HHS Public Access

Author manuscript

*DIS (Des Interact Syst Conf)*. Author manuscript; available in PMC 2025 January 13.

Published in final edited form as:

*DIS (Des Interact Syst Conf)*. 2022 June ; 2022: 881–897. doi:10.1145/3532106.3533560.

# Helping Helpers: Supporting Volunteers in Remote Sighted Assistance with Augmented Reality Maps

**Jingyi Xie***,
Pennsylvania State University, University Park, PA, USA

**Rui Yu***,
Pennsylvania State University, University Park, PA, USA

**Sooyeon Lee**,
Rochester Institute of Technology, Rochester, NY, USA

**Yao Lyu**,
Pennsylvania State University, University Park, PA, USA

**Syed Masum Billah**,
Pennsylvania State University, University Park, PA, USA

**John M. Carroll**
Pennsylvania State University, University Park, PA, USA

## Abstract

Remote sighted assistance (RSA) has emerged as a conversational assistive service, where remote sighted workers, *i.e.*, agents, provide real-time assistance to blind users via video-chat-like communication. Prior work identified several challenges for the agents to provide navigational assistance to users and proposed computer vision-mediated RSA service to address those challenges. We present an interactive system implementing a high-fidelity prototype of RSA service using augmented reality (AR) maps with localization and virtual elements placement capabilities. The paper also presents a confederate-based study design to evaluate the effects of AR maps with 13 untrained agents. The study revealed that, compared to baseline RSA, agents were significantly faster in providing indoor navigational assistance to a confederate playing the role of users, and agents' mental workload was significantly reduced—all indicate the feasibility and scalability of AR maps in RSA services.

### Keywords

People with visual impairments; blind; remote sighted assistance; computer vision; augmented reality; 3D map; navigation; camera; smartphone

jzx5099@psu.edu .

*Both authors contributed equally to this research.

## 1 INTRODUCTION

Remote sighted assistance (RSA) is an assistive service, where a blind user establishes a video connection with a remotely located sighted assistant (namely, *agent*), who interprets the user's camera feed in real-time and converses with them to achieve complex navigational tasks [61]. Examples of RSA services include VizWiz [23], BeSpecular [46], Crowdviz [45] from academic research; and TapTapSee [87], BeMyEyes [20], and Aira [13] from the tech industry.

Existing RSA services have varying popularity, availability, affordability, and user satisfaction. For example, BeMyEyes [20] is a free and affordable RSA service with over 0.3 million blind and low-vision users and 5 million sighted volunteers worldwide [21]. However, their nonprofit volunteering offers no guarantee of service availability or quality [17, 33, 35]—sighted volunteers only receive calls during the daytime [3] and are not required to have any training in orientation & mobility (O&M) [31, 66, 74]. In contrast, Aira [13], a paid RSA service that operates primarily in the US and other English-speaking countries, provides a higher quality service with trained sighted agents who are always available [4, 61], but is less affordable (i.e., costing $40 to $60 USD per hour). As a result, many blind users utilize free services like BeMyEyes for everyday tasks, such as asking for a dress color, reading text, finding household items [17, 31] and rely on a paid service like Aira for high-stake tasks, such as navigating airports [61, 91].

The performance of trained and untrained agents has been studied and compared in prior work. Avila et al. [17], for example, reported that users reacted negatively to assistance delay and the quality of wayfinding tasks in BeMyEyes services. Similarly, Kameswaran et al. [49] found that BeMyEyes volunteers were unreliable and unable to provide precise instructions in navigation tasks, which challenged users' safety and level of confidence.

Although the expertise of trained agents leads to more robust and preferable assistance, limited access to users' visual field through the smartphone's camera and location information can pose a substantial challenge to all agents, regardless of their training [50, 62]. To address these challenges, the current generation of RSA services leveraged Google Maps and GPS location and sometimes utilized a pair of specialized glasses with wide-angle lenses for the users to increase their visual field [49]. These measures enable agents to deliver directional information preemptively and subsequently improve the user experience in outdoor navigation [60]. However, providing navigational assistance indoors remains a key challenge for the agents for several reasons, such as weaker GPS signal, scarcity of indoor maps with finer details, and agents' lack of familiarity with the users' current physical environment [62, 91]. In addition, researchers have identified several open challenges for the agents common to providing navigational assistance outdoors and indoors. These challenges include the agent's difficulty in continuously tracking and orienting the users on the map and within their (users') surroundings; estimating objects' depth, detecting landmarks and obstacles in users' camera feed; and interpreting and delivering the visual information in real-time via conversations [50, 62]. Unfortunately, the current generation of RSA services is limited in addressing these challenges, which, however, impact the service quality of the agents, especially agents who are untrained or work as volunteers.

To address the above challenges, our prior work has envisioned assisting agents with computer vision (CV) technologies, particularly 3D map construction, object annotations, real-time localization, and video stream augmentation [32, 62, 91]. We also designed a series of low-fidelity prototypes (e.g., pictorials) illustrating how CV technologies can be adopted in RSA services [91]. Trained agents who evaluated our design found that, if implemented, the proposed design could enhance their ability to stay ahead and reduce their mental workload.

This paper first presents an interactive system demonstrating how to implement a high-fidelity, functional prototype of computer vision-mediated RSA service, as envisioned in our prior work [91], in a controlled lab environment. The system contains a pair of mobile apps, namely, *RSA-User* for the users and *RSA-Agent* for the agents. Both apps utilized iOS's augmented reality (AR) library, namely, ARKit [8], which is available on recent iPhone and iPad devices. A similar library is also available on Android devices. In order to reduce bootstrapping efforts, we created AR maps of the lab environment by scanning it with RSA-User app and later augmenting the maps with virtual elements (e.g., distance band and room numbers) offline (see Figure 1). During indoor navigation, RSA-User app can achieve accurate real-time localization on AR maps with iOS devices equipped with a LiDAR scanner[1]. RSA-Agent app accessed the same AR map, where the position and orientation of the user are updated in real-time. Moreover, to reduce communication latency and implementation complexity, we set up a private Wi-Fi network, leveraged an off-the-shelf video chat app (e.g., Skype) to establish an audio call, and mirrored the user's screen onto the agent's laptop to transmit the live, augmented camera feed.

Next, this paper presents a confederate-based study design [15] to evaluate the effects of CV technologies in RSA service. The design was informed by an expert blind advisor who collaboratively reviewed and refined the system as well as the study procedure and tasks. This blind advisor represents the target user of our system, and our collaboration with the advisor evoked shared empathy, which is essential in developing technology for people with disabilities [22]. During an early evaluation, the advisor observed that blind participants could find the study tasks laborious to perform due to the deterministic and repetitive nature of the tasks. In addition, the advisor indicated that blind participants could develop a mental map of the testing sites using their O&M skills, which could impact the study results. As such, we adopted a confederate-based design, where the confederate is an individual who participates in the experiment but is not observed by the researchers [47].

In our study, the confederate was a sighted participant who played the role of a user who had recently lost vision and had not acquired sufficient O&M skills. The confederate wore a blindfold during a task and was equipped with a white cane. Furthermore, we recruited 13 sighted participants to play the role of sighted volunteers or untrained remote-sighted assistants (i.e., agents).

Our study revealed that, compared to baseline RSA, the agents were significantly faster in providing indoor navigational assistance to the confederate, and the agents' mental workload

[1]Up to January 2022, the iPhone 12 & 13 Pro, Pro Max, 2020 and 2021 iPad Pro featured a 3D LiDAR scanner.

was significantly reduced—all are indicatives of the feasibility and scalability of AR maps in RSA services.

We note that blind users are not the direct consumer of AR maps in the proposed CV-mediated RSA service. However, the above benefits for the agents are likely to translate for blind users in faster call resolutions, richer conversation, and overall increased capacity to handle their service requests.

In summary, we make 3 contributions:

- An interactive system implementing a high-fidelity RSA prototype using augmented reality-based 3D maps in lab settings with manageable efforts.
- A controlled laboratory study based on confederate design and role-playing to evaluate the implemented system; and
- Preliminary evidence that 3D maps and localization benefit RSA agents, especially untrained sighted volunteers to provide indoor navigational assistance.

## 2 BACKGROUND AND RELATED WORK

### 2.1 Remote Sighted Assistance Services for People with Visual Impairments

The implementation of various RSA services differs in three key areas: (i) the communication medium between users and remote sighted assistants. Earlier prototypes used audio [72], images [23, 57], one-way video using portable digital cameras [30, 41], or webcams [30], whereas the recent ones are using two-way video with smartphones [13, 19, 20, 44]; (ii) the instruction form, e.g., via texts [59], synthetic speech [72], natural conversation [13, 19, 20], or vibrotactile feedback [34, 82]; and (iii) localization technique, e.g., via GPS-sensor, crowdsourcing images or videos [23, 58, 76, 93], fusing sensors [76], or using CV as discussed in the next subsection.

Researchers examined the feasibility of crowdsourced RSA services (e.g., TapTapSee [87], BeMyEyes [20]), and concluded that this is a promising direction to tackle navigation challenges for blind users [17, 24]. However, they commented on the issue of crowdworkers not being available at times [31]. Nguyen et al. [70] and Lee at al. [61] studied a paid RSA service, Aira [13]. They reported that unlike crowdworkers, Aira agents are always available and trained in communication terminology and etiquette. In this paper, we propose a new RSA prototype to study the feasibility of AR maps with CV techniques to improve existing RSA services for indoor navigation.

### 2.2 Use of Computer Vision in Navigation for People with Visual Impairments

Budrionis et al. [29] reported that CV-based navigation apps on smartphones are a cost-effective solution for indoor navigation. A major focus on CV-based approach is how to make visual information more accessible through recognizing objects [94], obstacles [75], color-codes or landmark (e.g., storefronts [81]), or through processing of tags such as barcodes [88], QR codes [37], and RFID [69]. Extending this focus, researchers have proposed indoor positioning and navigation systems [55, 63, 67]. However, Saha et al.

[81] concluded that for a deployable level of accuracy, using CV techniques alone is not sufficient yet. In this work, we use CV and 3D maps to assist sighted assistants rather than users with visual impairments, who could be vulnerable to inaccuracies of CV systems.

Another line of work is to develop autonomous location-aware pedestrian navigation systems. These systems combine CV with specialized hardware (e.g., suitcase [51], wearable CV device [65]) and sensors (e.g., Lidar [43], Bluetooth [42]), and support collision avoidance. However, their real-world adaptability is still questionable, as Banovic et al. [18] commented that navigation environments in real-world are dynamic and ever-changing.

Lately, researchers are exploring the potential of AR toolkit in indoor navigation. This toolkit is built into smartphones (e.g., ARKit [8] in iOS devices, ARCore [2] in Android devices), thus has the possibility of a widespread deployment [78]. Clew demonstrated the potential of using AR toolkit to localize blind users on a pre-recorded route with acceptable accuracy [92]. Verma et al. [90] reported that an AR-based navigation system could provide a better user experience than traditional 2D maps. Brata et al. [27] found that sighted participants rate 2D digital map over a location-absed AR interface for efficiency and dependability, but AR interface over 2D digital map for hedonic qualities and overall performance. Fusco et al. [38] also reported that with ARKit, users with visual impairments do not need to keep the camera towards a target to recognize it. Troncoso Aldas et al. [89] proposed an ARKit-based app, AIGuide, to help people with visual impairments to recognize and localize objects.

We proposed CV-mediated RSA service to assist sighted agents instead of blind users in our prior work [91], however, its implementation requires substantial bootstrapping efforts (e.g., creating and annotating 3D maps); specialized hardware (e.g., LiDAR scanner) and software technology (e.g., AR apps, CV detection models with high accuracy), infrastructure support (e.g., low-latency wireless communication and large-scale storage); and access to blind users and trained agents who can perform fixed tasks deterministically and repetitively. These requirements are major constraints for academic researchers to test new designs in RSA services. As a result, despite being promising, the implication of CV-mediated RSA design, in reality, is unknown. In this work, we relaxed certain constraints and evaluated the CV-mediated RSA in lab settings.

### 2.3 Collaboration between Human and AI

Despite the recent advancements in CV and automatic scene understanding, 3D reconstruction from video stream remains a challenge. The performance of existing systems is impacted by various factors, such as motion blur, change of light, scale, and orientation [48]. As a solution to this limitation, interactive, hybrid approaches that involve human-AI collaboration have been studied. Brady et al. [25] offered users several options for answer sources, including crowdsourced workers, to manually identify objects unrecognized by CV algorithms. An interesting variation of hybrid approaches is the human-in-the-loop framework. Branson et al. [26] used human responses to questions posed by the computer to drive up recognition accuracy while minimizing human effort. Some researchers studied interactive 3D modeling in which humans provide guidance by drawing simple outlines

or scribbles [56, 85]. Our approach is in line with the above research, but we explore the human-AI collaboration design space leveraging CV and 3D maps to support sighted human assistants.

### 2.4 Challenges of Evaluating Assistive Technologies

Prior research identified several challenges of evaluating assistive technologies. First, the evaluation can be complicated by small target user populations and insufficient number of participants [84, 86]. The variability between human participants with disabilities is also uncontrollable, such as age of onset, condition of vision impairment for participants with visual impairments [86].

Second, assistive technologies require substantial efforts to assess and adjust [71]. Those lack of implementation efforts will lead to abandonment. Causes for abandonment have many dimensions [68, 77], such as improper fit to target users, tasks [39] or users' changing requirements [73], and difficulties in configuring and modifying the settings [52]. In this study, several design decisions have been made to balance implementation efforts and feasibility of the prototype, such as recruitment of blind advisor, confederate design and role playing.

Third, functional testing, that is ensuring an assistive technology does what it is designed to do, is another challenge. For truly innovative technologies or treatment, deciding what to measure [84] and what alternative to be used as the control [86] are obstacles during the evaluation. Researchers also struggle with the evaluation settings, comparing and choosing between lab-based and field-based evaluations [53, 54]. Following recommendations in prior research [53, 54], we captured videos from different angles to provide high-quality data comparable to field studies.

## 3 COMPUTER VISION MEDIATED RSA PROTOTYPE SYSTEM

In this work, we aim to develop a high-fidelity CV-mediated RSA prototype system to study the feasibility and implication of designs for indoor navigation with 3D maps in our prior work [91]. To support the design, implementation, and experiment of our system, we recruited a blind advisor who is experienced in using RSA services for daily tasks (e.g., indoor navigation). With the guidance of the blind advisor, we developed a high-fidelity CV-mediated RSA prototype system with manageable implementation efforts in a lab setting.

### 3.1 Role of Blind Advisor in System Design

The recruitment of the blind advisor responded to the HCI community's call for empathy in technology development. It is principal especially in design for people with disabilities, arguing "being with" the target population rather than "being like" them [22]. The advisor (marked as A* in Table 1) in our project helped us build empathy for users with visual impairments.

The advisor was engaged in the system design process from beginning to end. In particular, the advisor helped us determine the design objectives, reviewed the implementation

decisions, run the prototype, and gave feedback from an expert blind user's perspective to ensure that we preserved the challenges in indoor navigation in our settings.

## 3.2 Design Objectives

According to the advice of the blind advisor and the findings from RSA professionals [91] and users with visual impairments [62], we identified the key functions of a CV-mediated RSA system for indoor navigation, which are listed as follows as the design objectives:

**3.2.1 Basic RSA functions.—**Existing RSA services (e.g., Aira, BeMyEyes) mainly support live video chat between the user and agent. Professional agents emphasize the importance of the live video feed and consider it as the *"lifeline of information"*. Although web-based information (e.g., maps and satellite images) is useful, live video and audio feed from blind users' mobile devices are the only resource of information that is real-time and guaranteed to be accurate and up-to-date [91]. The basic video chat function is common and can be executed with off-the-shelf tools. For example, Kamikubo et al. [50] realized the user-agent interaction with a video conferencing system.

**3.2.2 Interactive 3D maps.—**To address the lack of maps in indoor navigation, RSA agents expect to learn the environmental context of the users from 3D maps in an interactive manner similar to a web mapping platform (e.g., Google Maps) they have employed in existing RSA services for outdoor navigation [91]. The 3D maps also need to allow agents to change the scale and view via common interactions (e.g., zooming, panning, and rotating) to find both general and detailed information. In practice, 3D maps are usually saved as polygon meshes or point clouds in various formats (e.g., .obj, .dae, .ply), preferably with colored texture.

**3.2.3 Real-time localization.—**Real-time localization has been widely applied in wayfinding apps for people with and without visual impairments. These apps (e.g., Google Maps [5]) rely on GPS for localization, which is a mature technology for outdoor navigation. However, the weaker GPS signal strength and low accuracy render these apps unreliable in the indoor environment. As informed by prior study [91], continuously updating the user's position and orientation on a 3D map is desired by professional RSA agents for indoor navigation. To simplify the system, we prefer to realize real-time localization only using standalone mobile devices (phones or tablets) without extra environmental infrastructure (e.g., Bluetooth beacon [12], RFID tags [40]). 3D maps using AR technology have shown the potential to address this problem with smartphones [92]. With the rapid development of AR toolkits (e.g., Apple's ARKit, Google's ARCore) and more powerful sensors (e.g., LiDAR scanner) equipped in mobile devices, it is feasible to achieve reliable real-time localization with AR technology and mobile devices.

**3.2.4 AR Element 1: Room Numbers.—**Virtual annotations on 3D maps, such as room numbers in office buildings and aisles and sections in grocery stores, provide agents with more spatial details and inform them of possible paths. Agents considered annotations *"very helpful just to get a sense for the location and the store"* and help them to guide blind users more efficiently [91]. Moreover, supposing the destination is a specific room,

the annotation of the room number on 3D maps will greatly ease this challenging indoor navigation task.

**3.2.5 AR Element 2: Distance Bands.—**It is difficult for agents to estimate distance through a live video feed, especially considering differences in camera height and angle. Distance band is one of the methods to present distance information as a grid overlaid on maps. Agents gave positive feedback about this feature because it enhances their ability to stay ahead of blind users and prepare for approaching obstacles [91]. Compared with other distance measurements overlaid on the video feed, distance bands have the advantage of not obscuring the *"lifeline"* real-time video feed. In this prototype, we draw distance bands on the 3D maps with an interval of 10 feet.

**3.2.6 Dashboard with extra information.—**Professional agents indicated that they use split-screen or multiple monitors and *"sometime will have several maps open"* for reference to identify points of interest [91]. To support their habit, the RSA prototype will follow the split-screen dashboard design. Specifically, the 3D map as an additional information source will be displayed on a separate screen from the live video feed. The manipulation will mostly be performed on the interactive 3D map screen.

### 3.3 Implementation Trade-offs

In accordance with the design objectives, we developed a pair of iOS apps, namely RSA-User and RSA-Agent. Specifically, RSA-User app can reconstruct 3D maps by scanning the environment, save the 3D map, and conduct real-time localization; RSA-Agent app can load and display the 3D map, and receive real-time data (e.g., camera pose) sent from RSA-User app. We implemented RSA-User and RSA-Agent apps with Apple's ARKit [8], RealityKit [10], and SceneKit [11]. All the codes were developed in Swift and Xcode version 12.4. The RSA-User app only supports Apple mobile devices equipped with a LiDAR scanner.

We found it quite challenging to integrate all the design objectives into RSA-User and RSA-Agent apps. For example, it is arduous and time-consuming to develop a real-time video chat software (e.g., FaceTime, Skype) from scratch and integrate into our apps. Besides, we need to set up a server for data transmission.

To mitigate these challenges, we made several implementation decisions to fulfill the design objectives with limited resources and in a lab setting. First, we realized the basic RSA function by an existing video chat software (Skype). Since the RSA-User app occupied the camera for real-time localization, we made another implementation trade-off to mirror the user's screen to the agent's laptop via a private WiFi and only use Skype for voice chat. Second, we employed Apple's Multipeer Connectivity framework [9] for data transmission between RSA-User and RSA-Agent apps. The framework only works in a WiFi network and doesn't support out-of-range remote communication. So does the screen mirroring. This requires the agent shares the same WiFi with the user. We set up a private WiFi and covered the agents' eyes before leading them to the experimental area. When the agents gave instructions in a separate control room, the setting made them feel they were assisting a "remote" user. Third, AR elements (e.g., distance bands, landmark annotations) are design ideas in our prior work [91], which should appear as AR lines or characters upon the

live video feed. However, implementation of this function would take substantial effort. Instead, we created 3D maps with RSA-User app and added AR elements (distance bands and room annotations) to the maps with third-party tools (e.g., Blender [6]) as AR maps. These implementation decisions were reviewed and tested by the blind advisor, balancing the implementation efforts and feasibility of our prototype.

### 3.4 Prototype Workflow

Figure 1 shows the workflow of the implemented RSA prototype. The prototype consists of two phases: offline mapping and online RSA. In the offline mapping phase, we create 3D maps and save them in cloud storage. In the online RSA phase, the agent assists the user with both video chat and 3D maps. We briefly introduce the five steps in Figure 1 as follows and will detail the implementations in Section 3.5 and 3.6.

**Offsline Mapping:**

1. A sighted volunteer scans the indoor environments of interest by an iPad Pro with the RSA-User app.
2. The sighted volunteer saves the scanned 3D maps and ARWorldMap (containing mapping state and anchors for relocalization supported by Apple's ARKit [8]) files to cloud storage (e.g., Google Drive).
3. The sighted volunteer downloads the 3D map files to a laptop, adds annotations (e.g., room numbers and distance bands), converts the map files to SceneKit-readable format [11], and uploads to cloud storage.

**Online RSA:**

4. The user downloads the ARWorldMap file of the current surroundings to the RSA-User app for relocalization. The RSA-User app will continuously send the camera pose (position and orientation) to the RSA-Agent app.
5. The agent loads the 3D map of the user's current area and views the user's real-time location in the RSA-Agent app. Meanwhile, the user's iPad screen is mirrored to the agent's laptop as the live camera feed.

### 3.5 Implementation of Offline Mapping

**3.5.1 Build 3D Maps.—**There are two main reasons why we adopt Apple iOS devices for 3D mapping. First, Apple's ARKit facilitates camera relocalization by encapsulating the mapping state and anchors into an ARWorldMap object. This feature is crucial to the accurate real-time localization of RSA users in 3D maps. Second, the new Apple mobile devices featuring a LiDAR scanner, along with ARKit, support powerful 3D scene understanding capabilities and reconstructing high-quality polygon mesh as 3D maps.

In our experiment, the sighted volunteer used the RSA-User app to scan areas of interest with a 2020 iPad Pro. During scanning, the RSA-User app continuously constructed a 3D mesh and displayed it overlaying on surfaces in real scenes. Meanwhile, ARKit constantly updated an ARWorldMap object, which contains the space-mapping state and ARAnchor

objects for relocalization the next time. When completing the scanning, the sighted volunteer saved both the 3D mesh and ARWorldMap to cloud storage. The 3D mesh was extracted from the ARMeshAnchor object and exported to an .obj (Wavefront object) file with Model I/O [7]. ARMeshAnchor only contains an untextured mesh. In the preliminary test, we found the 3D maps without textures could not provide enough environmental context for indoor navigation, and the implementation of texture mapping was non-trivial. To quickly evaluate the proposed prototype, we adopted an open 3D scanner app [1] to generate textured mesh and manually aligned it to the untextured mesh with Blender [6]. Figure 2(a) and 3(a) show two 3D maps used in this study.

**3.5.2 Annotation and Format Conversion.—**By zooming in, we can easily recognize large objects (e.g., trash cans) in the generated 3D maps but not small signs (e.g., room numbers). While finding a specific room is a common task in indoor navigation, we propose to use Blender to annotate the room numbers on the 3D maps to facilitate path planning. As shown in Figure 2(a) and Figure 3(a), we add blue room numbers on the wall near the doors and the blank space behind the wall. We also use Blender to draw distance bands on the floor on 3D maps to help agents estimate distances. As shown in Figure 2(b) and Figure 3(b), the distance between two adjacent parallel white lines is 10 feet.

The exported 3D maps from Blender are in .dae (COLLADA) format. To directly load 3D maps from cloud storage in the RSA-Agent app, we converted the .dae file to .scn (SceneKit scene file) format and uploaded it to cloud storage to support online RSA services.

### 3.6 Implementation of Online RSA

**3.6.1 Real-time Localization.—**During RSA, the user relocalized the iPad after entering the map area. First, the user roughly scanned the environment with RSA-User app for about 10 seconds to create a new ARWorldMap object, and then loaded the previously saved ARWorldMap file from cloud storage. The RSA-User app would align the two ARWorldMap objects and transform the current world coordinate system to the previous one. During the navigation, RSA-User app was continuously performing scene understanding and maintaining the high-accuracy camera pose estimated from multiple sensors including a LiDAR scanner, RGB cameras, and an inertial measurement unit (IMU).

We employ Apple’s Multipeer Connectivity framework [9] to establish the connection between RSA-User and RSA-Agent apps. The user’s camera pose is ceaselessly sent from RSA-User to RSA-Agent via a private Wi-Fi. In RSA-Agent app, we use a magenta pyramid-shaped camera gizmo to represent the angle of view of the user’s camera, as shown in Figure 2(a) and Figure 3(a). After loading 3D AR maps to RSA-Agent app, the agent can observe the camera pose (position and orientation) of the user’s device on AR maps in real-time. The RSA-Agent interface supports standard touchscreen controls provided by SceneKit to browse the AR map: rotating the view with one-finger pan, translating the view with two-finger pan, zooming in/out with two-finger pinch or three-finger pan vertically, and resetting the view with double-tap. Figure 4 demonstrates how to find trash cans (Figure 4(a)) and water fountain (Figure 4(b)) in the AR maps from the top-down view with touchscreen gestures in RSA-Agent app.

**3.6.2 Live Video Chat.—**The RSA-User and RSA-Agent apps currently do not support video chat. To verify the prototype, we employ off-the-shelf videoconferencing solutions to realize the live video chat. Specifically, we use Skype voice calls for user-agent oral communication. In the experiments, the user hung the iPad Pro on the chest with a belt around the neck to capture the video in front. The user's iPad screen is mirrored to the agent's laptop via the private WiFi so that the agent can see the real-time camera feed from the laptop.

# 4 METHOD: CONFEDERATE DESIGN AND ROLE PLAYING

In this section, we will introduce the study method in terms of the role of the blind advisor, confederate design and role-playing, participants, environment, apparatus, task design, and procedure. We aim at investigating the feasibility of AR maps for RSA, compared to the 2D maps. We assess the performance of agents with task completion time and their self-evaluation. Our experiment is IRB-approved.

## 4.1 Role of Blind Advisor

Our research project was staffed by researchers responsible for conducting the user study as well as a blind advisor (A* in Table 1) who is experienced in using RSA services for daily tasks, including indoor navigation. Before the experiment, the advisor guided the confederate regarding how to use a white cane and instructed sighted participants (untrained agents) on visual interpretations, such as describing directional cues with contextual details and avoiding numerical terms. We consulted the advisor throughout the process of study design and data analysis. The advisor reviewed the task design and procedure, observed the experiment, and gave feedback from an expert blind user's perspective.

## 4.2 Confederate Design and Role Playing

We conducted a confederate-based, role-playing study to evaluate our prototype because, as suggested by the advisor, it is hard for blind users to conduct this repetitive task over time. In a dry run, we observed that the blind advisor could develop a mental map of indoor environments after completing a single task and she could return to the starting point without any instructions. Therefore, considering tasks need to be deterministic and repetitive in a controlled lab study, we cannot manipulate the learning effect if blind users engage in all four navigational tasks for each agent-user pair.

Instead, one sighted confederate (marked as U* in Table 1) played the role of a user, who newly turns to be visually impaired and doesn't know O&M skills. Before the evaluation, the blind advisor taught him to use a white cane. The confederate wore a blindfold and was equipped with a white cane throughout the evaluation.

Because we pay particular attention to the influence of our prototype on agents' performances, we need to minimize the confounding factors. Different users might have different O&M skills, walking paces, and preferences for more or less scenery description [91], which could significantly impact the agents' performances but provide less insightful design implications for our prototype. Therefore, we used the same confederate for all tasks.

In addition, we also trained the confederate to primarily rely on the instructions and make fewer action decisions by himself.

### 4.3 Participants

Because access to trained agents (e.g., Aira agents) is limited, we focus on more accessible untrained volunteers and leverage CV techniques to assist them in this study. We recruited 13 sighted participants on-site who were not trained for RSA previously. As shown in Table 1, the participants (P1 to P13) were college or graduate students, ranging from 18 to 30 years old (average 21.5). There were 6 females and 7 males (self-reported gender). The backgrounds of the participants were diverse, covering natural science and social science disciplines. To verify the feasibility of the AR maps in promoting environmental familiarity, all the participants had never been to the testing site before the experiments. Each participant was compensated for their time and effort.

### 4.4 Environment

The experiments were conducted in two specified office areas inside a campus building. Figure 2 and 3 show the maps of the two testing areas. In the experiments, each agent will be sitting in a control room that is inside the first testing area (the small room between E346 and E363 in Figure 2(d)). And the confederate will be walking in the hallways under the agent's instructions. Each side of the hallways looks similar, which makes the navigation tasks (see Section 4.6) complex enough for evaluation. The control room blocks all visual and verbal information from outside, so the agent cannot get any information about the confederate except from the prototype that supports online RSA.

### 4.5 Apparatus

The confederate was blindfolded, equipped with a white cane and an iPad with RSA-User app installed. The confederate hung the iPad around his neck with a strap and pointed the camera of the iPad in the direction of the environment. He was able to change the orientation of the camera when needed, such as lifting it and pointing it to the left or right. The agents were always equipped with a laptop that enables Skype and an iPad with RSA-agent app installed. Depending on the type of task, we provided the participants with a 2D or 3D map on the iPad. After their consent, an iPhone was used to record agents' activities during each task.

### 4.6 Task Design

We conducted a within-subject study in a lab setting, where we compare 2D maps with 3D maps. Using 2D maps in indoor navigation was considered as baseline and it was not provided with localization or orientation function for two reasons. First, GPS signal strength is weaker and unreliable indoors [64, 79]. Second, prior work [62, 91] indicated that conventional, static 2D maps are easily accessible resources and are used in currently available RSA services (e.g., BeMyEyes [20], Aira [13]). The task design was reviewed and refined by an expert blind advisor.

Our study had two conditions and two tasks, as described below. Each agent-user pair performed two tasks per condition. In total, each pair completed four tasks.

**C1:** use a 2D map, as shown in Figure 2(d) and 3(d).

**C2:** use a 3D map, as shown in Figure 2(a) and 3(a).

For each condition, we set up Skype for oral communication and mirrored screen for live video as described in Section 3.6. The only difference between C1 and C2 is substituting the 3D map and RSA-Agent app with a 2D map picture on the same iPad. We design two types of tasks for each condition:

**T1:** find a specific room with annotation on the map.

**T2:** find a landmark not annotated on the map.

All trials of T1 were performed in the first testing area as shown in Figure 2. We prepared various trials for T1 with different start points and target rooms but similar distances and difficulties. The agents are free to choose any possible routes to the destination. Figure 2(b) demonstrates an example of T1 starting from the green dot and ending with the green star (E351). In this example, the agent could plan the path in two different directions, either up-to-right or right-to-up. For T2, we provide two trials of finding a trashcan (Figure 2(c)) in the first testing area or a water fountain (Figure 3(c)) in the second area. There are two groups of trashcans in the first testing area, as marked as red stars in Figure 2(b). We regarded the task as accomplished when the confederate reached one of the trashcans. The start point was chosen from the two red dots in Figure 2(b). The second testing area was only used for T2 of finding the water fountain. Figure 3(b) shows the start point (red dot) and the position of the water fountain (red star). We assigned the two T2 trials to C1 and C2, respectively. To reduce the effect of task-specific variance, half of C1 (or C2) trials adopt finding trashcans and the other half finding water fountain.

To measure NASA-TLX separately for the two conditions, we conducted the experiments first in one condition and then in the other condition. To minimize the influence of the order, we randomly chose half of the agents to be tested in the order C1 & C2 and the other half in the reverse order C2 & C1. For each condition, one task is from T1, while the other task is from T2. Since there are overlaps between the testing areas of T1 and T2, we first performed T2 task and then T1 task. The reason is that the agents might come across the T2 targets (e.g., trashcans) during the T1 navigation, but not vice versa; because the agents would hardly notice the target room of T1 during the T2 process. Based on the path chosen by the agent in T2, we then specified the start point and destination of T1 in the unexplored area. Other confounding factors were also minimized for fair comparisons of performance. For example, we only conducted the experiments at the time when there were no other people in the testing areas in case of interruptions.

### 4.7 Procedure

There were two steps before the experiment. First, we hosted each agent outside the testing area, introduced the study, obtained their oral consent, and then led the agent to the control room. Given that the control room is located in the testing area, we covered the agent's eyes until arriving at the control room to make sure they will not get any spatial information before the experiment. Second, we briefed the agent on the process of the experiment and taught them how to use the prototype. Agents were given sufficient time (~10 min) and

instructions to familiarize themselves with the prototype. Referring to the expressions for giving guidance, we introduced agents with directional, numerical, descriptive, and status check instructions [50]. We also warned the agents about potential issues, such as latency (technical issue) and the possibility of the confederate walking out of the testing area (experimental issue). We emphasized that agents should be responsible for the confederate's safety throughout the task. To ensure the confederate's safety, we designated one researcher to quietly watch the experiments in the hallway.

Before each task, one researcher set up the apparatus and led the blindfolded confederate to a start point that was unknown to both the agent and the confederate. After arrival, the researcher assigned either T1 or T2 to the confederate. We defined that a task started when the agent and the confederate began communication and ended when the confederate successfully found the target location or object. With their consent, we filmed the agents' activities during each task. We also documented the completion time of each task and double-checked it with the recorded videos.

We aim at investigating and comparing the agents' experiences in two different conditions. Therefore, each agent was asked to perform two tasks in one condition first. After finishing the tasks, we measured the agent's perceived workload using the NASA-TLX. Next, the agent conducted the other two tasks in the other condition and assessed their workload with NASA-TLX. The NASA-TLX scores were recorded and used as a prompt in the following interview. Finally, we conducted a semi-structured interview with the agent. The interview questions were mainly about the experiences of guiding the confederate and comments on the two conditions. The interviews were audio-recorded and transcribed into text documents. Each session lasted for about 90 minutes.

# 5 FINDINGS

## 5.1 Quantitative Analysis

In this section, we will analyze two sets of data quantitatively, including (i) task completion time and (ii) subjective evaluations with NASA-TLX scores. One outlier was detected in the dataset of task completion time by both *z*-scores and interquartile ranges. P13 spent a much longer time finding an unannotated landmark with 3D maps than other participants. The reason was that he didn't realize the real-time localization and orientation in the first trial of using 3D maps *"so I [he] just did the task in the way I [he] did with the 2D map"*. After realizing the functionality of 3D maps, he reported that *"the last task I followed the red spot so I think I finish the task very quickly"*. Thus, we removed P13's task completion time. Considering that P13's unawareness of 3D map functionality could affect his self-evaluation, we removed his NASA-TLX scores as well. A paired, two-tailed *t*-test was computed on the remaining data sets.

**5.1.1 Task Completion Time.—**Figure 5(a) shows the mean completion time for (i) Task 1, finding a room with annotation on the map, (ii) Task 2, finding an unannotated landmark, and (iii) both tasks. The completion time of using 3D maps was significantly shorter than that of using 2D maps in Task 1 ($p = 0.001$), Task 2 ($p = 0.034$), and an average of both tasks ($p = 0.008$). The results indicate that 3D maps can significantly reduce the task

completion time in finding either annotated or unannotated targets. Especially in Task 2 of finding unannotated landmarks, the mean completion time of using a 2D map ($\bar{t} = 430$ sec) is more than twice that of using an 3D map ($\bar{t} = 211$ sec). It demonstrates the superiority of 3D maps in finding landmarks that are not annotated but can be recognized from the 3D structure and texture.

**5.1.2 Subjective Results.—**NASA-TLX scores elicit agents' experiences in terms of (i) overall experiences, (ii) mental demand, (iii) physical demand, (iv) temporal demand, (v) performance, (vi) effort, and (vii) frustration. Figure 5(b) shows that the mean NASA-TLX scores of using 3D maps were lower than that of using 2D maps in all measures, except for physical demand. In particular, the mean performance and frustration scores of 3D maps are only around 45% of that of 2D maps, indicating 3D maps can considerably improve the self-rated performance and reduce the frustration of agents in indoor navigation tasks. It is also understandable that 3D maps received slightly higher scores in physical demand ($\bar{x} = 24$) than 2D maps ($\bar{x} = 23$), because interactive 3D maps require extra gesture interactions compared with static 2D maps.

The results of $t$-test are consistent with Figure 5(b). 3D maps help the agents significantly reduce the mental demand ($p = 0.013$), temporal demand ($p = 0.024$), effort ($p = 0.011$), and frustration ($p < 0.001$). The agents also believed they performed significantly better with 3D maps than 2D version ($p = 0.002$). Moreover, there are no statistically significant differences between 2D and 3D maps in agents' physical demand ($p = 0.699$), suggesting that interactive 3D maps did not introduce additional physical burdens compared with 2D maps.

In sum, we found from the experiments that 3D maps could significantly improve the agents' performance in task completion time and by the subjective assessment, and reduce their mental demand, temporal demand, effort, and frustration while not markedly increasing the physical demand.

## 5.2 Qualitative Analysis

We conducted semi-structured interviews with the agents after they finished all the tasks. We elicited their experiences of instructing people's indoor navigation. We used thematic analysis [28] to understand such experiences. First, the authors went over all transcripts to identify data related to the agents' experiences in both conditions. For example, how did the agents use a 2D/3D map to identify where the confederate was. The authors open-coded such experiences without any pre-defined framework. Second, the authors held regular meetings to refine and organize the codes. The meetings focused on discovering the underlying relationship between codes; codes that were closely related would be grouped together. For instance, we identified that some codes were about agents' experiences of planning the path for the confederate; we then grouped such code together and named the group as "navigational strategy". We identified themes through the process; we also went back and forth to testify if the themes were mutually exclusive and applicable to most codes. Third, we came up with three themes and finalized their names as "navigational strategy", "efficient coordination with reducing unnecessary instructions", and "user experience of split-screen dashboard". We reported the themes in detail in the following subsections.

#### 5.2.1 Navigational Strategy.—

During the tasks, agents utilized navigational strategy to make sense of the spatial layout, path planning, and the confederate's location and orientation. Common examples of navigation strategies are leveraging real-time updates and detailed visualization of the environment.

**Real-time Localization and Update.:** During the experiments, real-time data was always captured and provided from the live camera feed. The other type of real-time data was represented as real-time localization and orientation, which was only available on the dynamic and interactive 3D maps but not on the static 2D maps.

Getting aware of the starting point, that is, where the confederate is at the beginning of the tasks, is the first step and challenge in navigation. The success and easiness of this step affect agents' subsequent decision-making and confidence. When navigating with 2D maps, the only source of real-time data was live camera feed captured and transmitted from the confederate's camera-based device. Thus, agents relied heavily on the confederate scanning around his surroundings, obtained visual cues of room numbers or fire extinguishers, and matched them with the corresponding signs on 2D maps for localization. This process was tedious, repetitive for both groups, requiring the confederate to go back and forth and adjust his camera to point to appropriate visual cues, and requiring agents to align the information from live camera feed with the one on 2D maps repeatedly.

> "I need to make sure… where he [confederate] is at. So I want him to look around and to gather information for me to know his location."
>
> (P8)

> "Because I'm not really familiar with this building. So I cannot really remember the location of each room, I have to ask him to turn around again and again to check if the number of the room is correct."
>
> (P10)

In contrast, seven agents (P1, P3, P5, P8, P9, P11, P13) indicated that real-time localization on 3D maps helped them make sense of the confederate's location throughout the navigation, which *"is the most helpful thing"* (P5) and *"makes everything easier"* (P8). P10 reported that the real-time localization gave her a birds-eye view of the environments, which was an extra help to get rid of the reliance on scanning and broke the restriction of limited camera view.

Orientation is another important factor in navigation. Even if agents localize the confederate correctly, failure to identify the orientation can result in navigating the confederate in the wrong direction and leading him away from the destination. Deducing the confederate's orientation is more challenging in homogeneous environments. As pointed out by P10, *"the roads, the walls, the hallways are kind of similar"* so she made a wrong judgment of direction without 3D maps. Identifying the confederate's orientation requires mental efforts because P8 *"need[ed] to put myself [himself] in a location and to check if that's the right direction"*. Additionally, he had to *"keep thinking every time he [the confederate] made a turn"*. Real-time update of the confederate's orientation on 3D maps alleviates this problem.

Benefiting from this feature, P1's cognitive load was reduced by not putting efforts to situate herself in the confederate's physical position and *"translate"* the confederate's orientation.

**Detailed Visualization of the Environment.:** Visualization of the environment is represented as the room numbers, fire extinguisher signs, and building layout on 2D maps, as well as the room numbers, distance bands, and detailed environmental information on 3D maps. Compared with the abstract spatial layout on 2D maps, 3D maps supplement more environmental details, such as the location, shape, and color of objects. These details are not limited to annotated targets as on the 2D maps, but all the objects in the scope of the testing area. This feature of 3D maps increased agents' familiarity with the environment because they *"can see everything on the map"*.

> "Because for the 3D map, you can see everything… like tables, desks, water fountain, and trash cans, you can see everything on the map. But for 2D map, not that many details."
>
> (P7)

This characteristic makes 3D maps not only more comprehensive but also more intuitive. As pointed out by P9, reading 2D maps is challenging because she needs to transform the spatial layout in her mind and match it to 3D scenes that she is familiar with. The detailed environmental information on 3D maps is *"natural"* as it is similar to what people absorb and experience in their daily lives.

> "I'm not good at reading [2D] map. So reading map itself is hard for me… I'll get lost in the map because I'm really scared of reading [2D] map… [But] the objects are all on the 3D map… I think it's just quite natural to use the tool [3D map]."
>
> (P9)

As mentioned above, agents aligned the real-time data from the live camera feed with the information on 2D maps for localization and orientation. In addition to these purposes, visualization of the environment on either 2D or 3D maps facilitates agents to find the destination, which is crucial in path planning and agents' decision-making. Ten agents (P2, P3, P5–12) believed that the detailed environmental information on 3D maps made the wayfinding easier, especially in the task of finding an unannotated object (e.g., trash can or water fountain). The reason was that they could be more familiar with the environment by identifying the object with its shape or color and get aware of its location on 3D maps. The determination of the destination accelerated the navigation process by contributing to accurate decision-making and reducing the possibility of giving wrong instructions. It decreased agents' mental demand and made them feel more confident and *"secure"* (P7).

> "The water fountain is actually on the 3D map. So I just look around the map and then just choose the one I want him to go to. And then just ask him to go there… Because you can scroll the [3D] map, you can just see where the direction really is. And you can just move the [3D] map to see the surrounding environment. So that may help me to make more accurate decisions."
>
> (P10)

In comparison, the lack of detailed environmental information on the 2D maps led to frustration when finding an unannotated object. Agents had limited environmental knowledge and familiarity in this task because they were unable to find the specific object on 2D maps and further determine the destination. The only strategy was to guide the confederate to randomly explore the testing area.

> "…he says he wants to use the water fountain but [it's] not shown on the map. So the only way I know where it is is just to get him to walk around until I find that. So that is like a big frustration. You don't know where the destination is."
>
> (P3)

**5.2.2 Efficient Coordination with Reducing Unnecessary Instructions.—**In addition to making strategies for navigation, agents further reported their experiences of coordination. Because the goal of the tasks was to help the confederate find a location or an object, agents needed to digest the information perceived through navigation strategies and convert it into feasible instructions for the confederate. Therefore, the navigation tasks can be considered as a type of coordination between the agent and the confederate. Especially for the agents, they needed to make sense of the environment and give clear, accurate, and timely feedback to the confederate. Such challenges made agents use the best of the split-screen dashboard to facilitate coordination. Overall, agents found the 3D map was more helpful than the 2D map while giving instructions to coordinate with the confederate.

Localizing and orientating the confederate was an indispensable part of giving instructions. With the 2D map, the only source of this information was the camera feed. In this condition, localization and orientation involved tedious work by asking the confederate to move back and forth and to adjust the angle of the camera, so that the agent could get a clearer view from the camera and get more accurate information. P8 complained that such work was effort-consuming. In contrast, 3D maps showed the location and orientation of the confederate, significantly reducing the agent's effort and time in instructing the confederate to scan the appropriate areas at the correct angle.

> "[When using 2D maps,] sometimes he's not facing directly to the number of the door. So that requires lots of work. I asked him to turn right, turn left, and then maybe slightly right, turn up, go back, check number. Yeah, this process is adding more effort for me to know the location. But 3D map shows where he's facing to… I didn't really ask him to turn to check the doors or something like that. So the 3D map saves lots of time by showing the location."
>
> (P8)

In addition to the localization and orientation, the environmental information on 3D maps provided a better condition for exploration. In some trials, agents were asked to find unannotated objects, such as a water fountain. Thus, agents had to explore the space to look for the objects. The exploration process required the agents to coordinate with the confederate with the help of the dashboard. The map became an important source of information for the exploration, and the information provided by the map determined the

efficiency of the coordination. P5 recalled her experiences of looking for something that was not annotated on both maps:

> "On the 2D map, I cannot find it [water fountain] because I don't know where the sign is. I instruct him to go the wrong way so I feel frustrated at that time. But on the 3D map, it is clear to see all the things on the map, so I think I don't make any wrong instructions. So I feel good."
>
> (P5)

The absence of environmental information on 2D maps negatively impacted P5's navigation because she guided the confederate in the wrong direction and felt bad about it. However, she noticed that, although the objects were not annotated on the 3D map as well, she could identify the water fountain from the detailed visualization of the environment displayed on the 3D map. The detailed environmental information on the 3D map made up for the shortcoming that not all objects were annotated. It also reduced the frustration of P5 and made her feel more confident when giving instructions.

Agents also illustrated the benefit of the 3D map in terms of monitoring the confederate's movement. P10 reported why she paid constant attention to the confederate's movement and how she did that with both the 2D and 3D maps:

> "The 3D map helped me to define his step and the distance between him and those other things. But when I was using a 2D map to guide him to a room, I just asked him to turn left at a time when I thought he should turn left, but actually he was not at the actual place. Because when he turned left, I saw he would hit the wall. So that is a problem. And I think the 3D map helped me to better define where he was."
>
> (P10)

The testing area consisted of many intersections, where the confederate needed to make a turn to get to the destination. To agents, deciding where and when to make turns was another important part of instructing the confederate. Therefore, agents should precisely measure how fast the confederate walked and how long the confederate needed to walk until the next turn. Failing to do so might lead the confederate to the wrong place and sometimes even expose him to risks. Due to the lack of a real-time update on the 2D map, agents were unable to get the accurate location and orientation of the confederate. P10 pointed out that it was difficult to measure the pace of the confederate through the camera. Thus, she complimented the real-time localization and environmental information on 3D maps. These features helped P10 better predict when the confederate would reach the intersection and then give timely directions to the confederate.

Lastly, the coordination involved correcting wrong directions. Our experiment took place in a part of a building, so it was likely that the agent would give wrong directions to the confederate and make him walk outside of the testing area. Therefore, agents were also responsible for identifying whether the confederate was out of the area and guiding him back if he was. P5 told us when she realized the confederate might have walked out of the testing area:

> "Because I need to ask him to use the camera to see the room number [when using 2D map]. But when I use the 3D [map], I can see from the map instead of asking him…I can know he goes the wrong way. So I can make a quick decision to ask him to go around or something like that."
>
> (P5)

Before the experiment, we trained each agent, telling them that the area was small and the confederate might walk out. The 3D map's localization function helped cope with such situations, said P5. It could instantly show that the confederate was going in the wrong direction. Then, the agent was able to give immediate instructions to let the confederate turn back to prevent him from walking out of the testing area. But when using 2D maps, P5 could not get such timely feedback; she had to ask the confederate to turn around and get information from the camera. If the confederate was near a room that was not on the maps, then he was out of the area. Such comparison indicated that the 3D map was better in terms of correcting and preventing mistakes.

**5.2.3 User Experience of Split-screen Dashboard.—**Agents were split on their experience of interacting with the split-screen dashboard. Some agents appreciated the interactivity of 3D maps, which made RSA *"more playable and less frustrated"* (P8). Other agents reported the distraction and usability issues of 3D maps, such as unfamiliarity with complicated manipulation and latency.

**Interactivity.:** The interactivity of 3D maps is one of the characteristics that distinguish 3D maps from 2D maps. P8 said that the interactivity of 3D maps made the RSA interaction *"more playable and less frustrated"* compared with 2D maps. He considered the navigation with 3D maps was *"a fun thing to do"* rather than a job because he can *"play with a map"*. More specifically, P2 thought the interactivity of 3D maps was represented as cuteness, where the sign of real-time localization and orientation was jumping and *"moving like a duck"*. This characteristic of 3D maps released her frustration and made her more willing to help the confederate. The reduction in frustration was also reflected in Section 5.1.2.

> "I think the 3D map is cute because I saw that red point moving like a duck. When I directed that person moving, I thought that point was like jumping… When I thought the point, I just couldn't help laughing. It just releases my frustration and I'm more willing to take time to help that person to get to the right place."
>
> (P2)

The interactivity split the agents, with negative responses from P11. Although she appreciated the function of localization and orientation, P11 reported that the real-time updated, bright spot distracted her and made her ignore the live video stream.

> "It did help me in the task like the movement, especially like the x, y, z axes, and the pink thing helped me locate where he's going but that sometimes distracts me, because I keep looking at that and I ignore the stream screen."
>
> (P11)

**Usability.:** Some agents were bothered by the complicated manipulation and connection issues when using 3D maps. Regarding the manipulation, P12 said that *"it's a little bit hard for the user who doesn't get used to those kinds of three-finger thing"* and *"to memorize all the steps"*. Some agents (P1, P2, P7) spent more time learning and familiarizing themselves with the manipulation. Because of the complicated manipulation, P4 paid too much attention to 3D maps to search for the destination, meanwhile, he still asked the confederate to proceed. Distracted by 3D maps, P4 failed to precaution risk so that the confederate hit obstacles on the way.

> "I wasn't too sure of where the room was so it took longer for me to find that [on 3D maps]. However, I still told him [the confederate] to walk…"
>
> (P4)

Latency is the other issue that affects the usability of the prototype. Five agents (P3, P5, P6, P10, P12) have encountered the connection issue during the tasks. P12 spent time and mental efforts in identifying the confederate's current location under this circumstance. Similarly, P6 indicated that the latency on 3D maps was *"misleading"*, which confused him about the confederate's location and made him nervous about delivering wrong instructions.

> "However, the problem is there's some latency on the map. I only notice once, it wasn't very representative where the person was on the map because it moved rapidly fast from one location to the next… It was a little bit misleading to me… I want to get rid of wrong commands. That's why I was nervous."
>
> (P6)

Three agents (P3, P5, P10) utilized the cues from the live camera feed to mitigate this challenge. They indicated that the real-time update from the live camera feed was accurate and reliable for localization and orientation, which could compensate for the latency on 3D maps.

## 6 DISCUSSION

In this section, we identified several advantages and opportunities of 3D maps. We discuss the usage of 3D map view and live video feed view and complementary design of these two sources of views, the promising applications of the proposed system to trained agents, and the evaluation of the RSA paradigm in a lab setting. Finally, we present the limitations and the directions for future work.

### 6.1 Complementary Design of 3D Map View and Live Video Feed View

Our findings elaborate that the functionalities of the 3D maps contribute to the reduction of the time and mental workload during guidance in three aspects. First, 3D maps give agents a birds-eye-view of the environments and synchronized confederate's location and orientation. Second, the detailed visualization of the environment increases agents' environmental familiarity as they *"can see everything on the map"*, including the location, shape, and color of both annotated and unannotated objects. Third, 3D maps enable efficient coordination between agents and confederate by reducing unnecessary instructions, such as asking the confederate to move back and forth and to adjust the angle of the camera. However,

even though the agents are equipped with 3D maps, the live video feed is essential and irreplaceable. For example, when latency occurs on 3D maps.

In our proposed system, these two views are displayed on two different screens: 3D maps on iPad and the live video feed on agent's laptop. The distraction of 3D map view is one issue in the current setting. As reported in Section 5.2.3, agents could pay more attention to 3D map view and thus ignore the live video feed. This phenomenon reflects the superiority of 3D map with real-time localization over live video feed in the navigation tasks of finding a room or landmark. However, when there is latency on 3D map, three agents (P3, P5, P10) utilized the live camera feed to infer the correct location, which demonstrates the significance of the live camera feed when map-based localization is inaccurate. Instead of comparing the importance of 3D map with live camera feed, combining these two views is our next phase of the prototype design, which can improve efficiency by agents not checking back and forth between two screens. To this end, we can display two sources of views on one device screen, present them in turn and supplement each other if the other one is disabled. It can be achieved by (i) switching these two views automatically or (ii) adding shortcut buttons for switching, which provides agents with more self-determination. One example is to alleviate latency on 3D map view. Latency is one limitation in our current implementation, encountered by five participants in experiments, but it is inevitable and manageable in early-stage research. If applying the combined-view design to this case, we can display 3D map view for default and switch to live video feed when the connections between RSA-Agent and RSA-User are unstable, or when the agents need to identify obstacles and notify users of dynamic environments.

### 6.2 Extending the Proposed System to Trained Agents

Indoor navigation is needlessly difficult, tedious, and cognitive-overload, requiring agents to orientate, localize users and understand their situations with indeterminate camera feed [50, 62]. Introducing the powerful 3D maps to this task empowers the volunteers by making the task more engaging, intriguing, and transparent to work with symbolic tools, and further addresses volunteer recruitment and retention [33]. As analyzed in Section 5.2.3, some agents appreciated the interactivity of 3D maps, which made the navigation *"more playable and less frustrated"* (P8).

3D maps can also potentially benefit trained agents and refine their skill set of indoor navigation by providing a birds-eye-view of indoor environments. Equipped with 3D maps, trained agents can take less practice in interpreting users' surroundings through camera and matching them with static 2D maps [62], as well as being able to do more with the snippet of the camera feed. Considering the expertise of trained agents, we believe that 3D maps can even make a bigger difference than untrained agents.

### 6.3 Evaluating RSA Paradigm in a Lab Setting

RSA is a highly dynamic paradigm in terms of agent-user collaboration and physical environments. We have encountered several challenges in simulating the RSA interaction and evaluating the prototype. Our strategies for alleviating these challenges provide insight into how the complex, dynamic RSA paradigm can be established in a controlled lab study.

First, an insufficient number of participants with visual impairments [84, 86], especially in geographical locations where population density is low, such as rural settings where our institution is located. This problem is more severe during the COVID-19 pandemic, a global crisis disproportionately affects people living with disabilities [16] and poses numerous challenges to visually impaired people's daily life [80]. To alleviate this problem, we engaged an expert blind user as an advisor throughout the entire study, including prototype design, experiment design, and data analysis. This process provided fruitful insights in terms of "being with" a representative from the target population and building empathy for visually impaired users.

Second, implementation efforts. High abandonment rate of assistive technologies has been reported in previous research (e.g., up to 78% for hearing aids [68, 83]) due to the substantial efforts required to develop, assess, and adjust these special-purpose technologies [71]. Thus, accessibility researchers need to balance implementation efforts and feasibility of their prototype before employing it extensively. In this study, several design decisions have been made to achieve this balance, such as recruitment of the blind advisor, confederate design and role-playing.

Third, functional testing, that is ensuring an assistive technology does what it is designed to do, is another challenging but essential part. Limited testing with the target population will lead to accessibility issues in final products, which are only flagged by the end-users [14]. Fortunately, for mobile technology, researchers reported that lab study could provide high-quality data comparable to field studies by (1) including mobility, (2) adding contextual features (e.g., scenarios and context simulations), and (3) supporting high-quality video data collection [53, 54]. Thus, we evaluated the prototype in a lab setting that is safer, controllable, leading to a high level of ecological validity. All of the requirements were fulfilled in our study by allowing role-playing users to move freely in the testing site, situating the scenarios of finding an office or a trash can, and capturing and video-recording the interaction between the user and the prototype.

### 6.4 Limitations and Future Work

**Participants.—**Our findings show that the 3D maps alleviate the navigational challenges for untrained agents by supplementing the environmental knowledge, real-time localization, and orientation. Trained agents are also bothered by these challenges [50]. Thus, we believe that 3D maps have the potential to be a boost for trained RSA agents. We acknowledge the limitation of no trained agents involved in the study. Our future direction is to review study footage with RSA professionals for expert evaluation.

We acknowledge that the blindfolded confederate may behave differently than visually impaired people, end-users of RSA. Engaging with the blind advisor, we found that people with visual impairments walk faster because they have more experience and are more confident in the situation, whereas the confederate is less used to walking around with a blindfold. Besides, visually impaired users probably store a different representation of spaces because they use more tactile rather than visual cues [36]. Notably, it is more challenging for visually impaired users to find dynamic objects (e.g., trashcan) than landmarks (e.g., stairs, doors). Thus, the corridor-like environment, which was used to

test the prototype, is likely easier for visually impaired users as they have experienced similar ones when walking down to doctors' offices, for example. The prototype is still in the research phase and it has not been tested to be fault-tolerant. We will engage visually impaired users in the future while ensuring their safety and lowering the risk.

**Prototype implementation.—**Besides the implementation trade-offs in Section 3.3, a notable limitation of the current implementation of our CV-mediated RSA prototype is latency and localization error. The issues of latency may arise due to unstable connections. Though ARKit-based localization is accurate enough in the experiments, one of the main adverse effects of latency was to display the real-time location of the user with lags. In our study, five participants (P3, P5, P6, P10, P12) encountered the latency issue during the experiments, but three of them (P3, P5, P10) overcame this challenge by associating 3D map with a live camera feed.

Our RSA prototype is also not ready to be deployed in large-scale scenarios. On one hand, the prototype only works in a WiFi environment and does not support cellular telecommunications. On the other hand, we only tested the apps with 3D maps of moderate size. When applied to large buildings (e.g., grand shopping malls), the current prototype may experience problems in processing the large-scale 3D map. We expect this issue will be addressed in the future with upgraded mobile devices.

Compared with 2D maps in baseline RSA, the increased processing and technology demands of 3D maps are a limitation in the short term. However, we expect advancement in hardware and computation platforms and a decrease in cost proposition for technology in the long term. Thus, exploring the possibilities of 3D maps in RSA paves the path for high-performance, faster navigation aids for visually impaired users.

Although our 3D map-based solution is more cost-effective and time-efficient than tag-based solutions (e.g., QR codes [37], Bluetooth beacons [42]), we still need a sighted volunteer to scan the environment to build an offline map in advance. Such an approach is suitable for use in public places frequented by people with visual impairments (e.g., shopping malls, airports). But for places that are not often visited, the requirement of scanning the environment by sighted volunteers is still demanding. As discussed in [62], one way to reduce setup requirements is to reconstruct on-the-fly maps while the visually impaired user is navigating the environment with a smartphone. Building an online map itself is not difficult – the SLAM algorithm embedded in ARKit is online. The main difficulty of implementation lies in data transmission. Due to a large amount of high-definition 3D map data, it is a great challenge to transmit and synchronize the map to the agent's device in real-time. A possible solution is to transfer only the 3D map layout, but this will lose a lot of map details and may not be able to address some key navigation problems. AR frameworks could support lightweight online SLAM in the future for real-time transmission of online maps and thus avoid the setup step by sighted volunteers.

## 7 CONCLUSION

We presented a system that implements a high-fidelity CV-mediated RSA prototype in lab settings and aims to address the indoor navigation challenges. In line with existing RSA services, our prototype is developed for visually impaired users' mobile devices without extra setups. Thanks to ARKit and LiDAR scanner, we are able to create high-quality AR maps and provide accurate localization with an iPhone or iPad. In a controlled lab study based on confederate design, role-playing, and empathy, we tested the CV-mediated system with 13 sighted volunteers. The results show that AR maps can significantly improve untrained agents' performance in indoor navigation tasks. The features of real-time localization, landmark annotations, and fine-grained AR map details are favored by most of the participants. In this study, the prototype implementation and empirical results provided concrete evidence supporting the feasibility of AR maps for RSA. We hope that our study can pave the way to enhance RSA systems with AR maps for indoor navigation.

## ACKNOWLEDGMENTS

We thank Dr. Zihan Zhou, Xinbing Zhang and Michelle McManus for their contribution to this work. We also thank the anonymous reviewers for their insightful comments. This research was supported by the US National Institutes of Health, National Library of Medicine (R01 LM013330).

## REFERENCES

[1]. 2022. 3d Scanner App™ on the Apple Store. Retrieved June 27, 2021 from https://apps.apple.com/us/app/3d-scanner-app/id1419913995

[2]. 2022. ARCore. Retrieved June 27, 2021 from https://developers.google.com/ar

[3]. 2022. Can I change my availability hours? - Be My Eyes Help Center. Retrieved February 12, 2022 from https://support.bemyeyes.com/hc/en-us/articles/360005892797-Can-I-change-my-availability-hours-

[4]. 2022. Frequently Asked Questions - Aira. Retrieved February 12, 2022 from https://aira.io/frequently-asked-questions/

[5]. 2022. Google Maps - Transit & Food. https://apps.apple.com/us/app/google-maps-transit-food/id585027354.

[6]. 2022. Home of the Blender project. Retrieved June 27, 2021 from https://www.blender.org

[7]. 2022. Model I/O. Retrieved June 27, 2021 from https://developer.apple.com/documentation/modelio

[8]. 2022. More to Explore with ARKit 5. Retrieved June 27, 2021 from https://developer.apple.com/augmented-reality/arkit/

[9]. 2022. Multipeer Connectivity. Retrieved June 27, 2021 from https://developer.apple.com/documentation/multipeerconnectivity

[10]. 2022. RealityKit. Retrieved June 27, 2021 from https://developer.apple.com/augmented-reality/realitykit

[11]. 2022. SceneKit. Retrieved June 27, 2021 from https://developer.apple.com/scenekit

[12]. Ahmetovic Dragan, Gleason Cole, Ruan Chengxiong, Kitani Kris, Takagi Hironobu, and Asakawa Chieko. 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 (Florence, Italy) (MobileHCI '16). Association for Computing Machinery, New York, NY, USA, 90–99. 10.1145/2935334.2935361

[13]. Aira. 2022. Aira. https://aira.io/.

[14]. Aizpurua Amaia, Arrue Myriam, Harper Simon, and Vigo Markel. 2014. Are users the gold standard for accessibility evaluation?. In Proceedings of the 11th Web for All Conference. 1–4.

[15]. Allen Mike. 2017. The SAGE Encyclopedia of Communication Research Methods. 4 (2017). 10.4135/9781483381411

[16]. Armitage Richard and Nellums Laura B. 2020. The COVID-19 response must be disability inclusive. The Lancet Public Health 5, 5 (2020), e257. [PubMed: 32224295]

[17]. Avila Mauro, Wolf Katrin, Brock Anke, and Henze Niels. 2016. Remote assistance for blind users in daily life: A survey about Be My Eyes. In Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments. 1–2.

[18]. Banovic Nikola, Franz Rachel L, Truong Khai N, Mankoff Jennifer, and Dey Anind K. 2013. Uncovering information needs for independent spatial learning for users who are visually impaired. In Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility. 1–8.

[19]. Baranski Przemyslaw and Strumillo Pawel. 2015. Field trials of a teleassistance system for the visually impaired. In 2015 8th International Conference on Human System Interaction (HSI). IEEE, 173–179.

[20]. BeMyEyes. 2022. Be My Eyes. https://www.bemyeyes.com/.

[21]. BeMyEyes. 2022. Be My Eyes. https://www.bemyeyes.com/about.

[22]. Bennett Cynthia L. and Rosner Daniela K.. 2019. The promise of empathy: Design, disability, and knowing the “other”. In The ACM Conference on Human Factors in Computing Systems (CHI). 10.1145/3290605.3300528

[23]. Bigham Jeffrey P, Jayant Chandrika, Miller Andrew, White Brandyn, and Yeh Tom. 2010. VizWiz:: LocateIt-enabling blind people to locate objects in their environment. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops. IEEE, 65–72.

[24]. Brady Erin, Bigham Jeffrey P, et al. 2015. Crowdsourcing accessibility: Human-powered access technologies. Foundations and Trends® in Human–Computer Interaction 8, 4 (2015), 273–372.

[25]. Brady Erin L., Morris Meredith Ringel, Zhong Yu, White Samuel, and Bigham Jeffrey P.. 2013. Visual challenges in the everyday lives of blind people. In 2013 ACM SIGCHI Conference on Human Factors in Computing Systems, Mackay Wendy E., Brewster Stephen A., and Bødker Susanne (Eds.). ACM, 2117–2126.

[26]. Branson Steve, Wah Catherine, Schroff Florian, Babenko Boris, Welinder Peter, Perona Pietro, and Belongie Serge J.. 2010. Visual Recognition with Humans in the Loop. In 11th European Conference on Computer Vision. 438–451.

[27]. Brata Komang Candra and Liang Deron. 2020. Comparative study of user experience on mobile pedestrian navigation between digital map interface and location-based augmented reality. International Journal of Electrical & Computer Engineering (2088–8708) 10 (2020).

[28]. Braun Virginia and Clarke Victoria. 2006. Using Thematic Analysis in Psychology. Qualitative Research in Psychology 3, 2 (2006), 77–101. 10.1191/1478088706qp063oa

[29]. Budrionis Andrius, Plikynas Darius, Daniušis Povilas, and Indrulionis Audrius. 2020. Smartphone-based computer vision travelling aids for blind and visually impaired individuals: A systematic review. Assistive Technology (2020), 1–17.

[30]. Bujacz M, Baranski P, Moranski M, Strumillo P, and Materka A. 2008. Remote guidance for the blind—A proposed teleassistance system and navigation trials. In 2008 Conference on Human System Interactions. IEEE, 888–892.

[31]. Burton Michele A, Brady Erin, Brewer Robin, Neylan Callie, Bigham Jeffrey P, and Hurst Amy. 2012. Crowdsourcing subjective fashion advice using VizWiz: challenges and opportunities. In Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility. ACM, 135–142.

[32]. Carroll John M., Lee Sooyeon, Reddie Madison, Beck Jordan, and Rosson Mary Beth. 2020. Human-Computer Synergies in Prosthetic Interactions. IxD&A 44 (2020), 29–52. http://www.mifav.uniroma2.it/inevent/events/idea2010/doc/44_2.pdf

[33]. Carroll John M, Shih Patrick C, Han Kyungsk, and Kropczynski Jess. 2017. Coordinating community cooperation: Integrating timebanks and nonprofit volunteering by design. International Journal of Design 11, 1 (2017).

[34]. Chaudary Babar, Paajala Iikka, Keino Eliud, and Pulli Petri. 2017. Tele-guidance based navigation system for the visually impaired and blind persons. In eHealth 360. Springer, 9–16.

[35]. Chaudary Babar, Pohjolainen Sami, Aziz Saima, Arhippainen Leena, and Pulli Petri. 2021. Teleguidance-based remote navigation assistance for visually impaired and blind people—usability and user experience. Virtual Reality (2021), 1–18.

[36]. Downs Roger M and Stea David. 1973. Cognitive maps and spatial behavior: Process and products. na.

[37]. Elgendy Mostafa, Herperger Miklós, Guzsvinecz Tibor, and Lanyi Cecilia Sik. 2019. Indoor Navigation for People with Visual Impairment using Augmented Reality Markers. In The 10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom). IEEE, 425–430.

[38]. Fusco Giovanni and Coughlan James M. 2020. Indoor localization for visually impaired travelers using computer vision on a smartphone. In Proceedings of the 17th International Web for All Conference. 1–11.

[39]. Galvin Jan C and Scherer Marcia J. 1996. Evaluating, Selecting, and Using Appropriate Assistive Technology. ERIC.

[40]. Ganz Aura, Schafer James M, Tao Yang, Wilson Carole, and Robertson Meg. 2014. PERCEPT-II: Smartphone based indoor navigation system for the blind. In 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, 3662–3665.

[41]. Garaj Vanja, Jirawimut Rommanee, Ptasinski Piotr, Cecelja Franjo, and Balachandran Wamadeva. 2003. A system for remote sighted guidance of visually impaired pedestrians. British Journal of Visual Impairment 21, 2 (2003), 55–63.

[42]. Guerreiro João, Ahmetovic Dragan, Sato Daisuke, Kitani Kris, and Asakawa Chieko. 2019. Airport accessibility and navigation assistance for people with visual impairments. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 1–14.

[43]. Guerreiro João, Sato Daisuke, Asakawa Saki, Dong Huixu, Kitani Kris M, and Asakawa Chieko. 2019. CaBot: Designing and evaluating an autonomous navigation robot for blind people. In The 21st International ACM SIGACCESS Conference on Computers and Accessibility. 68–82.

[44]. Holmes Nicole and Prentice Kelly. 2015. iPhone video link facetime as an orientation tool: remote O&M for people with vision impairment. International Journal of Orientation & Mobility 7, 1 (2015), 60–68.

[45]. Holton Bill. 2015. Crowdviz: Remote video assistance on your iphone. AFB AccessWorld Magazine (2015).

[46]. Holton Bill. 2016. BeSpecular: A new remote assistant service. Access World Magazine 17, 7 (2016).

[47]. Isen Alice M and Levin Paula F. 1972. Effect of feeling good on helping: cookies and kindness. Journal of personality and social psychology 21, 3 (1972), 384. [PubMed: 5060754]

[48]. Jafri Rabia, Ali Syed Abid, Arabnia Hamid R., and Fatima Shameem. 2014. Computer vision-based object recognition for the visually impaired in an indoors environment: a survey. The Visual Computer 30, 11 (2014), 1197–1222.

[49]. Kameswaran Vaishnav, Fiannaca Alexander J., Kneisel Melanie, Karlson Amy, Cutrell Edward, and Morris Meredith Ringel. 2020. Understanding in-situ use of commonly available navigation technologies by people with visual impairments. In The 22nd international ACM SIGACCESS conference on computers and accessibility. 1–12.

[50]. Kamikubo Rie, Kato Naoya, Higuchi Keita, Yonetani Ryo, and Sato Yoichi. 2020. Support strategies for remote guides in assisting people with visual impairments for effective indoor navigation. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 1–12.

[51]. Kayukawa Seita, Higuchi Keita, Guerreiro João, Morishima Shigeo, Sato Yoichi, Kitani Kris, and Asakawa Chieko. 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. 1–12.

[52]. Kintsch Anja and DePaula Rogerio. 2002. A framework for the adoption of assistive technology. SWAAAC 2002: Supporting learning through assistive technology (2002), 1–10.

[53]. Kjeldskov Jesper and Skov Mikael B. 2014. Was it worth the hassle? Ten years of mobile HCI research discussions on lab and field evaluations. In Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services. 43–52.

[54]. Kjeldskov Jesper, Skov Mikael B, Als Benedikte S, and Høegh Rune T. 2004. Is it worth the hassle? Exploring the added value of evaluating the usability of context-aware mobile systems in the field. In International Conference on Mobile Human-Computer Interaction. Springer, 61–73.

[55]. Ko Eunjeong and Kim Eun Yi. 2017. A vision-based wayfinding system for visually impaired people using situation awareness and activity-based instructions. Sensors 17, 8 (2017), 1882. [PubMed: 28813033]

[56]. Kowdle Adarsh, Chang Yao-Jen, Gallagher Andrew C., and Chen Tsuhan. 2011. Active learning for piecewise planar 3D reconstruction. In The 24th IEEE Conference on Computer Vision and Pattern Recognition. 929–936.

[57]. Kutiyanawala Aliasgar, Kulyukin Vladimir, and Nicholson John. 2011. Teleassistance in accessible shopping for the blind. In Proceedings on the International Conference on Internet Computing (ICOMP). The Steering Committee of The World Congress in Computer Science, Computer …, 1.

[58]. Lasecki Walter S, Murray Kyle I, White Samuel, Miller Robert C, and Bigham Jeffrey P. 2011. Real-time crowd control of existing interfaces. In Proceedings of the 24th annual ACM symposium on User interface software and technology. ACM, 23–32.

[59]. Lasecki Walter S, Wesley Rachel, Nichols Jeffrey, Kulkarni Anand, Allen James F, and Bigham Jeffrey P. 2013. Chorus: a crowd-powered conversational assistant. In Proceedings of the 26th annual ACM symposium on User interface software and technology. ACM, 151–162.

[60]. Lee Sooyeon, Reddie Madison, Gurdasani Krish, Wang Xiying, Beck Jordan, Rosson Mary Beth, and Carroll John M.. 2018. Conversations for Vision: Remote Sighted Assistants Helping People with Visual Impairments. arXiv:1812.00148 [cs.HC]

[61]. Lee Sooyeon, Reddie Madison, Tsai Chun-Hua, Beck Jordan, Rosson Mary Beth, and Carroll John M. 2020. The emerging professional practice of remote sighted assistance for people with visual impairments. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 1–12.

[62]. Lee Sooyeon, Yu Rui, Xie Jingyi, Billah Syed Masum, and Carroll John M. 2022. Opportunities for human-AI collaboration in remote sighted assistance. In 27th International Conference on Intelligent User Interfaces. 63–78.

[63]. Legge Gordon E, Beckmann Paul J, Tjan Bosco S, Havey Gary, Kramer Kevin, Rolkosky David, Gage Rachel, Chen Muzi, Puchakayala Sravan, and Rangarajan Aravindhan. 2013. Indoor navigation by people with visual impairment using a digital sign system. PloS one 8, 10 (2013).

[64]. Li Ki-Joune and Lee Jiyeong. 2010. Indoor spatial awareness initiative and standard for indoor spatial data. In Proceedings of IROS 2010 Workshop on Standardization for Service Robot, Vol. 18.

[65]. Liu Yang, Stiles Noelle RB, and Meister Markus. 2018. Augmented reality powers a cognitive assistant for the blind. ELife 7 (2018), e37841. [PubMed: 30479270]

[66]. Loomis Jack M, Klatzky Roberta L, Golledge Reginald G, et al. 2001. Navigating without vision: basic and applied research. Optometry and vision science 78, 5 (2001), 282–289. [PubMed: 11384005]

[67]. Manduchi Roberto, Kurniawan Sri, and Bagherinia Homayoun. 2010. Blind guidance using mobile computer vision: A usability study. In Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility. 241–242.

[68]. Martin Bob and McCormack Lisa. 1999. Issues surrounding Assistive Technology use and abandonment in an. Assistive technology on the threshold of the new millennium 6 (1999), 413.

[69]. McDaniel Troy, Kahol Kanav, Villanueva Daniel, and Panchanathan Sethuraman. 2008. Integration of RFID and computer vision for remote object perception for individuals who are blind. In Proceedings of the 2008 Ambi-Sys Workshop on Haptic User Interfaces in Ambient Media Systems, HAS 2008.

[70]. Nguyen Brian J, Kim Yeji, Park Kathryn, Chen Allison J, Chen Scarlett, Van Fossan Donald, and Chao Daniel L. 2018. Improvement in patient-reported quality of life outcomes in severely

visually impaired individuals using the Aira assistive technology system. Translational Vision Science & Technology 7, 5 (2018), 30–30.

[71]. Petrie Helen, Carmien Stefan, and Lewis Andrew. 2018. Assistive technology abandonment: research realities and potentials. In International conference on computers helping people with special needs. Springer, 532–540.

[72]. Petrie Helen, Johnson Valerie, Strothotte Thomas, Raab Andreas, Michel Rainer, Reichert Lars, and Schalt Axel. 1997. MoBIC: An aid to increase the independent mobility of blind travellers. British Journal of Visual Impairment 15, 2 (1997), 63–66.

[73]. Phillips Betsy and Zhao Hongxin. 1993. Predictors of assistive technology abandonment. Assistive technology 5, 1 (1993), 36–45. [PubMed: 10171664]

[74]. Ponchillia Paul E and Ponchillia Susan Kay Vlahas. 1996. Foundations of rehabilitation teaching with persons who are blind or visually impaired. American Foundation for the Blind.

[75]. Presti Giorgio, Ahmetovic Dragan, Ducci Mattia, Bernareggi Cristian, Ludovico Luca, Baratè Adriano, Avanzini Federico, and Mascetti Sergio. 2019. WatchOut: Obstacle sonification for people with visual impairment or blindness. In The 21st International ACM SIGACCESS Conference on Computers and Accessibility. 402–413.

[76]. Rafian Paymon and Legge Gordon E. 2017. Remote sighted assistants for indoor location sensing of visually impaired pedestrians. ACM Transactions on Applied Perception (TAP) 14, 3 (2017), 19.

[77]. Riemer-Reiss Marti L and Wacker Robbyn R. 1999. Assistive Technology Use and Abandonment among College Students with Disabilities, 3 (23). IEJLL: International Electronic Journal for Leadership in Learning (1999).

[78]. Rocha Sebastião and Lopes Arminda. 2020. Navigation based application with augmented reality and accessibility. In Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI EA '20). Association for Computing Machinery, New York, NY, USA, 1–9. 10.1145/3334480.3383004

[79]. Rodrigo Ranga, Zouqi Mehrnaz, Chen Zhenhe, and Samarabandu Jagath. 2009. Robust and efficient feature tracking for indoor navigation. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 39, 3 (2009), 658–671.

[80]. Rosenblum LP, Chanes-Mora P, McBride CR, Flewellen J, Nagarajan N, Nave Stawaz R, and Swenor B. 2020. Flatten inaccessibility: Impact of covid-19 on adults who are blind or have low vision in the united states. American Foundation for the Blind. Available online at: afb. org/FlattenInaccessibility (2020).

[81]. Saha Manaswi, Fiannaca Alexander J, Kneisel Melanie, Cutrell Edward, and Morris Meredith Ringel. 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. 222–235.

[82]. Scheggi Stefano, Talarico A, and Prattichizzo Domenico. 2014. A remote guidance system for blind and visually impaired people via vibrotactile haptic feedback. In 22nd Mediterranean Conference on Control and Automation. IEEE, 20–23.

[83]. Scherer Marcia Joslyn. 2005. Living in the state of stuck: How assistive technology impacts the lives of people with disabilities. Brookline Books.

[84]. Simpson Richard. 2011. Challenges to effective evaluation of assistive technology. In Design and Use of Assistive Technology. Springer, 51–56.

[85]. Sinha Sudipta N., Steedly Drew, Szeliski Richard, Agrawala Maneesh, and Pollefeys Marc. 2008. Interactive 3D architectural modeling from unordered photo collections. ACM Trans. Graph 27, 5 (2008), 159.

[86]. Stevens Robert D and Edwards Alistair DN. 1996. An approach to the evaluation of assistive technology. In Proceedings of the second annual ACM conference on Assistive technologies. 64–71.

[87]. TapTapSee. 2022. TapTapSee. https://taptapseeapp.com/.

[88]. Tekin Ender and Coughlan James M.. 2010. A mobile phone application enabling visually impaired users to find and read product barcodes. In Computers Helping People with Special

Needs, Miesenberger Klaus, Klaus Joachim, Zagler Wolfgang, and Karshmer Arthur (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 290–295.

[89]. Aldas Nelson Daniel Troncoso, Lee Sooyeon, Lee Chonghan, Rosson Mary Beth, Carroll John M, and Narayanan Vijaykrishnan. 2020. AIGuide: An Augmented Reality Hand Guidance Application for People with Visual Impairments. In The 22nd International ACM SIGACCESS Conference on Computers and Accessibility. 1–13.

[90]. Verma Prashant, Agrawal Kushal, and Sarasvathi V. 2020. Indoor navigation using augmented reality. In Proceedings of the 2020 4th International Conference on Virtual and Augmented Reality Simulations. 58–63.

[91]. Xie Jingyi, Reddie Madison, Lee Sooyeon, Billah Syed Masum, Zhou Zihan, Tsai Chunhua, and Carroll John M. 2022. Iterative Design and Prototyping of Computer Vision Mediated Remote Sighted Assistance. ACM Transactions on Computer-Human Interaction (TOCHI) 29, 4 (2022), 1–40.

[92]. Yoon Chris, Louie Ryan, Ryan Jeremy, Vu MinhKhang, Bang Hyegi, Derksen William, and Ruvolo Paul. 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. 210–221.

[93]. Zhong Yu, Lasecki Walter S, Brady Erin, and Bigham Jeffrey P. 2015. Regionspeak: Quick comprehensive spatial descriptions of complex images for blind users. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. ACM, 2353–2362.

[94]. Zientara PA, Lee S, Smith GH, Brenner R, Itti L, Rosson MB, Carroll JM, Irick KM, and Narayanan V. 2017. Third Eye: A shopping assistant for the visually impaired. Computer 50, 02 (feb 2017), 16–24. 10.1109/MC.2017.36

## CCS CONCEPTS

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

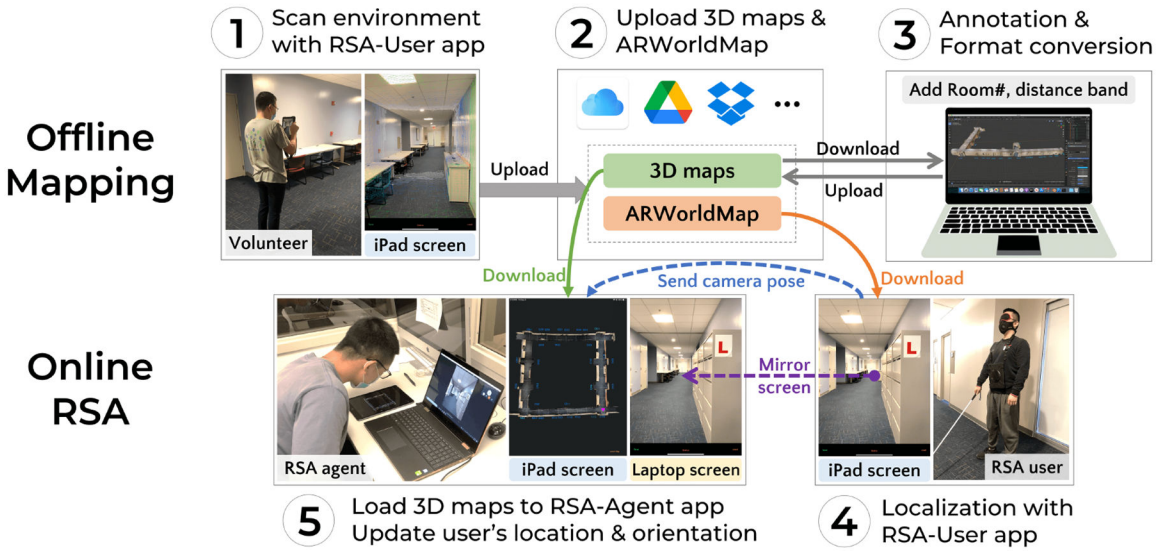
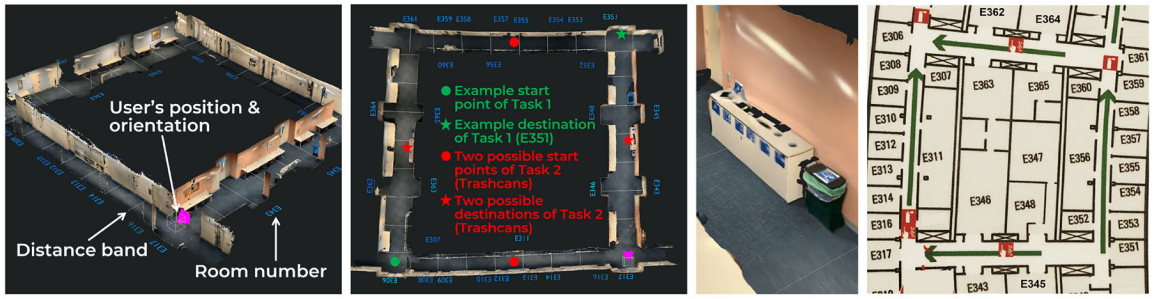


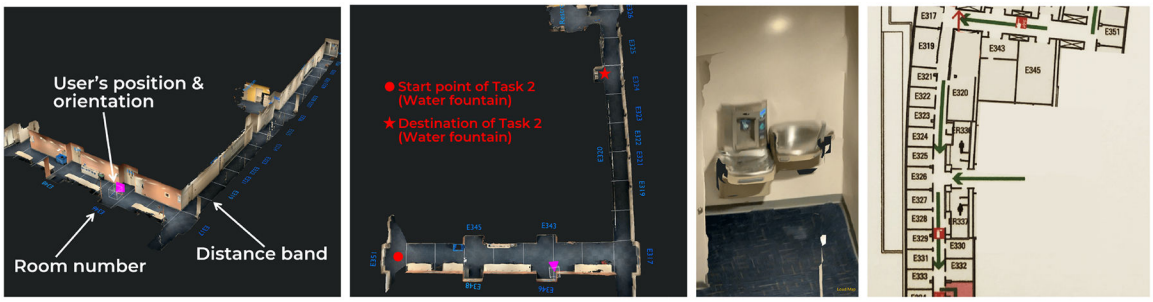

**Figure 1:**
Overview of our RSA prototype with 3D AR maps.




(a) 3D AR map (b) Top-down view (c) Trash cans (d) 2D map

**Figure 2:**

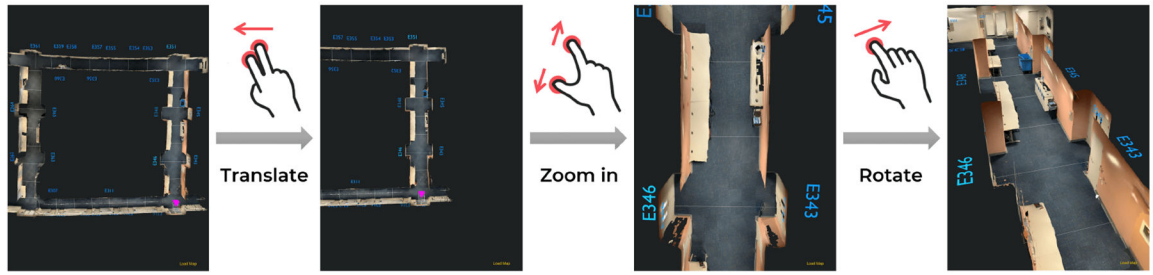
Maps of the first testing area: (a) 3D AR map with real-time localization; (b) Top-down view of AR map and the start points and destinations of Task 1 and Task 2; (c) Trash cans in AR map (zoom in); (d) 2D map.



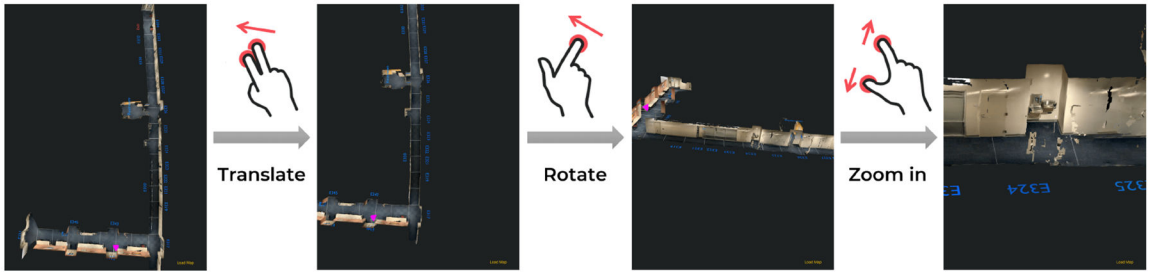

(a) 3D AR map (b) Top-down view (c) Water fountain (d) 2D map

**Figure 3:**

Maps of the second testing area: (a) 3D AR map with real-time localization; (b) Top-down view of AR map and the start points and destinations of Task 2; (c) Water fountain in AR map (zoom in); (d) 2D map.

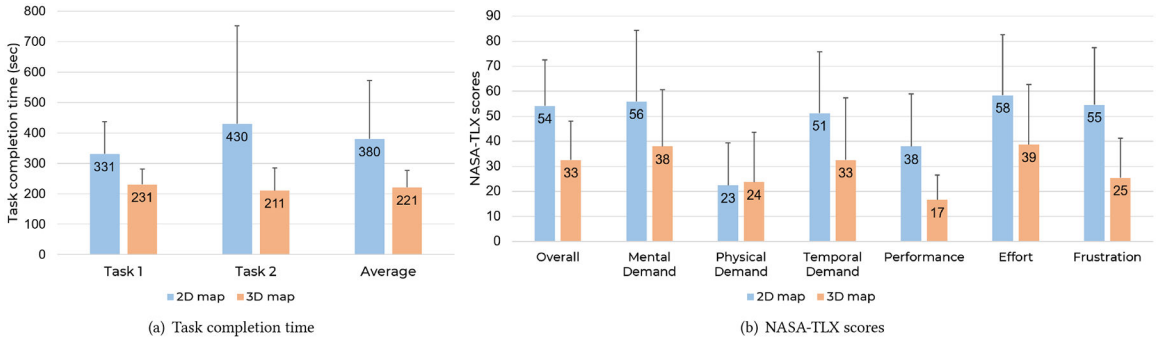



(a) Example gestures to find trashcans on the first 3D AR map.




(b) Example gestures to find a water fountain on the second 3D AR map.

**Figure 4:**
Gesture interactions with 3D AR maps in our RSA-Agent app.




(a) Task completion time

(b) NASA-TLX scores

**Figure 5:**

Comparisons between 2D and 3D map in (a) task completion time and (b) NASA-TLX scores.

**Table 1:**

Participants' demographics. P1-P13 were untrained remote sighted agents. U* was the confederate, and A* was the blind advisor.

| ID | Age | Gender | Major/Designation |
|---|---|---|---|
| P1 | 19 | F | Secondary English Education |
| P2 | 21 | F | Psychology |
| P3 | 19 | M | Computer Science |
| P4 | 18 | M | Mechanical Engineering |
| P5 | 21 | F | Psychology |
| P6 | 26 | M | Mechanical Engineering |
| P7 | 21 | M | Statistics |
| P8 | 21 | M | Data Science |
| P9 | 30 | F | Information Science |
| P10 | 20 | F | Advertising |
| P11 | 20 | F | Accounting |
| P12 | 19 | M | Division of Undergraduate Studies |
| P13 | 25 | M | Industrial Engineering |
| U* | 34 | M | Information Science |
| A* | 49 | F | Member of a local NFB Chapter |