NATIONAL CENTER FOR UNDERSTANDING FUTURE
TRAVEL BEHAVIOR AND DEMAND

Final Project Report

A Pilot Experimental Project for Predicting Pedestrian Flows Using Computer Vision and Deep Learning

BY

Chaeyeon Han
Email: chan303@gatech.edu

Shreya Chivilkar
Email: schivilkar3@gatech.edu

Subhrajit Guhathakurta
Email: subhro.guha@design.gatech.edu

School of City and Regional Planning
Georgia Institute of Technology
245 4th St NW, Atlanta, GA 30332

April 2025

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. N/A	**2. Government Accession No.** N/A	**3. Recipient's Catalog No.** N/A	
4. Title and Subtitle A pilot experimental project for predicting pedestrian flows using computer vision and deep learning		**5. Report Date** September 1, 2024 **6. Performing Organization Code** N/A	
7. Author(s) Chaeyeon Han, https://orcid.org/0000-0001-7311-3944 Shreya Chivilkar, https://orcid.org/0009-0001-8122-3112 Subhrajit Guhathakurta, Ph.D. https://orcid.org/0000-0002-2456-3284		**8. Performing Organization Report No.** N/A	
9. Performing Organization Name and Address School of City and Regional Planning Georgia Institute of Technology 245 4th St NW, Atlanta, GA 30332		**10. Work Unit No. (TRAIS)** N/A **11. Contract or Grant No.** 69A3552344815 and 69A3552348320	
12. Sponsoring Agency Name and Address U.S. Department of Transportation, University Transportation Centers Program, 1200 New Jersey Ave, SE, Washington, DC 20590		**13. Type of Report and Period Covered** Final Report, 2023-2024 **14. Sponsoring Agency Code** USDOT OST-R	
15. Supplementary Notes Link to ASPEDv.a Dataset: https://urbanaudiosensing.github.io/ASPEDa.html			
16. Abstract This study presents a pilot project for predicting pedestrian flows using video-based computer vision and deep learning techniques. Two custom datasets were collected on the Georgia Tech campus, featuring video recordings of multiple pedestrian pathways along with annotated pedestrian counts and directional flow information. A methodological framework was developed incorporating Convolutional Neural Networks (CNNs) for spatial pattern recognition, Graph Convolutional Networks (GCNs) to model spatial relationships among distributed sensors, and a temporal lag optimization procedure to account for delayed influence from distant recorders. Experimental results demonstrate that CNNs effectively estimate and predict short-term pedestrian flow with high accuracy in smaller, localized environments (ASPED v.a), particularly when pedestrian count data is included. In larger and more complex networks (ASPED v.c), GCNs yield improved predictive performance but also exhibit challenges such as over-smoothing. The study concludes with a discussion on the potential for future multimodal sensing deployments and outlines limitations related to generalizability and scalability of the proposed approach.			
17. Key Words Urban sensing; pedestrian detection; pedestrian flow prediction		**18. Distribution Statement** No restrictions.	
19. Security Classif.(of this report) Unclassified	**20. Security Classif.(of this page)** Unclassified	**21. No. of Pages** 40	**22. Price** N/A

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

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, under Grant No. 69A3552344815 and 69A3552348320 from the U.S. Department of Transportation's University Transportation Centers Program. The U.S. Government assumes no liability for the contents or use thereof.

ACKNOWLEDGMENTS

This research was partially supported by the National Center for Understanding Future Travel Behavior and Demand (TBD), a National University Transportation Center sponsored by the U.S. Department of Transportation (USDOT) under grant numbers 69A3552344815 and 69A3552348320. The authors would like to thank the TBD National Center, USDOT, and the National Science Foundation (Award #2203408) for their support of university-based research in transportation, particularly for the funding provided for this project. The authors also extend their thanks to Uijeong Hwang, Senior Transportation Engineer at the Atlanta Regional Commission for his valuable contributions to the work presented in this report.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

EXECUTIVE SUMMARY

This report presents the outcomes of a pilot experimental study focused on predicting pedestrian flows through the integration of computer vision and deep learning methods, using video data collected from the Georgia Tech campus. The project seeks to advance urban sensing methodologies by evaluating the predictive capabilities of video-based pedestrian count and flow data using Convolutional Neural Networks (CNNs) and Graph Convolutional Networks (GCNs), with implications for future multimodal sensing applications.

Two custom datasets, ASPED v.a and ASPED v.c, were collected and annotated for this study. ASPED v.a provides pedestrian count and directional flow annotations within a compact campus courtyard, while ASPED v.c extends the spatial scale across multiple campus intersections and pathways that have higher pedestrian volume. Pedestrian detection was performed using deep object detection models (Mask2Former and YOLOv8), and directional flow labels were annotated through zone-based transition tracking and re-identification algorithms.

The analysis involved training CNN models for both flow estimation within a given input window and short-term prediction tasks. Results demonstrate that CNNs can accurately predict flow using limited count data, with marginal improvement when increasing the number of recorder features. Importantly, the inclusion of pedestrian count data, particularly when combined with flow data, substantially improved short-term forecasting performance.

In the extended analysis using ASPED v.c, GCN models were implemented to capture spatial dependencies between distributed sensor nodes. Temporal lag optimization, based on geodesic distances and average walking speeds, was applied to account for delayed influence from distant recorder locations. Experiments showed that moderate lag configurations (e.g., 30 to 60 seconds) improved GCN performance, though the results exhibited over-smoothing effects in periods with high pedestrian flows.

The study establishes a reproducible framework for evaluating the predictive utility of count-based pedestrian data and demonstrates that non-directional count sensors, such as audio-based recorders, can serve as alternatives to video-based systems in specific use cases. While the modeling framework shows promise, generalizability and long-term deployment remain as challenges for further research. Future work should investigate temporal modeling enhancements (e.g., LSTM or GRU layers) and infrastructure adaptations to support continuous, real-time data collection.

These findings have practical implications for pedestrian mobility data collection and management, particularly in optimizing sensor placement, and enabling data-informed pedestrian flow management strategies in urban environments.

INTRODUCTION

As cities worldwide place increasing emphasis on walkability, understanding pedestrian flow dynamics has become an essential component of urban planning. Pedestrian flow data offers valuable insights into the volume of inflow and outflow between locations, aiding in the identification of key routes, origins, and destinations. Such data can inform the optimization of pathways, crosswalks, and overall connectivity, while also guiding the strategic placement of amenities and shaping land use and zoning decisions.

Beyond infrastructure design, pedestrian flow information supports congestion management by identifying bottlenecks where pedestrian and vehicular traffic may conflict, such as around transit hubs or event venues (American Planning Association, 1965). Analyzing pedestrian movement patterns also enhances public transportation planning, improves safety measures, and facilitates more effective emergency evacuation strategies (Ratti et al., 2006; Yabe et al., 2023). Moreover, pedestrian flow data provides valuable behavioral insights, such as peak travel times and responses to environmental conditions, offering a deeper understanding of how people navigate urban spaces.

Traditional methods of collecting pedestrian movement data, such as smartphone-based tracking, have faced significant privacy concerns, particularly under regulations like the European General Data Protection Regulation (Van Steen et al., 2022). The use of smartphone data, while effective in capturing trajectories and demographic details, is often restricted or prohibitively expensive to access from private companies. In response to these challenges, alternative urban sensing approaches, such as infrared cameras, video surveillance, and other sensor-based techniques, have been developed. However, these methods also come with limitations, including high costs, data storage requirements, and a lack of directionality in tracking pedestrian movements (Han et al., 2024). Moreover, it is yet understudied how multimodal sensors can be distributed to effectively capture pedestrian flows.

To address these limitations, this study explores the potential of using multimodal sensors by leveraging video-based pedestrian detection techniques. We present a framework for identifying optimal set of recorder features for predicting pedestrian flow at one location, that maximizes data utility while minimizing redundancy. Using deep learning techniques, such as Convolutional Neural Networks (CNNs) and Graph CNNs (GCNs), we demonstrate how pedestrian data derived from video and computer vision can serve not only as predictors of pedestrian movement but also as a methodological foundation for evaluating alternative sensing approaches. Specifically, the research objectives are to:

(1) Establish a replicable process/framework for identifying the optimal set of sensor locations to efficiently estimate or predict pedestrian flow at a certain location and sensor network.
(2) Experiment pedestrian flow estimation and prediction with video recordings collected from two areas with distinct scales in Georgia Tech Campus.
(3) Investigate the plausibility of predicting pedestrian flow based on pedestrian count data, which has no directional information.

While the experimental data and analysis in this study rely exclusively on video-based

recordings, the research also informs the future development of low-cost, scalable sensing systems, including audio-based recorders. Audio-based sensing offers several advantages: microphones are affordable, energy-efficient, and capable of capturing data over wide areas. Since sound waves can travel around obstacles, microphones can detect pedestrian movement in situations where other sensing methods may fail. The modeling framework and feature optimization strategies developed here provide valuable insights into how pedestrian count data, regardless of sensing mode, can be effectively used for pedestrian data collection. In this way, video-based experiments conducted in this project help illuminate the potential of audio sensors, particularly in settings where camera deployment may be restricted or constrained.

This research contributes to the field by demonstrating that urban sensing based on pedestrian counts, when coupled with video-based flow data and deep learning techniques, can be used to predict pedestrian flow. Our findings suggest that pedestrian counting sensors, despite lacking direct directionality information, can serve as a valuable tool for pedestrian flow prediction when integrated with appropriate machine learning models. This approach has the potential to significantly reduce costs and expand pedestrian sensing capabilities in urban environments.

LITERATURE REVIEW

The estimation and prediction of movement patterns in urban environments is a widely studied domain in transportation research. While vehicular flow estimation is a well-established domain, pedestrian detection and pedestrian flow modeling have rapidly evolved with the advancement of computer vision and deep learning techniques. This literature review highlights progress in three intersecting domains:

(1) traffic flow estimation, primarily from video and sensor-based systems,

(2) pedestrian detection models, with a focus on object recognition and tracking, and

(3) pedestrian flow prediction, which integrates spatial-temporal modeling and multimodal sensing.

(1) Traffic Flow Estimation

Traffic flow estimation has been an actively studied domain in intelligent transportation systems, which enables real-time monitoring, adaptive signal control, and infrastructure planning. A dominant strand of research uses computer vision-based techniques applied to traffic surveillance videos. For example, Algiriyage et al. (2021) deployed YOLOv4 for vehicle detection and the SORT algorithm for object tracking in New Zealand's road networks. The system accurately identified cars (mAP 96.94%) and estimated directional flows by tracking bounding boxes across user-defined boundaries (Algiriyage et al., 2021). However, its applicability was limited by its inability to handle multi-directional intersections, bias toward cars, and its unreliability of re-identification during night-times.

Similarly, several studies applied Faster Region-based-CNN (R-CNN) for vehicle detection combined with the tracking algorithms such as Camshift and Kalman filter (Ahmed et al., 2021; Xu et al., 2017; Y. Zhang et al., 2017). While Faster R-CNN is suggested as a promising method to detect and count vehicle flows as it is insensitive to traffic volume or illumination changes in the videos, their system still underperformed in complex scenes especially in the presence of diverse vehicle types, such as buses, trucks, motorcycles, and bicycles.

Beyond video-based approaches, researchers have explored alternative sensor modalities and hybrid models to improve flow estimation accuracy in diverse environments. For instance, (Li et al., 2021) utilized Floating Car Data (FCD) which consists of GPS-based travel durations from vehicles. They applied Gaussian Process Regression (GPR) models to infer vehicle flows from estimated travel times provided by platforms like Google Maps. To enhance accuracy and adaptability, the study introduced both single-model and multi-model configurations, the latter using clustering techniques, such as k-means and Support Vector Machines (SVMs), to segment travel profiles into distinct "types of days" for model specialization. This multi-model structure reduced RMSE (root mean square error) significantly and allowed the system to generalize better acriss varying traffic conditions. The findings suggest the utility of non-visual data for traffic flow modeling, particularly in environments where occlusion, low lighting, or privacy concerns constrain video surveillance.

Another growing area of interest is the use of fusion-based techniques that combine video with additional sensor data (e.g., radar, LiDAR, or loop detectors) to improve robustness. These multi-sensor systems often leverage deep learning models to learn complex temporal patterns

across modalities. However, challenges remain in aligning heterogeneous data streams, addressing latency, and maintaining scalability across regions with limited sensor infrastructure.

Overall, the field has matured significantly in its ability to detect and count vehicles with high precision under controlled conditions. Still, generalizing across locations with multimodal traffic patterns, and achieving speed and density estimation remain active areas of research. New directions include computing for real-time deployment and transfer learning to adapt pre-trained models to new locations with minimal calibration.

(2) Pedestrian Detection Models

Pedestrian detection using computer vision is particularly relevant to applications in autonomous driving, city surveillance, and human mobility analytics. The field has evolved significantly from early handcrafted-feature approaches to deep learning-based systems. Traditional models such as Histogram of Oriented Gradients (HOG) combined with Support Vector Machines (SVM) (e.g., (Dalal & Triggs, 2005) laid the groundwork for pedestrian detection by encoding local gradient information and leveraging discriminative classifiers. These approaches, while computationally efficient, often struggled with scale variation, occlusion, and cluttered backgrounds.

The introduction of Convolutional Neural Networks (CNNs) was a turning point. Models like Faster R-CNN (Ren et al., 2015), YOLO (Redmon et al., 2015), and SSD (W. Liu et al., 2016) dramatically improved detection accuracy and speed by learning hierarchical features directly from data. Faster R-CNN, for instance, incorporates region proposal networks (RPNs) for accurate object localization, while YOLO emphasizes real-time detection by framing detection as a single regression problem. These models have been widely adopted for pedestrian detection in surveillance footage and autonomous vehicles, often achieving high precision in urban environments.

Recent advancements have focused on handling specific challenges, such as occlusion, small-scale detection, and dense crowds. Techniques include anchor-free detectors (e.g., CenterNet[1], FCOS[2]), attention mechanisms, and multi-scale feature fusion (e.g., HRNet[3]). Datasets like Caltech Pedestrian (Dollar et al., 2009), CityPersons[4], and JAAD[5] have enabled benchmarking. Moreover, domain adaptation and synthetic data (Baul, 2021) are being explored to reduce the dependency on large, labeled datasets. Lightweight models and Transformer-based architectures (e.g., DETR[6], ViT[7] adaptations) are emerging for edge deployment and context-aware reasoning.

Overall, pedestrian detection remains an active research area with ongoing challenges in generalization across environments and robustness under occlusion. The integration of temporal context, multi-modal data, and self-supervised learning promises to further advance the field.

[1] https://github.com/xingyizhou/CenterNet
[2] https://github.com/tianzhi0549/FCOS
[3] https://github.com/leoxiaobin/deep-high-resolution-net.pytorch?tab=readme-ov-file
[4] https://www.kaggle.com/datasets/hakurei/citypersons
[5] https://data.nvision2.eecs.yorku.ca/JAAD_dataset/
[6] https://github.com/facebookresearch/detr
[7] https://github.com/lucidrains/vit-pytorch

(3) Pedestrian Flow Estimation and Prediction

Pedestrian flow analysis has lagged compared to vehicular traffic modeling due to the unique complexities of pedestrian movement. Pedestrian trips are often more diverse in purpose, less structured in route choice, and frequently involve multi-modal journeys that incorporate walking as part of a larger trip. Additionally, the pedestrian networks in urban environments are inherently more complex, with a myriad of accessible pathways, including informal routes unavailable to vehicles. Pedestrians also tend to make unrecorded stops or pauses, which further complicates accurate flow analysis and prediction.

With the rise of machine learning and deep learning, however, researchers have begun to apply these techniques to address the intricacies of pedestrian flow prediction. Recent advances in pedestrian flow modeling have shown increasing success using AI tools, particularly for spatial-temporal prediction challenges. For instance, (Kitano et al., 2019) demonstrated the potential of AI in capturing these complexities, while (Ai et al., 2019) leveraged the power of deep learning, particularly CNNs for spatial feature extraction and Recurrent Neural Networks (RNNs) for modeling temporal dynamics. These advancements have significantly improved the predictive accuracy of pedestrian movement patterns. (Deo & Trivedi, 2021) proposed a novel grid-based prediction approach for multimodal trajectory forecasting, utilizing maximum entropy inverse reinforcement learning (MaxEnt IRL). Their model enabled more accurate predictions by learning policies in a grid-based framework, further enhancing the ability to forecast pedestrian and vehicle trajectories in complex environments.

Graph Convolutional Networks (GCNs) have emerged as a powerful tool for modeling irregular spatial correlations, such as those present in pedestrian networks. GCNs excel in capturing the complex interactions between spatial nodes, where nodes represent areas and edges reflect road links or origin-destination trajectories, allowing for more effective processing of spatial relationships. Liu et al., (2021) applied GCNs to camera-detected pedestrian data, showcasing the model's ability to handle the spatial topology of road networks. (Xia et al., 2021) expanded on this work by developing a 3-dimensional GCN (3DGCN) model to address dynamic spatial-temporal graph prediction challenges, incorporating Point-of-Interest (POI) data to further improve prediction accuracy. (Sun et al., 2022) introduced a hybrid model combining GCN and fully connected neural networks with a multi-view fusion module, which adeptly captured both spatial correlations and crowd flow dynamics.

In addition to spatial modeling, researchers have begun integrating external factors such as weather conditions and inter-regional traffic to improve pedestrian flow prediction. Zhang et al. (2017) and (D. Zhang & Kabuka, 2018) emphasized the influence of weather and day-specific effects on pedestrian dynamics. (Lin et al., 2019) introduced the DeepSTN+ model, which integrated POIs and temporal factors to better model spatial dependencies. (J. Zhang et al., 2018) further developed the ST-ResNet model, incorporating both weather and time data to predict crowd inflow and outflow.

A variety of data sources have been employed to enhance pedestrian flow predictions, including mobile GPS data, social media, and data from bike-sharing and taxi systems (Zhang et al., 2017; Lin et al., 2019; Sun et al., 2022). However, accessing these data sources for effective pedestrian mobility prediction remains a significant challenge due to privacy concerns and the need for collaboration with private companies. Considering these challenges, the use of alternative data sources, such as the video-based pedestrian count dataset explored in this study, offers a

promising solution.

DATA

(1) Data Collection

We collected and utilized two versions of multimodal pedestrian detection dataset, named Audio Sensing for PEdestrian Detection (ASPED) v.a and v.c (Seshadri et al., 2024). This dataset contains audio, video recordings, and number of pedestrians passing by each designated locations at one frame per second rate, annotated based on the video recordings. The data collection sites are demonstrated in *Figure 1*.

For this study, we use a part of ASPED v.a, which focuses on a Cadell Courtyard site in Georgia Tech Atlanta campus. ASPED v.c is an expanded version of ASPED v.a, which comprises of larger pedestrian network and volume. We publicized the first set of data, ASPED v.a[8], in 2024 and plan to publicize the expanded version, ASPED v.c, later this year as it is still in the process of synchronizing between video and audio files.

Figure 1. Data Collection Sites for ASPED v.a & ASPED v.c,

Georgia Tech Campus in Atlanta

Since this project only utilizes the video files in the dataset, we only elaborate on the device settings of the cameras and the way we annotated pedestrians from videos. We used GoPro HERO9 Black cameras with USB pass-through doors connected to Anker PowerCore III Elite 26K power

[8] https://urbanaudiosensing.github.io/ASPEDa.html

banks for longer recording (see *Figure 2)*. 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. 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 to mark the exact time, aiding in syncing with the audio recorders. In larger areas, multiple whistle signals were used.

Figure 2. Video Camera Installed on a Campus Street Light

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., 2021) to detect people per frame at 1 frame per second. We use the specific implementation by OpenMMLab[9] 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 (*Figure 3*). Subsequently, every frame was annotated with the number of pedestrians detected if the bottom-center point of pedestrian bounding boxes intersected with the 9m buffer around the recorders; otherwise, it was labeled as no-pedestrian present.

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

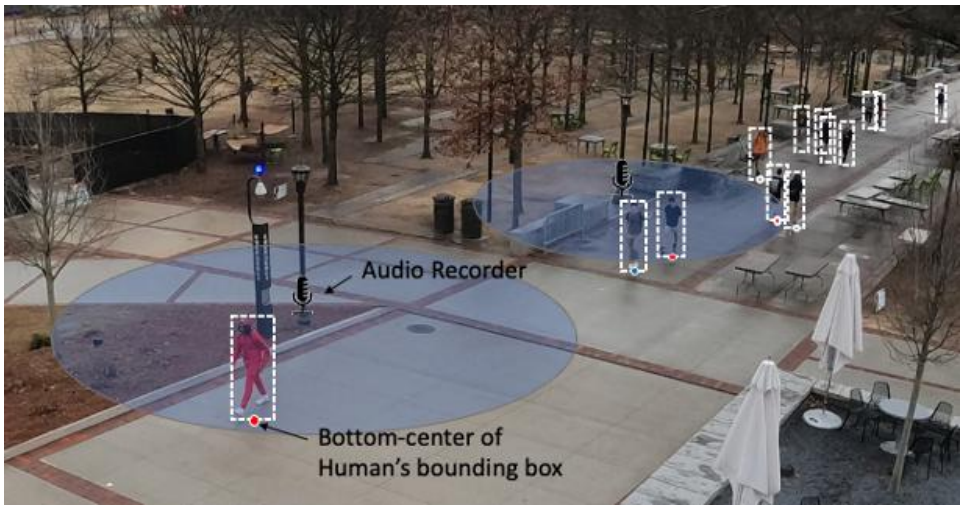


Figure 3. Pedestrian Count Detection using Mask2Former Model

(2) Summary of ASPED v.a

The ASPED v.a dataset consists of a total of 3,406,229 video frames, collected over five separate sessions, each lasting approximately two days. Figure 4 shows the distribution of detected pedestrian counts in each frame. Since our study focuses on estimating pedestrian flow from data without directional information (e.g., audio-based sensors), we restricted the dataset to include only frames recorded between 7 AM and 7 PM, which corresponds to the hours of peak pedestrian activity on campus. After filtering, the dataset contains 170,281 frames.

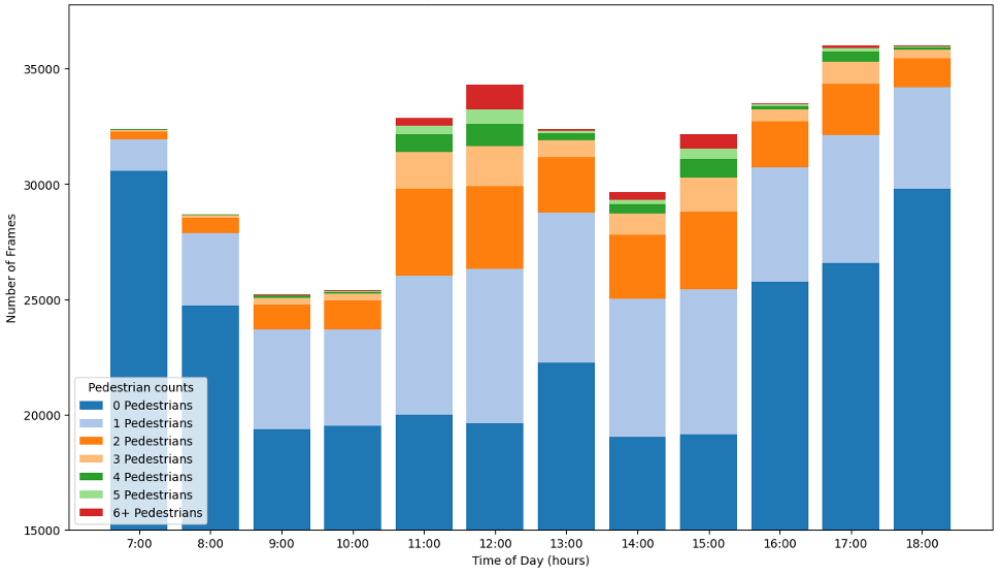


Figure 4. Hourly Distribution of Pedestrian Counts, ASPED v.a

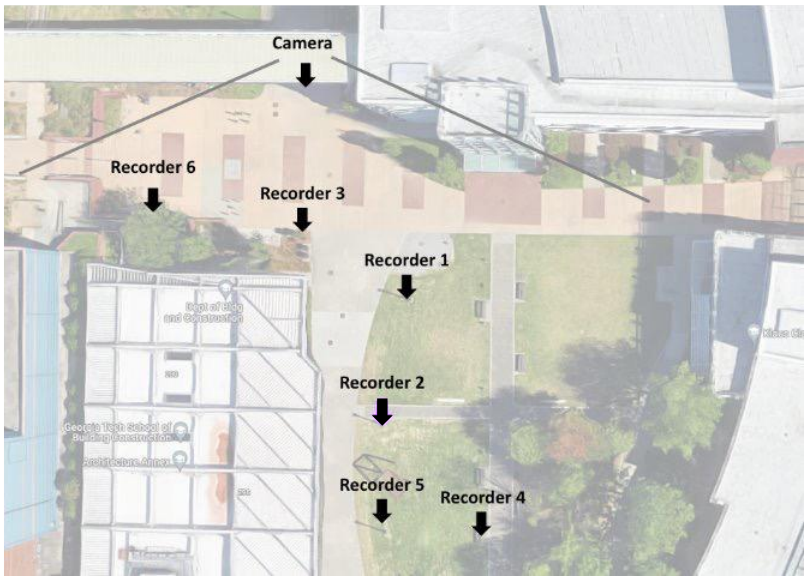


Figure 5. Data Collection Site (ASPED v.a, Cadell Courtyard)

While ASPED v.a provides pedestrian count data at specific locations, it lacks detailed pedestrian flow information, meaning it does not capture the number of pedestrians moving in different directions. To label pedestrian flow, we used the raw video recordings from the surveillance camera. Prior to processing, all video files were standardized using FFmpeg to remove audio tracks and ensure consistent formatting. For efficiency, we focused on predefined regions-of-interest (ROIs) by cropping the video frames to only include these areas. The video frames were then upscaled by a factor of 3 to enhance detection performance for pedestrians at greater distances.

Figure 6 Sample of upscaled ROI of a frame, where five pedestrians are detected moving upwards (yellow to blue polygon)

To estimate pedestrian flow, we employed a basic re-identification method, which involves tracking individuals across multiple frames. Similar to the pedestrian detection workflow, we processed the videos using OpenCV, with human detection performed using the YOLOv7 model, which provided bounding boxes for detected individuals in each frame. We calculated the bottom-center point of each bounding box and checked its position relative to the polygon masks to

determine if a person crossed one of the predefined regions (UP or DOWN). Specifically, the algorithm tracked individuals who crossed from the bottom ROI (yellow polygon in *Figure 6*) to the top ROI (blue polygon in *Figure 6*), labeling them as moving UP, or the opposite direction, labeling them as moving DOWN. The incremental (frame-by-frame) and cumulative counts of people moving up and down across all processed videos were stored in a CSV file.

Table 1. Data Description

		Min	Max	Mean	Median
Pedestrian Count within 9m Buffer	Recorder 1	0	19	0.35	0
	Recorder 2	0	13	0.19	0
	Recorder 3	0	20	0.33	0
	Recorder 4	0	12	0.06	0
	Recorder 5	0	12	0.19	0
	Recorder 6	0	17	0.20	0
Pedestrian Flow	Up	0	4	0.01	0
	Down	0	4	0.01	0

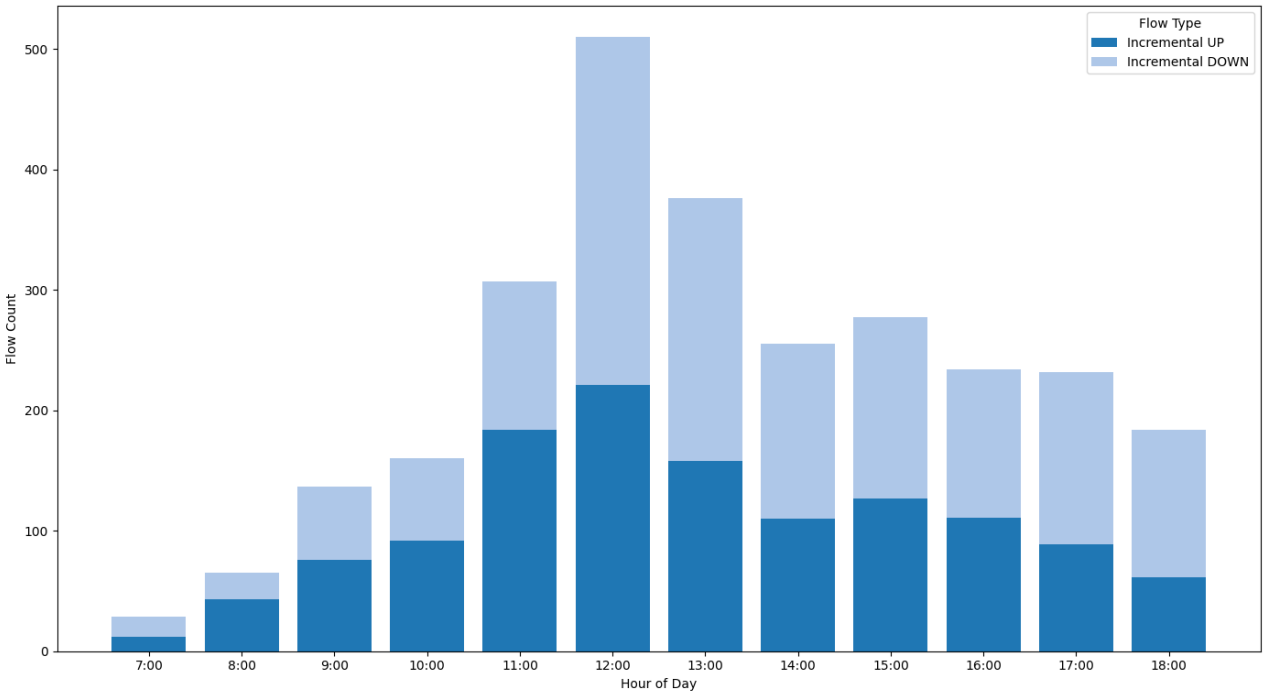


Figure 7. Hourly Up/Down Flow Distribution at ASPED v.a Pathway

(3) Summary of ASPED v.c

The recording devices were set up in five locations for four sessions on the Georgia Tech campus around the Tech Green area, in which we named as Intersection B, Intersection C, Intersection D, Path1 and Path2. For our experiments in this report, we specifically used the pedestrian counts within a 9-meter buffer (*Figure 8*) annotated based on video. Pedestrian counts were captured via a surveillance camera, with footage recorded at a frame rate of thirty frames per second.

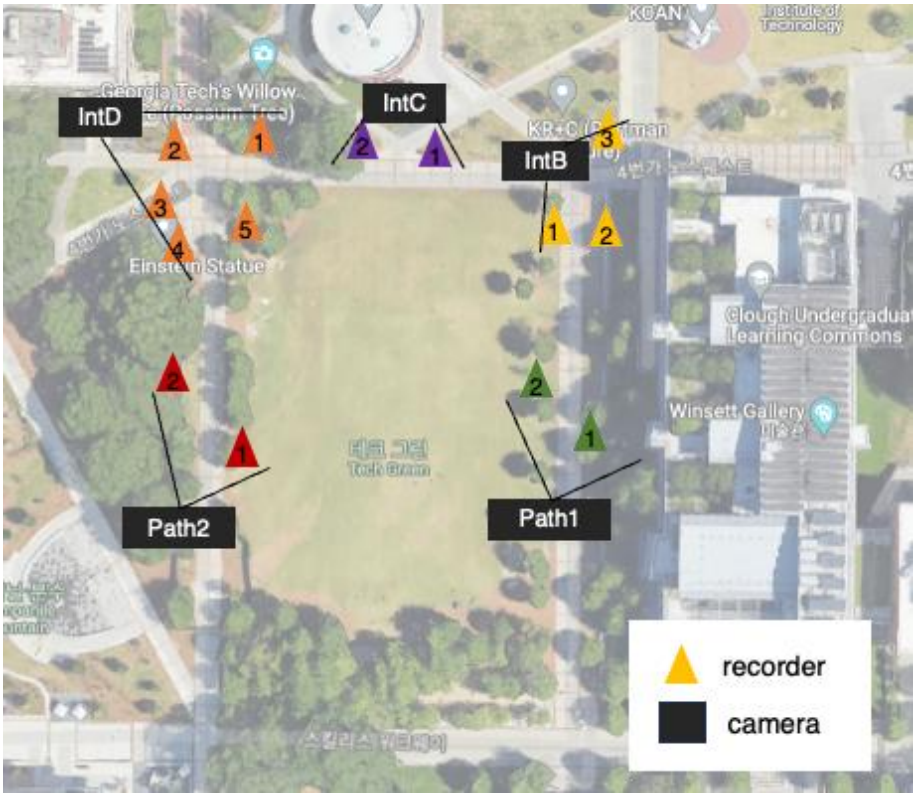


Figure 8. Data Collection Site (ASPED v.c, Tech Green)

Table 2: Data Collection for ASPED v.c Sessions

Session	Days
Session_02152024	15th February to 17th February 2024
Session_02292024	29th February to 2nd March 2024
Session_10222024	22nd October to 24th October 2024
Session_10292024	29th October to 31st October 2024

The total frames for each scene are mentioned in *Table 3,* where we have concatenated values from all the sessions for that scene.

Table 3 Data Summary for ASPED v.c Locations

Scenes	Total Frames
Intersection B	613,929
Intersection C	609,893
Intersection D	-
Path1	582,829
Path2	471,448

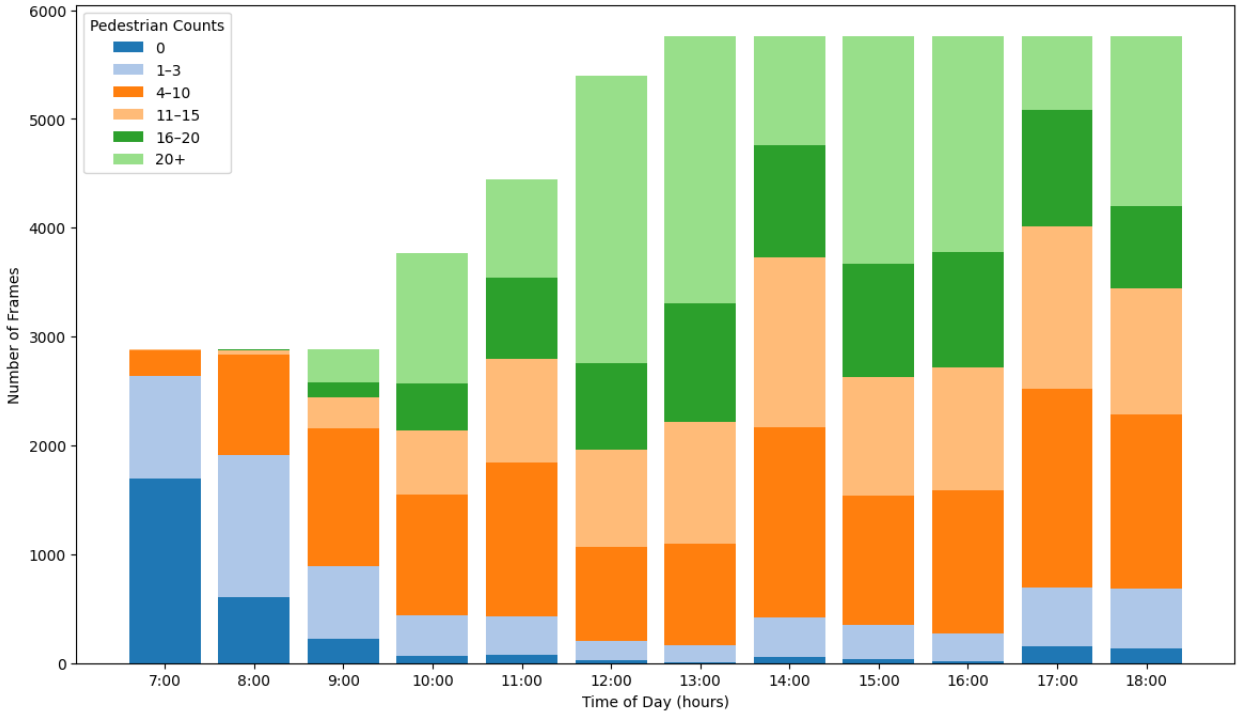


Figure 9. Hourly Distribution of Pedestrian Counts in Intersection B, ASPED v.c

Since the ASPED v.c dataset, like ASPED v.a, does not contain detailed directional pedestrian flow information, we used the video recordings to generate flow labels. The videos were preprocessed using the same standardization pipeline: audio tracks were removed, and consistent formatting was ensured using FFmpeg. To optimize detection, we cropped frames to focus only on predefined regions of interest (ROIs) and upscaled them to improve the visibility and detectability of pedestrians, especially those farther from the camera.

To detect pedestrians in video frames, we utilized the YOLOv8n model with a confidence threshold of 0.5. The model outputs bounding boxes around detected individuals, these are used to isolate regions of interest (ROIs) within each frame. These ROIs are passed to a feature extractor that helps with tracking. The feature extraction process involves cropping the detected bounding box regions from the frame, resizing them to 224×224 pixels, normalizing pixel values, and converting the cropped images into tensors. These preprocessed tensors are then passed through a ResNet-50 model (with the final classification layer removed) to generate high-dimensional feature embeddings that represent the visual characteristics of each pedestrian. The resulting

feature vectors are converted to NumPy arrays and used along with the bounding boxes for the DeepSort tracking algorithm. DeepSort leverages both these motion and appearance cues to maintain consistent identities for pedestrians across frames for efficient path tracking over time.

Further, to detect the flow of pedestrians: we divided each scene into zones and then tracked the motion of people moving from one zone to another (*Figure 10*). We then track pedestrian movements, or flows, between defined zones within each intersection. These zones were decided based on the visual layout and common movement paths observed in the footage. These zones were manually annotated using polygon coordinates to represent regions of interest. Once these zones were established, pedestrian flow was measured by tracking transitions between them.

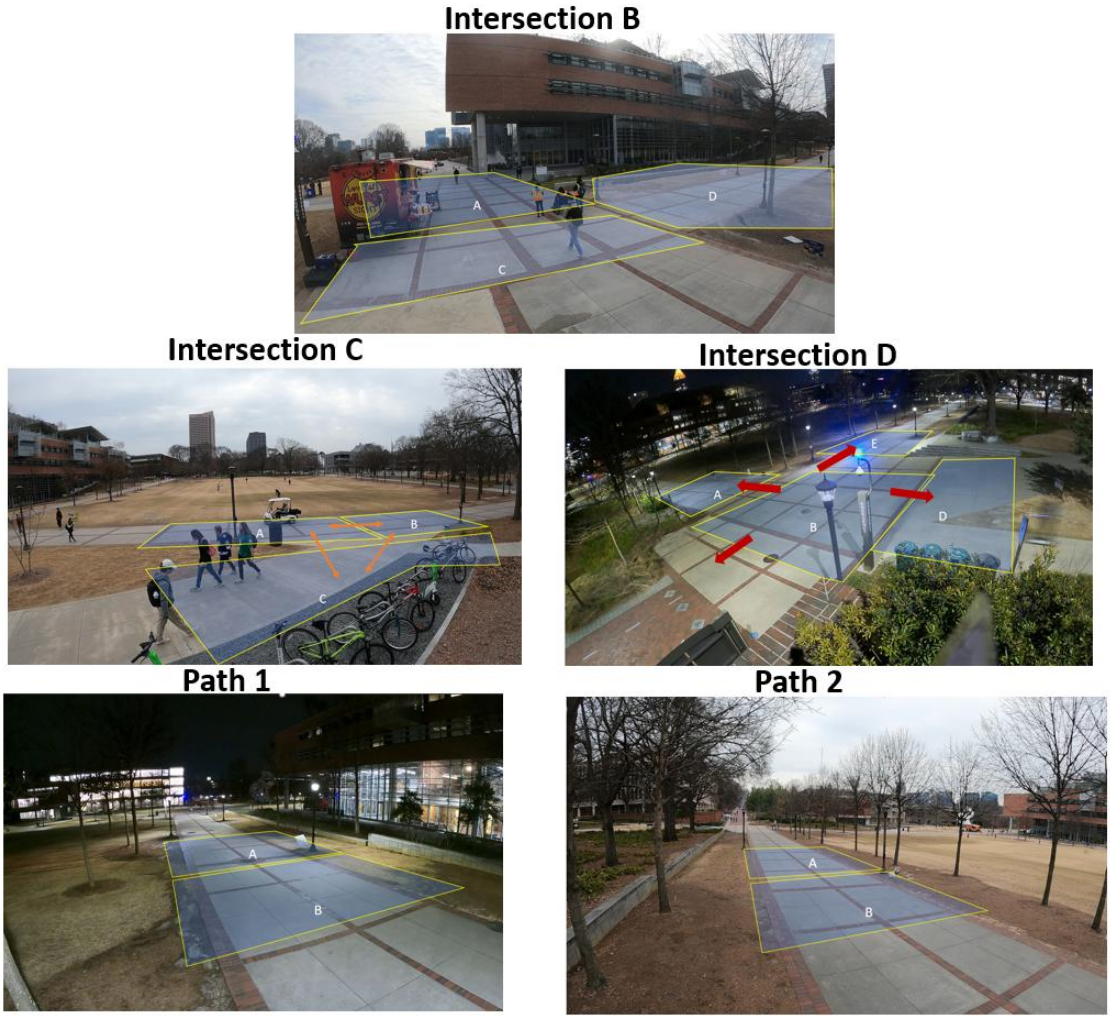


Figure 10: Zone labelling in each scene

For example, at Intersection B, six distinct pedestrian flows are identified based on directional movement between zones: from Zone A to Zone C, Zone C to Zone A, Zone A to Zone D, Zone D to Zone A, Zone C to Zone D, and Zone D to Zone C.

For each frame, once pedestrians were detected using the YOLOv8n object detection model they were tracked across time using the DeepSORT algorithm. Each detected person was assigned a unique ID, and their movement was continuously monitored across successive frames.

To determine which zone a pedestrian was in, the center point of their bounding box was computed and checked against the defined zone polygons using OpenCV's pointPolygonTest. This allowed us to identify whether a person was located in Zone A, Zone B, or neither at any given time. The previous zone associated with each track ID was stored, and if the current zone differed from the previous one, the system recorded a directional transition (e.g., from Zone A to Zone B). These transitions were used to increment corresponding flow counters; both for the current frame and the overall video session. This ensured that each transition was only counted once per unique movement. For example, in

Figure 11 a person with Track ID 367 was initially in Zone A and later detected in Zone B, this would be logged as an A_to_B transition. Transition A_B: count increased from 10 to 11 from frame 708 to frame 709.

This transition would not be recounted unless the same person moved back and forth again. The system visualized this information by color-coding bounding boxes, drawing motion trails, and displaying real-time transition tables on the output video. Additionally, all statistics such as zone occupancy, frame-wise transitions, and cumulative flows were logged to a CSV file for further analysis. Similar approach was used in all the scenes to determine the pedestrian movement across zones. This zone-based transition tracking framework enabled us to generate reliable directional flow labels for pedestrian movement in the absence of explicitly labeled data in the original ASPED.c dataset.

Figure 11. Example of Pedestrian Flow Detection in Path 1

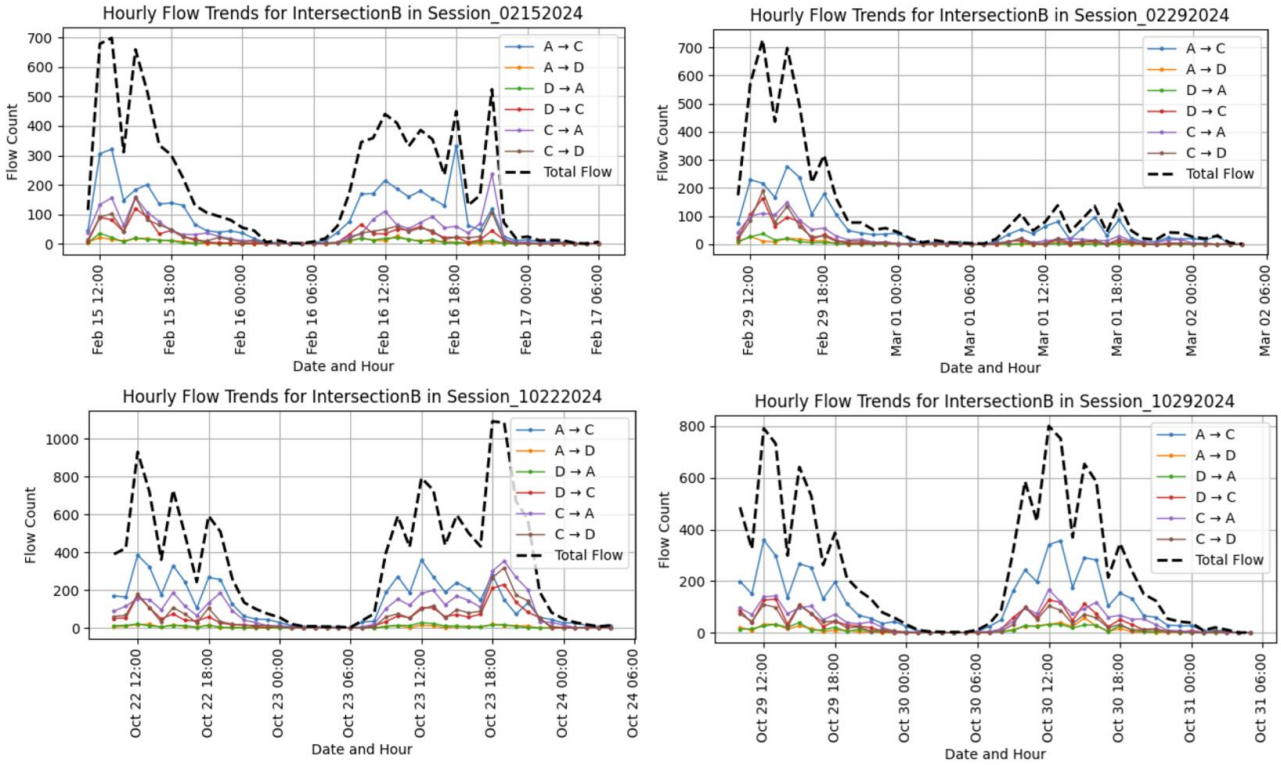


Figure **12**-14 indicate the hourly flow of pedestrians across the zones within different scenes. At *Intersection B*, the highest pedestrian activity is observed between 12 PM and 6 PM, with counts ranging from 300 to 1000. The peak consistently occurs around 12 PM, reaching a maximum of 1100 pedestrians on 23rd October at 6PM. In contrast, nighttime hours (12 AM to 6 AM) show negligible or no activity. Similarly, *Intersection C* experiences substantial flow between 12 PM and 6 PM, typically ranging from 50 to 600 pedestrians. Notably, on February 15th and 16th, the high-flow window extends until 8 PM, while the maximum pedestrian count of 600 is recorded on October 23rd at 6:15 PM. Night hours (12 AM to 8 AM) remain largely inactive in this region as well.

For *Path 1*, elevated pedestrian flow is generally seen between 12 PM and 8 PM, with volumes ranging from 20 to 400. The most significant surges occur on October 23rd at 7 PM (600 pedestrians) and October 30th at 12 PM (450 pedestrians). As with other locations, activity is minimal or absent between 12 AM and 6 AM. In the case of *Path 2*, high flow is typically recorded between 12 PM and 6 PM, except on October 23rd, where it persists until 8 PM. Pedestrian counts here range from 50 to 500, with peak activity observed on October 23rd and 30th, both reaching around 500 pedestrians.

Overall, *Intersection B* emerges as the most active region, consistently recording the highest pedestrian flow across all sessions when compared to other intersections and paths. Additionally, the evening of October 23rd stands out as a particularly busy period, with all scenes exhibiting peak pedestrian activity during that time.

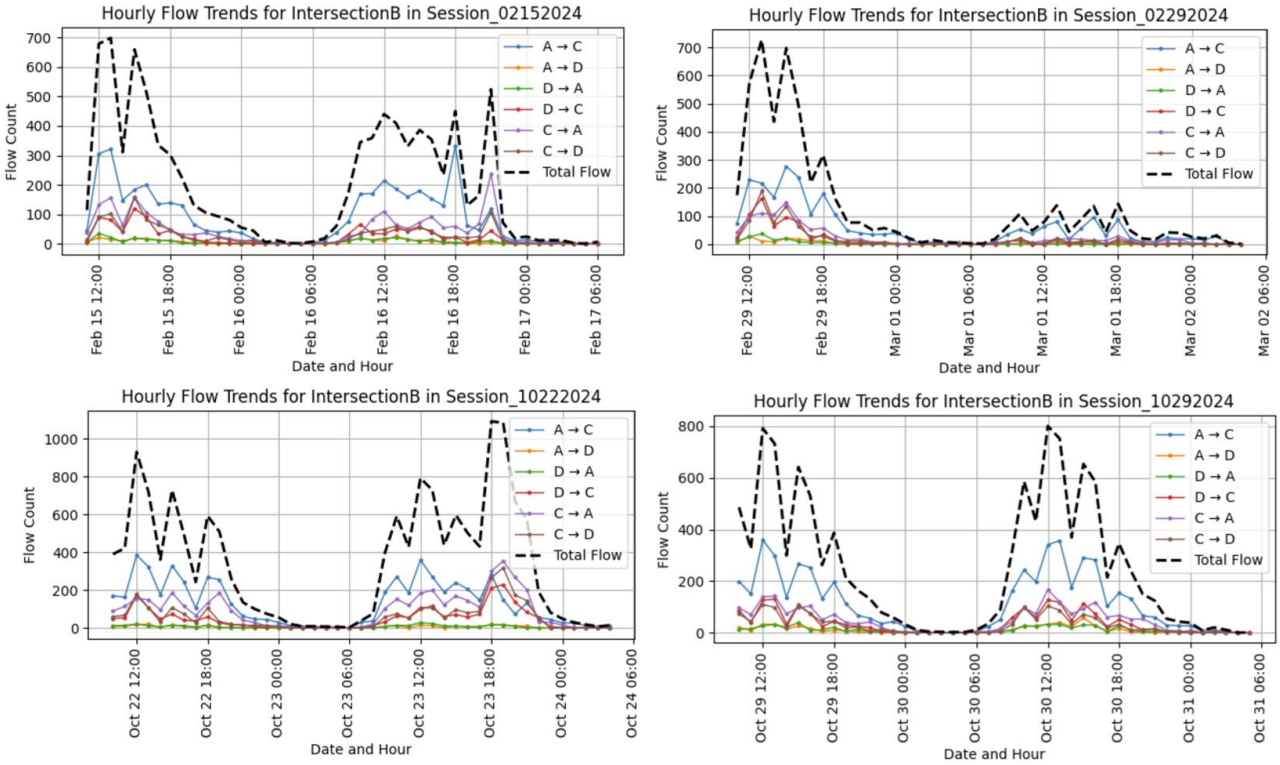


Figure 12: The hourly flow trends for Intersection B across different sessions

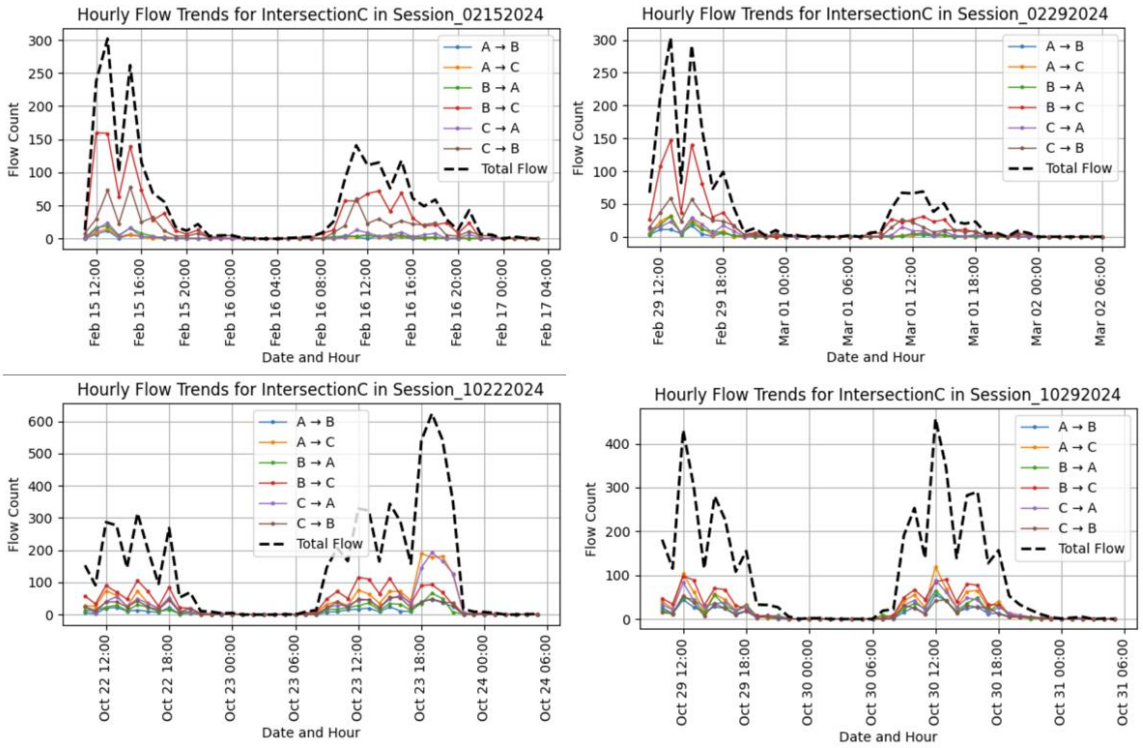


Figure 13: The hourly flow trends for Intersection C across different sessions

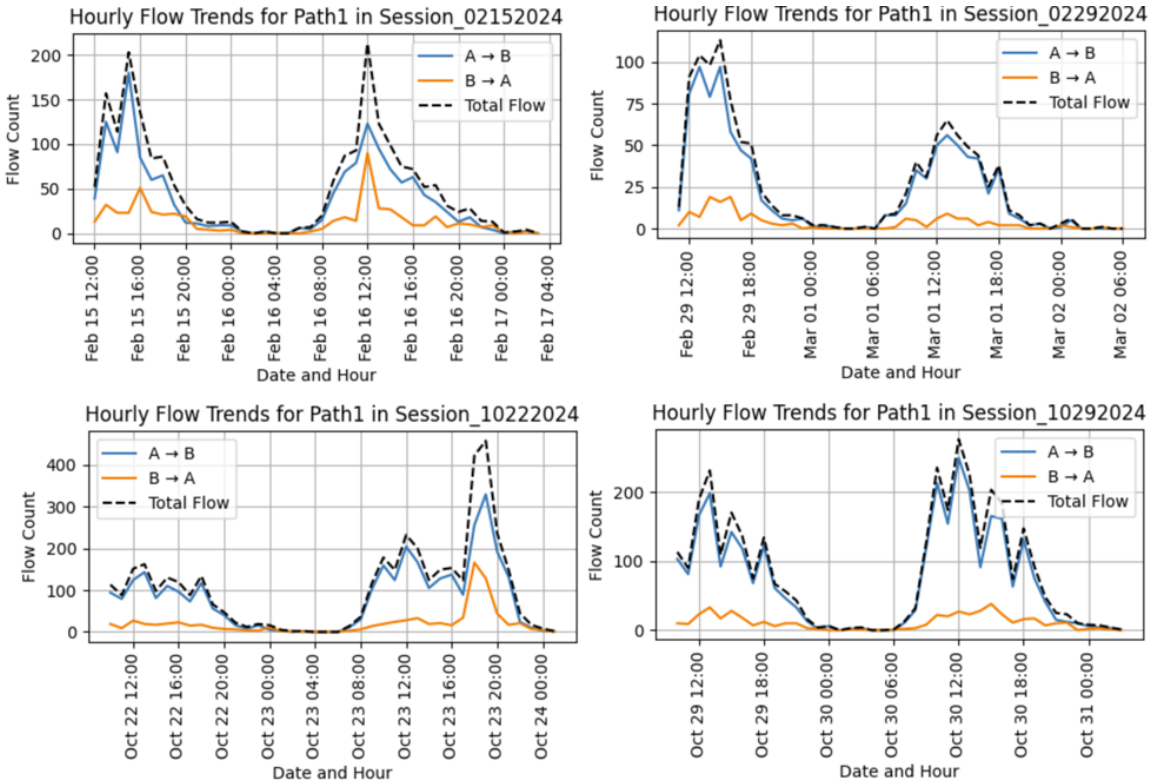


Figure 14:The hourly flow trends for Path1 across different sessions

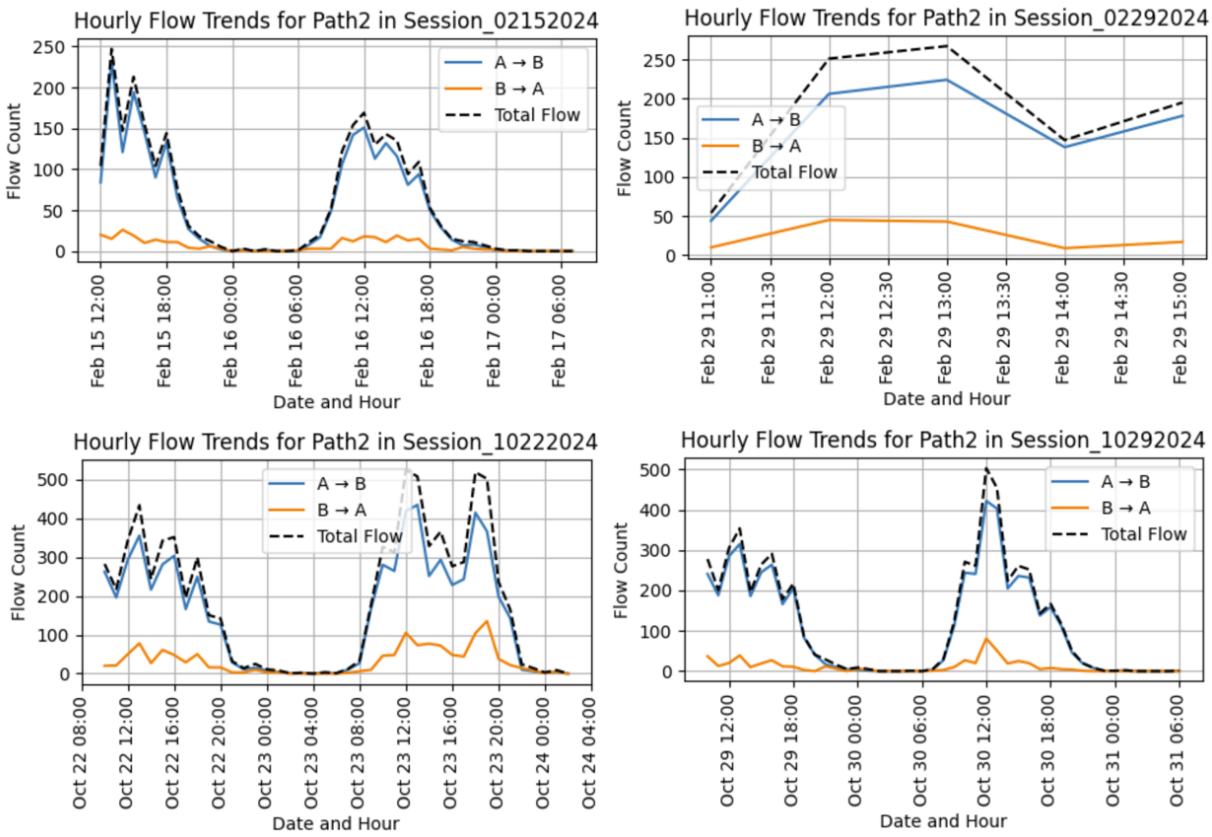


Figure 15:The hourly flow trends for Path 2 across different sessions

(4) Adjacency Matrix

We also collected coordinates of sensor locations and created an adjacency matrix to feed into the Graph Convolutional Network models. This matrix encodes spatial relationships within the physical sensor network by defining weights for different features based on pairwise distances between recorders.

Each entry A_{ij} in the matrix was calculated using a Gaussian kernel function:

$$A_{ij} = \exp\left(\frac{[dist(v_i, v_j)]^2}{\theta^2}\right),$$

Where $dist(v_i, v_j)$ denotes the Euclidean distance between sensors i and j, and θ represents the standard deviation of all pairwise distances in the network.

To sparsify the graph and reflect the practical limitations of physical influence, we applied a distance threshold, k, defining the final adjacency matrix as:

$$A_{ij} = \begin{cases} A_{ij} & A_{ij} > k \\ 0 & A_{ij} \leq k \end{cases}$$

$$\theta: std.deviation\ of\ distances\ between\ recorder\ i\ and\ j$$

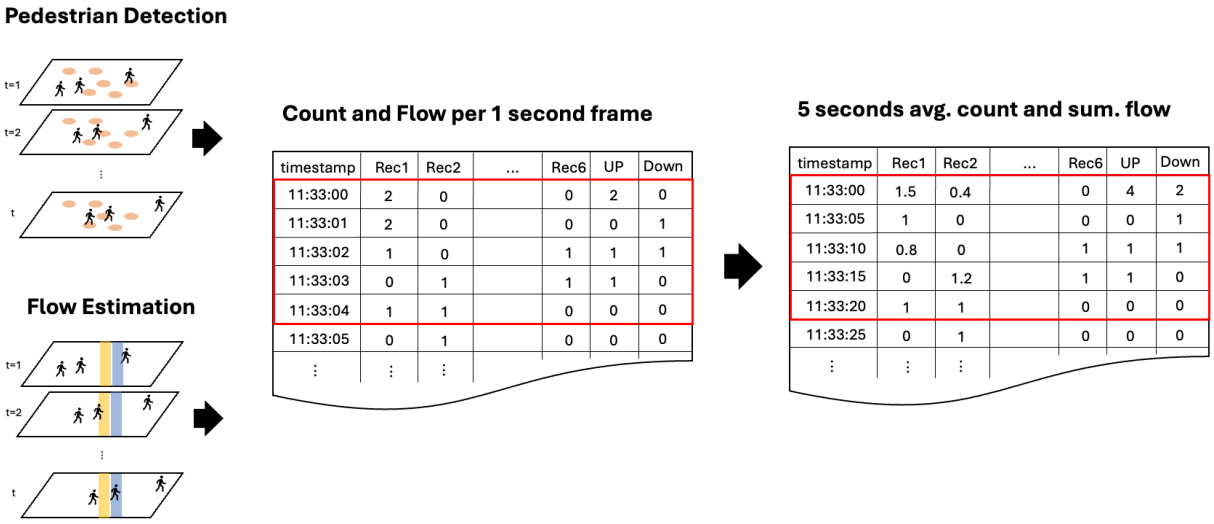
We experimented with different cut-offs, $k \in \{20, 30, 40, 50\}$. If $distance > k$, we assume that the edge does not exist or does not act as a link (M. Liu et al., 2021).

ANALYSIS

In this section, we introduce our experimental setup, framework, and step-by-step processes. The experiments are introduced in two subsections: (1) Pedestrian Flow Estimation and Prediction using ASPED v.a, and (2) Pedestrian Flow Estimation with ASPED v.c dataset.

Pedestrian Flow Estimation and Prediction using ASPED v.a

In the first experiment, we trained a series of models using data collected from six sensor locations in ASPED v.a. dataset. The following diagram illustrates our experimental workflow and methodological design (*Figure 17-17*).



timestamp	Rec1	Rec2	...	Rec6	UP	Down
11:33:00	2	0		0	2	0
11:33:01	2	0		0	0	1
11:33:02	1	0		1	1	1
11:33:03	0	1		1	1	0
11:33:04	1	1		0	0	0
11:33:05	0	1		0	0	0

timestamp	Rec1	Rec2	...	Rec6	UP	Down
11:33:00	1.5	0.4		0	4	2
11:33:05	1	0		0	0	1
11:33:10	0.8	0		1	1	1
11:33:15	0	1.2		1	1	0
11:33:20	1	1		0	0	0
11:33:25	0	1		0	0	0

Figure 16. 5-seconds Aggregation of Data

(1) Data Preprocessing

Pedestrian count data was aggregated into 5-second intervals, where the counts at each location were averaged, and the flow information was summed. This method accounts for the likelihood of overlapping pedestrian counts across frames, while ensuring that flow data is handled distinctly (*Figure 16*).

To capture temporal dependencies, the data was transformed into sequences. Each sequence was made up of 10 consecutive rows of features, representing 50 seconds of aggregated data. The target for each sequence was the sum of pedestrian flow over that 50-second window, enabling the model to estimate the flow based on the counts in that period.

The dataset was then split into training, validation, and test sets using a 6:3:1 ratio to allow for a robust evaluation of the models. After splitting, the training and validation sets were balanced to address the significant number of frames showing zero pedestrian flow. Without balancing, the model could become biased toward predicting zero flow to minimize error, because Cadell

Courtyard average flow is near 0 (*Table 1*). To mitigate this, sequences with zero pedestrian flow were reduced to half the size of those with non-zero flow, ensuring the model learned to predict instances of pedestrian movement accurately. This balanced approach improves the model's ability to predict non-zero pedestrian flow events, providing more meaningful predictions.

(2) Model Training.

To estimate pedestrian flow information (i.e., the number of pedestrians moving up and down) based on pedestrian counts within 9-meter buffers around recorders, we first employed a Convolutional Neural Network (CNN) model designed to capture spatial dependencies in pedestrian flow data. The CNN model processes spatial patterns by applying convolutional filters, which extract spatial features from the pedestrian count data, allowing the model to identify variations in flow across different locations.

We train the CNN model for two specific purposes:

(1) *Flow Estimation Within the Input Period*: This model estimates the pedestrian flow (up and down) based on pedestrian counts collected during the same 50-second input period. The CNN extracts spatial features from the count data to predict the flow directly within that interval.

(2) *Short-Term Flow Prediction*: This model predicts pedestrian flow for the next 5 seconds following the initial 50-second input period. The CNN utilizes the spatial features extracted from the pedestrian count data in the preceding 50 seconds to forecast the short-term pedestrian movement.

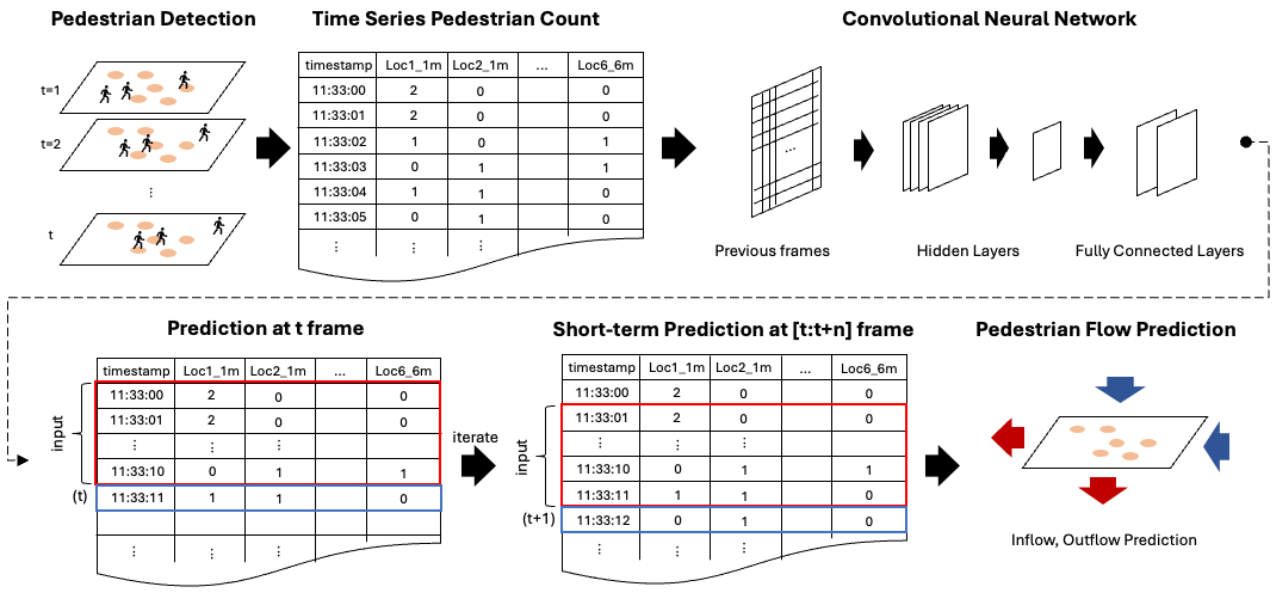


Figure 17. Workflow of Pedestrian Flow Prediction Using ASPED v.a

By utilizing CNN, we aim to balance spatial feature extraction with temporal dependency management in the short-term forecasting task. This approach highlights the effectiveness of CNNs in capturing the spatial relationships inherent in pedestrian count data, demonstrating the potential for these models to predict both immediate and short-term pedestrian flow efficiently.

Extended Experiments with ASPED v.c

(1) Model selection.

The ASPED v.c dataset includes a larger and busier pedestrian network. To better capture the spatial relationships among distantly placed sensors, we incorporated Graph Convolutional Networks (GCNs). These models use sensor adjacency matrices weighted by physical distance, enhancing the model's ability to learn from network topology.

(2) Temporal Lag Optimization for Distributed Sensors

Given the increased distances between sensors in ASPED v.c, we implemented a systematic procedure to determine appropriate temporal lags for incorporating information from sensors that are placed far away. Although the optimal temporal lag selection would be unique to each sensor and not be generalizable, our pipeline provides a replicable method for determining lags in other spatially distributed settings.

As a case study, we targeted Intersection B as the prediction location. Data from sensors at Intersection B were used with no lag. However, data from sensors at Intersection C and Path 1, located approximately 30 and 27 meters away respectively, were tested with lag windows of 10–20, 20–30, and 30–40 seconds. These lag values were derived based on average walking speeds and geodesic distances between locations.

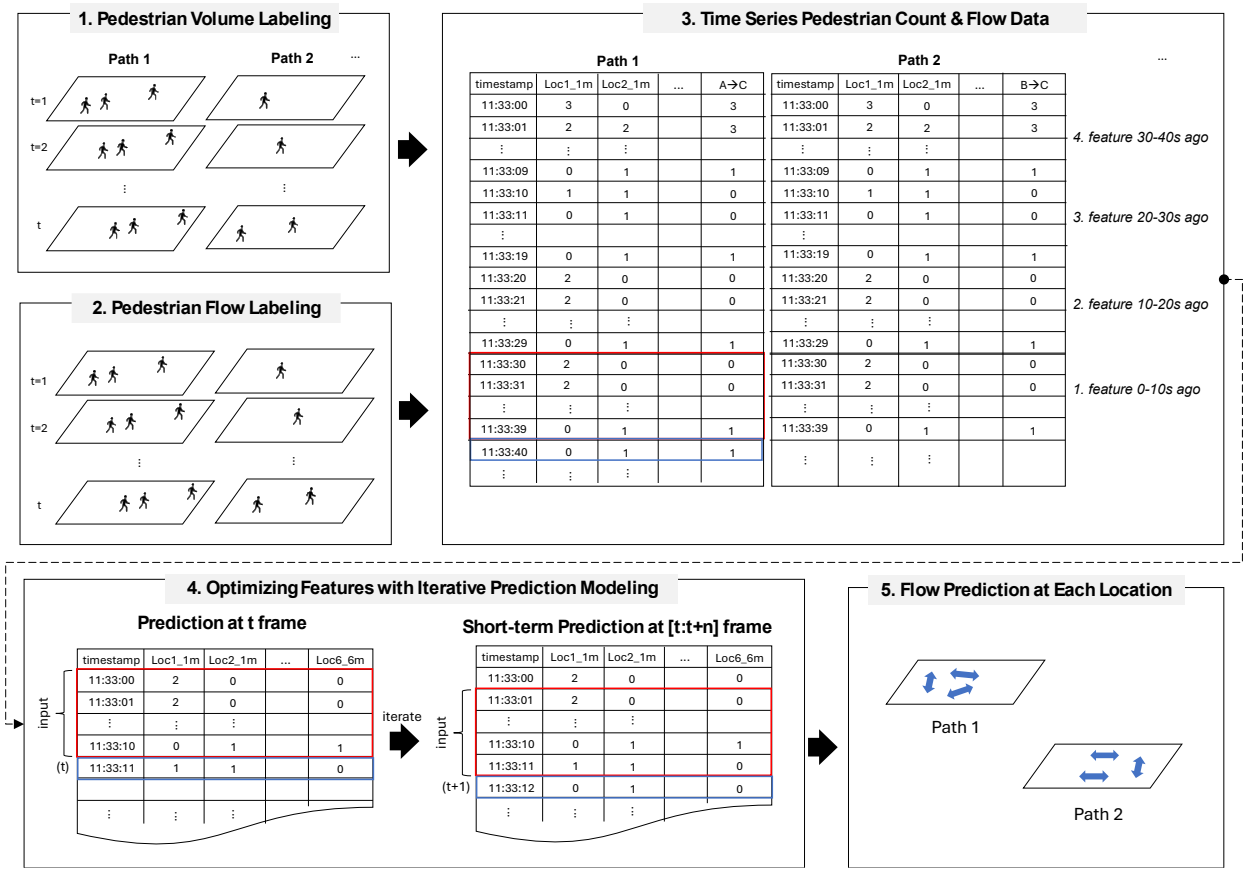


Figure 18. Workflow for Flow Estimation and Prediction Using ASPED v.c

Evaluation Metrics and Comparative Analysis.

Model performance is evaluated on the test set using standard regression metrics such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). We also consider R-squared (R^2) to assess the goodness of fit for each model. After training, we compare the predictive accuracy and efficiency of each model. This comparison highlights the effectiveness of CNN modeling components, as well as the impact of combining them in hybrid models.

By experimenting with these models, we aim to identify the best approach for predicting pedestrian flow based on both spatial and temporal patterns.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

When n is the number of observations, y_i is the actual value and $\hat{y}_i$ is the predicted value. MAE measures the average magnitude of the errors in prediction.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

RMSE is a root of MSE, which gives a sense of how large the errors are, in a more interpretable way, as it is in the same unit as the original data.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

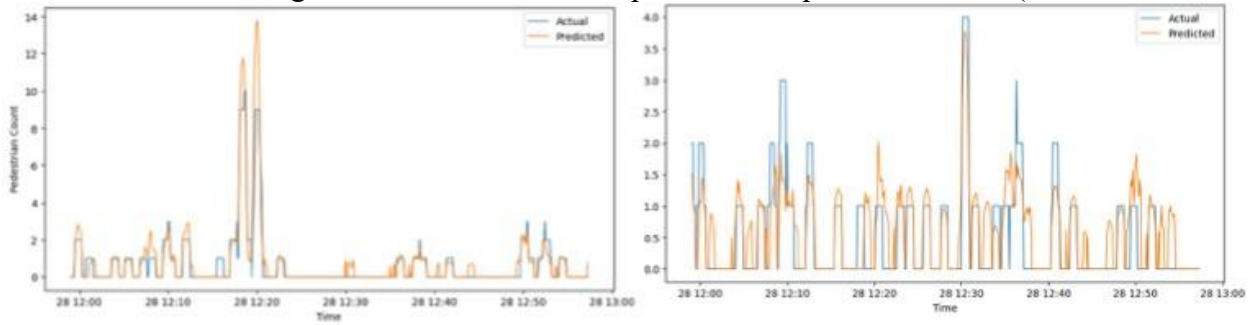
R-square measures how much of the variance in the data is explained by the model. A value close to 1 indicates a good fit.

RESULTS

Pedestrian Flow Estimation and Prediction with ASPED v.a

(1) Estimating Pedestrian Flow Using Pedestrian Count Data

The Convolutional Neural Network (CNN) model successfully estimated pedestrian flows (both up and down flows) based on pedestrian count data collected in 5-second intervals across 10 chunks during the input period (*Table 4, Figure* 19). The results indicate that even with a limited number of recorder locations, the model maintained a reasonable level of accuracy. In fact, expanding the sensor network to include more locations did not significantly enhance model performance (*Table 4)*. The highest R-squared value of 0.70 and the lowest MAE of 0.04 was achieved using data from three locations, indicating that additional recorders beyond this number may not substantially improve model accuracy. The MAE 0.04 can be reverse min-max scaled into an actual value of 0.01 persons, indicating a good model fit.



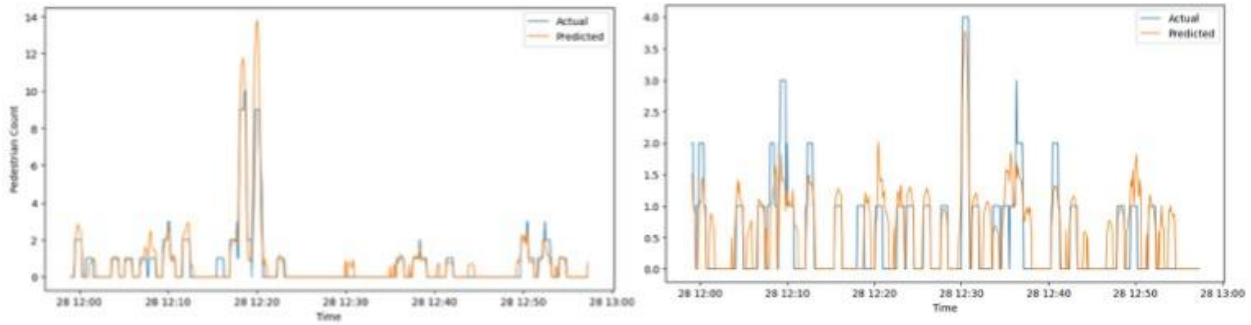


Figure 19. Estimating Pedestrian Flow at Peak time (12-1 PM) using Pedestrian Counts from Two Locations using Convolutional Neural Network

Table 4. Convolutional Neural Network Performance in Estimation Flow from Pedestrian Counts, with different number of features

	MAE	RMSE	R^2
2 Locations, 1 Time	0.09	0.14	0.61
3 Locations, 1 Time	0.04	0.10	0.70
4 Locations, 1 Time	0.07	0.10	0.63
6 Locations, 1 Time	0.06	0.10	0.64

(2) Short-Term Prediction of Pedestrian Flow

In this experiment, we predict the pedestrian flows based on the pedestrian count and flow data of the previous 10 chunks (i.e., predicting flows of the next five seconds based on the previous 50 seconds). We also examined whether short-term predictions of pedestrian flow could be improved by integrating pedestrian count data. The models tested included using only flow data, only count data, and a combination of both.

The findings show that when only flow data was used, the model produced a MAE of 0.05 and RMSE of 0.07 (*Table 5*). Adding pedestrian count data as an additional feature significantly improved the model's performance, reducing the MAE to 0.02 and the RMSE to 0.05. Moreover, combining both flow and count data further enhanced the model's predictive accuracy, resulting in the lowest MAE (0.01) and RMSE (0.04), and the highest R-squared value (0.29). This suggests that integrating pedestrian count data with flow information provides a more comprehensive and accurate short-term prediction of pedestrian flows.

Table 5. Convolutional Neural Network Performance in Short-Term Prediction of Pedestrian Flow

	MAE	RMSE	R^2
Flow Data	0.05	0.07	-
Count Data	0.02	0.05	0.16
Flow + Count Data	0.01	0.04	0.29

The results of the study indicate that the Convolutional Neural Network (CNN) model performed well in estimating pedestrian flows using pedestrian count data, showing reasonable accuracy even with a limited number of recorder locations. While expanding the number of sensors slightly improved the model's performance, it did not lead to a substantial increase in accuracy beyond a certain point. This suggests that a minimal, yet strategically placed network of sensors could be sufficient for pedestrian flow estimation in urban environments.

The short-term forecasting results revealed that integrating pedestrian count data with flow data significantly improved model performance. Short-term prediction models demonstrated that combining pedestrian count data with flow data provides a more comprehensive understanding of pedestrian dynamics. This indicates that using both data types together is more effective than relying on either alone.

Despite attempts to apply the Graph Convolutional Network (GCN) model, it did not outperform the CNN model. This outcome may be due to the nature of the ASPED v.a data. Unlike traditional transportation networks where nodes (intersections) are usually connected by edges (streets), the pedestrian count data collected on the ASPED v.a site does not inherently follow a network structure as it is collected in one courtyard. Recorder locations are relatively close to each other compared to existing studies where they found GCN useful, which makes the use of GCNs less effective compared to CNNs that can process spatially localized features independently of network connections.

While no existing study matches the exact setup of this experiment of estimating and predicting pedestrian flow using count data from multiple recorders, we can draw a partial comparison to a study conducted in Shenzhen, China by Liu et al. (2021). (Other related papers introduced in the literature review, except Liu et al. did not report comparable metrics, such as normalized RMSE or R-square. Most of them only reported the actual RMSE, which scale-dependent.) Liu et al. (2021) involved 25 surveillance cameras installed across a large, busy commercial district known as Dongmen walking street. Unlike our study, they focused solely on predicting pedestrian counts (not flows) for the next minute using aggregated data from the previous minute. Their reported GCN model performance varied by time of day, with R-squared values ranging from 0.67 (weekday mornings) to 0.94 (weekend evenings). These higher scores likely reflect the much larger pedestrian volumes and broader spatial extent of their study site.

Additionally, Liu et al. found that model performance significantly dropped during periods of lower pedestrian density, as indicated by a strong positive correlation between crowd size and RMSE. This finding aligns with our observation that predictive performance is more limited in lower-flow periods. Given the significantly smaller scale and volume of pedestrian activity at our site, the relatively modest R-squared values in our study are expected. Nonetheless, our results provide insights into flow estimation in smaller-scale or resource-constrained settings.

Pedestrian Flow Estimation with ASPED v.c; Case Study of Intersection B, Pedestrian Flow from Clough to Intersection C

We extend the experiments using ASPED v.c, a new dataset collected in the Tech Green area, characterized by a larger network and higher pedestrian volumes. In this stage, we experiment with different model configurations (CNN, GCN) and temporal lags, which are introduced to account for the varying spatial distances between sensors and the target prediction location.

(1) Convolutional Neural Network (CNN)

We first directly used the count values by grouping them into 10 second sequences. We conducted experiments using only flow data, only count data, and a combination of both. A CNN model was employed for these experiments, with Smooth L1 loss as the loss function. The analysis examined whether short-term predictions of pedestrian flow could be improved by integrating pedestrian flow data with the count data.

The findings show that when only flow data was used, the model produced a MAE of 0.19 and RMSE of 0.22 (*Table 6*). Given that the range of target flow data is [0, 26], the MAE could be reversed min-max scaled to 4.94 persons. However, the negative R-square value suggests that the prediction does not explain the variance of the target data very well.

Adding pedestrian count data as an additional feature significantly improved the model's performance, reducing the MAE to 0.14 and the RMSE to 0.18. Moreover, combining both flow and count data further enhanced the model's predictive accuracy, resulting in the lowest MAE (0.14) and RMSE (0.18), and the highest R-squared value (0.14). The 0.14 MAE is 3.64 persons. This suggests that integrating pedestrian count data with flow information provides a more comprehensive and accurate short-term prediction of pedestrian flows, reducing the MAE by 26.32%.

Table 6. Convolutional Neural Network Performance in Short-Term Prediction of Pedestrian Flow

	MAE	RMSE	R^2
Flow Data	0.19	0.22	-0.31
Count Data	0.14	0.18	0.09
Flow + Count Data	0.14	0.18	0.14

From this experiment, we observed that combining flow and count data yields better results compared to using them individually. This suggests that including pedestrian counting sensors combined with trajectory tracking can improve the pedestrian flow prediction accuracy.

We conducted further experiments using the combined flow and count incorporating lagged count values, considering that the recorders are positioned at a distance from the actual scene where flow predictions are being made. The amount of lag introduced was determined based on the relative distance of each recorder from the scene, with greater lag assigned to those farther away. For example, Path 1 and Intersection C are around 30 meters away from the Intersection B

(location where we are predicting flow at). Thus, we experimented with the pedestrian count and flow data collected from those areas from certain number of seconds previous to the frame that we are predicting.

Table 7. Convolutional Neural Network Performance in Short-Term Prediction of Pedestrian Flow with lagged recorder counts

P1 Recorder lag (seconds)	Intersection C recorder lag (seconds)	Training Results			Validation Results		
		MAE	RMSE	R^2	MAE	RMSE	R^2
10	10	0.140	0.175	0.17	0.206	0.245	-0.05
20	10	0.142	0.176	0.17	0.193	0.233	0.06
20	20	0.140	0.176	0.16	0.200	0.239	0.01
30	20	0.141	0.178	0.14	0.207	0.247	-0.06
30	30	0.147	0.183	0.10	0.216	0.257	-0.15
40	20	0.137	0.172	0.20	0.191	0.232	0.06
40	30	0.141	0.178	0.14	0.212	0.253	-0.11
40	40	0.139	0.174	0.17	0.202	0.242	-0.03
60	**30**	**0.128**	**0.162**	**0.28**	**0.184**	**0.226**	**0.11**
90	50	0.135	0.169	0.23	0.192	0.232	0.06

The results (*Table 7.* Convolutional Neural Network Performance in Short-Term Prediction of Pedestrian Flow with lagged recorder counts suggest that for predicting the pedestrian flow from Intersection B, setting temporal lag 30 seconds and 60 seconds to Intersection C and Path 1 respectively, performed the best with lowest MAE, RMSE, and highest R-square. After reverse min-max scaling the errors, the MAE and RMSE are 3.33 and 4.21 persons respectively. While this temporal lag will differ for each location and depend on the spatial distribution of sensors, this process of choosing the temporal lags for each location can be replicated in other areas.

(2) Graph Convolutional Neural Network (GCN)

Given the larger spatial network and higher pedestrian volumes in the ASPED v.c. area compared to the more compact Cadell Courtyard (ASPED v.a), we also evaluated Graph Convolutional Network (GCN) models. GCNs are well-suited for this task, as they can explicitly incorporate spatial relationships between sensors via a graph structure.

We tested two configurations. (1) Raw 1-second data, and (2) Aggregated 5-second data, which helps smooth out momentary fluctuations without sacrificing temporal granularity. The 5-second aggregation notably improved model stability and performance (*Table 8*). This is expected, as 1-second data tends to capture noisy transitions of pedestrians crossing zones. We also updated the loss function to MSE, as it brought better results compared to the SmoothL1 function used in the CNN modeling approaches. In addition, since the model performed best when we encoded 30- and 60-seconds temporal lags to recorders in Intersection C and Path 1, we additionally tested on that configuration.

With a 30- and 60-second temporal lag, the MAE decreased from 0.13 (equivalent to 3.38

persons) in the CNN model to 0.11 (2.86 persons) in the GCN model (*Table 7, Table 8*), reflecting an approximate 15.38% reduction in error This improvement is notably greater than the gains reported by Liu et al. (2021), who observed a 9.55% error reduction when moving from CNN to GCN, and a 10.38% reduction with the Spatio-Temporal GCN (STGCN) in predicting pedestrian counts in the next minute using pedestrian count data. This suggests that the proposed GCN model configuration yields more performance gains in pedestrian flow prediction under the current experimental conditions.

Table 8. Comparison of GCN Models with Disaggregated and Aggregated Input Data – Flow + Count Data

	Validation Results (Disaggregated)			Validation Results (5s Aggregated)		
Lag	MAE	RMSE	R^2	MAE	RMSE	R^2
No Lag	0.05	0.22	0.06	0.11	0.33	0.42
10 seconds lag for Path 1, Int C.	0.06	0.24	0.00	0.07	0.27	0.51
30 seconds lag for Int C, 60 seconds lag for Path 1	0.04	0.21	-0.36	0.11	0.34	0.40

Table 9. GCN Models with Temporal Lag Configurations, 5-seconds Aggregate Input

		Training Results			Validation Results		
Lag		MAE	RMSE	R^2	MAE	RMSE	R^2
No lag	Flow Data	0.10	0.32	0.27	0.14	0.37	0.29
	Count Data	0.09	0.27	0.47	0.13	0.36	0.34
	Flow + Count Data	**0.07**	**0.26**	**0.50**	**0.11**	**0.33**	**0.42**
2 sequences lag (10 seconds lag for recorders in P1 and Int C)	Flow Data	0.11	0.33	0.20	0.14	0.37	0.27
	Count Data	0.08	0.28	0.44	0.13	0.36	0.34
	Flow + Count Data	**0.07**	**0.26**	**0.51**	**0.11**	**0.34**	**0.41**
6 sequences/30 seconds lag for Int C, 12 sequences/60 seconds lag for Path 1	Flow Data	0.10	0.32	0.27	0.14	0.37	0.28
	Count Data	0.08	0.29	0.39	0.14	0.37	0.29
	Flow + Count Data	**0.08**	**0.28**	**0.40**	**0.11**	**0.34**	**0.40**

Then, we experimented GCN models with flow data, count data, and the combined flow and count data as input to see how much count data predicts or improves prediction of pedestrian flow (*Table 9*). The results suggest that optimal predictive performance, reflected by the lowest MAE and highest R^2 values, was achieved under configurations employing either no temporal lag or a two-sequence lag (i.e., 10 seconds) for features derived from recorders at Path 1 and Intersection C. After reverse min-max scaling this value, the MAE of 0.07 is actually 2.08 persons and RMSE 0.26 is 6.76 persons. Given the 5-second aggregation of the dataset and the use of 10

sequential input chunks, each model input already includes temporal information spanning the prior 50 seconds even in the absence of additional lag. Furthermore, across all lag configurations, the inclusion of pedestrian count features consistently enhanced model performance, with further gains observed when both count and flow features were jointly utilized.

To revisit results from Liu et al. (2021) for comparison, our R-square 0.51 is still very low, even compared to their lowest performing period which was during weekdays 7-8 AM (R-square of 0.67). However, because our task involves predicting flow rather than counts, a direct comparison may not be appropriate due to inherent differences in the target variable and task complexity.

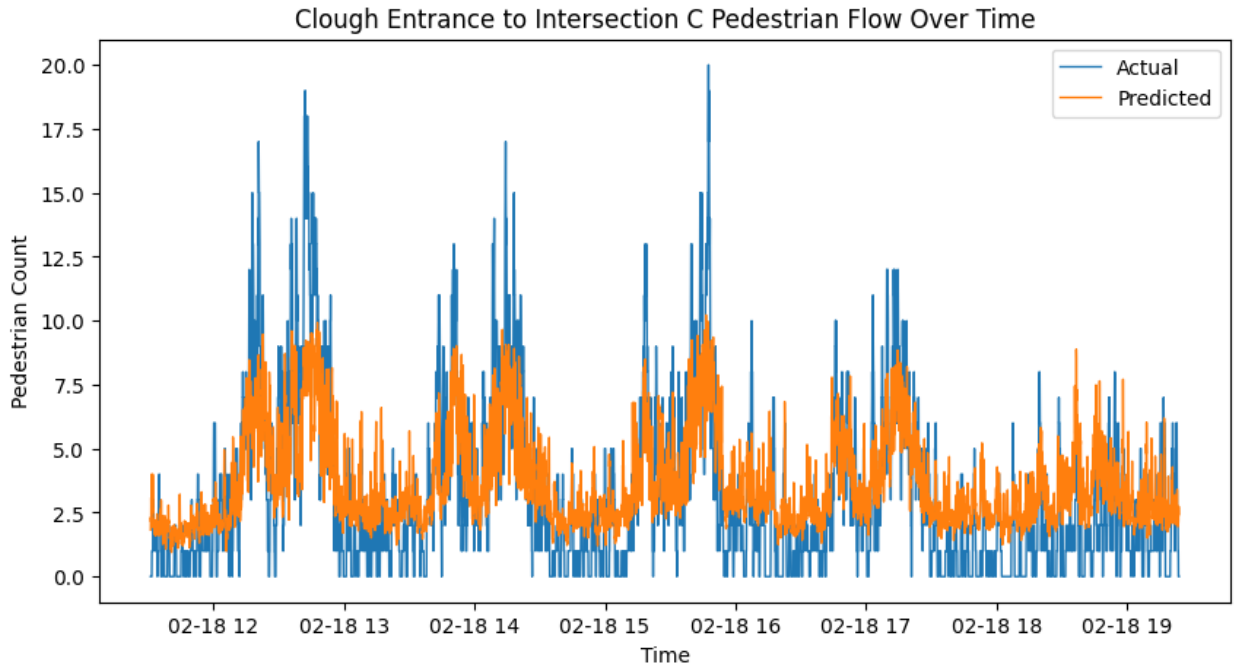


Figure 20. GCN Prediction Results, Intersection B: Clough to Intersection C

From the visualization of prediction results (*Figure 20.* GCN Prediction Results, Intersection B: Clough to Intersection C, we discovered that GCN prediction exhibits over-smoothing, where both extremely low and high values are compressed toward the mean. As suggested by Wong et al. (2024), this issue can potentially be mitigated by incorporating a temporal modeling layer such as LSTM or Gated Recurrent Unit (GRU). Therefore, future research should conduct an additional experiment using the same model configuration with an additional temporal layer.

CONCLUSIONS AND POLICY IMPLICATIONS

This study demonstrates the effectiveness of using video-based pedestrian detection techniques for estimating and predicting pedestrian flow. By leveraging deep learning approaches such as Convolutional Neural Networks (CNNs) and Graph Convolutional Neural Networks (GCNs), we developed a framework that identifies optimal feature selection that enables efficient modeling of pedestrian flow.

While the current experiments relied solely on video sensors, the methodological framework is replicable to other settings and the empirical findings offer a foundation for evaluating alternative sensing modalities, including audio-based recorders. In contexts where video deployment is limited by cost, visibility, or privacy constraints, audio sensors may serve as a viable complement or substitute, particularly for capturing general pedestrian activity levels. Ultimately, this research contributes to the advancement of urban pedestrian sensing strategies that support data-informed planning.

Limitations and Future Research Directions

Several limitations must be acknowledged regarding the generalizability and scalability of the proposed framework. First, the model was developed and tested in two areas within the Georgia Tech campus, and its effectiveness in different cities or urban contexts, such as areas with much larger pedestrian road networks, remains to be validated. Expanding the framework's application across various scale of settings is necessary to assess its adaptability.

The model also shows sensitivity to pedestrian volume and activity levels. Its performance may differ between high-traffic pedestrian corridors and deserted areas. Real-world environments may also present anomalies, such as festivals or spontaneous gatherings, potentially affecting prediction accuracy. Another limitation is the need to develop and calibrate separate models for each intersection. Because the spatial configuration, sensor layout, and pedestrian movement patterns vary significantly from one location to another, a single generalized model may not be feasible without substantial retraining and adaptation.

From a technical standpoint, the current sensor setup powered by battery boxes and reliant on SD card storage poses challenges for consistent long-term data collection. Stable model training and validation depend on uninterrupted sensor operation and reliable data access. In future deployments, integrating a consistent power source and remote data transmission settings will be critical to ensure data continuity and reduce operational burden.

Lastly, the use of pedestrian count data and potential integration with mobile sensing technologies raises ethical and privacy considerations. Although this study anonymized data sources and avoided collecting personal identifiers by automatically blurring torsos, future research and implementation efforts must strictly adhere to privacy regulations and ethical guidelines.

Practical Implications

This study has implications for pedestrian mobility planning, public infrastructure investment, and data governance. First, the results support the case for incorporating pedestrian counting sensors, especially video systems, into urban monitoring networks. Such data can inform evidence-based

decisions, such as crosswalk timing, sidewalk allocation, and accessibility improvements, particularly in walkable city initiatives.

Second, the framework developed here can aid transportation agencies in identifying high-traffic pedestrian zones that may require targeted interventions, such as signal timing adjustments, crowd control measures, or redesign of intersections to prioritize pedestrian flow and safety. Predictive pedestrian flow models also have value for emergency preparedness and crowd management during planned events.

However, the use of sensor-based pedestrian monitoring must be accompanied by clear policies on data governance and public transparency. Municipalities adopting such systems should establish protocols for anonymizing collected data, limiting its use to planning purposes, and communicating openly with the public about what is collected and why. In the case of future integration with audio or video-based sensing, adherence to privacy regulations and the development of individual consent-based frameworks will be important to maintain public trust.

REFERENCES

Ahmed, S. H., Raza, M., Mehdi, S. S., Rehman, I., Kazmi, M., & Qazi, S. A. (2021). Faster RCNN based Vehicle Detection and Counting Framework for Undisciplined Traffic Conditions. *2021 IEEE 18th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI (HONET)*, 173–178. https://doi.org/10.1109/HONET53078.2021.9615466

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

Algiriyage, N., Prasanna, R., E.H. Doyle, E., Stock, K., Johnston, D., Punchihewa, M., & Jayawardhana, S. (2021, May). Towards Real-time Traffic Flow Estimation using YOLO and SORT from Surveillance Video Footage. *WiP Paper - AI and Intelligent Systems for Crises and Risks*. 18th ISCRAM Conference, Blacksburg, VA, USA. https://idl.iscram.org/files/nilanialgiriyage/2021/2311_NilaniAlgiriyage_etal2021.pdf

American Planning Association. (1965). *The Pedestrian Count*. American Planning Association. https://www.planning.org/pas/reports/report199.htm

Baul, A. (2021). *Learning to Detect Pedestrian Flow in Traffic Intersections from Synthetic Data* [Master Thesis, The University of Texis Rio Grande Valley]. https://scholarworks.utrgv.edu/etd/829/

Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., & Girdhar, R. (2021). *Masked-attention Mask Transformer for Universal Image Segmentation*. https://doi.org/10.48550/ARXIV.2112.01527

Dalal, N., & Triggs, B. (2005). *Histograms of oriented gradients for human detection. 1*, 886–893.

Deo, N., & Trivedi, M. M. (2021). *Trajectory Forecasts in Unknown Environments Conditioned on Grid-Based Plans* (arXiv:2001.00735). arXiv. http://arxiv.org/abs/2001.00735

Dollar, P., Wojek, C., Schiele, B., & Perona, P. (2009). Pedestrian detection: A benchmark. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 304–311. https://doi.org/10.1109/CVPR.2009.5206631

Han, C., Seshadri, P., Ding, Y., Posner, N., Koo, B. W., Agrawal, A., Lerch, A., & Guhathakurta, S. (2024). Understanding pedestrian movement using urban sensing technologies: The promise of audio-based sensors. *Urban Informatics*, *3*(1), 22. https://doi.org/10.1007/s44212-024-00053-9

Kitano, Y., Kuwamoto, S., & Asahara, A. (2019). OD-network-based Pedestrian-path Prediction for People-flow Simulation. *2019 IEEE International Conference on Big Data (Big Data)*, 1656–1661. https://doi.org/10.1109/BigData47090.2019.9006314

Li, J., Boonaert, J., Doniec, A., & Lozenguez, G. (2021). Multi-models machine learning methods for traffic flow estimation from Floating Car Data. *Transportation Research Part C: Emerging Technologies*, *132*, 103389. https://doi.org/10.1016/j.trc.2021.103389

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

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

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). SSD: Single Shot MultiBox Detector. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer Vision – ECCV 2016* (Vol. 9905, pp. 21–37). Springer International Publishing. https://doi.org/10.1007/978-3-319-46448-0_2

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

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2015). *You Only Look Once: Unified, Real-Time Object Detection* (Version 5). arXiv. https://doi.org/10.48550/ARXIV.1506.02640

Ren, S., He, K., Girshick, R., & Sun, J. (2015). *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.1506.01497

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

Van Steen, M., Stanciu, V.-D., Shafaeipour, N., Chilipirea, C., Dobre, C., Peter, A., & Wang, M. (2022). Challenges in Automated Measurement of Pedestrian Dynamics. In D. Eyers & S. Voulgaris (Eds.), *Distributed Applications and Interoperable Systems* (Vol. 13272, pp. 187–199). Springer International Publishing. https://doi.org/10.1007/978-3-031-16092-9_12

Wong, V. (2024). *SPATIO-TEMPORAL REPRESENTATION LEARNING: APPLICATIONS TO MANUFACTURING PLANNING AND PEDESTRIAN CROWD ANALYSIS*. STANFORD UNIVERSITY.

Xia, T., Lin, J., Li, Y., Feng, J., Hui, P., Sun, F., Guo, D., & 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

Xu, Y., Yu, G., Wang, Y., Wu, X., & Ma, Y. (2017). Car Detection from Low-Altitude UAV Imagery with the Faster R-CNN. *Journal of Advanced Transportation*, *2017*, 1–10. https://doi.org/10.1155/2017/2823617

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

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

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

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

Zhang, Y., Wang, J., & Yang, X. (2017). Real-time vehicle detection and tracking in video based on faster R-CNN. *Journal of Physics: Conference Series*, *887*, 012068. https://doi.org/10.1088/1742-6596/887/1/012068