Urban Sensing for Human-Centered Systems: A Modular Edge Framework for Real-Time Interaction

Navid Salami Pargoo[1,*], Mahshid Ghasemi[2], Shuren Xia[1], Mehmet Kerem Turkcan[2], Taqiya Ehsan[1], Chengbo Zang[2], Yuan Sun[1], Javad Ghaderi[2], Gil Zussman[2], Zoran Kostic[2], Jorge Ortiz[1,*]
[1]WINLAB, Rutgers University, New Jersey, USA
[2]Columbia University, New York, USA
Email:{navid.salamipargoo,shuren.xia,taqiya.ehsan,ys820,jorge.ortiz}@rutgers.edu;
{mahshid.ghasemi,mkt2126,cz2678,jg3465,gil.zussman,zk2172}@columbia.edu

Abstract

Urban environments pose significant challenges to pedestrian safety and mobility. This paper introduces a novel modular sensing framework for developing real-time, multimodal streetscape applications in smart cities. Prior urban sensing systems predominantly rely either on fixed data modalities or centralized data processing, resulting in limited flexibility, high latency, and superficial privacy protections. In contrast, our framework integrates diverse sensing modalities, including cameras, mobile IMU sensors, and wearables into a unified ecosystem leveraging edge-driven distributed analytics. The proposed modular architecture, supported by standardized APIs and message-driven communication, enables hyper-local sensing and scalable development of responsive pedestrian applications. A concrete application demonstrating multimodal pedestrian tracking is developed and evaluated. It is based on the cross-modal inference module, which fuses visual and mobile IMU sensor data to associate detected entities in the camera domain with their corresponding mobile device. We evaluate our framework's performance in various urban sensing scenarios, demonstrating an online association accuracy of 75% with a latency of ≈39 milliseconds. Our results demonstrate significant potential for broader pedestrian safety and mobility scenarios in smart cities.

ACM Reference Format:
Navid Salami Pargoo[1,*], Mahshid Ghasemi[2], Shuren Xia[1], Mehmet Kerem Turkcan[2], Taqiya Ehsan[1], Chengbo Zang[2], Yuan Sun[1], Javad Ghaderi[2], Gil Zussman[2], Zoran Kostic[2], Jorge Ortiz[1,*] . 2025. Urban Sensing for Human-Centered Systems: A Modular Edge Framework for Real-Time Interaction. In *The 3rd International Workshop on Human-Centered Sensing, Modeling, and Intelligent Systems (HumanSys '25), May 6–9, 2025, Irvine, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3722570.3726890

1 INTRODUCTION

The rapid urbanization and rising population necessitate new approaches to pedestrian safety and mobility. In the U.S., 83% of the population resides in urban areas, reaching 89% by 2050 [18]. Globally, 56.9% of people lived in urban areas in 2022, with projections indicating an increase to 68% by 2050 [23]. These trends underscore the need for advanced urban safety and mobility solutions.

Urban environments pose complex challenges, particularly pedestrian safety at intersections. In 2022, 1,705 pedestrian were killed at intersections in the U.S., accounting for 23% of total pedestrian fatalities [6]. Many existing urban sensing systems lack real-time responsiveness, multimodal data integration, and privacy safeguards, limiting their effectiveness. Addressing these issues requires a novel approach that fuses advanced sensing, real-time processing, and privacy protection to improve urban mobility and safety.

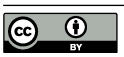
This paper introduces *Streetscape applications*, a new class of real-time, hyper-local urban intelligence solutions. A streetscape comprises the road, buildings, and public spaces that shape a street's character. Streetscape applications enhance pedestrian experiences by integrating diverse sensors—including cameras, mobile IMUs, wearables, and edge computing units—to monitor and respond to urban dynamics. These systems improve road safety, traffic efficiency, public security, accessibility, and environmental monitoring. Given the challenges of pedestrian safety and intersection risks, streetscape applications provide a novel solution by delivering real-time, hyper-local intelligence.

A robust sensing infrastructure integrating advanced networking and distributed sensing is essential for streetscape applications [16]. Existing solutions typically focus on single-modality sensing without integration of multiple sources, centralized processing with high latency and limited responsiveness, and rigid architectures with limited plug-and-play capabilities, while struggling to balance real-time data analytics and privacy. Our proposed sensing framework uniquely addresses these limitations by seamlessly integrating multiple sensing modalities into a coherent multimodal ecosystem, leveraging distributed edge computation and analytics to deliver low-latency responsiveness, providing a highly modular and flexible architecture driven by standardized APIs, and embedding edge-driven privacy-preserving mechanisms through anonymization at data capture and secure local data handling.

We demonstrate preliminary capabilities of our framework through an application for cross-modal matching between mobile IMU sensors and pedestrian camera detections. A multimodal model learns an affinity matrix to match pedestrians' mobile phones with their corresponding traces in the camera domain. This representative application enables diverse urban intervention scenarios such as extending crosswalk signals for mobility-impaired individuals or providing navigation assistance for visually impaired pedestrians.

Our framework is evaluated in multiple urban sensing scenarios, including a controlled parking lot and a city-scale mobile wireless

* Corresponding authors.



This work is licensed under a Creative Commons Attribution 4.0 International License.
HumanSys '25, Irvine, CA, USA
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1609-6/2025/05
https://doi.org/10.1145/3722570.3726890

testbed in New York City (COSMOS) [17]. By enabling multimodal pedestrian association, our approach enhances urban safety by enabling faster, more accurate pedestrian responses. Our results demonstrate an association accuracy of 75% with a latency of ≈39 milliseconds, highlighting the potential of our framework to improve pedestrian safety and mobility in smart cities.

2 RELATED WORK

2.1 Human-Centric Sensing and Well-Being in Urban Environments

Urban sensing technologies increasingly prioritize human well-being, using multimodal data to enhance pedestrian safety and mobility. Research highlights the benefits of walkability for public health, traffic safety, and urban livability [18]. Wearable sensors provide real-time insights into pedestrians' interactions with their environment [2], allowing city planners to design safer and more accessible spaces. Recent studies have leveraged multimodal urban analytics, such as fusing environmental, physiological, and behavioral data, to support human-centered smart cities [4, 25]. Our work extends this paradigm by integrating visual and IMU data to improve real-time pedestrian tracking and safety interventions.

2.2 Advanced AI Models for Human-in-the-Loop Applications

Human-in-the-loop AI systems combine algorithmic decision-making with user feedback, ensuring adaptability and trustworthiness in safety-critical urban environments. Interactive learning frameworks enable real-time adjustments based on pedestrian behavior, refining models for traffic control and safety [13]. Federated reinforcement learning has been proposed to balance personalization and fairness in urban applications [5], demonstrating effectiveness in pedestrian-aware mobility planning. Our framework aligns with these approaches by incorporating multimodal learning and on-device processing to enhance safety in streetscape scenarios.

2.3 Context-Aware and Wearable Devices for Urban Mobility

Wearable and smartphone-based sensors enable context-aware mobility solutions, helping pedestrians navigate urban spaces safely. For example, tactile feedback devices assist visually impaired users with real-time path guidance [9]. IMU-based pedestrian localization has also been explored for smartphone-assisted navigation in complex environments [19]. These studies demonstrate the potential of multimodal fusion for mobility enhancement, which our work leverages by integrating visual, IMU, and environmental data to improve pedestrian tracking accuracy.

2.4 Privacy and Ethical Considerations in Human-Centric Urban Systems

Privacy-preserving urban sensing remains a critical challenge, particularly for real-time pedestrian tracking. Differential privacy and cryptographic aggregation methods have been applied to protect personally identifiable information in crowd-sensing applications [8]. Federated learning approaches enable AI training across distributed devices without centralized data collection, reducing privacy risks while maintaining analytical utility. Ethical concerns surrounding biometric surveillance have also led to the adoption of on-device anonymization techniques [21]. Our system incorporates these principles by anonymizing visual data at the edge and restricting API access to processed outputs to ensure privacy-preservation.

2.5 Existing Streetscape & Mobility-Focused Solutions

Existing streetscape sensing solutions often rely on singular sensing modalities, such as cameras or mobile devices, resulting in fragmented situational awareness. Early systems, like MetroSense [2], pioneered the fusion of static and mobile sensors but faced scalability challenges in real deployments. Subsequent analyses [1, 27] highlight persistent fragmentation between infrastructure-based and mobile sensing approaches, limiting their collective utility. Architectural solutions like the Modular Sensor System ([26]) provided modular hardware interfaces but lacked software-level flexibility for diverse urban applications. Furthermore, existing architectures often rely on centralized processing, which significantly increases latency, limiting their real-time responsiveness for pedestrian safety scenarios [24]. Lastly, privacy-preserving mechanisms in these systems often rely on post hoc, gateway-level anonymization [2, 26], falling short of protecting data at the source.

Our framework addresses these limitations through a fully modular and distributed architecture that integrates visual and mobile sensors via synchronization protocols and fusion techniques. Distributed edge computing nodes within localized zones enable real-time processing with minimal latency. Clearly defined APIs support flexible, application-agnostic software composition, while anonymization and secure data handling ensure privacy preservation at the edge.

2.6 Summary of Gaps and Contributions

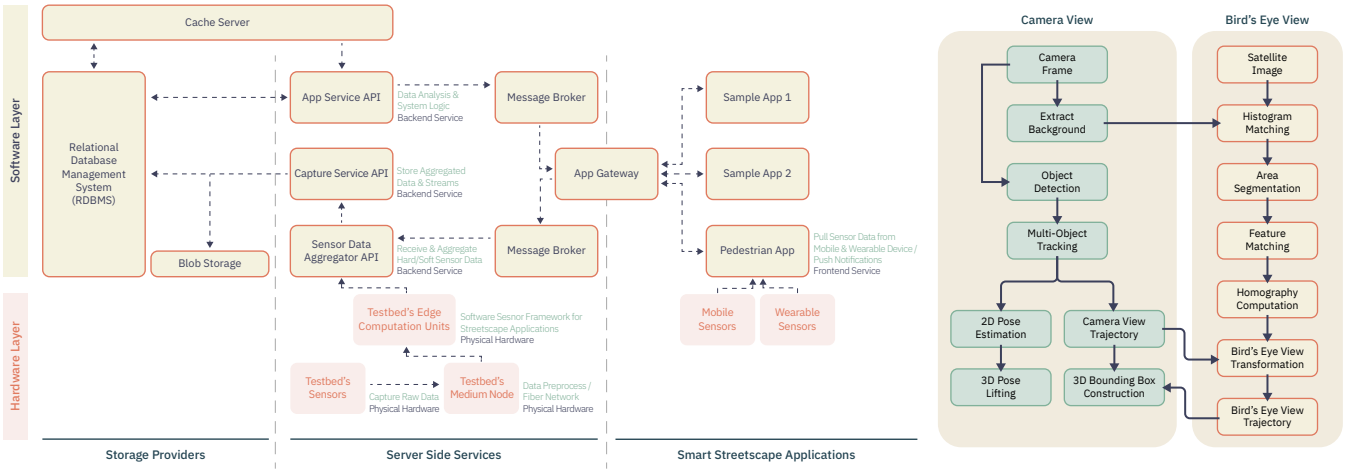
Despite advancements, current multimodal urban sensing frameworks face critical limitations:

- **Modality Fragmentation:** Existing systems rely on isolated sensor modalities, limiting integrated situational awareness.
- **High Latency:** Centralized data processing approaches hinder real-time responsiveness in dynamic urban environments.
- **Rigid Architectures:** Limited software modularity restricts adaptability and scalability across diverse urban applications.
- **Privacy Constraints:** Centralized anonymization methods inadequately address edge-level data protection requirements.

To overcome these limitations, our sensing framework integrates fixed and mobile modalities through standardized APIs, employs distributed edge processing to achieve low latency, provides modular software-defined architectures for rapid application development, and implements edge-level privacy preservation mechanisms. Our approach significantly advances multimodal urban sensing towards real-time, privacy-conscious, and scalable streetscape applications.

3 SYSTEM ARCHITECTURE AND DESIGN PRINCIPLES

We present a robust, scalable framework to support multimodal pedestrian-safety and mobility applications in urban environments.



(a) Sensing System Architecture tailored to Smart Streetscape Applications **(b) Software Sensor Framework**

Figure 1: System Architecture and Framework Design

This framework integrates cameras, mobile IMU sensors, and wearables, using edge computing for real-time data collection and analysis. By combining multiple data sources, the architecture delivers low-latency responses, supports continuous operation, and preserves user privacy.

3.1 Architectural Implications

Our experience with multimodal sensing applications on COSMOS testbed informs several key design requirements. Applications such as navigation or waypoint finding require fusing heterogeneous data (video, IMU, etc.) and performing cross-modal inference to link sensor streams accurately. Adaptive traffic signals and other responsive services demand continuous operation with minimal latency. We achieve this via redundant data pathways and distributed edge nodes. Handling sensitive pedestrian data necessitates sandboxing for privacy safeguards. We anonymize data at capture, transmit securely, and restrict direct API access to only processed outputs.

4 DESIGN

Our system architecture is built upon a flow from data capture to application output, ensuring efficient handling and processing of the information generated by the sensors, as depicted in Figure 1a.

4.1 Hardware Level

Cameras are installed at roadside, each paired with an edge node that performs initial tasks such as noise reduction, compression, and partial analytics. Processed video frames are then streamed over a fiber network to a shared pool of edge servers, where higher-level computer vision pipelines (e.g., detection and tracking) run. On the user side, mobile or wearable devices capture IMU data (e.g., motion, orientation) and, if available, physiological signals. These data streams are published via an app gateway and message broker so that relevant server-side services can subscribe to them.

4.2 Software Level

The Sensor Data Aggregator API is pivotal in our architecture, subscribing to a message broker to collect pedestrian app data and integrating it with both edge-processed soft sensor outputs and hard sensor data. This fusion produces a comprehensive dataset that represents the urban environment holistically.

The aggregated data is then forwarded to the Capture Service API, which persists it in both structured and unstructured formats. Structured data, comprising soft sensor outputs and mobile/wearable readings, is stored in an RDBMS, while unstructured data such as video frames is archived in blob storage. This segregation is critical for privacy preservation, as it ensures sensitive video content remains under stringent local controls. A caching layer further accelerates access to frequently requested data.

The App Service API interfaces with non-sensitive data from the RDBMS and cache, offering endpoints for application services. It supports plug-and-play integration of diverse machine learning models, logic modules, and analytical tools, enabling the system to dynamically host various smart urban applications. One such service (Section 5), demonstrate the architecture's versatility.

User requests are routed through an app gateway via the message broker to the App Service API. After processing, the results are sent back through the broker and delivered by the gateway to the user, ensuring seamless and responsive service.

4.3 Software Sensor Framework

Our soft sensor framework is a computer vision pipeline designed to extract features for smart urban applications. The process begins by transforming the camera view ($\check{I}$) into a Bird's Eye View (BEV, $\tilde{I}$) for precise spatial analysis.

4.3.1 Perspective Transformation. Traditional perspective transformation relies on knowing the camera's intrinsic and extrinsic parameters. The projection from 3D world coordinates $\mathbf{X}$ to 2D image coordinates $\mathbf{x}$ is given by $\mathbf{x} = \mathbf{K}[\mathbf{R} \mid \mathbf{t}]\mathbf{X}$, where $\mathbf{K}$ is the intrinsic matrix, $\mathbf{R}$ the rotation matrix, and $\mathbf{t}$ the translation vector:

Figure 2: Perspective transformation workflow using satellite and camera imagery. Left to right: (1) Original satellite image, (2) Original camera image, (3) Extracted background, (4) Area segmentation, (5) Histogram-equalized satellite image, (6) Histogram-equalized camera image, (7) Feature matching between satellite and camera images.

$$\mathbf{K}[\mathbf{R} \mid \mathbf{t}] = \begin{bmatrix} f_x & s & c_x & 0 \\ 0 & f_y & c_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\theta_c & -\sin\theta_c & 0 \\ 0 & \sin\theta_c & \cos\theta_c & -\frac{h_c}{\sin\theta_c} \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

yielding the transformation matrix:

$$\mathbf{T} = \mathbf{K}[\mathbf{R} \mid \mathbf{t}], \quad \mathbf{T} \in \mathbb{R}^{3\times4}.$$

Since acquiring exact camera parameters is often impractical, our system employs automatic inverse perspective mapping using satellite imagery ($\bar{I} \approx \tilde{I}$). By matching at least four corresponding feature points between $\check{I}$ and $\bar{I}$, we estimate a reduced-order planar transformation:

$$\mathbf{T} \in \mathbb{R}^{3\times3}, \quad \mathbf{T} : \check{I} \to \tilde{I}.$$

4.3.2 Pre-Calibration Enhancements. To improve feature matching between the camera view and satellite imagery, we perform:

- **Background Extraction:** We stabilize $\check{I}$ by iteratively estimating a static background $\check{B}^t$:
$$\check{B}^t = (1 - \alpha S(\check{I}^t))\check{B}^{t-1} + \alpha S(\check{I}^t)\check{I}^t,$$
where the confidence score is:
$$S(\check{I}^t) = \min\left(1, \tfrac{1-M^t}{\tau}\right),$$
with motion intensity:
$$M^t = \tfrac{|\check{I}^t - \check{B}^{t-1}|}{255}.$$
- **Histogram Equalization:** We adopt the methods in [14, 20] to perform color correlation-based histogram matching to align the intensity distributions of $\check{I}$ and $\bar{I}$ by solving:
$$\arg\min_M \textstyle\sum_k d\big(M(H^c_{\check{I}}(k)), H^c_{\bar{I}}(k)\big),$$
where $H^c_{\check{I}}(k)$ and $H^c_{\bar{I}}(k)$ denote channel histograms.
- **Area Segmentation:** We segment $\check{I}$ into regions $\{\check{I}_i\}_{i=1}^n$ and calibrate each with its own transformation $\mathbf{T}_i$:
$$\mathbf{T}_i \cdot \check{\mathbf{p}}_i = \bar{\mathbf{p}}_i.$$

4.3.3 Auto-Calibration. To ensure robust alignment without manual intervention, we apply the Affine Scale-Invariant Feature Transform (ASIFT) to detect keypoints in both $\check{I}$ and $\bar{I}$. Formally, we compute:

$$\{(\check{x}_i, \check{d}_i)\}_{i=1}^{\check{N}} = \text{ASIFT}(\check{I}) \quad \text{and} \quad \{(\bar{x}_i, \bar{d}_i)\}_{i=1}^{\bar{N}} = \text{ASIFT}(\bar{I}),$$

where $\check{x}_i$ and $\bar{x}_i$ are keypoint coordinates and $\check{d}_i$, $\bar{d}_i$ are descriptor vectors. Nearest-neighbor matching produces candidate pairs $\{(\check{x}_i, \bar{x}_{j(i)})\}$. To remove outliers, we use RANSAC to estimate the transformation matrix:

$$\mathbf{T} = \arg\max_{\mathbf{T}} \sum_i \delta\Big(\|\bar{x}_i - \mathbf{T}\check{x}_i\|\Big),$$

where $\delta(\cdot)$ equals 1 if the error is below a threshold and 0 otherwise. This matrix $\mathbf{T}$ encapsulates the homography between the camera and satellite views.

4.3.4 Object Detection and Tracking. For detection, we use YOLOv8 fine-tuned on public (COCO, VisDrone, SDD) and proprietary datasets to robustly detect pedestrians and vehicles under varying conditions. For tracking, we employ the OC-SORT algorithm [3] with an Observation-Centric Re-Update (ORU) mechanism:

$$\hat{z}_t = \text{Traj}_{\text{virtual}}(z_{t1}, z_{t2}, t), \quad t_1 < t < t_2.$$

Additional modifications, such as shadow tracking and adjusted hit thresholds, ensure continuous tracking despite occlusions.

4.3.5 Trajectory and 3D Bounding Box. Pedestrian positions in the camera view are calculated as:

$$\check{\mathbf{p}} = \left(\tfrac{x_1+x_2}{2}, y_2 - 0.1(y_2 - y_1)\right),$$

where (x_1, y_1) and (x_2, y_2) define the bounding box. The region i is determined by $i = f(\check{p})$ and the BEV position is computed as:

$$\tilde{\mathbf{p}} = \mathbf{T}_i \cdot \check{\mathbf{p}}.$$

Trajectories are recorded as:

$$\check{Traj}^{p_i} = \{(\check{p}^t, t) : t \in T_{p_i}\}, \quad \tilde{Traj}^{p_i} = \{(\tilde{p}^t, t) : t \in T_{p_i}\}.$$

A fixed-size ground plane bounding box (scaled by anthropometric ratios) is rotated using the pedestrian's heading angle:

$$\theta = \arctan 2(y_2 - y_1,\ x_2 - x_1).$$

The final 3D bounding box $\check{\mathbf{B}}_{3D}$ is generated by extruding the base to 90% of the original height.

4.3.6 Pose Estimation. We estimate 2D poses using RTMPose [10] on detected bounding boxes:

$$\mathbf{K}^i_{2D} = \mathcal{M}_{2D}(\mathbf{I}_t, \mathbf{b}_i),$$

and lift them to 3D using MotionBERT [28]:

$$\mathbf{K}^i_{3D} = \mathcal{M}_{3D}(\mathbf{K}^i_{2D}).$$

Kalman filtering is applied to refine 3D pose trajectories, enhancing robustness in real-world applications.

Figure 1b summarizes the software sensor framework and its key processing steps.

5 APPLICATION

We illustrate how our multimodal association approach supports real-time pedestrian safety and mobility in a streetscape environment. Specifically, we detail the methodology for cross-modal matching, implementation considerations, and evaluation of association accuracy.

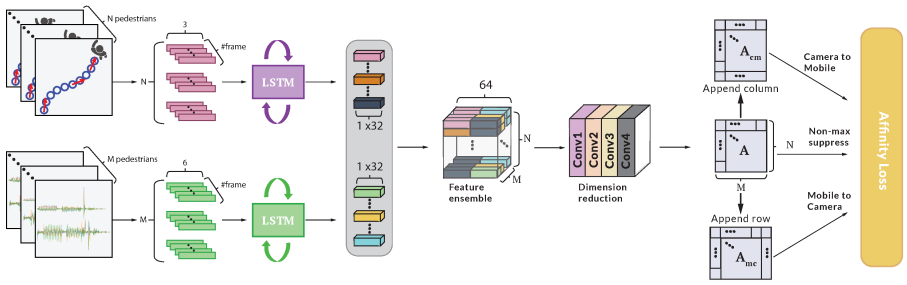


Figure 3: An overview of our multimodal association architecture, where LSTM-based feature extractors process phone (LSTM_P) and camera (LSTM_V) data. Their concatenated outputs form a permutation-based feature ensemble, which is compressed into an affinity matrix via a dimension-reduction module. Extra row and column are appended to handle unmatched entities, yielding A_r for robust phone–camera correspondence.

5.1 Methodology

Our approach fuses bird's eye view trajectories detected in the camera domain with phone IMU data to associate each detected pedestrian in the video stream with a corresponding mobile device. Inspired by [7, 11] and outlined in Figure 3, we use two bidirectional LSTMs to extract spatiotemporal embeddings: one from trajectory sequences, another from IMU signals. We concatenate these embeddings to form a permutation cubic, then apply a 1×1 CNN stack to produce a 2D affinity matrix, where each entry indicates the likelihood of a pedestrian–phone match. A final step appends extra rows and columns to handle unmatched entities, allowing for scenarios with varying pedestrian counts.

5.2 Implementation

We conducted experiments in a controlled parking lot simulating a small-scale city streetscape. An RTSP camera as part of the testbed overlooked the area. Five participants were recruited and instructed to carry their smartphones by hand while walking freely. Each smartphone logged accelerometer and gyroscope data at 100 Hz. The overhead camera recorded video at 30 fps, capturing each participant's bounding box and consequently bird's eye view trajectories as they moved. All data were resampled at 30 Hz.

Model Details. Our cross-modal matching model includes Bidirectional LSTMs (2 layers, hidden dimension = 32) for both camera bounding-box sequences and phone IMU streams. A CNN compression network consisting of 1×1 convolutions reduces the fused embeddings into a single-channel affinity matrix. Training used a batch size of 32, a learning rate of 0.001, for 100 epochs. We apply cross-entropy losses on the affinity matrix, augmented by consistency terms to penalize mismatches.

5.3 Evaluation

We measure association accuracy by comparing the predicted pedestrian–phone matches against ground-truth labels from manual annotation. Over a 3-second (90-frame) sliding window, our model achieves an online matching accuracy of 75%, effectively linking each participant's smartphone to its corresponding bounding box. The appended row/column mechanism also helps minimize false matches, especially when pedestrians leave or re-enter the camera's field of view. Overall, these results demonstrate the feasibility of real-time cross-modal matching in urban streetscape applications, enabling hyper-local interventions (e.g., adaptive crosswalks) that improve pedestrian safety and mobility.

To quantitatively substantiate the practical feasibility of our edge-driven framework for real-time streetscape applications, we explicitly measured several performance metrics. Our distributed edge processing pipeline achieved an average end-to-end latency of ≈39 *ms*, from sensor capture (camera frame acquisition and IMU measurement) through data processing to final association output. This represents a substantial improvement compared to prior centralized frameworks with latencies of ≈100 *ms* or higher [12, 15, 22]. Individual video frames were processed in ≈22 *ms* on average, comfortably supporting real-time responsiveness at 30 frames per second (fps) video capture rates. Our implementation demonstrated a throughput capability of handling at least 10 concurrent pedestrian–device associations per edge node without noticeable performance degradation, indicating scalability for realistic urban pedestrian densities.

6 CONCLUSION

In this paper, we presented a multimodal approach for cross-modal association in urban streetscapes, focusing on pedestrian–phone matching. Our framework fuses camera detections with mobile IMU data via bidirectional LSTMs and a CNN-based compression module, producing an affinity matrix that encodes match probabilities. Experiments in a controlled parking lot demonstrated a 75% online matching accuracy, highlighting the feasibility of real-time deployments for pedestrian safety and mobility in smart city environments. By integrating robust unmatched-dimension handling and end-to-end training, our approach scales to real-world conditions where participants move dynamically.

Future extensions of our framework will focus on advancing core capabilities to fully enable human-centered urban sensing. Specifically, we plan to provide services for multimodal data synchronization algorithms; implement robust multicamera fusion to address occlusion challenges; and expand our distributed edge computing capabilities to achieve high throughput with minimal latency. Additionally, we will integrate advanced privacy-preserving methods such as differential privacy directly at the edge, ensuring robust data security and regulatory compliance. Ultimately, this work lays

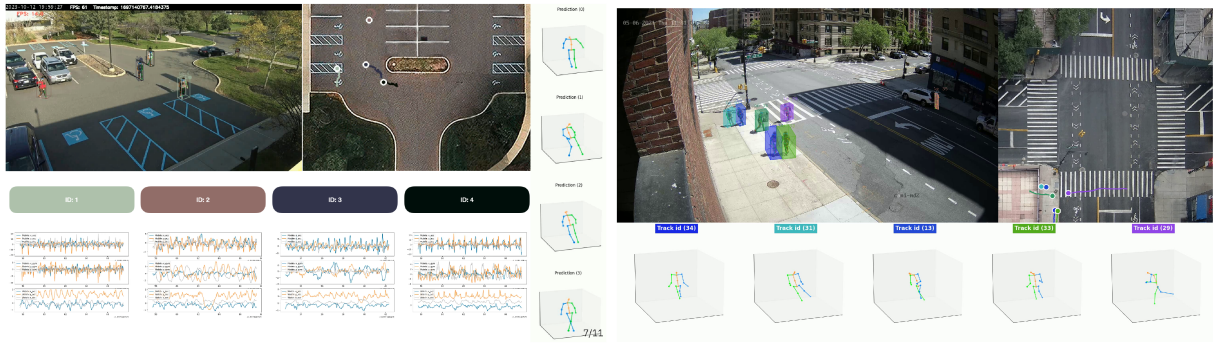


Figure 4: Visualization dashboard displaying live video, IMU streams, 3D bounding boxes, trajectories (camera and BEV views), and 2D/3D pedestrian poses, all processed via low-latency edge computing.

a foundation for scalable smart streetscape application development in real-world scenarios to deliver hyper-local intelligence for safer and more adaptive urban infrastructures.

Acknowledgments

This work was supported by the National Science Foundation (NSF) and Center for Smart Streetscapes (CS3) under NSF Cooperative Agreement No. EEC-2133516, NSF Grant CNS-2148128, NSF Grant CNS-2038984, and corresponding support from the Federal Highway Administration (FHWA).

References

[1] Djallel Eddine Boubiche, Muhammad Imran, Aneela Maqsood, and Muhammad Shoaib. 2019. Mobile crowd sensing–taxonomy, applications, challenges, and solutions. *Computers in Human Behavior* 101 (2019), 352–370.

[2] Andrew T Campbell, Shane B Eisenman, Nicholas D Lane, Emiliano Miluzzo, and Ronald A Peterson. 2006. People-centric urban sensing. In *Proceedings of the 2nd annual international workshop on Wireless internet*. 18–es.

[3] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. 2023. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9686–9696.

[4] NM Cusack, PD Venkatraman, U Raza, and A Faisal. 2024. Smart wearable sensors for health and lifestyle monitoring: commercial and emerging solutions. *ECS sensors plus* 3, 1 (2024), 017001.

[5] Salma Elmalaki. 2021. Fair-iot: Fairness-aware human-in-the-loop reinforcement learning for harnessing human variability in personalized iot. In *Proceedings of the International Conference on Internet-of-Things Design and Implementation*. 119–132.

[6] Insurance Institute for Highway Safety (IIHS). [n. d.]. Pedestrian Fatality Statistics. https://www.iihs.org/topics/fatality-statistics/detail/pedestrians. Accessed: 2024-06-24.

[7] Sachini Herath, Hang Yan, and Yasutaka Furukawa. 2020. Ronin: Robust neural inertial navigation in the wild: Benchmark, evaluations, & new methods. In *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE, 3146–3152.

[8] Chao Huang, Fengli Xu, Yong Li, Xinlei Chen, and Pei Zhang. 2018. Locally differentially private participant recruitment for mobile crowdsourcing. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*. 323–324.

[9] Gaurav Jain, Basel Hindi, Zihao Zhang, Koushik Srinivasula, Mingyu Xie, Mahshid Ghasemi, Daniel Weiner, Sophie Ana Paris, Xin Yi Therese Xu, Michael Malcolm, et al. 2023. StreetNav: Leveraging Street Cameras to Support Precise Outdoor Navigation for Blind Pedestrians. *arXiv preprint arXiv:2310.00491* (2023).

[10] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. 2023. Rtmpose: Real-time multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399* (2023).

[11] Hansi Liu, Abrar Alali, Mohamed Ibrahim, Bryan Bo Cao, Nicholas Meegan, Hongyu Li, Marco Gruteser, Shubham Jain, Kristin Dana, Ashwin Ashok, et al. 2022. Vi-fi: Associating moving subjects across vision and wireless sensors. In *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 208–219.

[12] Ivan Lujic, Vincenzo De Maio, Klaus Pollhammer, Ivan Bodrozic, Josip Lasic, and Ivona Brandic. 2021. Increasing traffic safety with real-time edge analytics and 5g. In *Proceedings of the 4th International Workshop on Edge Systems, Analytics and Networking*. 19–24.

[13] Robert Munro Monarch. 2021. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster.

[14] Huiqian Niu, Qiankun Lu, and Chao Wang. 2018. Color correction based on histogram matching and polynomial regression for image stitching. In *2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC)*. IEEE, 257–261.

[15] João Oliveira, Pedro Teixeira, Pedro Rito, Miguel Luís, Susana Sargento, and Bruno Parreira. 2024. Microservices in edge and cloud computing for safety in intelligent transportation systems. In *NOMS 2024-2024 IEEE Network Operations and Management Symposium*. IEEE, 1–7.

[16] Navid Salami Pargoo, Mahshid Ghasemi, Shuren Xia, Mehmet Kerem Turkcan, Taqiya Ehsan, Chengbo Zang, Yuan Sun, Javad Ghaderi, Gil Zussman, Zoran Kostic, et al. 2024. The Streetscape Application Services Stack (SASS): Towards a Distributed Sensing Architecture for Urban Applications. *arXiv preprint arXiv:2411.19714* (2024).

[17] Dipankar Raychaudhuri, Ivan Seskar, Gil Zussman, Thanasis Korakis, Dan Kilper, Tingjun Chen, Jakub Kolodziejski, Michael Sherman, Zoran Kostic, Xiaoxiong Gu, et al. 2020. Challenge: COSMOS: A city-scale programmable testbed for experimentation with advanced wireless. In *Proc. ACM MobiCom*.

[18] Hannah Ritchie, Veronika Samborska, and Max Roser. 2024. Urbanization. (February 2024). https://ourworldindata.org/urbanization Published on Our World in Data.

[19] Vivek Roy. 2022. Smartphone Localization for Indoor Pedestrian Navigation.

[20] Dori Shapira, Shai Avidan, and Yacov Hel-Or. 2013. Multiple histogram matching. In *2013 IEEE international conference on image processing*. IEEE, 2269–2273.

[21] Jing Shi, Rui Zhang, Yunzhong Liu, and Yanchao Zhang. 2010. Prisense: privacy-preserving data aggregation in people-centric urban sensing systems. In *2010 Proceedings IEEE INFOCOM*. IEEE, 1–9.

[22] Pedro Teixeira, Susana Sargento, Pedro Rito, Miguel Luís, and Francisco Castro. 2023. A sensing, communication and computing approach for vulnerable road users safety. *IEEE Access* 11 (2023), 4914–4930.

[23] UNCTAD. [n. d.]. Total and Urban Population. https://hbs.unctad.org/total-and-urban-population/. Accessed: 2024-06-24.

[24] Dingzhu Wen, Peixi Liu, Guangxu Zhu, Yuanming Shi, Jie Xu, Yonina C Eldar, and Shuguang Cui. 2023. Task-oriented sensing, computation, and communication integration for multi-device edge AI. *IEEE Transactions on Wireless Communications* 23, 3 (2023), 2486–2502.

[25] Tong Wu, Navid Salami Pargoo, and Jorge Ortiz. 2023. Multi-sensor Fusion for In-cabin Vehicular Sensing Applications. In *Proceedings of the 22nd International Conference on Information Processing in Sensor Networks*. 332–333.

[26] Wei-Ying Yi, Kwong-Sak Leung, and Yee Leung. 2017. A modular plug-and-play sensor system for urban air pollution monitoring: Design, implementation and evaluation. *Sensors* 18, 1 (2017), 7.

[27] Daqing Zhang, Dan Wu, Kai Niu, Xuanzhi Wang, Fusang Zhang, Jian Yao, Dajie Jiang, and Fei Qin. 2022. Practical issues and challenges in CSI-based integrated sensing and communication. In *2022 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 836–841.

[28] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. 2023. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15085–15099.