

# StreetNav: Leveraging Street Cameras to Support Precise Outdoor Navigation for Blind Pedestrians

Gaurav Jain  
Columbia University  
New York, NY, USA

Basel Hindi  
Columbia University  
New York, NY, USA

Zihao Zhang  
Columbia University  
New York, NY, USA

Koushik Srinivasula  
Columbia University  
New York, NY, USA

Mingyu Xie  
Columbia University  
New York, NY, USA

Mahshid Ghasemi  
Columbia University  
New York, NY, USA

Daniel Weiner\*  
Lehman College  
New York, NY, USA

Sophie Ana Paris\*  
New York University  
New York, NY, USA

Xin Yi Therese Xu\*  
Pomona College  
Claremont, CA, USA

Michael Malcolm\*  
SUNY at Albany  
Albany, NY, USA

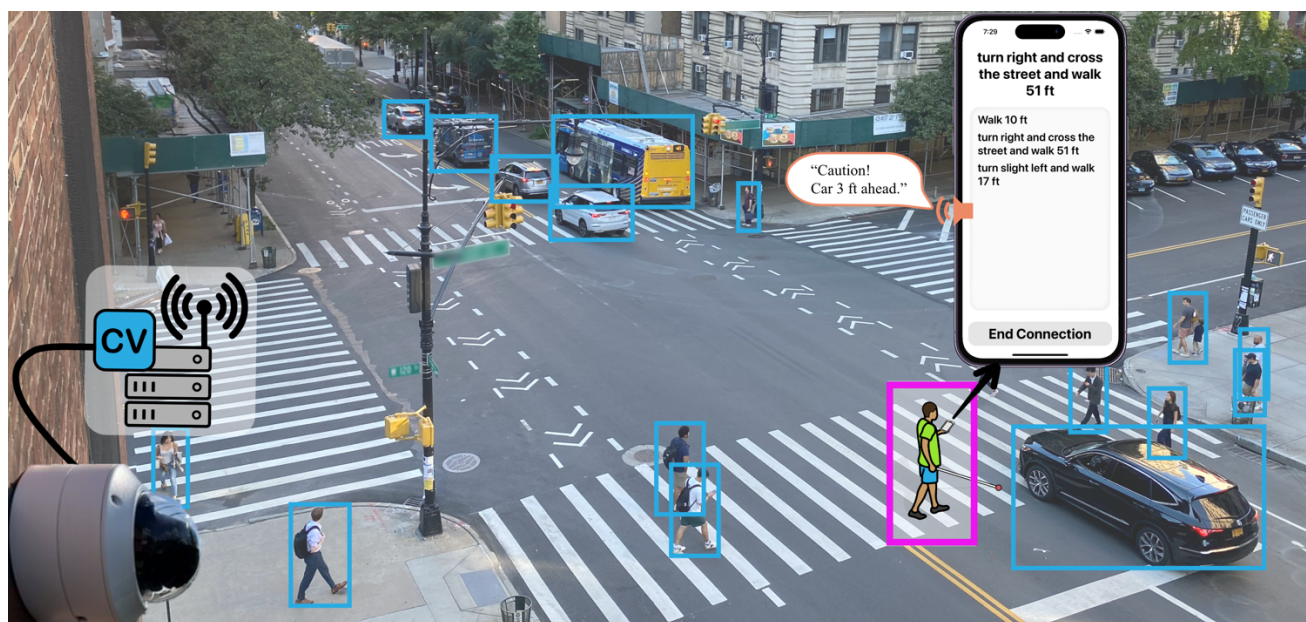
Mehmet Turkcan  
Columbia University  
New York, NY, USA

Javad Ghaderi  
Columbia University  
New York, NY, USA

Zoran Kostic  
Columbia University  
New York, NY, USA

Gil Zussman  
Columbia University  
New York, NY, USA

Brian A. Smith  
Columbia University  
New York, NY, USA



**Figure 1:** StreetNav is a system that explores the concept of repurposing existing street cameras to support precise outdoor navigation for blind and low-vision (BLV) pedestrians. It comprises two components: (i) a computer vision (CV) pipeline, and (ii) a companion smartphone app. The computer vision pipeline processes the street camera's video feeds and delivers real-time navigation feedback via the app. StreetNav offers precise turn-by-turn directions to destinations while also providing real-time, scene-aware assistance to alert them of nearby obstacles and facilitate safe street crossings.

\*Work done during internship at Columbia University.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
UIST '24, October 13–16, 2024, Pittsburgh, PA, USA

## ABSTRACT

Blind and low-vision (BLV) people rely on GPS-based systems for outdoor navigation. GPS's inaccuracy, however, causes them to veer off track, run into obstacles, and struggle to reach precise

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 979-8-4007-0628-8/24/10  
<https://doi.org/10.1145/3654777.3676333>

destinations. While prior work has made precise navigation possible indoors via hardware installations, enabling this outdoors remains a challenge. Interestingly, many outdoor environments are already instrumented with hardware such as street cameras. In this work, we explore the idea of repurposing *existing* street cameras for outdoor navigation. Our community-driven approach considers both technical and sociotechnical concerns through engagements with various stakeholders: BLV users, residents, business owners, and Community Board leadership. The resulting system, StreetNav, processes a camera's video feed using computer vision and gives BLV pedestrians real-time navigation assistance. Our evaluations show that StreetNav guides users more precisely than GPS, but its technical performance is sensitive to environmental occlusions and distance from the camera. We discuss future implications for deploying such systems at scale.

## CCS CONCEPTS

• **Human-centered computing** → **Accessibility systems and tools.**

## KEYWORDS

Visual impairments, outdoor navigation, street camera, computer vision

### ACM Reference Format:

Gaurav Jain, Basel Hindi, Zihao Zhang, Koushik Srinivasula, Mingyu Xie, Mahshid Ghasemi, Daniel Weiner, Sophie Ana Paris, Xin Yi Therese Xu, Michael Malcolm, Mehmet Turkcan, Javad Ghaderi, Zoran Kostic, Gil Zussman, and Brian A. Smith. 2024. StreetNav: Leveraging Street Cameras to Support Precise Outdoor Navigation for Blind Pedestrians. In *The 37th Annual ACM Symposium on User Interface Software and Technology (UIST '24)*, October 13–16, 2024, Pittsburgh, PA, USA. ACM, New York, NY, USA, 21 pages. <https://doi.org/10.1145/3654777.3676333>

## 1 INTRODUCTION

Outdoor navigation in unfamiliar environments is a major challenge for blind and low-vision (BLV) people. Among the many navigation systems that have been developed to assist BLV people outdoors, GPS-based systems are the most popular [30, 33, 44, 63, 68]. These systems, such as BlindSquare [44] and Microsoft Soundscape [30], guide users to a destination and notify them of surrounding points of interest (POIs). Despite GPS's undeniable impact in making outdoor environments navigable, its imprecision is a major limitation [61]. GPS precision can range from 5 meters at best to over tens of meters in urban areas with buildings and trees [23, 46, 69]. This imprecision causes BLV people to veer off track [53], run into unexpected obstacles [8, 54, 56], and struggle to reach precise destinations [61] when navigating outdoors.

Prior work on indoor navigation, on the contrary, has made precise navigation assistance possible for BLV people [2, 21, 36, 48, 62]. Most approaches do so by installing a dense network of Bluetooth [2] or WiFi [21] beacons. However, extending this approach for outdoor navigation is not feasible due to the vast scale and complex nature of outdoor spaces. Interestingly, many outdoor environments of interest, such as urban districts and downtown areas, are already instrumented with hardware that has the potential to help, including street cameras, traffic sensors, and other urban infrastructure components.

Street cameras, in particular, have the potential to support BLV pedestrians' outdoor navigation. The video feed from these cameras could be processed using computer vision to track BLV pedestrians and perceive their visual environment with greater precision and fidelity compared to GPS-based systems. The profound potential of street cameras for assistive technology is accompanied by significant challenges and concerns — both technical and sociotechnical.

On the technical front, there is a lack of understanding regarding the precise capabilities of street cameras to track BLV pedestrians and how camera-based systems should be designed to effectively support BLV people's outdoor navigation. Sociotechnically, a major concern revolves around privacy due to cameras' capability to collect pervasive data, not only affecting BLV users but also other pedestrians and vehicles in the vicinity [17]. Moreover, street cameras are often deployed by governments to force surveillance [4, 12, 20, 38, 42], which exacerbates people's privacy concerns. Limited work has been done to explore how camera-based technologies can respect people's privacy concerns and directly serve their interests, rather than solely serving government-defined purposes [17, 28, 41, 74].

In this work, we take a community-driven approach to explore the concept of leveraging street cameras to support outdoor navigation for blind pedestrians. To this end, we engage with various stakeholders including BLV users, local residents, local business owners, and Community Board leadership. We aim to address both the technical and sociotechnical aspects of this concept through the following research questions:

- RQ1.** What are stakeholders' privacy concerns toward camera-based assistive technology, and how might they be respected?
- RQ2.** How might a street camera-based navigation assistance system be designed?
- RQ3.** To what extent do street camera-based systems support BLV people's outdoor navigation?

To answer RQ1, we interviewed various stakeholders, including two BLV users, two local residents, a local business owner, and a Community Board leader. We discovered stakeholders' differing perspectives on privacy concerns towards camera-based assistive technology. All stakeholders expressed that repurposing *existing* cameras to help BLV people, rather than installing *new* cameras, significantly alleviates their privacy concerns. Participants also shared that regulating data storage, anonymization, and access policies could further enhance their sense of comfort around privacy.

To answer RQ2, we developed *StreetNav*, a system that leverages a street camera to support precise outdoor navigation for BLV pedestrians. StreetNav's design is informed by BLV people's outdoor navigation challenges (Section 3) and by various stakeholders' privacy concerns toward camera-based assistive technology (Section 4). As Figure 1 illustrates, StreetNav comprises two components: (i) *a computer vision pipeline*, and (ii) *a companion smartphone app*. The computer vision pipeline processes the street camera's video feed and delivers real-time navigation assistance to BLV pedestrians via the smartphone app. StreetNav offers precise turn-by-turn directions to destinations while also providing real-time, scene-aware assistance to prevent users from veering off course, alert them of nearby obstacles, and facilitate safe street crossings.

StreetNav supports BLV pedestrians’ outdoor navigation by repurposing an *existing* street camera. Through StreetNav, we explore the feasibility of street camera-based systems at a single street intersection as a first step. We chose to use a camera from the NSF PAWR COSMOS testbed [60, 72] because it is available to researchers after an approval process and IRB review. We considered other publicly available testbeds, such as Mobintel [45] and DataCity SMTG [14], but chose COSMOS due to its location in a major city (New York) with high pedestrian and vehicle traffic. Anonymized video samples from the COSMOS cameras, including the one used in this work, can be found online [13].

To answer RQ3, we conducted both a user evaluation and a technical evaluation of StreetNav. Our user evaluation involved eight BLV pedestrians who navigated routes with both StreetNav and BlindSquare [44], a popular GPS-based navigation app specially designed for BLV people. Our findings reveal that StreetNav offers significantly greater precision in guiding pedestrians compared to BlindSquare. Specifically, StreetNav guided participants to within an average of 2.9 times closer to their destination and reduced veering off course by over 53% when compared to BlindSquare. This substantial improvement was reflected in a forced ranking, where all participants unanimously preferred StreetNav over BlindSquare.

Despite an improved user experience, StreetNav’s technical evaluation exposes certain limitations. We found that although StreetNav tracks pedestrians with an 82% precision and 65% recall at 0.5 IOU threshold, the accuracy drops significantly as the pedestrian’s distance from the camera increases. The false negative rates goes up from 1% at a distance of 5 meters to 74% at a distance of 40 meters from the camera. Additionally, StreetNav’s performance is sensitive to occlusions and distance from camera. We discuss future implications of our findings in the context of deploying street camera-based navigation systems at scale.

In summary, we contribute (1) a study of various stakeholders’ privacy concerns toward camera-based assistive technology, (2) the StreetNav system through which we explore the concept of repurposing *existing* street cameras for precise outdoor navigation assistance, and (3) both a user and technical evaluation of StreetNav.

## 2 RELATED WORK

Our work builds on the following three main research threads: (i) outdoor navigation approaches, (ii) overhead camera-based robot navigation, and (iii) indoor navigation approaches.

**Outdoor Navigation Approaches.** Existing approaches for outdoor navigation primarily rely on GPS-based navigation systems for guiding users to the destination and providing information about nearby POIs [30, 33, 44, 63, 68]. BlindSquare[44], for instance, utilizes the smartphone’s GPS signal to determine the user’s location and then provides the direction and distance to the destination, gathered from Foursquare and Open Street Map. The GPS signal, however, offers poor precision with localization errors as big as tens of meters [2, 23, 46, 73]. The accuracy is lower in densely populated cities [70], which is even more concerning given that a disproportionately high percentage of BLV people live in cities [27]. Despite GPS-based systems’ undeniable impact on helping BLV people in outdoor navigation, their low precision and inability to provide real-time support for avoiding obstacles and veering off the path

limits their usability as a standalone navigation solution. Our work attempts to investigate street cameras’ potential as an alternative solution for providing precise and real-time navigation assistance.

Another approach for outdoor navigation has explored developing personalized, purpose-built, assistive devices that support BLV people with scene-aware aspects of outdoor navigation, such as crossing streets [26, 39, 66], recording routes [73], and avoiding obstacles [16, 18, 34, 40, 59, 71]. While these solutions address some of the precise and real-time aspects of BLV people’s outdoor navigation, support for point-to-point navigation is missing. Consequently, they do not offer a comprehensive, all-in-one solution for outdoor navigation. Furthermore, these systems place the burden of purchasing devices onto the BLV users. Our work, by contrast, explores the possibility of using existing street cameras to provide a comprehensive solution for outdoor navigation. We investigate repurposing existing hardware in outdoor environments to support accessibility applications, thus directly imbuing accessibility within the city infrastructure at no additional cost to the BLV user.

**Overhead Camera-based Robot Navigation.** A parallel research space to street cameras for blind navigation is robot navigation using overhead cameras. One common subspace within this field is sensor fusion for improved mapping. Research in this space focuses on fusing information between sighted “guide” robots and overhead cameras [11], fusing multiple camera views for improved tracking [11, 52, 55], and improving homography for robust mapping, independent of camera viewing angle [64, 65]. Another challenge tackled within this space is robot path planning. Research in this space aims to improve path planning algorithms [11, 52, 65], assign navigational tasks to robot assistants [11, 52], and address the balance between obstacle avoidance and path following [11, 65]. While prior work on robot navigation using fixed cameras explores the research space of automating “blind” robot navigation, our work explores how fixed cameras, specifically street cameras, could be repurposed to support navigation for blind pedestrians. Our preliminary work [31] explores an initial system concept that considers street cameras for blind navigation. This concept was not evaluated, however, nor were community issues considered. In this work, we perform both a technical and user evaluation to holistically explore the concept of leveraging street cameras for blind navigation. Moreover, we take a community-driven approach to consider both technical and sociotechnical challenges in developing street camera-based navigation systems, engaging with not only BLV users but also various stakeholders.

**Indoor Navigation Approaches.** Prior work in indoor navigation assistance has made significant progress through the utilization of various localization technologies, which usually relies on hardware like WiFi or Bluetooth beacons [2, 21, 36, 48, 62]. These solutions have proven highly effective within indoor environments. NavCog3 [2], for example, excels in indoor navigation by employing Bluetooth beacons for precise turn-by-turn guidance. Nakajima and Haruyama [48] exploit the use of visible lights communication technology, utilizing LED lights and a geomagnetic correction method to localize BLV users. However, extending these approaches to support outdoor navigation is not feasible. This is particularly evident when considering the substantial effort in hardware setup that these systems typically require, making them ill-suited for the

larger, unstructured outdoor environment. Furthermore, most of these methods lack the capability to assist with obstacle avoidance and to prevent users from veering off course — both of which are less severe issues indoors compared to outdoors [53]. Our exploration of using existing street cameras is better suited to address the largely unaddressed challenges of outdoor navigation. This approach has the potential to offer precise localization without requiring dense hardware installations. It can harness existing street cameras for locating a pedestrian’s position. Additionally, it holds the potential to tackle the distinctive challenges posed by the unstructured nature of outdoor environments, including real-time obstacle avoidance and safe street crossing.

### 3 BLV PEOPLE’S CHALLENGES IN OUTDOOR NAVIGATION USING GPS-BASED SYSTEMS

We conducted semi-structured interviews with six BLV participants to identify challenges that they face when navigating outdoors using GPS-based systems. Our interviews found three major challenges, C1: following routing instructions through complex environment layouts, C2: avoiding unexpected obstacles while using GPS-based systems, and C3: crossing streets safely. While these challenges are well-documented within existing literature [8, 53, 54, 56, 61], our findings highlight areas that could be prioritized for resolution through the implementation of a street camera-based navigation system. Appendix A provides additional detail on participant demographics, interview procedure, and interview findings.

### 4 STAKEHOLDERS’ PRIVACY CONCERNS TOWARD CAMERA-BASED SYSTEMS

We conducted five semi-structured interviews with various stakeholders from Harlem, New York City, where the COSMOS testbed is located. Harlem is a diverse community within a major city that has become sensitive to government surveillance and overreach. The interviews were with two BLV users (B1, B2), two local residents (R1, R2), a local business owner (O1), and a Community Board leader (CB1). Our objective was to understand stakeholders’ privacy concerns regarding camera-based assistive technology and explore ways to address these concerns (RQ1).

#### 4.1 Methods

**Participants.** Table 3 (Appendix B) reports participant demographics. Each interview lasted for about 45-60 minutes, except for the interview with the Community Board leader that lasted for 15 minutes. Three interviews (B1, B2, R1) were conducted online over Zoom, two (O1, R1) were conducted in person, and one (CB1) was conducted over phone. All participants, except for CB1, were compensated \$50 for their participation in this IRB-approved study. CB1 refused to accept the compensation.

**Procedure.** We began by giving participants a short presentation describing an initial system concept. The presentation illustrated how street cameras could capture street intersections, use computer vision to track pedestrians and vehicles, and deliver navigation instructions to BLV users via smartphones. We verbally described visuals to BLV participants during the study. We then asked participants questions about their perceived benefits and concerns,

preferences around data collection and use scenarios that may raise privacy concerns: e.g., *Does it matter to you who has access to the camera feed?* During interview with the Community Board leader, we inquired about the feasibility of such a system: e.g., *How feasible would it be to use street cameras for assistive technology purposes?* We concluded interviews by discussing strategies for how such systems might respect their privacy concerns.

**Interview Analysis.** We used thematic analysis [10] to analyze the interviews, similar to our methodology described in Section 3. This analysis involved three researchers independently generating initial sets of codes, which were then collaboratively iterated to identify emerging themes.

#### 4.2 Findings: Privacy Concerns

Our participants, irrespective of their stakeholder category, held differing perspectives on privacy concerns toward camera-based assistive technology. While some had no privacy concerns whatsoever, others felt uncomfortable with the concept of a camera monitoring them. When asked if there was anything that could satisfy their concerns, concerned participants identified two strategies: (i) regulating data storage, anonymization, and access policies; and (ii) repurposing *existing* cameras rather than installing *new* cameras to assist BLV people. The following sections detail our findings on stakeholders’ differing viewpoints on privacy and strategies that this assistive technology could employ to respect those viewpoints.

##### **Stakeholders’ differing perspectives on privacy concerns.**

Nearly half of the participants (B1, R2, O1) expressed no concerns about being recorded by the camera. In fact, they highlighted the added benefits of street cameras in enhancing public safety, particularly aiding in crime investigation. These participants expressed the willingness to sacrifice some privacy in exchange for societal benefits such as accessibility and public safety. This finding aligns with earlier findings by Profita et al. [57]. Additionally, B1 pointed out that complete privacy should not be expected in public spaces: *“If you’re on a public street, you pretty much could expect for anyone to see you at any time. So it’s no more invasive than anything else on a public street. A public street is pretty much fair game for anybody.”*

In contrast, other participants (B2, R1) expressed discomfort with cameras’ capability not only to track people’s movements but also to “*know what [they] look like*” (B2). R1 compared a camera’s presence to an “*overarching shadow that’s always looking over [and] monitoring their everyday moves.*” These participants voiced concerns against the use of such cameras for public safety purposes. They feared that the ability to determine individuals’ identities from the video feed could result in the targeting of marginalized groups such as people of color (R1) and BLV individuals (B2). As B2 stated: *“The fact that I’m being surveilled even more as a blind person, and knowing that police disproportionately target the disabled whenever things are going wrong, that just makes me feel even less safe.”*

##### **Regulating data storage, anonymization, and access policies.**

We inquired about participants’ preferences regarding the collection and storage of the video feed. Those without privacy concerns (B1, R2, O1) expressed indifference regarding the duration and form (e.g., anonymized vs. raw footage) of video footage storage, asserting they had “*nothing to hide*” (O1). B1 elaborated: *“It really doesn’t*



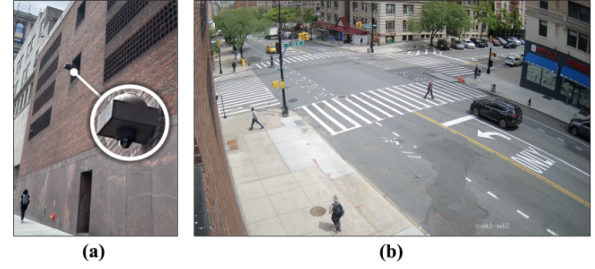
matter to me. I'm not the person who is going to commit the crime, so I don't care if they keep [the data] forever." Conversely, participants with privacy concerns (B2, R1) expressed discomfort with any long-term data storage if such storage was not necessary for the functionality of the assistive technology. They proposed anonymizing the video footage by techniques such as blurring faces (R1) or representing individuals with "dots" (B2) or avatars akin to those used in GPS-based applications. However, complete anonymization would negate safety benefits desired by some individuals. A compromise was reached in favor of limited storage duration, up to a week, alongside clear guidelines regarding access and legal use of the video footage. Most participants expressed greater trust in government entities than in corporations to manage these cameras, citing concerns about potential data exploitation by the latter.

**Repurposing existing cameras rather than install new cameras.** During the interview, R2 highlighted the ubiquity of cameras in urban areas: "It's New York, there's going to be a camera every other block. There's no way that these cameras can't pick you up." We pursued this observation with other participants and discovered that assisting BLV pedestrians with *existing* cameras rather than install *new* cameras significantly alleviated their privacy concerns. B2 affirmed this, saying, "I would be okay with that because, you know, it's a dual purpose thing. [The Dept. of Transportation] is already putting the speeding cameras there, so at least it does something nice for people while the camera is in place." The business owner, O1, consented to lending the cameras at their restaurant's entrance, overlooking the street, under two conditions: (i) it should be used solely and responsibly to assist people, and (ii) it should not record any views inside their restaurant.

We consulted with the Community Board leader (CB1) to understand the feasibility of repurposing existing street cameras. CB1 emphasized the need for collaboration among various government entities to effectively enable this technology. This collaboration would not only involve granting access to the cameras but also ensuring that they possess the necessary capabilities to support this application. CB1 identified several key institutions that could play vital roles in this effort: the Department of Transportation, responsible for providing camera access and relevant technical support; the Department of Buildings or the Metropolitan Transportation Authority (MTA), tasked with granting camera access and permissions to house any required computational resources; and the National Security Agency (NSA), tasked with ensuring that camera access maintains security protocols. Additionally, CB1 highlighted the importance of implementing processes to monitor the impact of this technology on local communities. For instance, public outreach initiatives would help the public understand the purpose of the technology, ensuring transparency and accountability throughout the deployment process.

## 5 THE STREETNAV SYSTEM

StreetNav is a system that explores the concept of repurposing *existing* street cameras to support outdoor navigation for BLV pedestrians (RQ2). The following sections describe StreetNav's design rationale (Section 5.1), the computer vision pipeline (Section 5.2), and the smartphone app's user interface (Section 5.3).



**Figure 2: Street camera used for StreetNav's development and evaluation. The camera is (a) mounted on the building's second floor and (b) faces a four-way intersection.**

### 5.1 StreetNav: Design Rationale

Our design and development of StreetNav considers prior work on navigation assistance, functions of traditional mobility aids, and insights gathered from our interviews with BLV people (Section 3) and with various stakeholders (Section 4)

To address challenges that BLV people face when navigating outdoors using existing GPS-based systems, StreetNav provides users precise turn-by-turn navigation instructions to destinations and prevents veering off track (C1); gain awareness of nearby obstacles (C2); and assist in crossing streets safely (C3). StreetNav enables these affordances through its two main components: (i) *computer vision pipeline*, and (ii) *companion smartphone app*. The computer vision pipeline processes the street camera's video feeds to give BLV pedestrians real-time navigation feedback via the app.

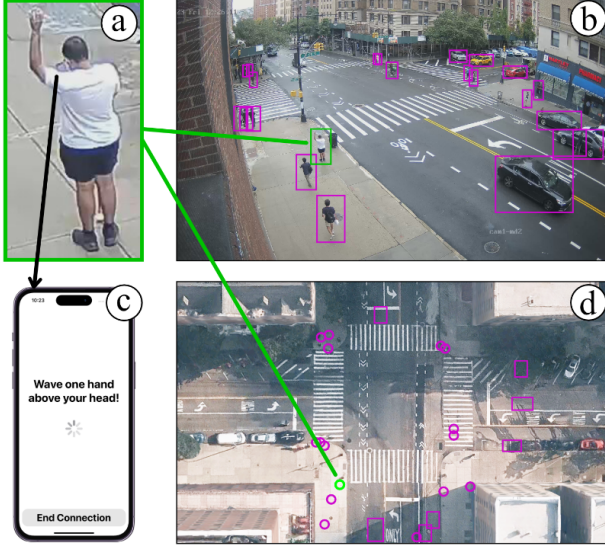
To ensure that StreetNav respects stakeholders' privacy concerns, we explore how an *existing* camera may be repurposed, rather than installing a *new* camera, to support BLV people's outdoor navigation (Section 4). For this reason, we chose a camera that faces a four-way street intersection—the most common type of intersection—and is mounted on a building's second floor, offering a typical street-level view of the intersection. Figure 2 shows the street camera and its view of the street intersection. StreetNav eliminates the requirement of *storing* any video data by processing the camera feed in real-time to generate navigation instructions.

Appendix C describes StreetNav's technical setup which enables the real-time navigation assistance.

### 5.2 StreetNav: Computer Vision Pipeline

StreetNav's computer vision pipeline processes the street camera's video feed in real time to facilitate navigation assistance. It consists of four components: (i) *localizing and tracking the user*: locating user's position on the environment's map; (ii) *planning routes*: generating turn-by-turn navigation instructions from user's current position to destinations; (iii) *identifying obstacles*: predicting potential collisions with other pedestrians, vehicles, and objects (e.g., trash can, pole); and (iv) *recognizing pedestrian signals*: determining when it is safe for pedestrians to cross (walk vs. wait) and calculating the duration of each cycle. Next, we describe the computer vision pipeline's four components in detail.

**Localizing and tracking the user.** To offer precise navigation assistance, a system must first determine the user's position from



**Figure 3: Gesture-based localization for determining a user's position on the map.** (a) A study participant (P1) is (c) prompted to wave one hand above their head, enabling the computer vision pipeline to distinguish them from other pedestrians in (b) the camera feed view and (d) the map.

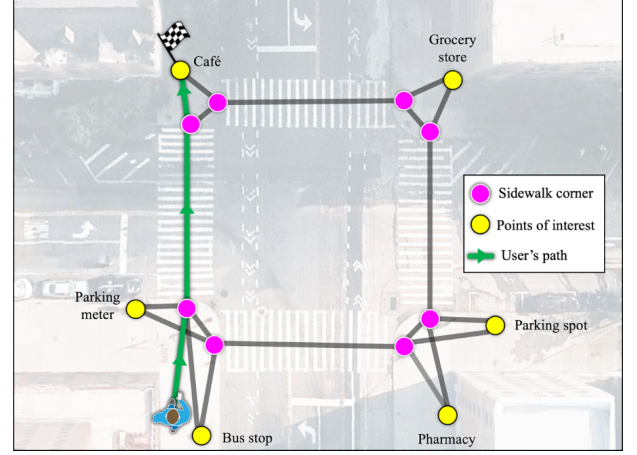
the camera view and then project it onto the environment's map. Figure 3d shows the map representation we used, which is a snapshot from Apple Maps' [5] satellite view of the intersection where the camera is deployed.

StreetNav tracks pedestrians from the camera's video feed using Nvidia's DCF-based multi-object tracker [49] and the YOLOv8 object detector [67]. The tracker detects all pedestrians and assigns them a unique ID. However, the system needs a way to differentiate between the BLV user and other pedestrians.

Figure 3 shows the *gesture-based localization* approach we introduced to address this issue. To connect with the system, BLV pedestrians must wave one hand above their head for 2–3 seconds (Figure 3a), enabling the system to determine the BLV pedestrian's unique tracker ID. We chose this gesture after discussions with several BLV individuals, including our BLV co-author, and most agreed that this single-handed action was both convenient and socially acceptable to them. Moreover, over-the-head gestures such as waving a hand can also be detected when users are not directly facing the street camera.

We implement hand gesture-based localization by first creating image crops of all detected pedestrians, then classifying them as 'waving' or 'walking' pedestrians using CLIP [58]. CLIP classifies each pedestrian by computing visual similarity between the pedestrian's image crop and two language prompts: 'person walking' and 'person waving hand.' We experimentally fine-tuned the confidence thresholds and these language prompts.

We estimate the user's feet position to be the mid-point of bounding box's bottom edge. Finally, we transform the user's feet position from the street camera view (Figure 3b) to the map (Figure 3d) using a simple feed-forward neural network trained on data that we



**Figure 4: StreetNav's internal graph representation for route planning.** The user's current position is added dynamically as a start node to the graph upon choosing a destination. The shortest path, highlighted in green, is then calculated as per this graph representation.

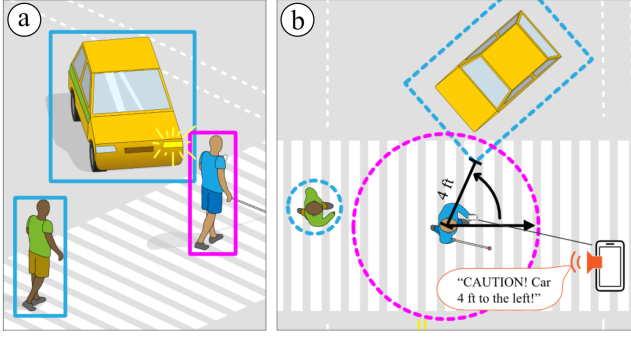
manually annotated. The network takes as input the 2D pixel coordinate from the street camera view and outputs the corresponding 2D coordinate on the map.

**Planning routes.** To plan routes, a street camera-based systems require a map of the environment, internally represented as a graph with waypoints and connections between them. For StreetNav, one of the researchers manually annotated a satellite view image of the street intersection to create this graph, a process that took roughly 10 minutes. This process could be automated by integrating with OpenStreetMap [51] map data in the future.

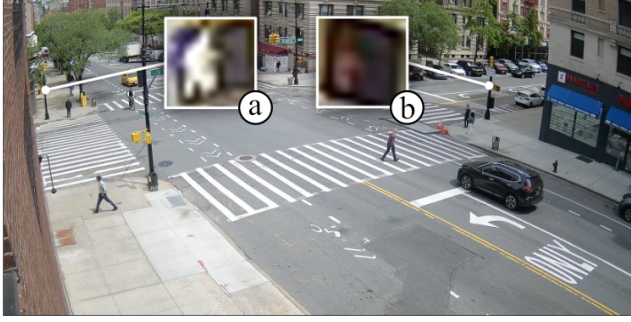
Figure 4 shows the internal graph structure that StreetNav uses for planning routes. Similar representations have been used in prior work on indoor navigation systems [2, 25, 62]. In the graph, nodes correspond to POIs and sidewalk corners, whereas edges correspond to walkable paths. Once the user chooses a destination from the POIs, StreetNav adds the user's current position as a start node to this graph representation and computes the shortest path to the chosen POI using A\* algorithm [15].

**Identifying obstacles.** StreetNav provides users with information about an obstacle's category and relative location. This gives users context on the size, shape, and location of an obstacle; enabling them to confidently apply their mobility skills to go around unexpected obstacles.

Figure 5 illustrates how the system identifies obstacles in the user's vicinity. StreetNav's multi-object tracker is used to track other objects and pedestrians. Examples of other objects include cars, bicycles, poles, and trash cans. The computer vision pipeline then projects the detected objects' positions onto the map. To identify obstacles in the BLV user's vicinity, StreetNav computes the distance and angle between the user and other detected objects with respect to the map (Figure 5b). Any object (or pedestrian) within a fixed radial distance from the BLV user is flagged as an obstacle. Through a series of experiments with our BLV co-author, we found



**Figure 5: Identifying obstacles in the user's vicinity.** (a) A vehicle turning left yields to the BLV pedestrian (detected in purple) crossing the street. (b) StreetNav identifies the obstacles' category and relative location on the map to provide real-time feedback via the app.



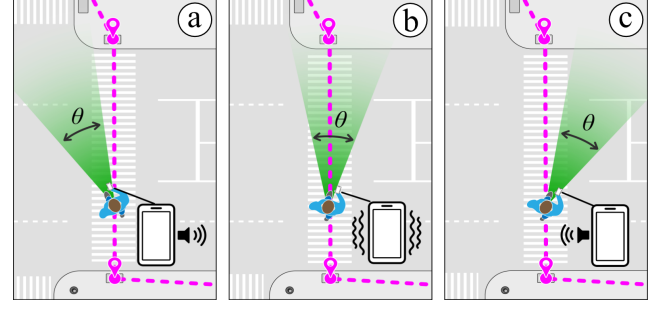
**Figure 6: Recognizing pedestrian signal states.** StreetNav compares the number of white and red pixels in the signal crops to determine its state: (a) *walk* vs. (b) *wait*.

that a 4 foot radius works best for StreetNav to provide users with awareness of obstacles in a timely manner.

**Recognizing pedestrian signals.** To determine the pedestrian signals' state (i.e., *walk* vs. *wait*), we leverage the fact that walk signals are always white and wait signals are always red in color. A street camera-based system would first need to detect pedestrian signals from the camera feed before detecting its state. For StreetNav's implementation, one of the researchers manually annotated the pedestrian signals' screens in the camera feed. Future iterations could scale this process by automatically detecting signals by training custom object detectors.

Figure 6 shows pedestrian signals in the camera's video feed. StreetNav applies pixel-thresholding onto the pedestrian signal crops to filter all white and red pixels. Then, it compares the number of white and red pixels to determine signal state: *walk* (Figure 6a) vs. *wait* (Figure 6b). We experimentally fine-tuned the thresholds to identify the signal state.

Our formative interviews revealed that BLV pedestrians struggle with pacing themselves while crossing streets (C3). To assist them, StreetNav informs users of the remaining crossing time. Its computer vision pipeline tracks signal cycle durations and maintains



**Figure 7: Audiohaptic cues for preventing users from veering off track.** Sample user trajectories showing feedback when users (a) veer to the left, (b) do not veer, and (c) veer to the right. When the user's heading coincides with the route to the destination, within a tolerance angle  $\theta$  (highlighted in green), users receive (b) subtle haptic vibrations to reinforce them. When they veer off the route, outside the tolerance angle  $\theta$ , they hear spatialized beeping sounds that are rendered from the (a) right speaker when veering left, and from the (c) left speaker when veering right.

a timer that records signal state changes. By observing full cycles, StreetNav accurately monitors signal states and timings. Periodic timer updates ensure adaptability to changes in signal durations due to traffic management.

### 5.3 StreetNav App: User Interface

The StreetNav iOS app interacts with the computer vision pipeline to allow BLV pedestrians to choose a destination and receive real-time navigation feedback that guides them to it. BLV users first initiate a connection request through the app, which activates the gesture-based localization (Section 5.2) in the computer vision pipeline. The app prompts the user to wave one hand over their head (Figure 3b), enabling the system to begin tracking their precise location on the map (Figure 3d). BLV users can then select a destination from nearby POIs and begin receiving navigation feedback through the app.

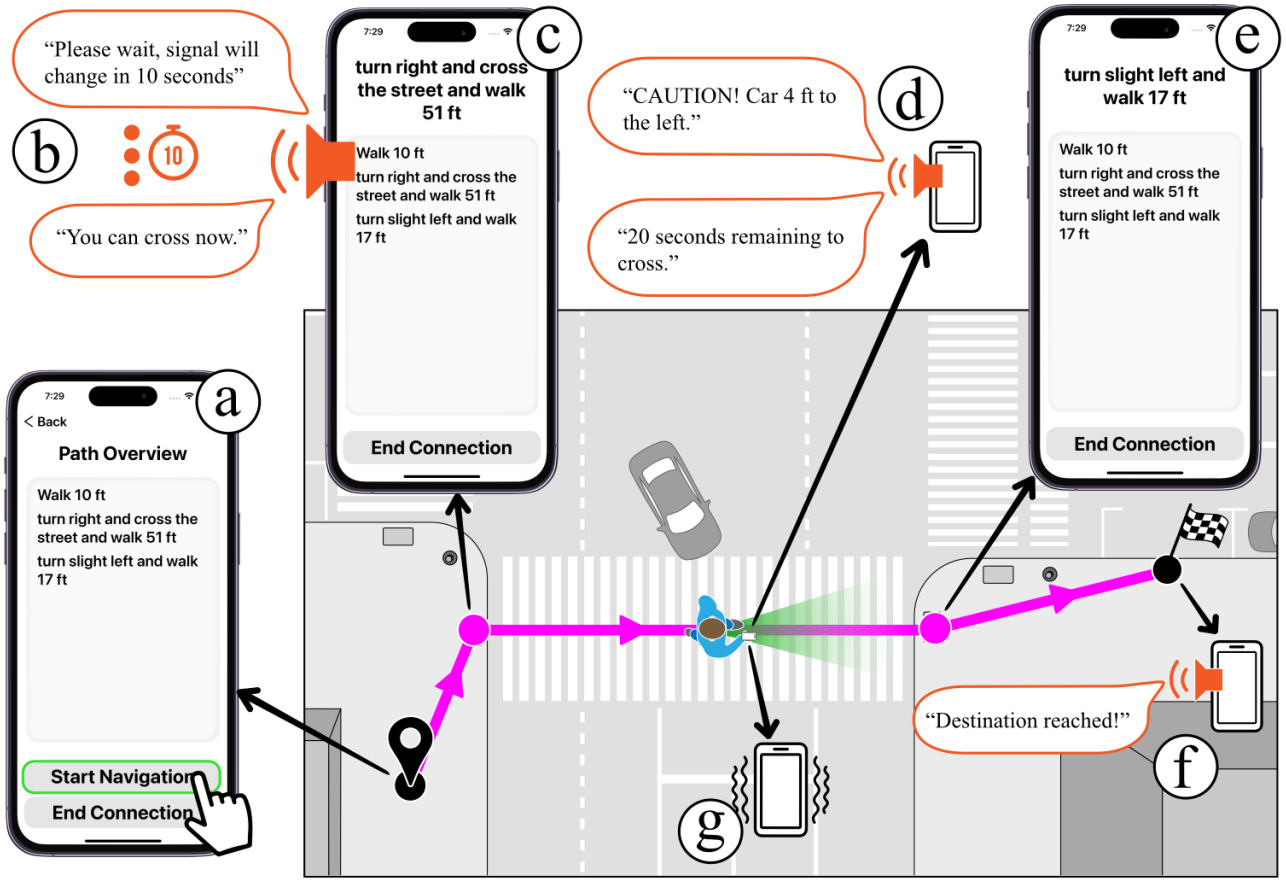
Figure 8 shows the StreetNav app's user interface, which uses audiohaptic cues for (i) providing routing instructions, (ii) preventing veering off track, (iii) notifying about nearby obstacles, and (iv) assisting with crossing streets. Upon reaching the destination, the app confirms their arrival. The following sections describe the app's interface in detail.

**Providing routing instructions.** The app provides routing instructions to users by offering a route overview before they start walking, as shown in Figure 8a. This helps users prepare for their journey [1, 24, 32]. During navigation, the app announces instructions based on the user's location and provides continuous audiohaptic feedback to guide them.

Figure 8b–f show how the app dynamically updates instructions based on the user's location. Users can access the path overview and current instructions on demand via VoiceOver [7].

Figure 7 illustrates the app's audiohaptic feedback. Based on the user's position, heading, and destination, StreetNav computes





**Figure 8: The StreetNav App’s user interface.** It provides routing instructions to their destination via (a) a path overview and (c, e) real-time feedback that updates their current instruction based on their location. Upon reaching a sidewalk, (b) the app informs the user about when it is safe to cross and (d) how much remains for them to cross over. It also (d) notifies the user of a nearby obstacle’s category and relative location to help them avoid it. The app (f) confirms the user’s arrival at the destination. Throughout the journey, the app provides (g) continuous audiohaptic feedback to prevent users from veering off track.

the direction and extent of veering. We initially used the Kalman filter to predict the user’s heading based on their trajectory, but this proved inaccurate due to noisy tracking data. Instead, we used the smartphone’s compass, offset by a fixed value to align its zero with the map’s horizontal direction, allowing us to perform all heading computations relative to the map’s frame of reference.

For directional guidance, we used stereo sound: beeping from the right speaker when users veer left (Figure 7a) and from the left speaker when users veer right (Figure 7c). The frequency of beeps increases with the extent of veering, allowing users to navigate effectively without headphones. To prevent overwhelming users with continuous audio feedback, a tolerance angle ( $\theta$ ) of 50 degrees was introduced. Within this angle, subtle haptic vibrations guide users in the correct direction, while beeping sounds indicate veering, balancing audio as negative reinforcement and haptic feedback as positive reinforcement.

**Notifying about nearby obstacles.** Figure 8d shows how StreetNav alerts the user of obstacles nearby. The app announces the obstacle’s category, distance, and relative location. For example,

when a car approaches the user, the app announces: “*Caution! Car, 4 ft. to the left.*” Similar to veering feedback, the relative location is computed using both the computer vision pipeline’s outputs and the smartphone’s compass reading.

We tried feedback formats with varying granularity to convey the obstacle’s relative location. First, we experimented with *clock-faced directions*: “*Car, 4 ft. at 1 o’clock.*” Clock-faced directions are commonly used in many GPS-based systems such as BlindSquare to convey directions. We learned from pilot evaluations with our BLV co-author that this feedback format was too fine-grained, as it took them a few seconds to decode the obstacle’s location. This does not fare well with moving obstacles, such as pedestrians, that may have already passed the user before they are able to decode the location. Moreover, StreetNav’s goal with obstacle awareness is to give users a quick idea that something is nearby them, which they can then use to circumnavigate via their mobility skills. To address this, we tried the more coarse format with just four directions: left, right, front, and back. This was found to give users a quick intimation, compared to the clock-faced directions.

**Assisting with crossing streets.** The StreetNav app helps users cross streets by informing them *when* to cross and how much time remains before the signal changes.

Figure 8b and Figure 8d illustrate the feedback. Upon reaching a sidewalk corner, the app checks for the signal state recognized by the computer vision pipeline. If the signal is ‘wait’ when the user arrives, the app informs the user to wait along with the time remaining before the signal changes. If the signal is ‘walk’ when the user arrives, the app informs the user to begin crossing only if the time remaining is sufficient for crossing. For the intersection used in our user studies, this was experimentally found to be 15 seconds. Otherwise, the user is advised to wait for the next cycle. Once the user begins crossing on the ‘walk’ signal, the app announces the time remaining for them to cross over. This feedback is repeated at fixed intervals until the user reaches the other sidewalk corner. We experimentally fine-tuned this interval with feedback from our BLV co-author. We tried several intervals, such as 5, 10, and 15 seconds, and found that shorter intervals overwhelmed the users, whereas longer intervals practically would not be repeated enough times to give them meaningful information. We settled on repeating the feedback every 10 seconds for our implementation.

## 6 USER STUDY

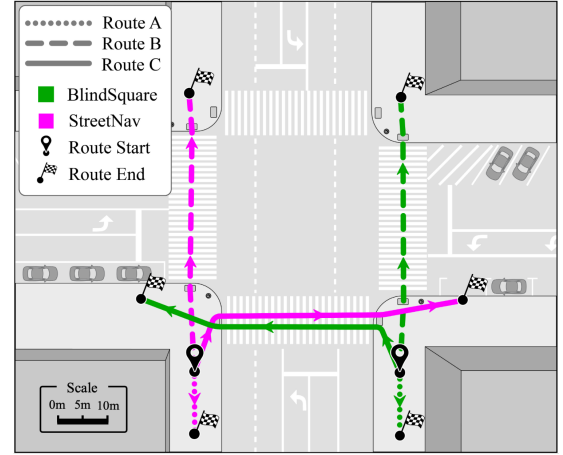
Our user study had two goals, related to RQ3. First, we wanted to evaluate the extent to which StreetNav addressed BLV pedestrians’ challenges in navigating outdoor environments when using existing GPS-based systems (Section 3). Second, we wanted to analyze BLV pedestrians’ experience of navigating outdoors using StreetNav compared to existing GPS-based systems.

### 6.1 Study Description

**Participants.** We recruited eight BLV participants (five males, three females; aged 24–52) by posting to social media platforms and by snowball sampling [22]. Participants identified themselves with a range of racial identities (Asian, Black, White, Latino, and Mixed), and all of them lived in a major city in the US. Participants also had diverse visual abilities, onset of vision impairment, and familiarity with assistive technology (AT) for navigation.

Table 2 summarizes participants’ information. All but three participants (P1, P7, and P8) reported themselves as being moderately–extremely experienced with AT for navigation (3+ scores on a 5-point rating scale). Only P3 reported minor hearing loss in both ears and wore hearing aids. All participants except two (P2, P9) used white cane as their primary mobility aid. P2 did not use any mobility aid, while P9 primarily used a guide dog for navigation. The IRB-approved study lasted for about 120 minutes, and participants were compensated \$75 for their time. We obtained informed consent from all study participants.

**Experimental Design.** In the study, participants completed three navigation tasks at a street intersection in two conditions: (i) StreetNav and (ii) BlindSquare [44], a popular GPS-based navigation app specially designed for BLV people. We selected BlindSquare as the baseline because it emerged as one of the most frequently used apps among our BLV participants for outdoor navigation, as identified during the formative interviews (Section 3). We evaluated the two systems via their respective iOS apps on an iPhone 14 Pro. Both



**Figure 9: The routes used in the navigation tasks. (A) 12 meters, stationary person to avoid on the sidewalk. (B) 30 meters, cross street, and moving person to avoid on the sidewalk. (C) 38 meters, a 90° turn, cross street, and moving person to avoid on the crosswalk. To mitigate learning effects, routes for the two conditions are symmetrically designed, situated on opposite sides of the street.**

systems’ apps seamlessly integrated with VoiceOver, and all eight participants had a high level of familiarity with using iPhones and VoiceOver, with ratings of 3 or higher on a 5-point scale.

Note that our study objective is to compare StreetNav against BLV people’s current navigation methods using GPS-based systems. Since such apps, including BlindSquare, do not offer any assistance with obstacle awareness or crossing streets, the comparison effectively becomes StreetNav vs. participants’ own abilities with mobility aids and non-visual senses.

Our study followed a within-subjects design, in which participants tested the two navigation systems in a counter-balanced order to minimize potential order bias and learning effects. In each condition, participants were tasked with completing three distinct navigation challenges corresponding to three specific routes. Figure 9 illustrates these three navigation routes. We deliberately chose the routes to lie within the street camera’s field of view and include a range of difficulty levels for each task: (A) a short route, 12 meters, that involved avoiding a stationary person on the sidewalk; (B) a long route, 30 meters, that involved crossing a street and avoiding a moving person on the sidewalk; and (C) a complex route, 38 meters, that involved making a 90 degree turn, crossing a street, and avoiding a moving person on the crosswalk. For each of these tasks, one of the researchers assumed the role of an obstacle. None of the participants were familiar with the study location.

Given that participants navigated the same intersection in both conditions, the potential for learning effects as a confounding factor was carefully considered. To address this concern, we took deliberate measures by creating distinct routes for each condition. Specifically, we designed the routes in both conditions to be symmetric—rather than being identical—with the starting and ending points of each route strategically positioned on opposite sides of the street intersection, as illustrated in Figure 9. The symmetry of routes



ensured that participants encountered the same challenges in both conditions. To ensure participants' safety, the researchers accompanied them at all times during the study, prepared to intervene whenever necessary.

**Procedure.** We began each study condition by giving a short tutorial of the respective smartphone app for the system. During these tutorials, participants were taught how to use the app and how to interpret the various audiohaptic cues it offered. To accommodate potential challenges arising from ambient noise at the street intersection, participants were given the option to wear headphones during the study. Only two participants, namely P3 and P5, exercised that option; rest of the participants relied on the smartphone's built-in speaker to hear the audiohaptic cues.

After completing the three navigation tasks for each condition, we administered a questionnaire comprising four distinct parts. These parts were designed to assess participants' experiences around challenges faced by BLV pedestrians in outdoor navigation, specifically addressing the following aspects: routing to destination (C1), veering off course (C1), avoiding obstacles (C2), and crossing streets (C3). It included questions about how well each system assisted with the challenges, if at all. Participants rated their experience on a 5-point rating scale, where a rating of "1" indicated "not at all well," and a rating of "5" indicated "extremely well." After each part of the questionnaire, we asked follow-up questions to gain deeper insights into the reasons behind their ratings and their overall experiences.

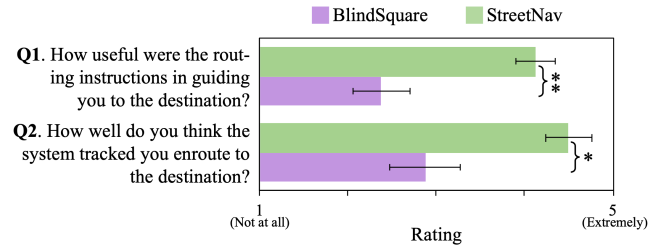
Following their experience with both navigation systems, participants were asked to complete a post-study questionnaire. This questionnaire required them to rank the two navigation systems in terms of their preference for outdoor navigation. Subsequently, we directed our discussion toward StreetNav, engaging participants in a conversation about potential avenues for improvement. We also inquired about the specific scenarios in which they envision using this system in the future.

In addition to questionnaires capturing participants' subjective experiences, we also analyzed system usage logs and video recordings to assess participants' actual performance in the navigation tasks. We note that willingness to be video-recorded was completely voluntary. All eight participants still agreed to be video-recorded, providing us with written consent to do so.

**Analysis.** We report participants' spontaneous comments that best represent their overall opinions, providing further context on the quantitative data we collected during the study. We analyzed the transcripts for participants' quotes and grouped them according to the (i) questionnaire's four parts: routing to destination, veering off course, avoiding obstacles, and crossing streets; (ii) overall satisfaction and ranking preferences, and (iii) how users' individual experiences influenced their preferences.

## 6.2 Results

Our results show that StreetNav guided participants to destinations with greater precision and reduced veering, improved obstacle awareness, and increased confidence in street crossing. For the statistic analysis of each measure, we first used a Kolmogorov-Smirnov test to determine if the data was parametric. Then, for



**Figure 10: Results for participants' experience with routing to the destination. Participants rated the (1) usefulness of routing instructions, and (2) the system's ability to track them en route to the destination. Participants found StreetNav's turn-by-turn instructions significantly more useful and precise than BlindSquare's "as the crow flies"-style routing instructions. Pairwise significance is depicted for  $p < 0.01$  (\*) and  $p < 0.05$  (\*\*). The error bars indicate standard error.**

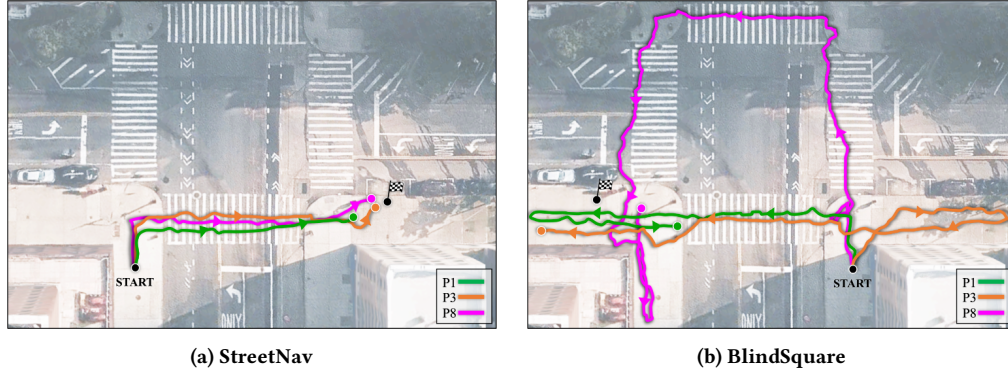
parametric data, we used a paired t-test to compare the two conditions. Additionally, we analyzed video recordings, annotating routes that participants took during the study. We report key results below, with additional findings in Appendix D.

**Routing to Destination.** Figure 10 shows participants' average rating for their experience following routes to the destination in each condition. The mean ( $\pm$  std. dev.) rating for participants' perceived usefulness of the routing instructions in guiding them to the destination was 4.13 ( $\pm 0.64$ ) for StreetNav and 2.38 ( $\pm 0.91$ ) for BlindSquare. The condition had a significant main effect ( $p = 0.014$ ) on participants' experience reaching destinations with the routing instructions. The mean ( $\pm$  std. dev.) rating for participants' experience with the system's ability to track them was 4.50 ( $\pm 0.76$ ) for StreetNav and 2.88 ( $\pm 1.13$ ) for BlindSquare. The condition had a significant main effect ( $p = 0.001$ ) on participants' perception of how well the system tracked them en route to the destination. This indicates that participants found StreetNav more useful than BlindSquare for guiding them to the destination.

Figure 11 illustrates our analysis of the video recordings, plotting the typical paths taken by participants in the third route across both conditions. We computed various metrics from their paths, that provide insights into participants' self-reported ratings.

We found that when using BlindSquare, participants covered greater distances to reach the same destinations compared to when using StreetNav. On average, participants traveled a distance approximately 2.1 times longer than the shortest route when relying on BlindSquare. In contrast, when using StreetNav, they covered a distance of only about 1.1 times the shortest route to their destination. This represents a 51% reduction in the unnecessary distance traveled with StreetNav in comparison to BlindSquare. Figure 11b shows how participants using BlindSquare often exhibited an oscillatory pattern near their destinations (P1, P8) before eventually reaching close to them.

Additionally, StreetNav's routing instructions displayed a notably higher level of precision, guiding participants to their destinations with 2.9 times greater accuracy than BlindSquare. Figure 11 clearly shows this trend for the third route. On average, across



**Figure 11: Comparison of paths traveled by three participants (P1, P3, P8) for route ‘C’ using (a) StreetNav, and (b) BlindSquare. StreetNav’s routing instructions consistently guided participants to the destination via the shortest path. BlindSquare, however, caused participants to take incorrect turns (P1, P3, P8), oscillate back and forth near destinations (P1, P8), and even go around the whole intersection before getting close to the destination (P8).**

the three study routes, participants using StreetNav concluded their journeys within a tighter radius of 12.53 feet from their intended destination. In contrast, participants relying on BlindSquare concluded their journeys within a radius of 35.94 feet from their intended destination. Two study participants, P4 and P5, even refused to navigate to the destination in two of the three tasks with BlindSquare. This was primarily attributed to BlindSquare’s low precision in tracking the participants and often guiding them to take incorrect turns. Figure 11b highlights how BlindSquare caused P8 to go around the intersection before finally getting close the destination.

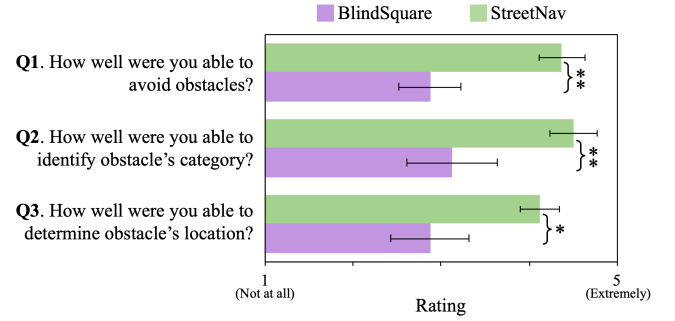
Participants preferred StreetNav over BlindSquare for its audio-haptic cues for turn-by-turn navigation instructions, which they found to be more useful and precise than BlindSquare’s “as the crow flies”-style clock face and distance-based instructions. P3’s comment encapsulates this sentiment:

*“When it’s time for me to turn right and walk a certain distance, [StreetNav] is very, very, very precise.” –P3*

Although all participants preferred StreetNav’s routing feedback over BlindSquare’s, distinct patterns emerged in their preference and utilization of these cues. StreetNav delivers a combination of audiohaptic and speech feedback for routing, and participants adopted varying strategies for utilizing this feedback. Some individuals placed greater reliance on the veering haptic feedback as their primary directional guide, while reserving speech feedback as a fallback option. Conversely, some participants prioritized the speech feedback, assigning it a higher level of importance in their navigation process compared to audio-haptic cues.

Maintaining a straight walking path is crucial for effective routing. Thus, we separately analyzed the extent to which each system prevented veering, with findings reported in Appendix D.1.

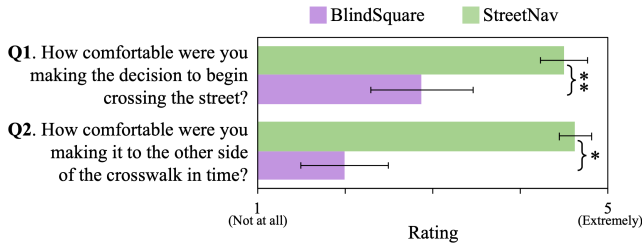
**Obstacle Awareness.** Figure 12 shows participants’ average rating for their perceived awareness of obstacles across the two conditions. Specifically, participants rated their ability to (1) avoid obstacles, (2) identify its category (e.g., person, bicycle, trash can), and (3) determine its relative location. The mean ( $\pm$  std. dev.) rating for participants’ perceived ability to avoid obstacles was 4.38 ( $\pm 0.74$ ) for



**Figure 12: Results for participants’ perceived obstacle awareness. Participants rated their ability to (1) avoid obstacles, (2) identify its category (e.g., person, bicycle), and (3) determine its relative location; on a scale of 1–5. StreetNav significantly improved participants’ awareness of nearby obstacles during navigation. Pairwise significance is depicted for  $p < 0.01$  (\*) and  $p < 0.05$  (\*\*). The error bars indicate standard error.**

StreetNav and 2.88 ( $\pm 0.99$ ) for BlindSquare, to identify its category was 4.50 ( $\pm 0.76$ ) for StreetNav and 3.13 ( $\pm 1.46$ ) for BlindSquare, and to determine obstacle’s relative location was 4.13 ( $\pm 0.64$ ) for StreetNav and 2.88 ( $\pm 1.25$ ) for BlindSquare. A paired t-test revealed that the condition had a significant main effect on participants’ perceived ability to avoid obstacles ( $p = 0.030$ ), identify its category ( $p = 0.037$ ), and relative location ( $p = 0.004$ ). This suggests that StreetNav offered users a heightened awareness of nearby obstacles compared to the baseline condition of BlindSquare.

With StreetNav, participants had the option to use obstacle avoidance audio feedback in conjunction with their conventional mobility aids. However, in the case of BlindSquare, the system itself did not offer any obstacle-related information. Consequently, participants primarily relied on their traditional mobility aids in this condition, as is typical when using GPS-based systems. Our analysis of the video recordings found that in both experimental conditions, participants encountered no instances of being severely hindered by



**Figure 13: Results for participants’ perceived comfort in crossing streets. Participants rated their perceived comfort in (1) making the decision on when to begin crossing the street, and in (2) pacing themselves when crossing. Participants were significantly more comfortable crossing streets with StreetNav in comparison to BlindSquare. Pairwise significance is depicted for  $p < 0.01$  (\*) and  $p < 0.05$  (\*\*). The error bars indicate standard error.**

obstacles. Instead, they adeptly navigated around obstacles with the assistance of their white canes or guide dogs.

Although participants generally had a positive perception of obstacle avoidance when using StreetNav, their opinions on the utility of obstacle awareness information varied. Some participants found this information beneficial, emphasizing its role in preventing “awkward bumping into people” (P2) and boosting their confidence, resulting in greater “speed in terms of walking” (P3). Conversely, participants who felt confident avoiding obstacles with their mobility aids regarded StreetNav’s obstacle information to be extraneous. P8 also expressed concerns about the potential information overload it could cause in dense urban areas:

*“To know where people are, is a bit of overkill, because, especially in a city like this, if you turn this thing on in Times Square, it would have your head go upside down... If I’m around a lot of people, I’m not really thinking about avoiding them. I have a cane for a reason. They can see and I can’t, so I’m relying on them to see me and get out of my way.” –P8*

Many participants proposed an alternative use case for StreetNav’s obstacle awareness information, highlighting its potential for providing insights into their surroundings. They suggested that this information could unlock environmental affordances, including the identification of accessible light signals and available benches for resting: “knowing there was a bench was top-notch for me” (P8). Therefore, StreetNav’s obstacle awareness information served a dual purpose, aiding in both obstacle avoidance and environmental awareness, allowing users to “know what’s around” (P8) them.

**Crossing Streets.** Figure 13 shows participants’ average rating for their perceived comfort in crossing streets. The mean ( $\pm$  std. dev.) rating of participants’ perceived comfort in making the decision on when to begin crossing the street was 4.50 ( $\pm 0.76$ ) for StreetNav and 2.88 ( $\pm 1.64$ ) for BlindSquare. The mean ( $\pm$  std. dev.) rating of participants’ perceived comfort in safely making it through the crosswalk and reach the other end was 4.63 ( $\pm 0.52$ ) for StreetNav and 2.00 ( $\pm 1.41$ ) for BlindSquare. A paired t-test showed that the condition had a significant main effect on participants’ comfort in

beginning to cross streets ( $p = 0.029$ ) and in safely making it to the other side ( $p = 0.001$ ).

As BlindSquare does not provide feedback for crossing streets, participants reported relying on their auditory senses by listening for the surge of parallel traffic. However, during the semi-structured interviews, some participants highlighted challenging scenarios that can make this strategy less reliable. P4, for instance, pointed out that ironically, less traffic can complicate street crossings:

*“I don’t always know when to cross because it’s so quiet. And sometimes two, three light cycles go by, and I’m just standing there.” –P4*

This issue has been exacerbated by the presence of electric cars, which are difficult to hear due to their quiet motors. For P3, their hearing impairments made it challenging to listen for traffic. Thus, most participants appreciated StreetNav’s ability to assist with crossing streets:

*“When it’s quiet, I would cross. But now with hybrid cars, it’s not safe to do that. [StreetNav] app telling you which street light is coming on is really helpful.” –P7*

Participants made decisions to cross the streets by combining StreetNav’s feedback with their auditory senses. Many participants emphasized that having information about the time remaining to cross significantly boosted their confidence, especially when this information aligned with the sounds of traffic: “I thought it was great because I could tell that it matched up” (P8). This alignment between the provided information and their sensory perception inspired confidence in participants:

*“Relying on my senses alone feels like a gamble about 90 percent of the time, so a system like [StreetNav] that accurately displays the amount of time I have to cross the street is great.” –P2*

P4 compared StreetNav with the Oko App [9]. While P4 found Oko effective in identifying signal state, they appreciated StreetNav’s seamless integration, which does not require pointing the camera at the pedestrian signal.

## 7 TECHNICAL EVALUATION

We evaluate StreetNav’s technical performance to compare its effectiveness against the status quo of GPS-based systems. StreetNav’s main advantage is its precise user localization. Thus, this evaluation aims to answer the question: *How precisely does StreetNav localize the user, and what factors impact this precision?*

In comparing overall accuracy, StreetNav’s localization error was 0.41 ( $\pm 1.49$ ) meters in estimating the user’s feet position and an additional 0.65 ( $\pm 0.26$ ) meters in transforming this position from the camera view to the map. This error is significantly lower than GPS, which achieves localization errors in excess of 10-15 meters in urban areas [23, 46, 69].

We independently analyzed technical performance of the three steps in StreetNav’s computer vision pipeline for user localization: (i) CLIP-based gesture recognition, (ii) pedestrian feet position estimation, and (iii) camera to map-view transformation. Key results for each step are reported here, with detailed discussion in Appendix E.

StreetNav’s CLIP-based gesture recognition achieves 83% accuracy in identifying the hand-waving gesture, with a 24% false

positive rate and a 10% false negative rate. For pedestrian feet position estimation, the root mean squared error ( $\pm$  std.) is 0.41 ( $\pm$  1.49) meters. Although StreetNav detects pedestrians with 82% precision and 65% recall at a 0.5 IOU (intersection over union) threshold, accuracy decreases as the pedestrian's distance from the camera increases, with the false negative rate rising from 1% at 5 meters to 74% at 40 meters. The root mean squared error ( $\pm$  std.) for transforming points from camera view to map view is 0.65 ( $\pm$  0.26) meters.

Appendix E elaborates on the evaluation procedure and provides additional detail on factors that impact performance for each step.

## 8 DISCUSSION

Our goal with StreetNav was to explore the idea of repurposing existing street cameras to support precise outdoor navigation for BLV pedestrians. We reflect upon our findings to discuss how street camera-based systems might be deployed at scale, privacy concerns with camera-based assistive technology, implications of a street camera-based navigation approach for existing GPS-based navigation systems, and the affordances enabled by precise, real-time outdoor navigation assistance.

### *Deploying street camera-based navigation systems at scale.*

StreetNav demonstrates that street cameras have the potential to be repurposed for supporting precise outdoor navigation for BLV pedestrians. Our study results show that street camera-based navigation systems can guide users to their destination more precisely and prevent them from veering off course (Figure 11). Our results also show that street camera-based systems can support real-time, scene-aware assistance by notifying users of nearby obstacles (Figure 12) and giving information about when to cross streets (Figure 13). These benefits of a street camera-based approach over existing GPS-based systems underscore the need for deploying such systems at scale. Although StreetNav was deployed at a single intersection, we learned insights on potential challenges and considerations that must be addressed to deploy street camera-based systems at scale.

Several internal and external factors need to be considered before street cameras can be effectively leveraged to support blind navigation at scale. External factors, including lighting conditions and occlusions on the street, may affect system performance. For instance, we noticed that StreetNav's ability to track pedestrians was affected severely in low-light conditions (e.g., at night) and by occlusions due to the presence of large vehicles (e.g., trucks, buses) and the installation of scaffolding for construction (Figure 17d). Such challenges affect the reliability of street camera-based systems and may limit its operational hours. Internal factors, including the positioning of cameras, their field of view, and variability in resolution, may affect the extent to which such systems can promise precise navigation assistance. For instance, the visibility of the pedestrian signals from the camera feed could affect how much such systems can assist users with crossing streets. With StreetNav, we observed a drop in tracking accuracy as pedestrians moved further away from the camera.

Therefore, deploying street camera-based systems at scale would require future work to investigate the extent to which both external factors (e.g., lighting, occlusions) and internal factors (e.g., camera resolution) affect system performance and reliability. To address

some of the technical limitations around tracking performance and field of view limitations, future research could explore integrating multiple cameras at various elevations and viewing angles. Prior work on robot navigation has explored the fusion of multiple cameras to improve tracking performance [11, 52, 55]. Future work could also explore an ecosystem of accessible street cameras that can share information to automatically manage hand-offs across street intersections, providing users with a seamless experience beyond a single street intersection. Such ecosystems, which span beyond one intersection to a whole district or city, could enable new affordances, such as automatically sensing pedestrian traffic to inform traffic signals and vice versa [37].

### *Privacy concerns with camera-based assistive technology.*

Privacy is a significant consideration for the practical deployment of street camera-based assistive technology. Our study with various stakeholders (Section 4) revealed differing perspectives on privacy and identified strategies for respecting those perspectives. Recall from Section 4 the two strategies that our stakeholders identified: (i) regulating data storage, anonymization, and access policies; and (ii) repurposing *existing* cameras rather than installing *new* ones. Concerning the first strategy, StreetNav's implementation does not necessitate any data storage for facilitating outdoor navigation assistance. The video feed is processed in real-time on a local server, and only navigation instructions are shared with the BLV user's smartphone. Furthermore, StreetNav employs a map view representation—as depicted in Figure 3d—for computing routes and identifying obstacles, inherently enabling data anonymization. The questions regarding who should have access to these cameras and for what other purposes, including public safety, they might be used for, still require further investigation. As for the second strategy, although StreetNav repurposes a camera from an existing publicly available testbed, the feasibility of securing camera access and resources of already existing street cameras at scale remains an open question. From our interview with the Community Board leader (Section 4), collaboration among different government entities emerged as a potential next step. Future research could investigate the roles of different government entities and the implementation of policies that ensure responsible and transparent use of street cameras.

**Implications for GPS-based navigation systems.** When cameras are available, and conditions align favorably, street camera-based systems offer BLV individuals a valuable source of fine-grained, high-precision information, significantly enhancing their navigational experience and environmental awareness. These capabilities are currently beyond the reach of conventional GPS-based systems. All eight study participants unanimously chose StreetNav over BlindSquare as their preferred navigation system due to its precise, scene-aware navigation assistance (Section D.2). However, it's important to acknowledge that street camera-based systems have their own set of limitations. The widespread availability of street cameras is not yet a reality, and ideal conditions may not always be met for their effective use. In contrast, GPS-based systems, while lacking in precision and environmental awareness, are universally accessible and resilient in varying conditions, including low light. A harmonious integration of these two approaches



is a promising solution. Users can tap into street-camera information when conditions permit, seamlessly transitioning to GPS data when necessary. This can be facilitated through sensor fusion or information hand-offs, creating a synergy that ensures a smooth and reliable navigational experience. Future approaches could explore how these two systems can effectively complement each other, addressing their respective limitations and enhancing overall performance.

**Affordances of precise outdoor navigation assistance for BLV people.** Previous research in indoor navigation has demonstrated the advantages of accurately pinpointing users' locations [2, 36, 62] and providing scene-aware navigational information [25, 35]. However, achieving such precision has remained a challenge in outdoor environments, primarily due to the limited accuracy of GPS technology [23]. StreetNav's approach of leveraging existing street cameras demonstrates that precise outdoor navigation support for BLV pedestrians is possible. Our study reveals the advantages of precise, fine-grained navigation for BLV individuals. These benefits include a substantial reduction in instances of veering and routing errors, such as deviation from the shortest path or missing intended destinations, as well as augmented environmental awareness.

StreetNav offered our participants a glimpse into the potential of precise outdoor navigation. Several participants desired even greater precision, including the ability to discern the exact number of steps remaining before reaching a crosswalk's curb. Future research could delve into exploring how to best deliver such granular feedback to BLV users, alongside the necessary technological advancements needed to achieve this level of precision. These advantages, as our findings suggest, extend beyond merely improving navigation performance. Participants shared insights into how precise navigation could enhance their independence when navigating outdoors. It could empower BLV people to venture outdoors more frequently, unlocking new travel opportunities, as exemplified by P3's newfound confidence in using public transportation with StreetNav-like systems:

*"I don't really use the city buses, except if I'm with somebody, but [StreetNav] would make me want to get up, go outside, and walk to the bus stop." –P3*

This newfound confidence is particularly noteworthy, considering the unpredictable nature of outdoor environments. Future research could explore new affordances that street camera-based systems can enable for people, in general.

## 9 LIMITATIONS

Our work revealed valuable insights into the benefits and effectiveness of a new approach that uses existing street cameras for outdoor navigation assistance. At the same time, we acknowledge that our work has several limitations.

StreetNav was developed using a camera from an existing cloud-networked testbed that is publicly available to the researchers [13, 60, 72], situated at a specific street intersection. It is important to note that our development process may not have encountered all potential technical challenges and design considerations, given the constraints of this setup. Additionally, StreetNav's use of the testbed camera instead of a regular security camera may yield slightly different performance due to factors like camera perspective, resolution,

availability, and even the layout of the intersection. Future research could expand upon our design and investigate how street camera-based systems can be adapted to different environments.

Furthermore, to ensure the safety of participants and to fit the user study within a 120-minute timeframe, we designed the study routes to be less complex and dangerous. Real-world outdoor environments can vary significantly across regions, and our study location may not fully capture the diversity of scenarios BLV people encounter when navigating outdoors.

Lastly, it is important to note that our design of StreetNav was guided by interviews with six BLV individuals, six stakeholders from New York City, and was evaluated in a study with only eight BLV individuals. While our participants' insights are valuable, their preferences may not represent the general population's perspectives on BLV people's navigation challenges and various stakeholders' privacy concerns. There could be additional challenges and design possibilities that we did not explore because of the cultural and regional context. Future research should consider a more extensive and diverse participant pool to gain a more comprehensive understanding of BLV people's challenges and privacy preferences of various stakeholders.

## 10 CONCLUSION

We explored the idea of leveraging *existing* street cameras to support precise outdoor navigation for BLV pedestrians. Our resulting system, StreetNav, investigates both technical and sociotechnical concerns with a street camera-based navigation system. Our evaluations revealed StreetNav's potential to guide users more precisely to destinations compared to existing GPS-based systems. It also demonstrated camera-based system's ability to offer real-time, context-aware navigation assistance, aiding in obstacle avoidance and safe street crossings. However, we also identified challenges and opportunities for deploying street camera-based navigation systems at scale. These challenges suggest areas for future research to enhance system robustness and reliability while addressing privacy concerns. Our work highlights the potential of embedding accessibility into urban infrastructure using existing resources like street cameras. We envision a future where these systems seamlessly integrate into urban environments, providing BLV people with safe, precise navigation capabilities and empowering them to navigate their surroundings confidently.

## ACKNOWLEDGMENTS

We thank Lindsey Tara Weiskopf for literature review, Arjun Nichani for initial prototyping, and Chloe Tedjo and Josh Bassin for help with formative interviews. We thank our study participants for participating in the study. This work was supported in part by the National Science Foundation (NSF) and Center for Smart Streetscapes (CS3) under NSF Cooperative Agreement No. EEC-2133516, ARO Grant No. W911NF1910379, NSF Grant No. CNS-1827923, NSF Grant No. CNS-2038984, and corresponding support from the Federal Highway Administration (FHWA). Daniel Weiner and Xin Yi Therese Xu were supported by the Columbia–Amazon SURE Program. Sophie Ana Paris was funded by the NSF Grant No. 2051053 and 2051060.



## REFERENCES

- [1] Nazatul Naquiah Abd Hamid and Alistair D.N. Edwards. 2013. Facilitating route learning using interactive audio-tactile maps for blind and visually impaired people. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems on - CHI EA '13*. ACM Press, Paris, France, 37. <https://doi.org/10.1145/2468356.2468364>
- [2] Dragan Ahmetovic, Cole Gleason, Chengxiong Ruan, Kris Kitani, Hironobu Takagi, and Chieko Asakawa. 2016. NavCog: A navigational cognitive assistant for the blind. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '16)*. Association for Computing Machinery, New York, NY, USA, 90–99. <https://doi.org/10.1145/2935334.2935361>
- [3] Dragan Ahmetovic, Roberto Manduchi, James M. Coughlan, and Sergio Mascetti. 2017. Mind Your Crossings: Mining GIS Imagery for Crosswalk Localization. *ACM Transactions on Accessible Computing* 9, 4 (Dec. 2017), 1–25. <https://doi.org/10.1145/3046790>
- [4] Amnesty International. 2021. Surveillance City: NYPD can use more than 15,000 cameras to track people using facial recognition in Manhattan, Bronx and Brooklyn. <https://www.amnesty.org/en/latest/news/2021/06/scale-new-york-police-facial-recognition-revealed/>
- [5] Apple Inc. 2023. Apple Maps. <https://www.apple.com/maps/>
- [6] Apple Inc. 2023. Swift. <https://developer.apple.com/swift/>
- [7] Apple Inc. 2023. VoiceOver. <https://www.apple.com/accessibility/vision/>
- [8] Mauro Avila and Limin Zeng. 2017. A Survey of Outdoor Travel for Visually Impaired People Who Live in Latin-American Region. In *Proceedings of the 10th International Conference on Pervasive Technologies Related to Assistive Environments (PETRA '17)*. Association for Computing Machinery, New York, NY, USA, 9–12. <https://doi.org/10.1145/3056540.3064953>
- [9] AYES Inc. 2023. Oko: The AI-powered Navigation App for People with a Disability. <https://www.ayes.ai/>
- [10] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. <https://doi.org/10.1191/1478088706qp0630a>
- [11] Wen-Chung Chang, Chia-Hung Wu, Wen-Ting Luo, and Huan-Chen Ling. 2013. Mobile robot navigation and control with monocular surveillance cameras. In *2013 CACS International Automatic Control Conference (CACS)*, 192–197. <https://doi.org/10.1109/CACS.2013.6734131>
- [12] Coco Feng. 2019. China the most surveilled nation? The US has the largest number of CCTV cameras per capita. <https://www.scmp.com/tech/gear/article/3040974/china-most-surveilled-nation-us-has-largest-number-cctv-cameras-capita>
- [13] COSMOS Project. 2023. Hardware: Cameras. <https://wiki.cosmos-lab.org/wiki/Hardware/Cameras>
- [14] DataCity SMTG. 2019. DataCity Smart Mobility Testing Ground. <https://cait.rutgers.edu/datacity/>
- [15] František Duchoň, Andrej Babinec, Martin Kajan, Peter Beňo, Martin Florek, Tomáš Fico, and Ladislav Jurišica. 2014. Path Planning with Modified a Star Algorithm for a Mobile Robot. *Procedia Engineering* 96 (2014), 59–69. <https://doi.org/10.1016/j.proeng.2014.12.098>
- [16] Ping-Jung Duh, Yu-Cheng Sung, Liang-Yu Fan Chiang, Yung-Ju Chang, and Kuan-Wen Chen. 2020. V-eye: A vision-based navigation system for the visually impaired. *IEEE Transactions on Multimedia* 23 (2020), 1567–1580.
- [17] Pardis Emami-Naeini, Joseph Breda, Wei Dai, Tadayoshi Kohno, Kim Laine, Shwetak Patel, and Franziska Roesner. 2023. Understanding People's Concerns and Attitudes Toward Smart Cities. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–24. <https://doi.org/10.1145/3544548.3581558>
- [18] Alexander Fiannaca, Ilias Apostolopoulos, and Eelke Folmer. 2014. Headlock: a wearable navigation aid that helps blind cane users traverse large open spaces. In *Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility (ASSETS '14)*. Association for Computing Machinery, New York, NY, USA, 323–324. <https://doi.org/10.1145/2661334.2661344>
- [19] John C. Flanagan. 1954. The critical incident technique. *Psychological Bulletin* 51, 4 (1954), 327–358.
- [20] Frank Hersey. 2017. China to have 626 million surveillance cameras within 3 years. <https://technode.com/2017/11/22/china-to-have-626-million-surveillance-cameras-within-3-years/>
- [21] Thomas Gallagher, Elyse Wise, Binghao Li, Andrew G. Dempster, Chris Rizos, and Euan Ramsey-Stewart. 2012. Indoor positioning system based on sensor fusion for the Blind and Visually Impaired. In *2012 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*. 1–9. <https://doi.org/10.1109/IPIN.2012.6418882>
- [22] Leo A. Goodman. 1961. Snowball Sampling. *The Annals of Mathematical Statistics* 32, 1 (1961), 148–170. <https://www.jstor.org/stable/2237615> Publisher: Institute of Mathematical Statistics.
- [23] GPS.gov. 2022. GPS Accuracy. <https://www.gps.gov/systems/gps/performance/accuracy/>
- [24] João Guerreiro, Daisuke Sato, Dragan Ahmetovic, Eshed Ohn-Bar, Kris M. Kitani, and Chieko Asakawa. 2020. Virtual navigation for blind people: Transferring route knowledge to the real-World. *International Journal of Human-Computer Studies* 135 (March 2020), 102369. <https://doi.org/10.1016/j.ijhcs.2019.102369>
- [25] João Guerreiro, Daisuke Sato, Saki Asakawa, Huixu Dong, Kris M. Kitani, and Chieko Asakawa. 2019. CaBot: Designing and Evaluating an Autonomous Navigation Robot for Blind People. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 68–82. <https://doi.org/10.1145/3308561.3353771>
- [26] Richard Guy and Khai Truong. 2012. CrossingGuard: exploring information content in navigation aids for visually impaired pedestrians. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Austin Texas USA, 405–414. <https://doi.org/10.1145/2207676.2207733>
- [27] David L Harkey, Daniel L Carter, Janet M Barlow, and Billie Louise Bentzen. 2007. Accessible pedestrian signals: A guide to best practices. *National Cooperative Highway Research Program, Contractor's Guide for NCHRP Project* (2007).
- [28] Hemment, Drew and Townsend, Anthony. 2013. Smart Citizens.
- [29] Peiyun Hu and Deva Ramanan. 2017. Finding Tiny Faces. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Honolulu, HI, 1522–1530. <https://doi.org/10.1109/CVPR.2017.166>
- [30] Microsoft Inc. 2018. Microsoft Soundscape - Microsoft Research. <https://www.microsoft.com/en-us/research/product/soundscape/>. (2018).
- [31] Gaurav Jain, Basel Hindi, Mingyu Xie, Zihao Zhang, Koushik Srinivasula, Mahshid Ghasemi, Daniel Weiner, Xin Yi Therese Xu, Sophie Ana Paris, Chloe Tedjo, Josh Bassin, Michael Malcolm, Mehmet Turkcan, Javad Ghaderi, Zoran Kostic, Gil Zussman, and Brian A. Smith. 2023. Towards Street Camera-based Outdoor Navigation for Blind Pedestrians. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '23)*. Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/3597638.3614498>
- [32] Gaurav Jain, Yuanyang Teng, Dong Heon Cho, Yunhao Xing, Maryam Aziz, and Brian A. Smith. 2023. "I Want to Figure Things Out": Supporting Exploration in Navigation for People with Visual Impairments. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (April 2023), 63:1–63:28. <https://doi.org/10.1145/3579496>
- [33] Hernisa Kacorri, Sergio Mascetti, Andrea Gerino, Dragan Ahmetovic, Valeria Alampi, Hironobu Takagi, and Chieko Asakawa. 2018. Insights on Assistive Orientation and Mobility of People with Visual Impairment Based on Large-Scale Longitudinal Data. *ACM Transactions on Accessible Computing* 11, 1 (April 2018), 1–28. <https://doi.org/10.1145/3178853>
- [34] Robert K Katzschnmann, Brandon Araki, and Daniela Rus. 2018. Safe local navigation for visually impaired users with a time-of-flight and haptic feedback device. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 26, 3 (2018).
- [35] Seita Kayukawa, Keita Higuchi, João Guerreiro, Shigeo Morishima, Yoichi Sato, Kris Kitani, and Chieko Asakawa. 2019. BBeep: A Sonic Collision Avoidance System for Blind Travellers and Nearby Pedestrians. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press, Glasgow, Scotland Uk, 1–12. <https://doi.org/10.1145/3290605.3300282>
- [36] Jee-Eun Kim, Masahiro Bessho, Shinsuke Kobayashi, Noboru Koshizuka, and Ken Sakamura. 2016. Navigating visually impaired travelers in a large train station using smartphone and bluetooth low energy. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing (SAC '16)*. Association for Computing Machinery, New York, NY, USA, 604–611. <https://doi.org/10.1145/2851613.2851716>
- [37] Zoran Kostic, Alex Angus, Zhengye Yang, Zhuoxu Duan, Ivan Seskar, Gil Zussman, and Dipankar Raychaudhuri. 2022. Smart City Intersections: Intelligence Nodes for Future Metropolises. *Computer* 55, 12 (Dec. 2022), 74–85. <https://doi.org/10.1109/MC.2022.3206273>
- [38] Laura Griffin. 2020. Surveillance Cameras are Everywhere. And They're Only Going To Get More Ubiquitous. <https://crimereads.com/surveillance-cameras-are-everywhere-and-theyre-only-going-to-get-more-ubiquitous/>
- [39] Xiang Li, Hanzhang Cui, John-Ross Rizzo, Edward Wong, and Yi Fang. 2020. Cross-Safe: A Computer Vision-Based Approach to Make All Intersection-Related Pedestrian Signals Accessible for the Visually Impaired. In *Advances in Computer Vision*, Kohei Arai and Supriya Kapoor (Eds.). Vol. 944. Springer International Publishing, Cham, 132–146. [https://doi.org/10.1007/978-3-030-17798-0\\_13](https://doi.org/10.1007/978-3-030-17798-0_13)
- [40] Yimin Lin, Kai Wang, Wanxin Yi, and Shiguo Lian. 2019. Deep Learning Based Wearable Assistive System for Visually Impaired People. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE, Seoul, Korea (South), 2549–2557. <https://doi.org/10.1109/ICCVW.2019.00312>
- [41] Line Kristin Haug. 2019. Is Smart city sustainable? A case study of Bodø municipality.
- [42] Liza Lin, Newley Purnell. 2019. A World With a Billion Cameras Watching You Is Just Around the Corner. <https://www.wsj.com/articles/a-billion-surveillance-cameras-forecast-to-be-watching-within-two-years-11575565402>
- [43] Sergio Mascetti, Dragan Ahmetovic, Andrea Gerino, and Cristian Bernareggi. 2016. ZebraRecognizer: Pedestrian crossing recognition for people with visual impairment or blindness. *Pattern Recognition* 60 (Dec. 2016), 405–419. <https://doi.org/10.1016/j.patcog.2016.05.002>
- [44] MIPsoft. 2016. BlindSquare. <https://www.blindsquare.com/>
- [45] Mobintel. 2019. Mobility Intelligence Platform. <https://about.mobintel.org/>

- [46] Marko Modsching, Ronny Kramer, and Klaus ten Hagen. 2006. Field trial on GPS Accuracy in a medium size city: The influence of built-up. In *3rd workshop on positioning, navigation and communication*, Vol. 2006. 209–218.
- [47] MQTT. 2022. MQTT: The Standard for IoT Messaging. <https://mqtt.org/>
- [48] Madoka Nakajima and Shinichiro Haruyama. 2012. Indoor navigation system for visually impaired people using visible light communication and compensated geomagnetic sensing. In *2012 1st IEEE International Conference on Communications in China (ICCC)*. 524–529. <https://doi.org/10.1109/ICCCChina.2012.6356940>
- [49] Nvidia. 2023. Nvidia DeepStream GStreamer Plugin: NvDCF Tracker. [https://docs.nvidia.com/metropolis/deepstream/dev-guide/text/DS\\_plugin\\_gst-nvtracker.html#nvdcf-tracker](https://docs.nvidia.com/metropolis/deepstream/dev-guide/text/DS_plugin_gst-nvtracker.html#nvdcf-tracker)
- [50] NVivo. 1997. NVivo. <https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home>
- [51] OpenStreetMap. 2023. OpenStreetMap. <https://www.openstreetmap.org/>
- [52] Petr Ošćádal, Daniel Huczala, Jan Bém, Václav Kryš, and Zdenko Bobovsky. 2020. Smart Building Surveillance System as Shared Sensory System for Localization of AGVs. *Applied Sciences* 10, 23 (Nov. 2020), 8452. <https://doi.org/10.3390/app10238452>
- [53] Sabrina A. Paneels, Dylan Varenne, Jeffrey R. Blum, and Jeremy R. Cooperstock. 2013. The Walking Straight Mobile Application: Helping the Visually Impaired Avoid Veering. (July 2013).
- [54] Jagannadh Pariti, Vinita Tibdewal, and Tae Oh. 2020. Intelligent Mobility Cane - Lessons Learned from Evaluation of Obstacle Notification System using a Haptic Approach. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–8. <https://doi.org/10.1145/3334480.3375217>
- [55] Roman Pflugfelder and Horst Bischof. 2010. Localization and Trajectory Reconstruction in Surveillance Cameras with Nonoverlapping Views. *IEEE transactions on pattern analysis and machine intelligence* 32 (April 2010), 709–21. <https://doi.org/10.1109/TPAMI.2009.56>
- [56] Giorgio Presti, Dragan Ahmetovic, Mattia Ducci, Cristian Bernareggi, Luca Ludovico, Adriano Barate, Federico Avanzini, and Sergio Mascetti. 2019. WatchOut: Obstacle Sonification for People with Visual Impairment or Blindness. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, Pittsburgh PA USA, 402–413. <https://doi.org/10.1145/3308561.3353779>
- [57] Halley Profta, Reem Albaghli, Leah Findlater, Paul Jaeger, and Shaun K. Kane. 2016. The AT Effect: How Disability Affects the Perceived Social Acceptability of Head-Mounted Display Use. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, San Jose California USA, 4884–4895. <https://doi.org/10.1145/2858036.2858130>
- [58] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. <http://arxiv.org/abs/2103.00020> arXiv:2103.00020 [cs].
- [59] Lisa Ran, Sumi Helal, and Steve Moore. 2004. Drishti: an integrated indoor/outdoor blind navigation system and service. In *Second IEEE Annual Conference on Pervasive Computing and Communications*.
- [60] Dipankar Raychaudhuri, Ivan Seskar, Gil Zussman, Thanasis Korakis, Dan Kilper, Tingjun Chen, Jakub Kolodziejcki, Michael Sherman, Zoran Kostic, and Xiaoxiong Gu. 2020. Challenge: COSMOS: A city-scale programmable testbed for experimentation with advanced wireless. In *Proc. ACM MobiCom*.
- [61] Manaswi Saha, Alexander J. Fiannaca, Melanie Kneisel, Edward Cutrell, and Meredith Ringel Morris. 2019. Closing the Gap: Designing for the Last-Few-Meters Wayfinding Problem for People with Visual Impairments. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, Pittsburgh PA USA, 222–235. <https://doi.org/10.1145/3308561.3353776>
- [62] Daisuke Sato, Uran Oh, João Guerreiro, Dragan Ahmetovic, Kakuya Naito, Hironobu Takagi, Kris M. Kitani, and Chieko Asakawa. 2019. NavCog3 in the Wild: Large-scale Blind Indoor Navigation Assistant with Semantic Features. *ACM Transactions on Accessible Computing* 12, 3 (Sept. 2019), 1–30. <https://doi.org/10.1145/3340319>
- [63] Sendero Group. 2019. Seeing Eye GPS. <http://www.senderogroup.com/products/seeingeyegps/index.html>
- [64] Jae Shim and Young Cho. 2016. A Mobile Robot Localization via Indoor Fixed Remote Surveillance Cameras. *Sensors* 16, 2 (Feb. 2016), 195. <https://doi.org/10.3390/s16020195>
- [65] Jae-Hong Shim and Young-Im Cho. [n. d.]. A Mobile Robot Localization using External Surveillance Cameras at Indoor. 56 ([n. d.]), 502–507. <https://doi.org/10.1016/j.procs.2015.07.242>
- [66] Hojun Son, Divya Krishnagiri, V. Swetha Jeganathan, and James Weiland. 2020. Crosswalk Guidance System for the Blind. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*. 3327–3330. <https://doi.org/10.1109/EMBC44109.2020.9176623> ISSN: 2694-0604.
- [67] Juan Terven and Diana Cordova-Esparza. 2023. A comprehensive review of YOLO: From YOLOv1 to YOLOv8 and beyond. *arXiv:2304.00501* (2023).
- [68] The Royal Institution for the Advancement of Learning (McGill University). 2017. Autour. <http://autour.mcgill.ca/en/>
- [69] F. Van Diggelen and P. Enge. 2015. The world’s first GPS MOOC and worldwide laboratory using smartphones, Vol. 1. 361–369.
- [70] Charles Viecek, Patricia McLain, and Michael Murphy. 1993. GPS/dead reckoning for vehicle tracking in the "urban canyon" environment. In *Proceedings of VNIS'93-Vehicle Navigation and Information Systems Conference*. IEEE, 461–34.
- [71] Hsueh-Cheng Wang, Robert K. Katzschmann, Santani Teng, Brandon Araki, Laura Giarre, and Daniela Rus. 2017. Enabling independent navigation for visually impaired people through a wearable vision-based feedback system. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, Singapore, Singapore, 6533–6540. <https://doi.org/10.1109/ICRA.2017.7989772>
- [72] Shiyun Yang, Emily Bailey, Zhengye Yang, Jonatan Ostrometzky, Gil Zussman, Ivan Seskar, and Zoran Kostic. 2020. COSMOS smart intersection: Edge compute and communications for bird’s eye object tracking. In *Proc. SmartEdge*.
- [73] Chris Yoon, Ryan Louie, Jeremy Ryan, MinhKhang Vu, Hyegi Bang, William Derksen, and Paul Ruvolo. 2019. Leveraging Augmented Reality to Create Apps for People with Visual Disabilities: A Case Study in Indoor Navigation. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, Pittsburgh PA USA, 210–221. <https://doi.org/10.1145/3308561.3353788>
- [74] Yuejin Zhang, Fangyao Liu, Zhiqiang Gu, Zhengxin Chen, Yong Shi, and Aihua Li. 2019. Research on Smart City Evaluation Based on Hierarchy of Needs. *Procedia Computer Science* 162 (2019), 467–474. <https://doi.org/10.1016/j.procs.2019.12.012>

## APPENDIX

### A FORMATIVE INTERVIEWS WITH BLV PEOPLE

We provide details on the semi-structured interviews with BLV participants that we conducted to identify challenges that they face when navigating outdoors using GPS-based systems.

#### A.1 Methods

**Participants.** We recruited six BLV participants (three males, three females; aged 29–66) by posting on social media platforms and snowball sampling [22]. Table 1 summarises the participants’ information. All interviews were conducted over Zoom and lasted about 60 minutes. Participants were compensated \$25 for this IRB-approved study. We obtained informed consent from all study participants.

**Procedure.** To identify the specific challenges that BLV people face when navigating outdoors, we used a recent critical incident technique (CIT) [19], in which we asked participants to recall and describe a recent time when they navigated outdoor environments using GPS-based assistive technology (AT). For example, we first asked participants to name the AT they commonly use and then asked them to elaborate on their recent experience of using it: “So, you mentioned using BlindSquare a lot. When was the last time you used it?” Then, we initiated a discussion by establishing the scenario for them: “Now, let’s walk through your visit from the office to this restaurant. Suppose, I spotted you at your office. What would I observe? Let’s start with you getting out of your office building.” We asked follow-up questions to gain insights into what made the aspects of outdoor navigation challenging and what additional information could help address them.

**Interview Analysis.** To analyze the interviews, we first transcribed the study sessions in full and then performed thematic analysis [10] involving three members of our research team. Each researcher first independently went through the interview transcripts and used NVivo [50] to create an initial set of codes. Then, all three iterated on the codes together to identify emerging themes.

**Table 1: Self-reported demographics of our participants. Gender information was collected as a free response; our participants identified themselves as female (F) or male (M). Participants rated their assistive technology (AT) familiarity on a scale of 1–5.**

PID	Age	Gender	Race	Occupation	Vision ability	Onset	Mobility aid	AT familiarity (1–5)
F1	29	Female	White	Claims expert	Totally blind	At birth	White cane	3: Moderately familiar
F2	61	Female	White	Retired	Light perception only	Age 6	Guide dog	1: Not at all familiar
F3	66	Female	White	Retired	Totally blind	Age 58	Guide dog	2: Slightly familiar
F4	48	Male	Black	Unemployed	Light perception only	Age 32	White cane	3: Moderately familiar
F5	27	Male	Mixed	Unemployed	Totally blind	At birth	White cane	3: Moderately familiar
F6	38	Male	White	AT instructor	Totally blind	At birth	White cane	5: Extremely familiar

## A.2 Findings

We found three major themes around challenges that BLV pedestrians face when navigating outdoors using GPS-based systems.

**C1: Routing through complex environment layouts.** Participants reported difficulties in following routing instructions provided by GPS-based systems. These instructions, as explained by the participants, often did not match their current location. Many participants cited problems such as making wrong turns into unexpected “alleyways” (F1, F2, F4) that landed them in dangerous situations with “cars coming through” (F2). Participants cited examples of how these instructions caused them to veer off course—a common issue for BLV individuals in open, outdoor spaces [53]—and end up in the middle of the streets. This problem was particularly pronounced in complex environment layouts, as F3 recalled: *“I didn’t know if crosswalks were straight or curved or if they were angled. [It was hard] to figure out which way you needed to be to be in the crosswalk.”* Since *“not everything is organized in the ideal grid-like way”* (F1), participants were hesitant to act on the navigation instructions without a clear understanding of the layout.

**C2: Avoiding unexpected obstacles while using GPS-based systems.** BLV people’s challenges relating to obstacles during navigation are well researched [54, 56]. However, we found specific nuances in their difficulties, particularly when they rely on their conventional mobility aids in conjunction with GPS-based navigation systems. Participants commonly reported the use of mobility aids like white canes alongside GPS systems for guidance. During this combined navigation process, they encountered difficulties in maintaining their focus on avoiding obstacles, often resulting in collisions with objects that they would have otherwise detected using their white canes. For instance, F2 shared an incident where they remarked, *“there were traffic cones [and] I tripped over those”* while following directions from BlindSquare [44]. Notably, moving obstacles such as pedestrians and cars, as well as temporarily positioned stationary obstacles like triangle sandwich board signs, posed significant challenges for navigation. F4 expressed this sentiment, stating, *“You know how many times I’ve walked into the sides of cars even though I have the right of way. Drivers have gotten angry, accusing me of scratching their vehicles. It can spoil your day [and make] you feel insecure.”*

**C3: Crossing street intersections safely.** Consistent with prior research [3, 26, 43], our study participants highlighted that crossing

streets remained a significant challenge for them. Since GPS-based systems do not help with street-crossing, most participants relied on their auditory senses and apps like Oco [9]. Regarding the use of auditory senses, they mentioned the practice of listening to vehicular sounds to gauge traffic flow on streets running parallel and perpendicular to their position. This auditory technique helped them assess when it was safe to cross streets. However, participants also reported instances where this method proved inadequate due to external factors: *“yeah, it can be tricky, because [there may be] really loud construction nearby that can definitely throw me off because I’m trying to listen to the traffic”* (F1). Furthermore, their confidence in street-crossing decisions was affected by their inability to ascertain the duration of pedestrian signals and the length of the crosswalk. This uncertainty led to apprehension, as they expressed a fear of becoming stranded mid-crossing, as exemplified by one participant’s comment: *“I don’t want to be caught in the middle [of the street]”* (F4). Regarding the use of Oco [9], participants found it cumbersome to point their phone’s camera toward a pedestrian signal and to switch between this app and others during navigation.

## B PARTICIPANT DEMOGRAPHICS

Table 3 summarizes demographics of various stakeholders we interviewed (Section 4), and Table 2 summarizes our user study participant demographics (Section 6).

## C STREETNAV: TECHNICAL SETUP

Figure 2 shows the street camera we used for developing and evaluating StreetNav. The camera is part of the NSF PAWR COSMOS wireless edge-cloud testbed [60, 72], and is available to researchers after an approval process and IRB review. We considered other publicly available testbeds such as Mobintel [45] and DataCity SMTG [14], but chose COSMOS due to its location in a major city (New York) with high pedestrian and vehicle traffic. Anonymized video samples from the COSMOS cameras, including the one used in this work, can be found online [13]. StreetNav’s computer vision pipeline takes the real-time video feed from the camera as input. For this purpose, we deployed the computer vision pipeline on one of the testbed servers, which captures the camera’s video feed in real time. This server runs Ubuntu 20.04 with an Intel Xeon CPU@2.60GHz and an Nvidia V100 GPU.

StreetNav’s two components—the computer vision pipeline and the app—interact with each other via a cloud server, sharing information using the MQTT messaging protocol [47]. Since MQTT is

**Table 2: Self-reported demographics of our user study participants. Gender information was collected as a free response. Participants rated their familiarity with assistive technology (AT) on a scale of 1–5.**

PID	Age	Gender	Occupation	Race	Vision ability	Onset	Mobility aid	AT familiarity (1–5)
P1	24	Male	App developer	Asian	Low vision	Age 19	White cane	2: Slightly familiar
P2	28	Male	Data manager	White	Low vision	At birth	None	3: Moderately familiar
P3	48	Male	Not employed	Black	Totally blind	Age 32	White cane	3: Moderately familiar
P4	46	Female	Social worker	Latino	Totally blind	Age 40	White cane	4: Very familiar
P5	43	Female	Not employed	Asian	Totally blind	At birth	White cane	4: Very familiar
P6	52	Male	Mgmt. analyst	Mixed	Light perception only	Age 9	White cane	5: Extremely familiar
P7	26	Female	Writer	Mixed	Low vision	At birth	White cane	2: Slightly familiar
P8	51	Male	Not employed	Black	Light perception only	Age 26	Guide dog	3: Moderately familiar

**Table 3: Self-reported demographics of our formative interviews with various stakeholders.**

PID	Stakeholder Category	Gender	Age	Notes
B1	BLV individual	Female	62	Light perception only
B2	BLV individual	Gender Neutral	41	Limited vision in only left eye
R1	Local resident	Female	29	Lived in Harlem for 12+ years
R2	Local resident	Female	35	Lived in Harlem for 13+ years
O1	Local business owner	Male	58	Running for 7+ years
CB1	Community Board leader	Male	53	Serving as leader in Harlem

a lightweight messaging protocol, it runs efficiently even in low-bandwidth environments. The computer vision pipeline only sends processed navigation information (e.g., routing instructions, obstacle’s category and location) to the app, rather than sending video data. This alleviates the privacy concerns around streaming the video feed to the users and avoids any computational bottlenecks that may happen due to smartphones’ limited processing capabilities. The StreetNav app’s primary purpose is to act as an interface between the user and the computer vision pipeline. We developed StreetNav’s iOS App using Swift [6], enabling us to leverage VoiceOver [7] and other built-in accessibility features.

## D ADDITIONAL USER STUDY RESULTS

### D.1 Results for Veering Prevention

Figure 14 shows participants’ average rating for their perceived ability to (1) maintain a straight walking path, i.e., prevent veering off course, and (2) intuitiveness of the feedback they received regarding direction to move in. The mean ( $\pm$  std. dev.) rating of participants’ perceived ability to maintain a straight walking path with StreetNav was 4.63 ( $\pm 0.52$ ) and with BlindSquare was 2.75 ( $\pm 1.17$ ). The condition had a significant main effect ( $p = 0.001$ ) on participants’ perceived ability to prevent veering off course. The mean ( $\pm$  std. dev.) rating for intuitiveness of the feedback that helped them know which direction to move in was 4.63 ( $\pm 0.52$ ) for StreetNav and 3.00 ( $\pm 0.76$ ) for BlindSquare. The condition had a significant main effect ( $p = 0.006$ ) on intuitiveness of feedback that helped participants prevent veering off path.

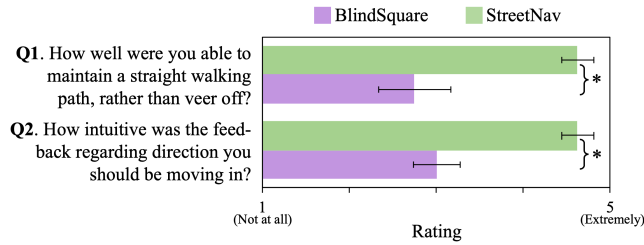
Our examination of the video recordings aligns closely with participants’ ratings. It reveals that StreetNav minimized participants’ deviations from the shortest path to the destinations in comparison to BlindSquare. Over the course of the three routes, participants displayed an average deviation from shortest path, that was reduced by 53% when using StreetNav as opposed to BlindSquare.

With BlindSquare, many participants reported difficulty maintaining awareness of their surroundings, including both obstacles and navigation direction, which frequently led to deviations from their intended paths. For instance, P2 reported challenges in maintaining their orientation with the need to avoid obstacles:

*“[BlindSquare] basically demanded me to keep track of my orientation as I was moving, which is pretty difficult to do when you’re also trying to keep other things in mind, like not bumping into things.” –P6*

In contrast, StreetNav effectively addressed this challenge by providing continuous audiohaptic feedback for maintaining a straight walking path, instilling a sense of confidence in participants. P3, who tested StreetNav before BlindSquare, reflected on their desire for a similar continuous feedback mechanism within BlindSquare, akin to the experience they had with StreetNav:

*“[with BlindSquare] even though I couldn’t see the phone screen, my eyes actually went towards where I’m holding the screen. It is almost as if on a subconscious level, I was trying to get more feedback. With [StreetNav] I had enough feedback.” –P3*



**Figure 14: Results for participants’ perceived ability to prevent veering off path. Participants rated their ability to (1) maintain a straight walking path, and (2) intuitiveness of the feedback regarding direction they should be moving in; on a scale of 1–5. StreetNav’s audiohaptic feedback was significantly more intuitive than BlindSquare’s in preventing participants from veer off path. Pairwise significance is depicted for  $p < 0.01$  (\*). The error bars indicate standard error.**

Many participants appreciated StreetNav’s choice of haptic feedback for veering. Some participants envisioned the haptic feedback to be especially useful in environments with complex layouts:

*“In the [areas] where the streets are very slanted and confusing, I think haptic feedback will be especially helpful.” –P5*

Other participants highlighted the advantage of haptic feedback in noisy environments where audio and speech feedback might be less effective.

However, both P4 and P6 exclaimed that StreetNav’s haptic feedback would only work well when holding the phone in their hands. This meant that hands-free operation of the app may not be possible, which is important for BLV people since one of their hands is always occupied by the white cane. P4 proposed integrating the app with their smartwatch for rendering the haptic feedback to enable hands-free operation.

## D.2 Forced Ranking Results

All eight participants unanimously chose StreetNav over BlindSquare as their preferred navigation assistance system. We asked participants to also rank their preferred type of routing instructions. All eight participants strongly preferred StreetNav’s turn-by-turn routing instructions compared to BlindSquare’s “as the crow flies,” direction and distance-style routing instructions.

In the semi-structured interview, participants were asked to elaborate on their rankings. Participants pointed out multiple navigation gaps in BlindSquare, with P2 summarizing participants’ sentiment:

*“If you’re only getting somebody 90 percent of the way there, you’re not really achieving what I would consider to be the prime functionality of the system.” –P2*

In contrast, participants praised StreetNav for its precision and real-time feedback, emphasizing the importance of granular and holistic information to support all facets of navigation. However, participants did acknowledge occasional “glitchiness” (P7) with StreetNav, which occurred when they moved out of the camera’s field of view or were occluded by other pedestrians or vehicles, resulting in lost

tracking. Nevertheless, participants still regarded StreetNav as a significant enhancement to their typical navigation experiences, expressing increased confidence in exploring unfamiliar outdoor environments in the future.

*“It would encourage me to do things that I would not usually... It would make me more confident about going out by myself.” –P4*

Participants also appreciated StreetNav’s ability to identify them in near real-time:

*“What I found very interesting about the connection part is how quickly it identifies where I am, as soon as I waved my hand, it senses me.” –P3*

Participants also provided suggestions for improving StreetNav. Some participants wanted a hands-free version that would allow them to hold a white cane in one hand while keeping the other free. Additionally, while they found the gesture of waving hands for connecting with the system socially acceptable, they acknowledged that it might be perceived as somewhat awkward by others in the street.

*“[Waving a hand] may seem kind of weird to people who don’t understand what is going on. But for me personally, I have no issue.” –P3*

Some participants highlighted that waving a hand might be misinterpreted by others on the street as a call for help, and may even cause security issues if a malicious person becomes aware that they were blind. P1 highlighted the role of public education in addressing this concern:

*“If [others] see someone with a white cane, they know that’s a blind person traveling. But if they see someone with their hand raised, they might think someone needs help or hailing a cab. So, I think it’s going to be education to other people as much as to the person who is using this navigation system.” –P1*

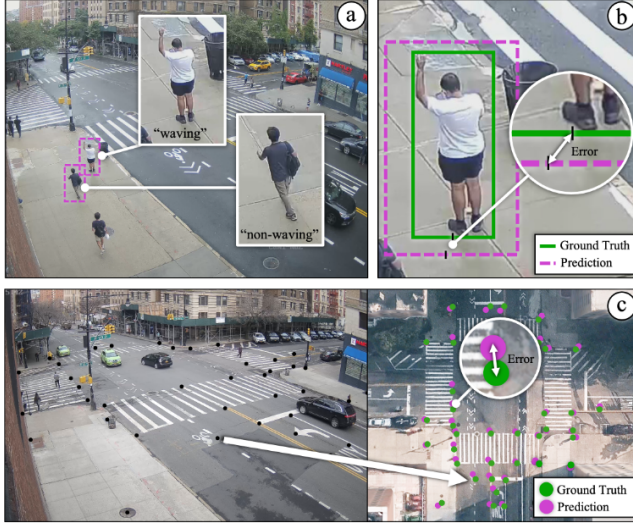
## D.3 How Individual Experiences Influenced Participants’ Preferences

Throughout the study, participants offered feedback based on their unique backgrounds. We observed distinct patterns in their preferences, affected by their (i) onset of vision impairment, (ii) level of vision impairment, and (iii) familiarity with assistive technology.

**Onset of vision impairment.** Participants with early onset blindness preferred nuanced, concise feedback with an emphasis on environmental awareness. They used the system as an additional data point without complete reliance. In contrast, participants with late onset blindness trusted the system more and relied heavily on its feedback.

**Level of vision impairment.** Totally blind participants appreciated the veering feedback, while low-vision users, having more visual information, relied on their senses and needed less assistance with veering. Low-vision participants preferred the street-crossing feedback to interpreting pedestrian signals across the street. Totally blind participants primarily listened for parallel traffic, their usual method, using StreetNav’s feedback for confirmation.





**Figure 15: Illustration of StreetNav’s localization steps analyzed in the technical evaluation: (a) CLIP-based gesture recognition, (b) pedestrian feet position estimation, (c) camera to map-view transformation.**

**Familiarity with assistive technology (AT).** We noticed that participants who commonly use AT for navigation quickly adapted to StreetNav, while those with less experience hesitated in trusting StreetNav’s feedback and had a slightly steeper learning curve. Still, all participants mentioned feeling more comfortable with StreetNav as the study progressed. Both groups also expressed increased confidence in exploring new areas with StreetNav.

## E TECHNICAL EVALUATION

We independently analyzed the technical performance of each of the three steps that enable StreetNav’s computer vision pipeline to localize the user. Figure 15 illustrates the three steps: (i) CLIP-based gesture recognition (Figure 15a), (ii) pedestrian feet position estimation (Figure 15b), and (iii) camera to map-view transformation (Figure 15c). Recall from Section 5.2, StreetNav first distinguishes the BLV pedestrians from other pedestrians by recognizing the hand-waving gesture, then estimates their feet position as the midpoint of bounding box’s bottom edge, and finally transforms their feet position from the camera view to the map.

### E.1 Procedure

We recorded a 15-minute evaluation video (22500 frames) from the camera feed to perform the technical evaluation. While recording this video, researchers posed as users navigating through the street intersection and played out different scenarios, such as waving hands and crossing streets. We also analyzed the errors for each of the three steps, revealing factors that impact StreetNav’s ability to precisely determine a user’s position.

### E.2 Results

**CLIP-based gesture recognition.** To evaluate the first step, we randomly sampled a balanced dataset of 140 image crops from the

evaluation video. Figure 15a highlights the pedestrian image crops from each class. The CLIP-based gesture recognition module classifies each crop as waving or non-waving (i.e., walking, standing) pedestrian.

Prediction \ Ground Truth	Waving	Non-waving
Waving	63	17
Non-waving	7	53

Count

60  
50  
40  
30  
20  
10

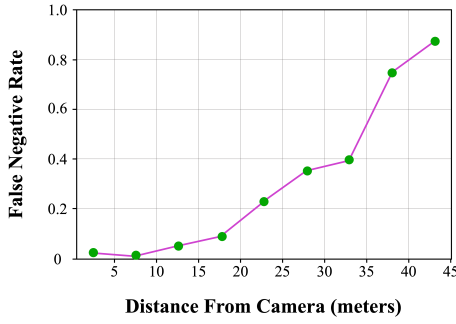
**Figure 16: Confusion matrix for StreetNav’s CLIP-based gesture recognition module. StreetNav distinguishes waving pedestrians from non-waving (i.e., walking, standing) ones with an 83% accuracy.**



**Figure 17: Failure cases in StreetNav’s CLIP-based gesture recognition module. False positives occur when other pedestrians perform actions similar to waving their hand, such as (a) talking over phone or (b) casually resting their hand on forehead. False negatives occur when (c) users are too far from the camera and (d) due to foreground occlusions and background overlaps with vehicles, scaffolding, or other pedestrians.**

Figure 16 shows the confusion matrix for CLIP-based gesture recognition module’s performance. StreetNav achieves an 83% accuracy in recognizing the hand-waving gesture, with a false positive rate of 24% and a false negative rate of 10%. We analyzed the failure cases to identify specific scenarios that lead to the errors.

Figure 17 shows instances of the most common scenarios leading to false positives and false negatives. The false positives occur when other pedestrians perform actions similar to waving their hand, such as talking over a phone (Figure 17a) or casually resting their hand on their forehead (Figure 17b). The false negatives occur when users are too far from the camera (Figure 17c) or due to foreground occlusions and background overlaps such as vehicles, scaffolding, and other pedestrians (Figure 17d). While false negatives may result in users needing to wave their hands for a longer duration until recognized, false positives can lead them to follow incorrect instructions based on another pedestrian’s location. StreetNav’s approach to mitigating false positives is to announce the relative location of the detected pedestrian (e.g., ‘southwest corner’), providing users



**Figure 18: False negative rate (FNR) for pedestrian detection over distance from the camera in meters. StreetNav’s error rates in detecting pedestrians increases significantly as they get further away from the camera. The FNR goes up from 1% at 5 meters to 74% at 40 meters distance from the camera.**

with additional contextual information to confirm whether they were recognized. The idea is that if this information does not align with the user’s perception, they could then choose to re-establish the connection. Fine-tuning the CLIP model for this purpose could potentially enhance accuracy even further.

**Pedestrian feet position estimation.** To evaluate the second step, we manually annotated the ground truth pedestrian bounding boxes for 250 frames, randomly sampled from the evaluation video. Figure 15b shows the ground truth bounding box and StreetNav’s predicted bounding box for a pedestrian. We report the root mean square errors between the feet positions estimated using the ground truth and predicted bounding boxes.

The root mean squared error ( $\pm$  std.) in estimating pedestrians’ feet position is 0.41 ( $\pm$  1.49) meters. The pixel distances were converted to physical distances to obtain the error in meters. We observed larger error rates for scenarios where pedestrians are occluded by other pedestrians or objects such as trash cans and fire hydrants. Future approaches could explore filtering abrupt changes in pedestrians’ bounding boxes, caused by occlusions, to reduce this error.

While analyzing the feet positions from the bounding boxes, we also noticed a trend in StreetNav’s pedestrian detection pipeline.

Recall from Section 5.2, StreetNav uses Nvidia’s DCF-based multi-object tracker [49] and the YOLOv8 object detector [67] for tracking pedestrians. We found that although StreetNav detects pedestrians with an 82% precision and 65% recall at 0.5 IOU (intersection over union) threshold, the accuracy drops significantly as the pedestrian’s distance from the camera increases. This is attributed to the relatively smaller size of pedestrians, low resolution, and high chances of occlusion as pedestrians move further away from the camera.

Figure 18 shows the false negative rate over distance from the camera. The false negative rate increases from 1% at a distance of 5 meters from the camera to 74% at a distance of 40 meters from the camera. Note that the distances were calculated between the pedestrian’s feet estimations and the camera position’s projection on the ground. Future approaches could combine detections from multiple cameras, such as two cameras positioned diagonally across a street intersection, to address this drop in accuracy. Alternatively, using training strategies that can detect both small and large pedestrians could also improve performance [29].

**Camera to map-view transformation.** To evaluate the third step, we selected a dataset of 50 points in the camera view and we manually annotated their corresponding position on the map. We chose these specific points for evaluation as they correspond to visual landmarks on the street and are evenly spread across the street intersection. For example, we selected points on the crosswalk edges and road signs. As a result, annotating their ground truth position on the map view could be done with reasonable accuracy by simply comparing the camera and map view images. For these 50 points, we also generated StreetNav’s predicted transformations from the camera view to the map view. Figure 15c shows the points we selected and their corresponding ground truth and predicted transformations. We computed the root mean square errors between the transformed ground truth positions and StreetNav’s predicted positions.

The root mean squared error ( $\pm$  std.) in transforming points from the camera view to the map view, averaged across the points shown in Figure 15c, is 0.65 ( $\pm$  0.26) meters. The pixel distances were converted to physical distances to obtain the error in meters. These errors occur due to the curvature in the camera lens.