Audio-Based Pedestrian Detection in the Presence of Vehicular Noise

Yonghyun Kim[1], *Chaeyeon Han*[2], *Akash Sarode*[3], *Noah Posner*[2], *Subhrajit Guhathakurta*[2], *Alexander Lerch*[1]

[1]Music Informatics Group, Georgia Institute of Technology, USA
[2]Center for Urban Resilience and Analytics, Georgia Institute of Technology, USA
[3]College of Computing, Georgia Institute of Technology, USA

***Abstract*—Audio-based pedestrian detection is a challenging task and has, thus far, only been explored in noise-limited environments. We present a new dataset, results, and a detailed analysis of the state-of-the-art in audio-based pedestrian detection in the presence of vehicular noise. In our study, we conduct three analyses: (i) cross-dataset evaluation between noisy and noise-limited environments, (ii) an assessment of the impact of noisy data on model performance, highlighting the influence of acoustic context, and (iii) an evaluation of the model's predictive robustness on out-of-domain sounds. The new dataset is a comprehensive 1321-hour roadside dataset. It incorporates traffic-rich soundscapes. Each recording includes 16 kHz audio synchronized with frame-level pedestrian annotations and 1 fps video thumbnails.**

***Index Terms*—Audio databases, Sound event detection, Urban sound analysis, Pedestrian detection, Vehicular noise**

1. INTRODUCTION

Pedestrian volume data offer valuable insights into urban activity patterns, which support planning efforts such as evaluating sidewalk improvements, assessing land use changes, and identifying areas needing investments in safety and walkability [1]. These data also support optimizing street connectivity and accessibility [1].

The widespread adoption of smartphones has brought new opportunities for automated human mobility sensing, particularly through mobile GPS data. However, growing privacy concerns, particularly under frameworks like the General Data Protection Regulation (GDPR) in the European Union, have placed restrictions on using mobile location data to track individuals [2]. In parallel, smart city initiatives have adopted the deployment of IoT-based sensors to monitor activity in urban environments. These efforts have largely focused on vision-based systems, such as computer vision and infrared cameras [3], although other sensing technologies have also been tested.

Urban sound offers a promising alternative. Microphones are affordable, energy-efficient, and effective in visually occluded environments. They can complement or replace cameras in contexts where installation is impractical, such as shaded areas, narrow corridors, or locations or scenarios where the costs of cameras are prohibitive. The general feasibility of using microphone recordings for the detection of pedestrians has been shown recently for a vehicle-free courtyard on a university campus [4].

This study addresses two key gaps in existing work. First, the generalizability of audio-based models remains unclear. Given the variability in urban soundscapes, shaped by traffic, land use, and average pedestrian activity levels, it is necessary to evaluate model performance across data collected from different settings, particularly in the presence of typical urban noise. Second, existing studies lack information on interpretability; it is unclear which sound characteristics existing models rely on for detecting pedestrians.

Thus, the main contributions of this study are

(i) a new publicly available[1] dataset for audio-based pedestrian detection in the presence of vehicular noise,
(ii) an investigation into how vehicular noise affects pedestrian detection performance, and
(iii) insights into the acoustic features that enable pedestrian detection.

[1]https://huggingface.co/datasets/urbanaudiosensing/ASPEDvb

2. RELATED WORK

2.1. Automated Pedestrian Detection Techniques

Urban pedestrian sensing technologies have evolved over decades, with video cameras and infrared sensors being the most widely deployed to date [3], [5], [6]. Video-based systems, now commonly augmented with computer vision and deep learning techniques, offer high spatial precision but can suffer from limitations in occluded or low-light environments. Furthermore, such systems often raise privacy concerns [7], [8]. Infrared counters, including active, passive, and target-reflective types, are less intrusive but tend to undercount pedestrians, particularly in high pedestrian volume scenarios [5], [6]. More sophisticated but cost-prohibitive options, such as radar, piezoelectric strips, and inductive loops, are limited in spatial scalability [9]. In contrast, audio-based pedestrian sensing remains underexplored, albeit with promising low-cost deployment, resilience to visual obstructions, and potential privacy advantages. As demonstrated by Seshadri et al. [4], audio-based systems can detect the presence of pedestrians by using advances in acoustic scene analysis and deep learning, although challenges persist in signal separation, data imbalance, and generalizability across urban soundscapes.

The generalizability of pedestrian detection models has been explored only recently. Rasouli et al. assessed seven state-of-the-art detection algorithms under varying real-world conditions using the JAAD dataset and found that model performance deteriorates in changed contexts, such as different weather conditions, pedestrian behaviors, or occlusion [10]. They emphasized the importance of incorporating diverse training data, showing that general-purpose object detection models trained on broader datasets tend to generalize better than those trained narrowly on pedestrian-focused inputs. More recently, Hasan et al. conducted a cross-dataset evaluation of pedestrian detectors and similarly found that traditional models generalized poorly because their training source usually does not contain dense pedestrian volume [11]. Interestingly, general-purpose object detectors, not trained for pedestrian detection, showed better cross-dataset performance, suggesting that varied training sources can improve model transferability. Although these studies do not focus on audio-based models, they emphasize that testing generalizability across datasets is crucial. In the context of audio-based sensing, this implication is particularly relevant, as urban soundscapes can vary considerably depending on the surrounding environment.

2.2. Audio-based Urban Sensing

Urban sound has emerged as a rich source of information for understanding city life, complementing traditional visual or spatial data. Early urban noise studies primarily emphasized environmental health and policy, focusing on the quantification of noise pollution from road traffic, railways, and industrial sources [12]–[14]. These works led to the development of standardized noise maps and public

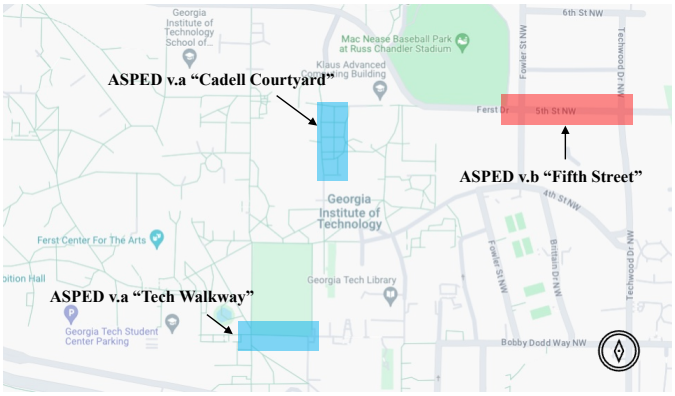

Fig. 1: Data collection sites on the Georgia Tech campus in Atlanta.

health guidelines (e.g., [15]). However, beyond its value as a nuisance, urban sound is increasingly recognized as a medium that implies information about human activity, mobility patterns, and the social vibrancy of public spaces [16], [17].

Recent advances in sensing technologies and machine learning enable granular, automated analysis of urban soundscapes. Projects such as SONYC (Sounds of New York City) [18] have established a baseline for classifying general urban sounds, using annotations for broad event categories that include speech-oriented human sounds. Extending this scope of urban audio analysis further, Han et al. and Seshadri et al. introduced audio-based methods for detecting pedestrian presence [4], [19]. Their approach utilized a new large-scale dataset with pedestrian-focused annotations. This dataset is composed of continuous recordings from real-world walking environments, enable models to learn from the full range of implicit acoustic cues (both speech and non-speech) that signal pedestrian presence. Their results highlight the potential of microphone-based sensing as a low-cost, privacy-preserving, and scalable complement to camera-based systems.

Despite recent progress, the generalizability of models across diverse urban environments and the interpretability of these models remain underexplored. Understanding the level of generalizability and audio cues that trigger models to predict pedestrian presence is —given the variety of urban soundscape— crucial for building robust and interpretable systems.

3. DATASET

This study builds on the previously published ASPED dataset [4], which includes annotated audio and video data collected in a vehicle-free courtyard environment and will be referred to in the following as ASPED v.a. This dataset provides the foundation for our pedestrian detection framework and is described in detail by Seshadri et al. [4]. The recorder setup and preprocessing steps are identical to those used in ASPED v.a.

In this study, we introduce an additional dataset, ASPED v.b, collected close to a road with vehicular traffic. Figure 1 highlights the recording location on the Georgia Tech campus in red. The vehicular noise primarily consists of engine sounds and intermittent shuttle buses operating at slow speeds. The proportion of frames containing at least one vehicle detected is 9.16%, 29.00%, 36.43%, and 42.91% for radii of 1 m, 3 m, 6 m, and 9 m, respectively.

The ASPED v.b dataset contains 1,321 hours of audio from 4 different sessions. Each session takes place over a time frame of approximately 40 hours and has audio data collected by 4 to 8 recorders spread along a street. The recording areas are monitored

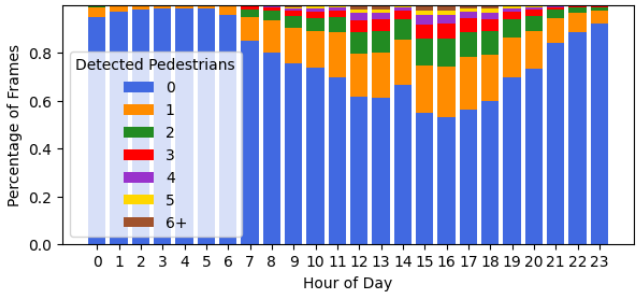

Fig. 2: Percentage of frames containing pedestrians by hour of day.

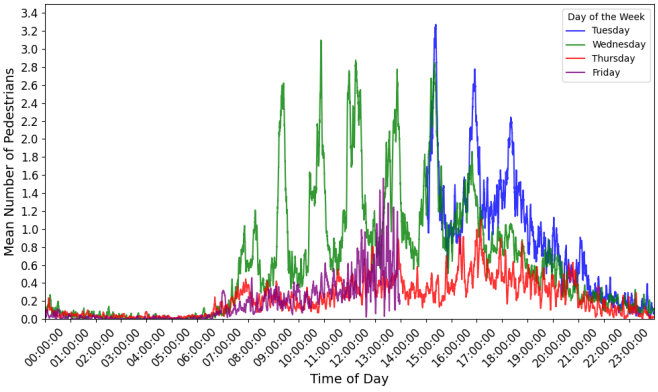

Fig. 3: Time-series distribution of pedestrian counts.

by 6 GoPro cameras, which captured 1 fps video recordings totaling 2,946,513 frames across all cameras.

Figure 2 illustrates general pedestrian patterns derived from the labels of the ASPED v.b dataset. The figure shows the ground truth number of pedestrians detected from video recordings at a specific timestamp, visualized for the recording zone with a 6 m radius. Pedestrian activity peaks between 3 PM and 5 PM and declines considerably at night. This class imbalance reflects the ecological validity of the dataset, capturing realistic periods of low activity.

The average number of pedestrians walking the street on a specific day of the week and time by taking the rolling average of the number of pedestrians detected across all cameras is shown in Fig. 3. The peaks align with the times that classes end on campus, demonstrating how pedestrian traffic on campus is closely tied to the class schedule.

Lastly, 2.9% of total frames were obstructed by buses, preventing the video-based pedestrian annotation from producing reliable labels. Therefore, these frames were flagged and discarded in modeling.

4. EXPERIMENTAL SETUP

The goal of this research is to provide new insights into audio-based pedestrian detection that might facilitate new approaches with enhanced performance. We conduct three key experiments to explore these aspects: (i) a cross-dataset evaluation to assess the generalization capabilities of models trained on noisy and noise-free sections of the datasets (v.a and v.b), (ii) evaluating the effect of vehicle presence in training data on the performance with vehicle-controlled test sets, and (iii) an analysis of the acoustic cues that the model associates with pedestrian and non-pedestrian instances.

For the experiments, we reproduced the model proposed by Seshadri et al. [4]. This model processes 10-second 16 kHz mono audio inputs by first computing power spectrograms using STFT (window: 25 ms, hop: 10 ms). These are then converted to 64-bin mel spectrograms

Table 1: Cross-dataset evaluation balanced accuracy (%).

Train Dataset	Test Dataset	
	ASPED v.a ↑	ASPED v.b ↑
ASPED v.a	71.74	66.48
ASPED v.b	64.77	69.15

(125–7500 Hz) and normalized via standard scaling. The mean and standard deviation values for normalization were obtained from the implementation provided in the GitHub repository of the ASPED v.a model.[2] The resulting log-mel spectrograms are fed into the VGGish backbone, pre-trained on AudioSet [20], to extract a sequence of 10 acoustic embeddings, each corresponding to a snippet of 1 s of the input. A Transformer encoder (1 layer, 4 attention heads, 128 hidden dimension), with added positional encoding, processes these embeddings to capture temporal dependencies. Finally, a linear projection layer with ReLU activation, followed by another linear layer and a sigmoid activation function, outputs a binary classification probability for each 1 s snippet, resulting in 10 predictions for one 10 s-input, using a batch size of 256.

4.1. Exp. 1: Cross-dataset evaluation

Following previously established methodology [4], the two datasets, ASPED v.a and v.b, were randomly partitioned into train, test, and validation subsets with an 80/10/10 split, respectively.

To address the inherent class imbalance, we employed weighted batch sampling and a variable weighted loss during training. Model inference results are reported using the checkpoint that yielded the lowest validation loss after 20 epochs.

4.2. Exp. 2: Impact of vehicle presence for training

A key difference between the previously existing dataset ASPED v.a and the new data lies in the presence of vehicle sounds in the audio recordings. In this experiment, we investigate whether this factor in the training data influences model performance on test environments with (VP: Vehicle-Present) and without vehicles (VA: Vehicle-Absent). To this end, we create two distinct test splits of ASPED v.b, controlled for vehicle presence and analyze the results for the models trained on v.a and v.b (cf. Sect. 4.1), respectively.

Furthermore, to assess the models' propensity for false positives, we sampled vehicle-related categories from the nonhuman sounds section of the FSD50K [21] dataset. FSD50K is an open dataset of human-labeled sound events containing 51,197 Freesound[3] clips unequally distributed in 200 classes drawn from the AudioSet Ontology. For this and Section 4.3, we downsampled the audio to 16 kHz and categorized it into human or non-human sounds by following the given ontology[4]. All classes in FSD50K are represented in AudioSet, except *Crash cymbal* (non-human), *Human group actions* (human), *Human voice* (human), *Respiratory sounds* (human), and *Domestic sounds, home sounds* (non-human). Only single-tagged audio samples were included in this analysis, and we filtered the dataset to include only categories containing at least 10 distinct files. The resulting refined dataset comprised 21 human sound categories (989 files) and 133 non-human sound categories (8,097 files). The probability of class 1, which the model was trained to associate with 'pedestrian' presence, was used to determine the model's response.

[2] https://github.com/urbanaudiosensing/Models/blob/main/data_utils/transforms.py, last access date: September 19, 2025

[3] https://freesound.org/, last access date: September 19, 2025

[4] https://research.google.com/audioset/ontology/human_sounds_1.html, last access date: September 19, 2025

Table 2: Impact of vehicle presence in training data — balanced accuracy (%) on ASPED v.b subsets. (VP: Vehicle-Present, VA: Vehicle-Absent)

Train Dataset	Test Dataset (ASPED v.b)	
	VP ↑	VA ↑
ASPED v.a	65.16	67.87
ASPED v.b	67.49	71.01

4.3. Exp. 3: Model sensitivity to different sound categories

To gain insights into "what the models are listening to," we analyze the sensitivity of the ASPED-trained models to various audio categories by classifying inputs from the FSD50K human and non-human sound ontologies. More specifically, we investigate which human-generated sound categories were most frequently detected as 'pedestrian.' Furthermore, we conduct a post-hoc analysis to determine if any non-human sound categories are consistently misclassified as 'pedestrian.'

A crucial consideration for this analysis is the difference in both audio characteristics/recording setup and labeling paradigms. The ASPED dataset labels are based on the presence of individuals within a certain amount of radius of the recording device (in this study, 6 m). In contrast, FSD50K annotations do not consider spatial proximity; for this evaluation, we operated under the assumption that all human sounds represent the 'pedestrian' class and all non-human sounds represent the 'non-pedestrian' class.

We further investigate the specific categories of human-related sounds that our model reliably detects or struggles to recognize. Additionally, we examine non-human sounds that are erroneously classified as pedestrian-related, leading to false positive errors.

5. RESULTS

5.1. Exp. 1: Cross-dataset evaluation

Table 1 presents the balanced accuracy, calculated as the average of sensitivity and specificity, achieved when models trained on one dataset version were tested on the other.

The results indicate a performance drop when models are tested on a dataset different from their training set, which indicates limited generalization across the two recording setups.

These cross-dataset results highlight the complex interplay between the presence of specific types of background noise, such as vehicular traffic, and model generalization. Further investigation into domain adaptation techniques may be beneficial to improve the robustness of pedestrian detection systems in real-world scenarios with varying acoustic environments.

5.2. Exp. 2: Impact of vehicle presence for training

To investigate the specific impact of vehicle presence in the training data, we evaluate the v.a-trained and v.b-trained models on subsets of v.b that were controlled for the presence or absence of vehicle sounds (VP: Vehicle-Present, VA: Vehicle-Absent), as shown in Table 2.

The results show that —as expected— the presence or absence of vehicle sounds in the test set impacts performance. Even though the v.b-trained model was exposed to vehicle sounds during training, predicting pedestrian presence in the absence of these potentially confounding sounds is simpler.

To assess whether the v.b-trained model exhibits a reduced tendency to misclassify common vehicle sounds as pedestrians compared to the v.a-trained model, we compared the average predicted probability of the 'pedestrian' class for a curated set of vehicle-related non-human sound categories from FSD50K (Table 3).

The v.a-trained model generally exhibited considerably higher average predicted probabilities for classifying vehicle sounds as

Table 3: Avg. prob. of pedestrian class for vehicle-related FSD50K categories.

Category	v.a-trained ↓	v.b-trained ↓
Race car, auto racing	0.86	0.69
Car	0.69	0.62
Vehicle	0.75	0.56
Vehicle horn, car horn, honking	0.74	0.62
Car passing by	0.71	0.59
Motor vehicle (road)	0.72	0.60

Table 4: Average probability for FSD50K human sound categories for models trained on ASPED v.a and v.b.

Category	v.a-trained ↑	v.b-trained ↑
Female singing	0.95	0.74
Speech	0.94	0.65
Crying, sobbing	0.92	0.63
Laughter	0.91	0.65
Singing	0.90	0.71
Human voice	0.89	0.64
Yell	0.89	0.66
Cheering	0.88	0.66
Chatter	0.87	0.66
Child speech, kid speaking	0.86	0.65
Human group actions	0.83	0.63
Speech synthesizer	0.82	0.59
Conversation	0.79	0.63
Burping, eructation	0.78	0.56
Male speech, man speaking	0.77	0.56
Whispering	0.76	0.52
Applause	0.76	0.48
Chewing, mastication	0.76	0.57
Hands	0.74	0.56
Run	0.74	0.56
Walk, footsteps	0.70	0.58

'pedestrian' compared to the v.b-trained model. This suggests that the absence of traffic noise during training in v.a might lead the model to erroneously associate vehicle sounds with human presence, increasing false positives. Conversely, the v.b-trained model trained with traffic noise was more effective at distinguishing pedestrian presence from vehicle sounds as indicated by fewer false alarms.

5.3. Exp. 3: Model sensitivity to different sound categories

To understand the models' sensitivity to different acoustic cues, we investigated the impact of signal energy and of different (human and non-human) sounds on the pedestrian detection accuracy.

5.3.1. Comparison with RMS energy: The Pearson correlation between the audio's RMS energy and the model's output logit is low for models trained on ASPED v.a and v.b ($r \approx 0.14$ and $r \approx 0.29$, respectively), confirming that the learned representations are more effective than a simple energy measurement.

5.3.2. Evaluation on FSD50K Human Sounds: Table 4 presents the human categories as a subset of the FSD50K dataset. On the right, we list the corresponding average predicted probability of the 'pedestrian' class for both models. The model trained on the ASPED v.a dataset demonstrates greater confidence when classifying human sounds as 'pedestrian' compared to its counterpart trained on the traffic-noise-rich ASPED v.b dataset. While speech-related sounds generally exhibited higher probabilities across both models, subtle performance variations in the ranking of specific categories might indicate that background noise during training influences the model's sensitivity to different types of human sounds. The overall lower average probabilities for the v.b-trained model likely reflect the masking effect of traffic noise on the acoustic features crucial for human sound identification. Notably, categories intuitively associated with pedestrian movement, such as *Walk, footsteps* and *Run*, were ranked relatively low within the broader set of human sound categories for both models. These findings underscore the impact of the training environment's acoustic characteristics on the learned representations and the subsequent generalization to out-of-domain human sounds. It should be noted, however, that the majority of signals in this dataset are very different from the typical urban sound recording; thus, these results should be interpreted carefully.

Table 5: Top and bottom 3 non-human sound categories by avg. prob.

Category	v.a-trained ↓	v.b-trained ↓
Top 3		
Harp	0.94 ± 0.07	0.71 ± 0.13
Trumpet	0.94 ± 0.14	0.80 ± 0.11
Plucked string instrument	0.93 ± 0.07	0.66 ± 0.15
Bottom 3		
Cricket	0.42 ± 0.33	0.48 ± 0.16
Chirp, tweet	0.51 ± 0.31	0.45 ± 0.15
Bicycle bell	0.52 ± 0.36	0.48 ± 0.14

5.3.3. Evaluation on FSD50K Non-Human Sounds: To understand the models' sensitivity to other sounds, we evaluated their predictions on a subset of 133 (categories with at least 10 samples) non-human sound categories from AudioSet. Table 5 displays the top 3 and bottom 3 categories, determined based on the v.a-trained model's average predicted probability of the 'pedestrian' class. The evaluation on non-human sounds reveals that the model trained on v.a data has a higher tendency to misclassify certain musical instruments as 'pedestrian' compared to the v.b-trained model. This may be due to such sound categories being particularly infrequent or entirely absent in the ASPED datasets. Interestingly, the bottom-ranked categories reveal greater prediction variability in the v.a-trained model compared to the v.b-trained model. This higher standard deviation suggests that the v.a model is less certain when classifying sounds that are dissimilar to human presence.

6. CONCLUSION

This research investigated the impact of the acoustic environment on pedestrian detection using a novel pedestrian detection dataset with vehicular noise. Our cross-dataset evaluation revealed a performance drop when models were trained on different environments, indicating limited domain generalization capability. Furthermore, the presence of vehicle sounds in the test set considerably influenced performance, with models showing varying sensitivities based on their training data's acoustic characteristics. Evaluation on out-of-domain FSD50K data highlighted that models trained in v.a exhibited higher confidence in identifying human sounds but were also more prone to false positives for non-pedestrian sounds. Conversely, models trained with traffic noise demonstrated more cautious predictions. However, the notable issue of false positives across various non-human sound categories warrants further attention. These findings underscore the critical role of the acoustic environment in training robust pedestrian detection systems. The limited generalization observed suggests that future work should focus on domain adaptation techniques to bridge the gap between different acoustic domains. Specifically, exploring methods to enhance the model's ability to filter out irrelevant background noise, such as vehicular traffic, while retaining sensitivity to subtle pedestrian-related cues is crucial. Additionally, we plan to investigate the integration of multi-modal information (e.g., visual cues) to increase robustness in challenging scenarios. Finally, a more comprehensive analysis of the model's failure cases, particularly the misclassification of specific non-human sounds, could inform the design of more discriminative acoustic features or robust model architectures.

REFERENCES

[1] American Planning Association, "The pedestrian count," https://www.planning.org/pas/reports/report199.htm, 1965.

[2] GDPR.eu, "General data protection regulation (gdpr) compliance guidelines," 2019, last access date: May 7, 2025. [Online]. Available: https://gdpr.eu/

[3] H. Li, Z. Wu, and J. Zhang, "Pedestrian detection based on deep learning model," in *International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE, 2016, pp. 796–800.

[4] P. Seshadri, C. Han, B.-W. Koo, N. Posner, S. Guhathakurta, and A. Lerch, "Asped: An audio dataset for detecting pedestrians," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 406–410.

[5] H. Yang, K. Ozbay, and B. Bartin, "Investigating the performance of automatic counting sensors for pedestrian traffic data collection," in *World Conference on Transport Research (WCTR)*, vol. 1115, 2010, pp. 1–11.

[6] ——, "Enhancing the quality of infrared-based automatic pedestrian sensor data by nonparametric statistical method," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2264, no. 1, pp. 11–17, 2011.

[7] A. Brunetti, D. Buongiorno, G. F. Trotta, and V. Bevilacqua, "Computer vision and deep learning techniques for pedestrian detection and tracking: A survey," *Neurocomputing*, vol. 300, pp. 17–33, 2018.

[8] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 34, no. 4, pp. 743–761, 2011.

[9] E. Ozan, S. Searcy, B. C. Geiger, C. Vaughan, C. Carnes, C. Baird, and A. Hipp, "State-of-the-art approaches to bicycle and pedestrian counters," North Carolina Department of Transportation, Tech. Rep., 2021.

[10] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "It's not all about size: On the role of data properties in pedestrian detection," in *European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 210–225.

[11] I. Hasan, S. Liao, J. Li, S. U. Akram, and L. Shao, "Generalizable pedestrian detection: The elephant in the room," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 11 323–11 332.

[12] J. Rulff, F. Miranda, M. Hosseini, M. Lage, M. Cartwright, G. Dove, J. Bello, and C. T. Silva, "Urban rhapsody: Large-scale exploration of urban soundscapes," *Computer Graphics Forum*, vol. 41, no. 3, pp. 209–221, 2022.

[13] M. S. Hammer, T. K. Swinburn, and R. L. Neitzel, "Environmental noise pollution in the united states: developing an effective public health response," *Environmental Health Perspectives (EHP)*, vol. 122, no. 2, pp. 115–119, 2014.

[14] H. J. Jariwala, H. S. Syed, M. J. Pandya, and Y. M. Gajera, *Noise Pollution & Human Health: A review*, 2021, last access date: May 7, 2025. [Online]. Available: https://www.researchgate.net/profile/Hiral-Jariwala/publication/319329633_Noise_Pollution_Human_Health_A_Review/links/59a54434a6fdcc773a3b1c49/Noise-Pollution-Human-Health-A-Review.pdf

[15] World Health Organization, "Environmental noise," in *Compendium of WHO and other UN guidance on health and environment*, 2022, ch. 11, last access date: May 7, 2025. [Online]. Available: https://cdn.who.int/media/docs/default-source/who-compendium-on-health-and-environment/who_compendium_noise_01042022.pdf?sfvrsn=bc371498_3#:~:text=For%20average%20noise%20exposure%2C%20the,dB%20LAeq%2C%2024h%20%E2%80%A2%20weekly

[16] A. Radicchi, P. Cevikayak Yelmi, A. Chung, P. Jordan, S. Stewart, A. Tsaligopoulos, L. McCunn, and M. Grant, "Sound and the healthy city," *Cities & Health*, vol. 5, no. 1-2, pp. 1–13, 2021.

[17] L. M. Aiello, R. Schifanella, D. Quercia, and F. Aletta, "Chatty maps: constructing sound maps of urban areas from social media data," *Royal Society Open Science*, vol. 3, no. 3, p. 150690, 2016.

[18] J. P. Bello, C. Silva, O. Nov, R. L. Dubois, A. Arora, J. Salamon, C. Mydlarz, and H. Doraiswamy, "Sonyc: A system for monitoring, analyzing, and mitigating urban noise pollution," *Communications of the ACM*, vol. 62, no. 2, pp. 68–77, 2019.

[19] C. Han, P. Seshadri, Y. Ding, N. Posner, B. W. Koo, A. Agrawal, A. Lerch, and S. Guhathakurta, "Understanding pedestrian movement using urban sensing technologies: the promise of audio-based sensors," *Urban Informatics*, vol. 3, no. 1, p. 22, 2024.

[20] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[21] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k: an open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 30, pp. 829–852, 2022.