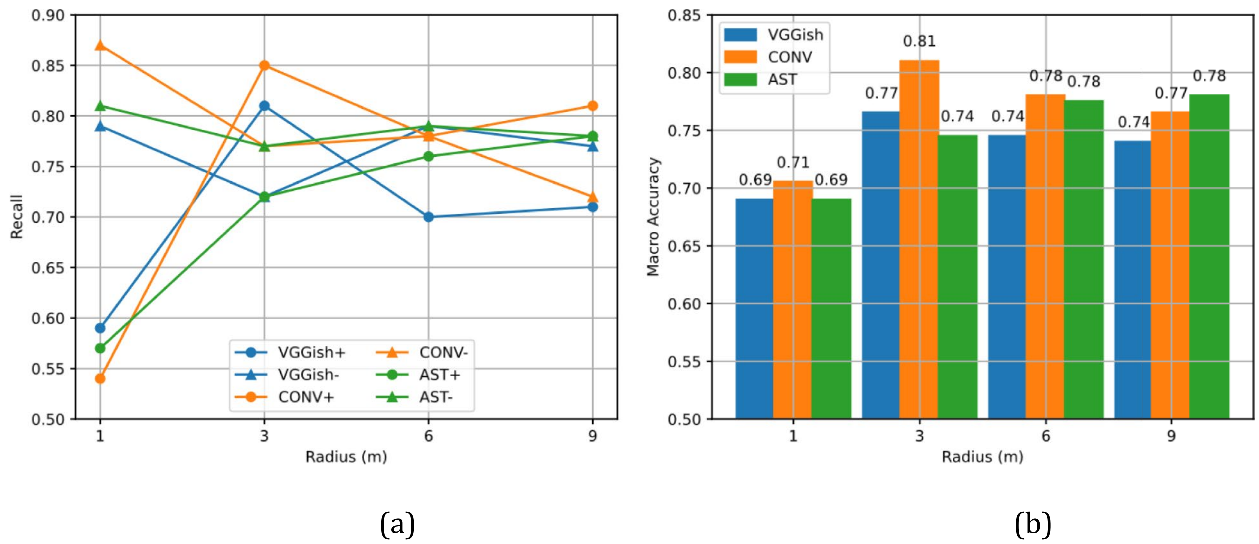


Fig. 6 **a** Recall for each class over each model and recording radius. Positive and negative classes are denoted by "+" and "-", respectively. **b** Macro average accuracy using the VGGISH, CONV, and AST models. Source: (Seshadri et al., 2024)

descent (SGD) optimizer for 25 epochs with a learning rate of 0.001.

5 Results

5.1 Pedestrian detection

Figure 6a and b present the model and radius comparisons for Experiments 1 and 2. We make the following observations. Generally, we find that audio encoders trained on our source task (CONV, AST) generally outperform the model not fine-tuned to pedestrian detection (VGGISH). CONV achieves the highest performance with middle radii 3 and 6, while VGGISH and AST are relatively constant from radii 3 to 9. We also observe that negative class recall slightly outperforms positive class recall, as expected, due to data distributions. The mitigation by loss weighting and oversampling seems to work, however, as the effect is considerably less dramatic than the dataset imbalance itself.

We note that while the test sets for each radius are identical in audio content, they are not identical in labels. Thus, the class distributions differ, and the results are not directly comparable. We find the most balanced

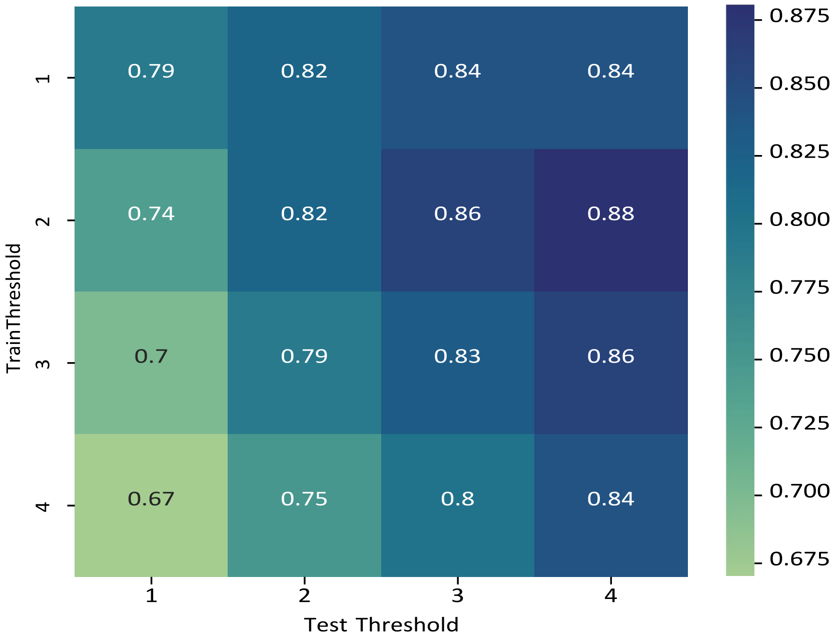


Fig. 7 Macro average accuracy over each train and test pedestrian count threshold for radius $r = 6$ m. Source: (Seshadri et al., 2024)

performance using middle radii 3 and 6 m, while radii 1 and 9 m see a slight decline. Proximity to the microphone and shifting data distributions (higher count of pedestrian events as the radius increases) likely explain the change in performance among different radii.

For Experiment 3, Fig. 7 visualizes the macro accuracy for each permutation of combinations of testing and training threshold in pedestrian count, with $N_{\text{Threshold}} \in [1,2,3,4]$. We can make the following observations: First, during testing, we see a greater proportion of correctly classified samples as the pedestrian count increases across all permutations. This is unsurprising since a larger count of pedestrian activity likely correlates to stronger and more easily detected signals. Second, while increasing the threshold during training, we find that performance generally decreases, which implies that the classifier training benefits from more difficult samples. Overall, we find optimal performance when trained with low pedestrian counts, classifying samples with high pedestrian counts (upper right box of Fig. 7).

5.2 Pedestrian flow detection

In predicting the number of pedestrians around the four levels of radii (1 m, 3 m, 6 m, and 9 m), the prediction accuracy is presented in Table 1. The CNN demonstrated high accuracy in predicting pedestrian numbers within a 1 m radius of each recorder, likely due to a prevalence of zero values in this range. Its performance was similarly strong for a 3 m radius, though with a notable decrease in accuracy at recorder location 3, possibly due to its complex positioning at a 4-way intersection. The accuracy for a 6 m radius remained above 90% for most locations, even exceeding 95% for recorders 1, 4, and 6. However, as the radius increased to 9 m, accuracy dropped to around 80% across all locations, with recorder 4 maintaining high performance, potentially due to fewer pedestrians passing by.

Table 1 Prediction accuracy by target boundary size

Target Recorder	Prediction Accuracy by Target Boundary Size (%)			
	1 m	3 m	6 m	9 m
1	99.90	99.58	95.45	82.89
2	99.90	99.55	93.18	86.78
3	99.66	93.97	86.52	79.20
4	99.65	98.68	96.78	95.73
5	99.99	99.90	91.86	85.31
6	99.91	99.80	96.28	80.16

Recorders 1, 2, 5, and 6 saw a significant decrease in accuracy at the 9 m radius, reflecting the complexity added by the larger area, as the prediction difficulty scales with the square of the radius. Despite these variations, the CNN did not overfit and maintained consistent performance on the test set.

Our results indicate that, given the high accuracy for predicting the next frame based on the previous 11 frames, a sliding window method could be effective for predicting pedestrian flow over a short period. To illustrate, using the data from frames $t-11$ to $t-1$ to predict frame t, we can then shift the window to predict frame $t+1$ using frames $t-10$ to t, and so on. However, the reliability of this method is limited to a short time frame, as it doesn't account for the potential influx of new pedestrians over longer periods.

For our next steps on large-scale pedestrian flow prediction, we aim to strategically position sensors during our next data collection. We will gather comprehensive data on pedestrian movement in various directions, essential for effectively training our model to predict the inflow and outflow. Additionally, we plan to implement GCNs. This approach is similar to CNNs, but it involves transforming our input data into a graph structure rather than a 2D matrix. In this graph format, sensor locations will be represented as nodes, and edges will symbolize the relative influence factors between nodes, including aspects such as topography, physical distances, and types of road segments.

6 Discussion and conclusion

In this study, we conducted experiments to assess the efficacy of pedestrian sensing using audio sensors. As a result, the CONV and AST models outperformed the VGGISH model in pedestrian detection, with the most consistent accuracy observed at radii of 3 to 6 m. The AST model showed a balanced performance across classes, and training methods effectively addressed data class imbalances. Accuracy varied across different pedestrian count thresholds. The best performance was achieved when models were trained on lower pedestrian counts and tested on higher counts, indicating better detection of stronger pedestrian signals. The results demonstrate promising potential for achieving reliable pedestrian prediction outcomes solely using audio sensors.

Moreover, we conducted a street-scale pilot experiment on short-term pedestrian flow prediction. Utilizing CNN effectively predicted pedestrian counts within varying radii, though we noted a decrease in accuracy as the radius increased. The experiment also highlights the potential of the sliding window method for short-term prediction using recent data for continuous forecasting.

These insights lay the groundwork for our following discussion, where we aim to address challenges associated with the dataset, identify future directions for improving the reliability of audio sensing, and examine the role of pedestrian flow prediction in urban planning, with a particular focus on the contribution of audio sensing to this field.

6.1 Towards intricate, robust and safe audio sensing

While our audio sensing system shows promising results, the performance of these basic algorithms is not yet comparable to video-based systems. Consequently, our immediate goal is to refine the accuracy of audio sensing. A potential limitation is that some of our audio features are trained on audio event detection, which is a more coarse-grained task than pedestrian detection. The challenge is compounded by the subtle nature of pedestrian sounds, influenced by diverse factors such as location, road surface, walking speed, and footwear, making the task particularly complex for the current audio features. Thus, we plan to develop a model specifically tailored to recognize pedestrian sounds by identifying distinctive patterns within the data.

Additionally, our dataset, primarily recorded in a campus setting, lacks the complexities of urban environments, such as vehicular noise. For the system to function effectively in a city, it must accurately detect pedestrians across various scenarios. To generalize our system to cover such scenarios, we could collect more data that includes different road conditions or use data augmentation during our training, which manually adds possible noise into our data to imitate some complex use cases.

Furthermore, privacy concerns arise from the nature of audio data, which may capture people's voices and potentially fragments of conversations containing sensitive information. To address this, we employed the Whisper speech-to-text model from OpenAI[3] to analyze our dataset for audible conversations. Our findings showed that in the initial dataset, conversations were either absent or not sufficiently loud and continuous for transcription. Typically, only isolated words were captured, not forming coherent sentences, likely because individuals were either consistently walking or not sufficiently close to the recorders. However, there remains a concern for scenarios where individuals decide to stop near the recorders and engage in substantive conversations.

To mitigate the leaking of private conversations in such instances, we propose modifying the audio segments where voices are clear enough for transcription. By employing a source separation algorithm to remove voices, we can obscure any identifiable or private dialogue while maintaining the integrity of the data for our model. This approach ensures we respect individuals' privacy while benefiting from the potential benefits of audio-based pedestrian detection.

6.2 Audio sensing for pedestrian flow prediction

Considering the difficulties researchers often face in accessing mobile GPS data for urban planning, our publicly accessible dataset can be a crucial resource for urban crowd management and pedestrian infrastructure

[3] https://openai.com/research/whisper, last access date Sep 5, 2023

investments. We especially anticipate it will significantly aid pedestrian flow prediction projects, ultimately leading to smarter, more effective urban design strategies.

By forecasting pedestrian flow patterns, urban planners can optimize street and road network designs to meet the dynamic needs of the city's inhabitants in each street segment at different times. Such knowledge allows for an informed categorization of streets and targeted interventions, distinguishing those bustling with activity from the underused sections. For example, areas identified as high pedestrian traffic zones can be prioritized for safety improvements, amenities, or aesthetic enhancements, while underutilized areas can be reimagined to better serve the community. Furthermore, pedestrian flow prediction has become a critical component in urban scenario planning, particularly in forecasting the impact of proposed changes in the built environment on foot traffic (Sevtsuk et al., 2021). For example, when there is a change in land use in a neighborhood or a new public infrastructure in an area, pedestrian flow prediction models could aid urban planners in assessing and simulating the impacts of these interventions.

In this context, integrating audio sensing into pedestrian flow prediction methodologies offers an exciting avenue for enhancing the accuracy and reliability of these predictions. Unlike video surveillance, audio sensing devices maintain their effectiveness across a range of lighting conditions and can provide valuable data in areas where traditional pedestrian sensing setups might fall short. This technology's unique capabilities, including its resilience to varying light conditions and its effectiveness in densely built-up areas, position it as a complementary tool that could significantly advance our understanding and forecasting of pedestrian movement patterns, ultimately contributing to the development of more livable and responsive urban environments.

Acknowledgements

The authors acknowledge the assistance of Vishnu Mano and Jing Luan in conducting this research.

Authors' contributions

Chaeyeon Han: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation. Pavan Seshadri: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation. Yiwei Ding: Writing – review & editing, Validation, Methodology, Formal analysis. Noah Posner: Sensor installation, data curation, data archiving and management, Bon Woo Koo: Formal analysis, Methodology. Animesh Agrawal: Formal analysis, Methodology. Alexander Lerch: Project administration, Investigation, Conceptualization, Supervision, Writing-review & editing. Subhrajit Guhathakurta: Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Investigation, Funding acquisition, Conceptualization.

Funding

This research was supported by the U. S. National Science Foundation [grant number CMMI-1854684].

Availability of data and materials

The ASPED data used for the audio sensing models reported in this paper can be accessed from https://urbanaudiosensing.github.io/ASPED.

Declarations

Ethics approval and consent to participate

This study received an "Exempt" status recommendation from the Institutional Review Board (IRB).

Consent for publication

All co-authors have approved this manuscript for consideration for publication in Urban Informatics.

Competing interests

The authors declare the following financial interests/personal relationships, which may be considered as potential competing interests: Subhrajit Guhathakurta and Alexander Lerch report financial support from the U.S. National Science Foundation.

Received: 6 February 2024 Revised: 9 June 2024 Accepted: 13 June 2024
Published online: 10 July 2024

References

Ai, Y., Li, Z., Gan, M., Zhang, Y., Yu, D., Chen, W., & Ju, Y. (2019). A deep learning approach on short-term spatiotemporal distribution forecasting of dockless bike-sharing system. *Neural Computing and Applications*, *31*(5), 1665–1677. https://doi.org/10.1007/s00521-018-3470-9. Retrieved 2024-01-27, from http://link.springer.com/10.1007/s00521-018-3470-9

Baker, S., Scharstein, D., Lewis, J. P., Roth, S., Black, M. J., & Szeliski, R. (2011). A database and evaluation methodology for optical flow. *International Journal of Computer Vision*, *92*(1), 1–31.

Barron, J. L., Fleet, D. J., & Beauchemin, S. S. (1994). Performance of optical flow techniques. *International Journal of Computer Vision*, *12*(1), 43–77.

Brunetti, A., Buongiorno, D., Trotta, G. F., & Bevilacqua, V. (2018). Computer vision and deep learning techniques for pedestrian detection and tracking: A survey. *Neurocomputing*, *300*, 17–33. https://doi.org/10.1016/j.neucom.2018.01.092. Retrieved 2021-09-08, from https://www.sciencedirect.com/science/article/pii/S092523121830290X

Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R. (2022). Maskedattention mask transformer for universal image segmentation. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. IEEE. Retrieved from https://doi.org/10.1109/CVPR52688.2022.00135

Deo, N., & Trivedi, M.M. (2021). Trajectory forecasts in unknown environments conditioned on grid-based plans. arXiv. Retrieved 2024-01-27, from http://arxiv.org/abs/2001.00735 (arXiv:2001.00735 [cs]).

Dollar, P., Wojek, C., Schiele, B., Perona, P. (2009). Pedestrian detection: A benchmark. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 304–311). (ISSN: 1063-6919).

Dollar, P., Wojek, C., Schiele, B., & Perona, P. (2012). Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(4), 743–761. https://doi.org/10.1109/TPAMI.2011.155

Fei-Fei, L., Fergus, R., & Perona, P. (2006). One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(4), 594–611.

Fields, B. (2012). *Active transportation measurement and benchmarking development: New Orleans pedestrian and bicycle count report, 2010-2011. (Tech. Rep.)*. Gulf Coast Research Center for Evacuation and Transportation Resiliency.

Figliozzi, M., Monsere, C., Nordback, K., Johnson, P., Blanc, B. (2014). Design and implementation of pedestrian and bicycle-specific data collection methods in Oregon (Tech. Rep. No. FHWA-OR-RD-14-15. SPR 754). Retrieved from https://pdxscholar.library.pdx.edu/cenginfac/345

Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., ... & Ritter, M. (2017). Audio set: An ontology and human-labeled dataset for audio events. *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 776–780).

Gong, Y., Chung, Y.-A., Glass, J. (2021). *AST: Audio spectrogram transformer*. Interspeech.

González, M. C., Hidalgo, C. A., & Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, *453*(7196), 779–782. https://doi.org/10.1038/nature06958. Retrieved 2024-01-27, from https://www.nature.com/articles/nature06958

Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., ... & Wilson, K. (2017). CNN architectures for large-scale audio classification. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 131–135). Institute of Electrical and Electronics Engineers (IEEE). (ISSN: 2379-190X).

Hughes, R.G., Huang, H., Zegeer, C.V., Cynecki, M.J. (2001). *Evaluation of automated pedestrian detection at signalized intersections (Tech. Rep.)*. United States. Federal Highway Administration.

Jiang, S., Ferreira, J., Jr., & Gonzalez, M. C. (2017). Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore. *IEEE Transactions on Big Data*, *3*(2), 208–219. Retrieved from http://hdl.handle.net/1721.1/120769

Jones, M.G., Ryan, S., Donlon, J., Ledbetter, L., Ragland, D.R., Arnold, L.S. (2010). *Seamless travel: Measuring bicycle and pedestrian activity in San Diego County and its relationship to land use, transportation, safety, and facility type (Tech. Rep.)*.

Kitano, Y., Kuwamoto, S., Asahara, A. (2019). OD-networkbased pedestrian-path prediction for people-flow simulation. In *2019 IEEE International Conference on Big Data (Big Data)* (pp. 1656–1661). IEEE. Retrieved 2024-01-27, from https://ieeexplore.ieee.org/document/9006314

Lee, J. H., Hancock, M. G., & Hu, M.-C. (2014). Towards an effective framework for building smart cities: Lessons from Seoul and San Francisco. *Technological Forecasting and Social Change*, *89*, 80–99. https://doi.org/10.1016/j.techfore.2013.08.033. Retrieved 2018-02-26, from http://linkinghub.elsevier.com/retrieve/pii/S0040162513002187

Lerch, A. (2023). *An introduction to audio content analysis: Music information retrieval tasks and applications* (2nd ed.). Wiley-IEEE Press. Retrieved 2022-11-04, from https://ieeexplore.ieee.org/servlet/opac?bknumber=9965970

Li, H., Wu, Z., Zhang, J. (2016). Pedestrian detection based on deep learning model. In *2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)* (pp. 796–800).

Li, G., Yang, Y., & Qu, X. (2020). Deep learning approaches on pedestrian detection in hazy weather. *IEEE Transactions on Industrial Electronics*, *67*(10), 8889–8899. https://doi.org/10.1109/TIE.2019.2945295. (Conference Name: IEEE Transactions on Industrial Electronics).

Lin, Z., Feng, J., Lu, Z., Li, Y., & Jin, D. (2019). DeepSTN+: Context-aware spatial-temporal neural network for crowd flow prediction in metropolis. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 1020–1027. https://doi.org/10.1609/aaai.v33i01.33011020. Retrieved 2024-01-27, from https://ojs.aaai.org/index.php/AAAI/article/view/3892

Liu, M., Li, L., Li, Q., Bai, Y., & Hu, C. (2021). Pedestrian flow prediction in open public places using graph convolutional network. *ISPRS International Journal of Geo-Information*, *10*(7), 455. https://doi.org/10.3390/ijgi10070455. Retrieved 2024-01-27, from https://www.mdpi.com/2220-9964/10/7/455

Martin, D. R., Fowlkes, C. C., & Malik, J. (2004). Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(5), 530–549. (Publisher: IEEE).

Mathews, E., & Poigné, A. (2009). Evaluation of a "smart" pedestrian counting system based on echo state networks. *EURASIP Journal on Embedded Systems*, *2009*, 1–9.

Minge, E., Falero, C., Lindsey, G., Petesch, M., & Vorvick, T. (2017). *Bicycle and pedestrian data collection manual (Tech. Rep.)*. Minnesota Department of Transportation.

Oliver, N., Lepri, B., Sterly, H., Lambiotte, R., Deletaille, S., De Nadai, M., … & Vinck, P. (2020). Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle. *Science Advances*, *6*(23), eabc0764. https://doi.org/10.1126/sciadv.abc0764. Retrieved 2024-01-27, from https://www.science.org/doi/10.1126/sciadv.abc0764

Ozan, E., Searcy, S., Geiger, B.C., Vaughan, C., Carnes, C., Baird, C., Hipp, A. (2021). *State-of-the-art approaches to bicycle and pedestrian counters* (p. 84).

Ozbay, K., Bartin, B., Yang, H., Walla, R., Williams, R. (2010). *Automated pedestrian counter: final report (Tech. Rep. No. FHWA-NJ-2010-001)*. Retrieved from https://rosap.ntl.bts.gov/view/dot/17680

Rahman, M., Islam, M., Calhoun, J., & Chowdhury, M. (2019). Real-time pedestrian detection approach with an efficient data communication bandwidth strategy. *Transportation Research Record*, *2673*(6), 129–139.

Ratti, C., Frenchman, D., Pulselli, R. M., & Williams, S. (2006). Mobile landscapes: Using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, *33*(5), 727–748. https://doi.org/10.1068/b32047. Retrieved 2024-01-27, from http://journals.sagepub.com/doi/10.1068/b32047

Ryus, P., Ferguson, E., Laustsen, K.M., Prouix, F.R., Schneider, R.J., Hull, T., Miranda-Moreno, L. (2014). *Methods and technologies for pedestrian and bicycle volume data collection*. Citeseer.

Scharstein, D., & Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, *47*(1), 7–42. (Publisher: Springer).

Seshadri, P., Han, C., Koo, B.-W., Posner, N., Guhathakurta, S., Lerch, A. (2024). ASPED: An audio dataset for detecting pedestrians. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Institute of Electrical and Electronics Engineers (IEEE). Retrieved 2023-12-14, from http://arxiv.org/abs/2309.06531

Sevtsuk, A., Basu, R., & Chancey, B. (2021). We shape our buildings, but do they then shape us? A longitudinal analysis of pedestrian flows and development activity in melbourne. *PLoS One*, *16*(9), e0257534. https://doi.org/10.1371/journal.pone.0257534

Sun, J., Zhang, J., Li, Q., Yi, X., Liang, Y., & Zheng, Y. (2022). Predicting citywide crowd flows in irregular regions using multi-view graph convolutional networks. *IEEE Transactions on Knowledge and Data Engineering*, *34*(5), 2348–2359. https://doi.org/10.1109/TKDE.2020.3008774. Retrieved 2024-01-27, from https://ieeexplore.ieee.org/document/9139357/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Polosukhin, I. (2017). Attention is all you need. In I. Guyon, et al. (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wolter, S.A., & Lindsey, G. (2001). *Summary report. Indiana Trail Study: A summary of trails in 6 Indiana cities*. Eppley Institute for Parks and Public Lands at Indiana University. Funded by: Indiana Department of Transportation, Indiana Department of Natural Resources, National Park Service.

Xia, T., Lin, J., Li, Y., Feng, J., Hui, P., Sun, F., & Jin, D. (2021). 3DGCN: 3-dimensional dynamic graph convolutional network for citywide crowd flow prediction. *ACM Transactions on Knowledge Discovery from Data*, *15*(6), 1–21. https://doi.org/10.1145/3451394. Retrieved 2024-01-27, from https://dl.acm.org/doi/10.1145/3451394

Yabe, T., Tsubouchi, K., Shimizu, T., Sekimoto, Y., Sezaki, K., Moro, E., Pentland, A. (2023). Metropolitan scale and longitudinal dataset of anonymized human mobility trajectories. https://doi.org/10.48550/ARXIV.2307.03401. Retrieved 2024-01-27, from https://arxiv.org/abs/2307.03401 (Publisher: arXiv Version Number: 1).

Yang, H., Ozbay, K., Bartin, B. (2010). Investigating the performance of automatic counting sensors for pedestrian traffic data collection. In *Proceedings of the 12th world conference on transport research* (Vol. 1115, pp. 1–11).

Yang, H., Ozbay, K., & Bartin, B. (2011). Enhancing the quality of infrared-based automatic pedestrian sensor data by nonparametric statistical method. *Transportation Research Record: Journal of the Transportation Research Board*, *2264*(1), 11–17. https://doi.org/10.3141/2264-02. Retrieved 2021-09-07, from http://journals.sagepub.com/doi/10.3141/2264-02

Zhang, D., & Kabuka, M. R. (2018). Combining weather condition data to predict traffic flow: A GRU-based deep learning approach. *IET Intelligent Transport Systems*, *12*(7), 578–585. https://doi.org/10.1049/iet-its.2017.0313. Retrieved 2024-01-27, from https://onlinelibrary.wiley.com/doi/10.1049/ietits.2017.0313

Zhang, J., Zheng, Y., Qi, D. (2017). Deep spatio-temporal residual networks for citywide crowd flows prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1). https://doi.org/10.1609/aaai.v31i1.10735. Retrieved 2024-01-27, from https://ojs.aaai.org/index.php/AAAI/article/view/10735

Zhang, J., Zheng, Y., Qi, D., Li, R., Yi, X., & Li, T. (2018). Predicting citywide crowd flows using deep spatio-temporal residual networks. *Artificial Intelligence*, *259*, 147–166. https://doi.org/10.1016/j.artint.2018.03.002. Retrieved 2024-0127, from https://linkinghub.elsevier.com/retrieve/pii/S0004370218300973

Zhuang, Y., Kang, Y., Fei, T., Bian, M., & Du, Y. (2024). From hearing to seeing: Linking auditory and visual place perceptions with soundscape-to-image generative artificial intelligence. *Computers, Environment and Urban Systems*, *110*, 102122. https://doi.org/10.1016/j.compenvurbsys.2024.102122. Retrieved from https://www.sciencedirect.com/science/article/pii/S0198971524000516

Publisher's Note