Bridging Physical and Virtual Spaces for Hosting Hybrid Conferences

MALLESHAM DASARI, EDWARD LU, MICHAEL W FARB, NUNO PEREIRA, ANTHONY ROWE

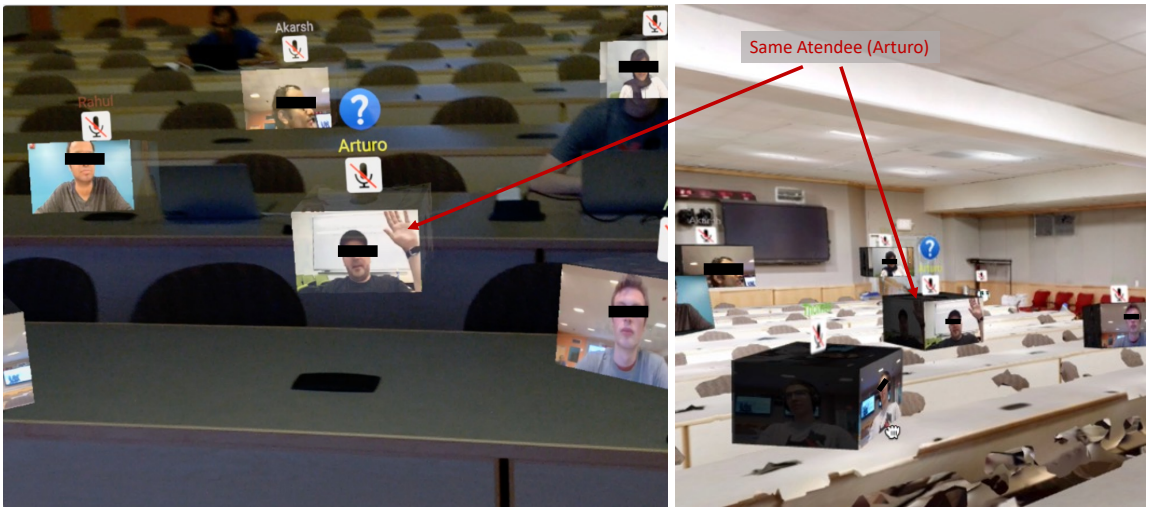


Fig. 1. Same conference attendees viewed in AR and VR. AR on a Magic Leap headset (left); Desktop VR (right).

Catalyzed by a shift towards remote work and the proliferation of consumer Virtual and Augmented Reality (XR) platforms, we are seeing increased interest in so-called "hybrid" telepresence approaches that seamlessly connect virtual and physical spaces. This paper proposes a web-based hybrid conference platform that connects remote VR users with local AR users in real-time. We use an accurately scaled 3D scan of the conference venue as the backdrop environment for VR users with optical anchors to stage AR interactions for in-person users. Remote participants can use VR in a browser (on a Desktop or VR headset) to navigate the scene and interact with other users with video and spatial audio in real-time. In-person participants can use AR headsets, mobile AR through WebXR browsers or portals placed in the environment that pass audio and video between local and virtual spaces. As the platform evolved over the last two years, having hosted many events, we discuss what features worked well and many open challenges. In this process, we define a taxonomy of various hybrid conferencing techniques that we evaluate in two user-studies: (1) a capability video comparison of four major telepresence modalities to see which people prefer for various tasks, and (2) a live user-experience evaluation of our system. Our study suggests immersive technologies can improve the ability to explore spaces, meet new people, or network, in a less fatiguing manner than 2D conference options. For more focused tasks like giving and watching a talk, 2D video options still prevail.

ACM Reference Format:
Mallesham Dasari, Edward Lu, Michael W Farb, Nuno Pereira, Anthony Rowe. 2022. Bridging Physical and Virtual Spaces for Hosting Hybrid Conferences. 1, 1 (April 2022), 15 pages. https://doi.org/XXXXXXX.XXXXXXX

Author's address: Mallesham Dasari, Edward Lu, Michael W Farb, Nuno Pereira, Anthony Rowe.

Manuscript submitted to ACM

1 INTRODUCTION

In the face of a global pandemic, the majority of remote conferences have used traditional video conferencing systems (e.g. Zoom, Teams, Hangouts [8, 10, 14]) or Collaborative Virtual Environments (CVEs; e.g. Hubs, VRChat [13, 16]). For reasons beyond just the pandemic like reduced travel expense, a shift toward remote work, and the desire to reduce carbon footprint, some elements of remote conferencing interaction are here to stay. Unfortunately, none of our current conferencing options allow for rich interactions between a mixture of remote users and local attendees. CVEs address some of these limitations by more closely capturing natural conversations with concurrency among different groups allowing them to explore and interact in virtual worlds [4].

Unfortunately, current CVEs don't make it easy to capture real-world spaces, especially with attendees in the physical space. This is challenging for several reasons. First, we need a workflow and mechanisms that allow the event hosts to capture a 3D model of the real-world space that is mapped and correctly scaled with hooks for device localization. Second (and related), we need easy to configure and use mechanisms for localizing devices, including wearables, tablets, and fixed sensing infrastructure within the physical space that map to the 3D model [11, 19, 21, 25]. Third, we need platforms at the event venue that enable interaction with its virtual counterpart. These could be wearables, tablets, 360 cameras, volumetric capture systems, or various types of portals [3, 15, 17, 23]. Finally, we need a networking platform that can host interactions with reasonable (human perception-level) latency and scale to hundreds of users. A tremendous amount of prior work was done on each of these elements individually, and many are still open sub-domains in the XR community.

This paper presents our experiences developing a system that takes many example solutions to the challenges above and integrates them into an end-to-end demonstrator. Through experiences from dozens of hosted events over the last two years, our system attempts to capture the best elements of CVEs coupled with in-person users in real-time. We leverage a WebXR-based platform that allows remote users to access a 3D version of the space in a standard browser window or in an immersive mode with a headset. Remote users can pass their audio and video feed into the virtual world to interact with each other using spatial audio and video mapped onto a cube above their avatar. We use off-the-shelf laser scanners (either TLS or from a mobile phone) in a space that has been pre-mapped with optical tags for registration. These optical tags allow in-person users to localize mobile phones using pass-through AR or AR headsets to see and hear digital participants in their correct location within the space. The tags also serve to register two-way audio and video windows that connect the virtual and physical spaces. From VR, users see a live camera feed at the correct pose in their virtual scene and people in the physical space see a window in the live VR world. We experimented with different modalities like 360 camera bubbles and various audio technologies. We support a geometry property for marking objects within the scene to be visible in AR and/or VR serving platform relevant content (i.e. hide background scene model in AR, etc). The platform we used supports a number of conference-specific features like the ability to share a computer screen mapped onto the actual projector screen location in VR and a virtual laser pointer that works across AR and VR.

To illustrate the conference experience proposed, in Section 3, we provide some insight into the setup of a 3D environment for the venue and describe a few common conference scenarios. Our hybrid conference experience was created using the open-source ARENA platform [20] (described in Section 2.3). We refer to the combination of tools and custom extensions used to support conference experiences as "our system" throughout this text for brevity. To evaluate the system, we performed a user study based on recorded explanatory videos across commonly used platforms to rate interaction preferences. Though much of the differences might come down to implementation, we probed into which platform modalities people preferred and how long it took to get used to them. Unsurprisingly, we see that

Table 1. Sample services presented in increasing order of immersion for a hybrid conference.

Taxonomy	Example	User Representation				3D Features				Advanced			
		Isometric 3d Video Audio	Rendered 3d Position Rotation			Parallel sessions Video Avatar Avatar	Multiple Rooms Distance audio Positional audio		VR 2d window VR 3d Headset AR 2d window	AR 2d passthrough Volumetric video Virtualized 3d audio			
Flat Video	Zoom/JitsiMeet	● ● -	- - -			◐ - ◐	● - -		- - -	- - -			---
Interactive (parallel sessions)	Gather.town	● ● ●	- ● -			● - ●	● ◐ -		- - -	- - -			---
Avatar, spatial audio	Hubs/VRChat	● ◐ -	● ● ●			● ◐ ◐	● ● ●		- ◐ ●	● - -			---
Spatial audio, video	ARENA/Vatom Spaces	● ● -	● ● ●			● ● ●	● ● ●		● ● ●	● - -			---

● *=provides property*; ◐ *=partially provides property*; − *=does not provide property*;

for exploration and social engagement CVE style systems excel. However, we did notice that for more informational tasks like giving a talk people preferred more traditional telepresence (many people asked for a full-screen slide mode in VR). We faced challenges with many of the wearable and mobile platforms in terms of comfort and issues around echo suppression with mobile microphones and speakers in the physical space (echo suppression is more difficult in environments where microphone locations are dynamic).

In summary, this paper makes the following contributions:

- We detail the design and evolution of a hybrid conference solution that connects virtual and physical participants with audio and video. This includes a workflow for scanning spaces and relocalizing a variety of devices.
- We present a set of user studies based on videos and real-world tests that show user preferences for various common conference interaction modalities.
- We discuss the effectiveness of a few innovative features, including: screen mirroring between virtual and physical, poster positioning and spatial sound configuration, virtual laser pointers, and sharing collaborative XR content.
- We present a simple taxonomy of operating points within the hybrid conference domain.
- We discuss pain points, limitations and open problems in the hybrid conference space.
- Provide open-sourced implementation of our XR experiences: event navigation, attending a talk, poster session, and extensions to the ARENA platform described in Section 3.1[1].

2 RELATED WORK

The idea of combining real and virtual spaces can be traced back to 1965, when Ivan Sutherland [22] proposed a trajectory toward virtual and augmented reality. Many attempts at creating this vision can be found in the literature, and [4] provides an interesting review. Focusing on currently available technology to implement a collaborative experience between local and remote attendees, we can identify two large categories: video conferencing systems and Collaborative Virtual Environments (CVEs), which we will review in the following subsections.

[1]Links to open-source implementation omitted in the spirit of double-blind review

Our study focuses on three representative video conferencing/CVE platforms: Zoom [14], Gather [5], and Hubs [16], which we compare with our XR experience built using ARENA components. These are described next, and Table 1 presents a comparison of available features in each platform.

2.1 Video conferencing

Video conferencing systems, such as Zoom, Teams, or Hangouts [8, 10, 14] provide a popular solution to collaborate between local and remote participants. However, they don't allow the serendipity of exploring the venue, meeting participants, and forming conversation groups. Recent studies have also identified known issues in these platforms due to prolonged close-up eye contact, lack of mobility, and higher cognitive load [1]. In this study, we selected Zoom as a representative video conferencing solution due to its popularity.

2.2 Collaborative Virtual Environments (CVEs)

Collaborative Virtual Environments (CVEs) are made of virtual worlds that multiple users can explore and interact with. In CVEs, users are represented through virtual avatars that can convey identity, location, and movement to other users. CVEs are also not new [4], but the increased interest in remote collaboration tools motivated an explosion of readily available platforms. One example, Gather [5] provides customizable 2D worlds where users can roam around represented by an avatar. As users get into each other's proximity, they can initiate a conversation supported by audio/video streaming. Gather also allows screen-sharing, chat, and other standard conferencing functionalities. While Gather offers more than the usual video conference, its 2D world still lacks in the user's sense of presence. It is also geared towards desktop experience, neglecting modalities made possible by mobile devices or headsets.

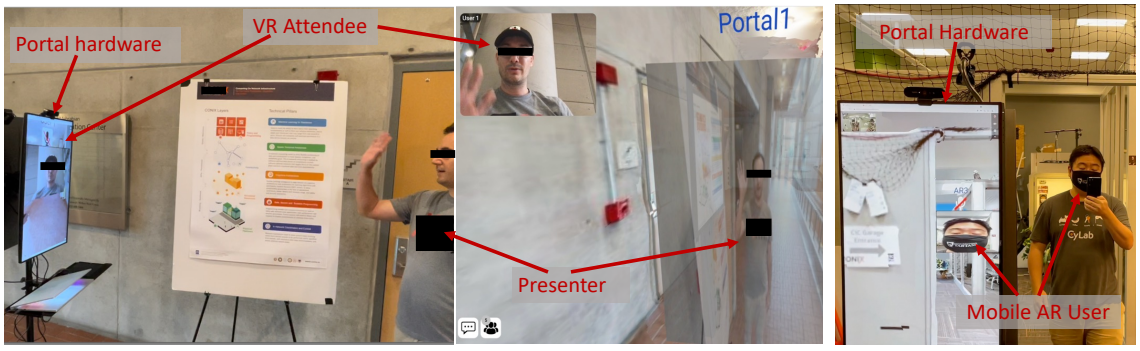
Many other CVEs create 3D virtual worlds that provide a better sense of presence and take advantage of mobile devices and headsets. AltspaceVR [9], Facebook Spaces [7], VRChat [13], Vatom Spaces [12] and Mozilla Hubs [16] are just a few examples of popular CVEs. While they have some differences in their target experience and audience (for example, VRChat focuses on highly customizable avatars and has a thriving community, whereas Hubs focuses on simpler avatars and smaller-scale events) they all offer a 3D customizable world in which virtual avatars can form conversation circles using spatialized audio. Due to their similarity, in this study, we focus on Hubs as the representative of 3D-world-based CVEs.

2.3 The ARENA

We built our hybrid conference experience using the open-source ARENA Platform [20]. This platform simplifies designing and deploying multi-user cross-platform mixed reality applications with built-in support for geographic content lookup, accurate relocalization, access control, the ability to host "hot-loadable" programs and manage user audio and video feeds at scale.

AR/VR Support. ARENA provides a programmable and network connected scene graph that can be displayed in AR or VR using a headset or within a 2D projection of a 3D environment in a desktop browser window. Figures 1,2, or 3, for example, present several examples of the same 3D environment displayed in both AR and VR. Users in the physical space can see the same 3D content anchored to the physical world in AR, and the properties of all objects (and state of users) are networked to provide a consistent real-time view from any device.

3D Exploration. Users can move through the 3D environment with the mouse, keyboard, or touchscreen swipes, long presses, and accelerometer rotations for mobile devices and VR headsets. In addition, in VR, users can effectively 'fly' (unlock their movement height and travel high above or below the ground plane) and 'teleport' (jump to near a user



(a) Photo of poster session with Portal hardware setup and poster presenter (left); Desktop VR (right).

(b) Mobile user in AR and Portal view into VR with user referenced in the real world.

Fig. 2. Double-sided AR/VR Portal examples.

or location). AR users physically move through the environment tracked by a headset relocalization system or with referenced optical markers on mobile phones and tablets.

User Presence. By default, users appear represented in the 3D environment as a static avatar, which provides an indication of the pose of their camera to other users. Users can enable or disable their microphone, video, or facial landmarks used to rig an avatar. Sound projecting from each user is spatial, so nearby users can be heard louder than users far away (there are controls for the sound drop-off characteristics). When turned on, a video-based avatar of the user is created by texture-mapping the live video onto the surfaces of a 3D cube, with the front side of the box highlighted, providing an important visual indication to other users about the direction the user is facing. On the user side, a video preview box shows the user what camera view will be transmitted. Users can also be represented by 3D deformable, rigged facial avatars that track users' facial movements. It translates a user's real facial expressions into the same 3D model expressions.

3D Content and Programs. ARENA makes it easy to load and display arbitrary 3D models from a shared filestore. We leverage this feature to connect scanned models of real-world spaces for our hybrid conferences. Figure 1 (right) shows an imported scanned model of a real-world space (the talk room). Large 3D models, scanned real-world spaces, and panoramic 360-degree photography are all supported depending on the desired degree of augmented reality or simulated virtual reality. Users can also share their screens to present slides or other material. The shared screen can be mapped onto one or multiple 3D surfaces. It is also possible to create programs to manipulate the 3D environment and create highly customized interactive environments.

3 THE HYBRID CONFERENCE

In support of hybrid conference experiences, we designed the following scenarios from both an AR and VR vantage point.

Event Navigation: Figure 3 shows a way-finding application that helps attendees navigate to a conference session in VR and AR. The AR component provides an incentive to connect people that might be interacting in the VR space.

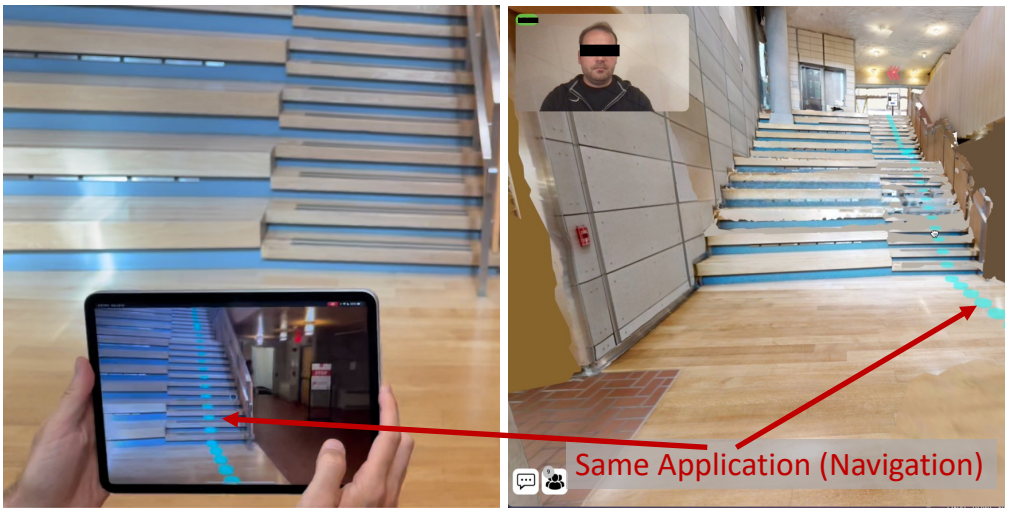


Fig. 3. Way-finding application viewed from AR and VR. Photo of AR view on an iPad (left); Desktop VR (right).

Fig. 4. Content (model of statue and screenshare) viewed from AR and VR. View from an iPad (left); Desktop VR (right).

Attending a Talk: In Figure 4, we see a speaker presenting slides with screenshare and addressing a 3D interactive model of a statue. The speaker can see both AR and VR users in the audience (shown in Figure 1). As mentioned below, it is possible to use a virtual laser pointer from AR and VR for the talk and audience in a Q/A mode.

Poster Session: We find that having a double-sided *AR/VR portal*, as shown in Figure 2, is a natural way to interact across the virtual/physical boundary. Remote participants can explore the poster session by seeing digital versions of the posters overlaid in the 3D model of the venue. They can peer through portals that allow participants to interact with speakers and see a glimpse into what is going on at the physical location. The virtual and real world geometry are aligned since the portal is registered with the 3D scan.

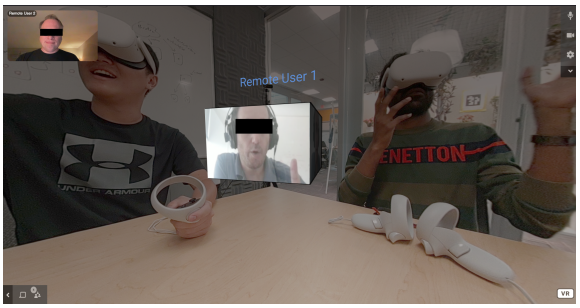


Fig. 5. Remote user view inside camera bubble where the image texture maps the background.

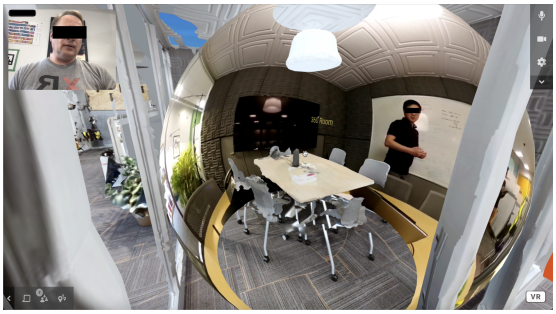


Fig. 6. Remote user view from outside of the camera bubble placed in the middle of a conference room.

In Section 3.1 we discuss a number of custom interactions designed specifically for conference environments. We also detail an important setup process to create the model of the venue and reference it in the real world (Section 3.2).

3.1 Custom Interactions

The applications created include a simple laser pointer, interactive manipulation of 3D models, simple games, and more. We extended the ARENA platform to support stand-alone cameras in the environment that map video onto large planes that act as AR/VR Portals as shown in Figure 2.

Laser Pointer: A simple laser pointer application can be loaded into the scene, letting users point to objects in a way that is visible to other users.

Manipulation of 3D models Pointer: We created an application that allows us to specify a 3D model that can be manipulated in 3D by using clicks or gestures. Figure 4.d presents an example of this application displaying a model of a statue.

XR Games: We also created a couple of games that can be used during social hours of the conference. **Treasure Hunt**: where finding hidden 3D models places a badge for each model on each avatar encouraging participants to share findings. **Pinata**: where a laughing pinata lounges in the sky asking for clicks, and after every 10 clicks the pinata loudly explodes until it respawns to taunt you again.

Portal Presence: AR/VR Portals (Figure 2) allow users in the real world to see a model aligned view into the VR space on a monitor. The opposite direction is then mapped to a live camera feed of the space for VR users to see real world participants.

Questions Session: We wanted users to be able to raise their hands in a hybrid conference, so we wrote an ARENA app to allow users to press a Question button on their screen to toggle the appearance of a Question icon above their avatar head.

360 Camera Bubble: We also included bubbles for a live camera feed of a conference room mapped onto a video sphere, which was placed in the same position within the 3D model (Figures 5 and 6). We wanted to give remote users the experience of entering a full live-streamed room, while physical users in the room could view the virtual visitors.

3.2 Venue Model

Creating a 3D model of the venue is an important step that allows remote participants to have a better sense of presence. The latest (2020) iPad and iPhone Pro models have laser scanners convenient for quick scans. For higher fidelity models,

Table 2. Experiment 1: Hybrid conference preferences questionnaire: sections, questions, and question IDs.

Section	Question	Question ID
Introduction	I have watched the video above (introduction video about the platforms).	Q1.0
2*Exploring Environments	Given the choice, which platform would you use to virtually explore a new environment?	Q1.1
	Given the choice, which platform would you use to virtually meet new people?	Q1.2
2*Live Presentations	Given the choice, which platform would you use to give presentations?	Q1.3
	Given the choice, which platform would you use to listen to/interact with presentations?	Q1.4
2*Private Conversations	Given the choice, which platform would you prefer if you were to enter a virtual/hybrid breakout room for a private conversation?	Q1.5
	Given the choice, which platform would you prefer when going to a virtual/hybrid poster session?	Q1.6
3*Social Interactions	If you wanted to virtually meet up with a group of colleagues from work for a social event, which platform would you prefer?	Q1.7
	If you wanted to virtually meet up with a group of close friends, which platform would you prefer?	Q1.8
	If you wanted to virtually meet up with teammates to work on a project, which platform would you prefer?	Q1.9

we use the Leica BLK360, a terrestrial laser scanner (TLS) with registered 360 color images. Using the BLK360 and 3D reconstruction software (e.g. Leica's Cyclone FIELD 360, or Matterport), we create 3D models of physical spaces that can easily be imported into ARENA. We place the laser scanner at different spots around the venue that are merged to create the final model. Each scan can take 30 seconds, or up to 2 minutes, depending on the resolution of the scan (we often use the fastest setting). The scan density might vary significantly from space to space, where large open spaces can be captured with few scans, and more complex areas require more scans due to occlusion. As an example, a model with over 400 square meters took less than 30 minutes to capture. Before loading the model, manual adjustments to simplify the model can be made using, e.g., Blender. A typical venue model (two floors, over 400 square meters) is less than 100MB, larger models can be split and loaded dynamically using a Level-of-Detail (LOD) [2] mechanism that allows for increasingly more detailed models as a user approaches.

Registering with the Real World: The underlying platform (ARENA) provides several mechanisms to streamline the management and sharing of anchor data and simplify the process of combining multiple tracking technologies into a uniform coordinate system. One simple, infrastructure-free, way of registering 3D content is using AR markers (such as AprilTags [24]) that can be placed in the venue and set as static or dynamic to determine if clients should use them for relocalization or to provide location information for the tag. For example, a 3D environment might contain several AR markers that have GPS coordinates and local coordinates referenced from the origin of the 3D virtual environment. ARENA's current client can decode AprilTags in browsers that allow camera access (e.g. Mozilla WebXR Viewer, and soon Chrome). If the client decodes a static marker, it uses the location data to compute the pose of the device's camera.

4 EXPERIMENTS

We used our prototype system in support of dozens of events, and have showcased it to multiple industry and academic collaborators. Events included research project meetings, poster sessions (at least 10), student project presentations, and brainstorming sessions, among others, and ranged from low 10s of participants up to 80. The experience generally received highly positive feedback, and it was particularly useful in capturing interactions across attendees in AR and VR. Some users noted that they prefer our experience to explore and interact with users in more social settings, such as poster sessions, but for more one-sided talks, they would prefer a video conferencing solution. We therefore wanted to examine more closely which solutions were more suited to the most common conference activities in hybrid mode and designed a survey (Experiment 1) to explore this. As we designed and used our experience, we also noticed users might

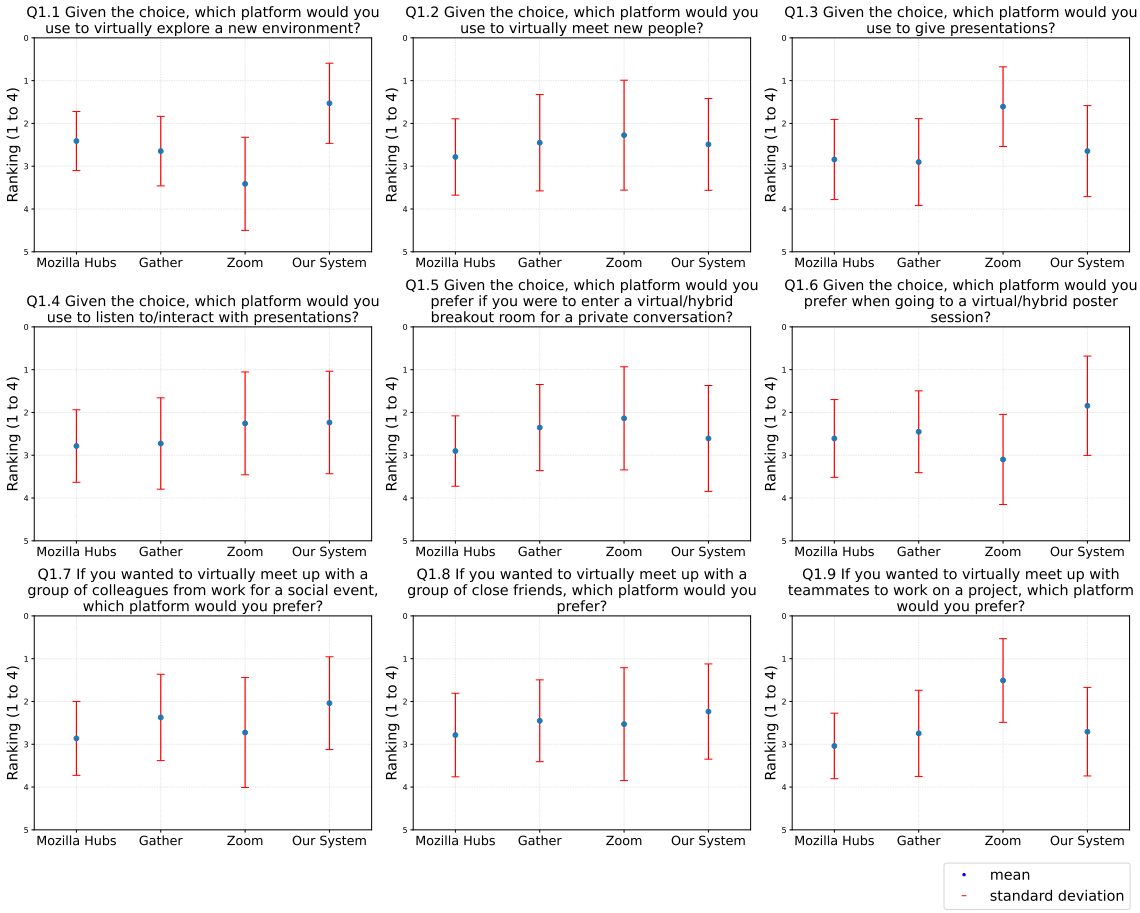


Fig. 7. Experiment 1: Hybrid conference preferences questionnaire results.

have different preferences when it comes to how they are represented in the virtual world (e.g. with a video stream, or an avatar) and thus, we created a survey (Experiment 2) to evaluate user presence preferences.

Finally, we realized that the conference experience was significantly different across different hardware platforms (e.g. AR versus VR headset) that had very distinct interaction modalities. Thus, we conducted a separate experiment to evaluate the users' workloads when performing XR tasks across four different hardware platforms (AR headset, VR headset, AR-capable tablet, and Desktop).

4.1 Experiment 1: Hybrid Conference Preferences

In order to study what kinds of video conferencing/CVE platforms a typical user prefers across a variety of hybrid conferencing scenarios, we ran an online survey comparing the experiences granted by four systems: Zoom, Gather.town, Mozilla Hubs, and our system.

4.1.1 Participants. The participants consisted of 51 students and non-students, inside and outside of the university, aged 18 and older. Participants had no prior knowledge of the experiment to which they contributed voluntarily and

signed a consent form. All participants had experience using Zoom and some had exposure to the other three platforms. No compensation was provided and no participant had hearing or vision impediments. The experiment was approved by the university's ethical committee.

4.1.2 Questionnaire. The questionnaire was composed of five main sections as follows.

Introduction: The first part of the study contained a short introduction video of the four platforms.

Exploring environments: The second section studied how each of the platforms handles the exploration of a new environment and the interactions of users. Participants were asked to rank their platform preference for the two scenarios.

Live presentations: Here, participants were asked to rank the platforms for their ability to allow users to give and listen to live presentations.

Private conversations: This part of the survey asked participants to rank which platform they would prefer when going into a private conversation/breakout room and a poster session.

Social interactions: The final part of the study asked for platform preferences in a variety of social scenarios (i.e. meeting with colleagues, teammates, and close friends).

See Table 2 for the full questionnaire.

4.1.3 Procedure. Participants were asked to complete the survey by email. The email contained a link to the questionnaire with the sections detailed above [2]. Each section of the questionnaire included a short video displaying the platforms and their relevant features. After watching the video, participants were asked to rank the platforms using a 4-point scale (1: preferred 4: least preferred) in order of their preference to perform different activities as shown in Table 2. A final section of this survey pertained to Experiment 2. After a brief introduction, participants would proceed to answer questions related to Experiment 2 (Section 4.2).

4.1.4 Results. We present the mean and standard deviation of rankings for each platform for each question in Figure 7. Note that in the y-axis, ranking 1 (higher in the plot) is better. To compare participant's preferences, we employed a pair-wise Wilcoxon Signed Rank test, as the datasets of rankings per platform per question were found to not be normally distributed, according to a Shapiro-Wilk test (all datasets had a p-value of less than 0.05).

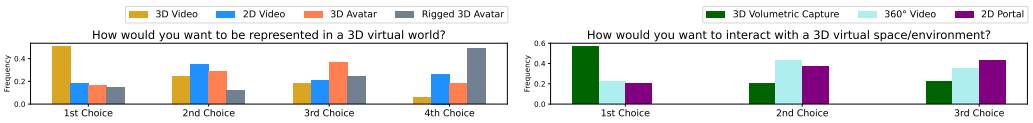
Exploring environments. Participants had a strong preference for our system to explore new environments (Q1.1). Comparing the platform rankings pair-wise with the Wilcoxon Signed Rank test indicated there was a statically significant difference ($p < 0.05$) in the rankings between our system and the rankings of all other platforms. This means the preference for the usage of our system over other systems to explore new environments is statically significant. To virtually meet new people (Q1.2), participants had a preference for Zoom, whose rankings had a statically significant difference compared to those of other platforms. Our system and Gather had similar rankings, with no statically significant difference ($p \geq 0.05$).

Live presentations. To give live presentations (Q1.3), participants had a strong, statistically significant preference for Zoom over other platforms ($p < 0.05$). The rankings of Gather and Hubs were similar, and did not have a statically significant difference ($p \geq 0.05$). Interestingly, to listen to/interact with presentations (Q1.4), participants, on average, ranked the experience of our system very similarly to Zoom, with both systems having a similar preference for users. We confirmed the difference in rankings of our system compared to Zoom was not statistically significant ($p \geq 0.05$).

[2]Reviewers can analyse a version of the survey redacted for anonymity (the original version had contact information and school policies) here: https://forms.gle/drHd31kdMsoJz8Mc6

Table 3. Experiment 2: XR Presence and World Preferences questionnaire: sections, questions, and question IDs.

Section	Question	Question ID
XR Presence Type	How would you want to be represented in a 3D virtual world? (among 3D video, 2D video, 3D avatar, rigged 3D avatar)	Q2.0
XR World Representation	How would you want to interact with a 3D virtual space/environment? (among 2D portal, 360 video, 3D volumetric capture)	Q2.1



(a) Preferred representation in a 3D virtual world.

(b) Preferred interaction in a 3D virtual space

Fig. 8. Experiment 2: XR presence and world preferences.

Private conversations: Participants had a slight preference for Zoom when it comes to private conversations (Q1.5) and a strong preference for our system for hosting hybrid poster sessions (Q1.6). We confirmed the difference in rankings between all platforms for this question was statistically significant ($p < 0.05$).

Social interactions: Participants had a preference for our system to meet up with colleagues and friends (Q1.7 and Q1.8), and we found the difference in rankings between our system compared to those of the others to be statistically significant ($p < 0.05$). When it comes to working meetings (Q1.9), participants had a statistically significant ($p < 0.05$) preference for Zoom, while our system had a similar preference with Gather, with no statistically significant ($p \geq 0.05$) difference in rankings.

4.2 Experiment 2: XR Presence and World Preferences

Our second online study asked participants to rank their preferences of XR presence while in a hybrid conference.

4.2.1 Participants. The participants in this survey are the same as in Experiment 1.

4.2.2 Questionnaire. The questionnaire was composed of two main sections as follows (see Table 3 for the full questionnaire).

XR Presence Type: We asked participants to rank their preferred presence modality in XR (3D live video, 2D live video, 3D avatar model, and rigged 3D avatar with facial landmarks translations).

XR World Representation: We asked participants to rank their preferred representations of a 3D environment (2D portal, 360 video, 3D volumetric capture).

4.2.3 Procedure. The questions in this experiment were presented to participants after answering Experiment 1's questions [3]. We presented each participant with four different user representations (3D live video, 2D live video, 3D avatar model, and rigged 3D avatar with facial landmarks translations) and asked them to rank their preference using a 4-point scale (1: preferred 4: least preferred) in order of their preference. Finally, we presented three representations of a 3D environment in our system (2D portal, 360 video, 3D volumetric capture).

4.2.4 Results. Figure 8 presents the presence and interaction preferences of the participants. We can observe a clear preference for 3D video as the preferred representation and 3D volumetric capture as the favorite interaction.

[3]Reviewers can analyse a version of the survey redacted for anonymity (the original version had contact information and school policies) here: https://forms.gle/drHd31kdMsoJz8Mc6

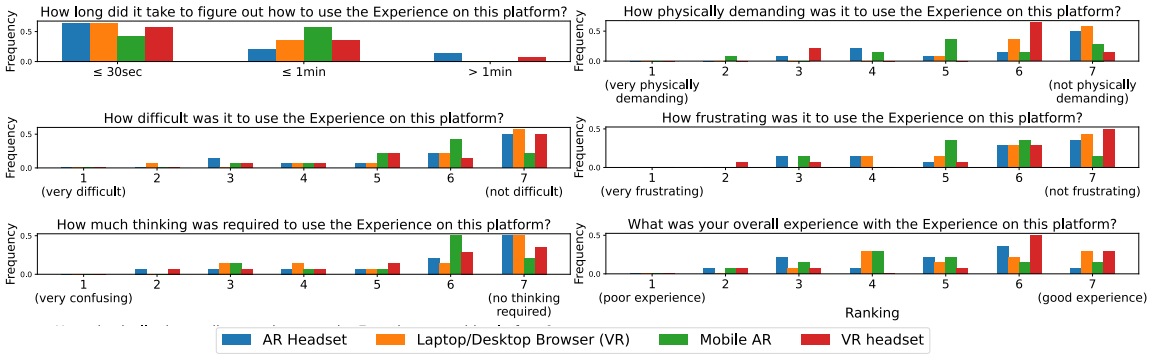


Fig. 9. Experiment 3: XR tasks work load questionnaire results.

4.3 Experiment 3: Workload of XR Tasks

We also studied the user workload when performing tasks (playing an XR game) in our system across four distinct XR hardware platforms: an AR headset, a mobile AR device, a desktop environment, and a VR headset.

4.3.1 Participants. To conduct this study, we gathered 14 university students who have some experience with at least two XR platforms (of the four used: AR headset, mobile AR device, desktop environment, VR headset). Participants had no prior knowledge of the experiment to which they contributed voluntarily. No compensation was provided and no participant had hearing or vision impediments.

4.3.2 Apparatus. We used four distinct XR hardware platforms: HoloLens 2 (AR headset), iPad (mobile AR device), a MacBook Pro (a desktop environment), and Oculus Quest 2 (a VR headset). The iPad contained an experimental XR Browser, capable of running WebXR. The HoloLens and iPad were both running a WebXR immersive AR session, while the Quest was running a WebXR immersive VR session. VR devices ran an XR environment containing a 3D volumetric scan of our Laboratory, while AR devices had users physically enter the Laboratory in real life. The scene included two interactive games: a tic-tac-toe application and a collaborative piñata hitting application.

4.3.3 Questionnaire. The survey contained 6 questions per device. Each question corresponded to a subscale of NASA's Task Load Index (TLX) [6]. Participants were asked to rank on a scale from 1 (worst) to 7 (better) on how temporally, mentally, and physically demanding the usage of the XR environment was on each device, as well as how difficult and frustrating the experience was. Then, they were asked to rank the overall experience on each device.

4.3.4 Procedure. Participants entered the same XR environment using each of the four devices and were given some guidance on how to use the environment on each device. Then, they were instructed to move around the environment and interact with the applications and other users. After using all four of the devices, the participants were given a survey asking them about their experience.

4.3.5 Results. Figure 9 details the results of this study. We observe most participants took less than 30 seconds to be comfortable using the XR environment on the AR headset, laptop, and VR headset, while the interface on the mobile AR device took longer to understand. Results also show that the VR and AR headsets were more physically demanding, particularly the VR headset. This can be explained by the Quest 2 being the most front-heavy of the devices. The laptop experience, despite being the easiest and most accessible to use, was underwhelming compared to the experience

provided by the VR and AR headsets. Overall, the experiences across the four devices was ranked positively by the participants.

5 LESSONS LEARNED

We now discuss some lessons learned over the past two years of using and developing our system to host a variety of both planned and impromptu virtual and hybrid conferences.

5.1 Impact of Event Size and Scaling

We learned that the scale of space and ratio of users has a significant impact on the social dynamic, and the use of the system for events evolved over time. Initially our poster sessions were quite sparse, spreading out posters in even rows, seeming like monoliths. Then, an accidental ad-hoc social gathering suggested more people in tighter spaces with better audio tuning substantially promotes discussion. Intuitively, short transmission range audio isolates poster sessions, but extended audio invites openings for social meeting points. We imagine that in many contexts, this should be dynamic across spaces.

Spatial Sizing. Since navigation in 3D systems use a default navigation mesh to help new users navigate it is useful to keep text, especially small text, closer to the ground plane. In our experience, formatting posters in landscape orientation provides easier reading than the same scale in portrait orientation. Physically spacing these posters at least 15m apart allows enough space to configure audio for good social dynamics.

Audio for Social Dynamics. Since the underlying platform supports spatial sound, we found it useful to configure audio settings beyond the default A/V cutoff distance of 20m. Inspired in part by other poster session experiences [18], we configured multiple audio qualities with an example in (**bold**) that provided us with a more natural audio experience:

- **A/V Cutoff:** the maximum distance between cameras/users until audio and video are cut off (**10m**).
- **Audio Distance Model:** an algorithm to reduce the volume of the audio source as it moves away from the listener (exponential, **inverse**, linear).
- **Volume Level:** what is nominal for all users in a scene (**1**).
- **Audio Reference Distance:** at which the volume reduction starts taking effect (**3m**).
- **Audio Rolloff Factor:** or how quickly the volume is reduced as the source moves away from the listener (**5**).

Social Experiences. We found that many entering a 3D video conference for the first time can be socially awkward just as in physical-only events. To help break the ice, we used the programmability of the underlying platform to add fun interactions to draw people in and we described some XR games we created in Section 3.1.

5.2 Challenges

Some challenges remain in the hybrid conference space that we think may be unique to these experiences.

Audio Cross-talk and Feedback When several AR-devices (phones, tablets, headsets) are co-located in a hybrid scene, the system is not yet well equipped to handle the cross-device bleed from multiple broadcast-style speakers. The digital and analog boundaries are not well synchronized, resulting in waves of echo feedback to remote participants and ugly feedback squelches for physical participants. The only immediate remedy being selective speaker and microphone disabling, albeit clumsy. Alternatively, everyone could wear headphones but this is awkward for the in-person event. Room-scale microphone and speaker arrays might be better suited to solve this issue if they can be well integrated into our system, deterministically sampling and synthesizing physical spatial audio which can be aligned to physical and remote participants as appropriate. Additionally, a system may need to encourage users to become more immersed in

audio, by detecting and discouraging broadcast speakers, in favor of stereo headphones mounted to wearables like AR glasses.

Virtual and Physical Clique-Forming. We have observed that people tend to form cliques between virtual and physical modes, habituating to similar sensory quality of experience. It may be helpful to democratize audio for physically co-located participants as in the previous paragraph to help reduce this effect. Also an immersive 360 portal bridge may be formed to melt the boundaries between worlds. This scenario is a hybrid form of the 2-Way Portal from earlier. It requires a room with a large, flat screen, and a mounted 360 camera stream. The flatscreen is used to present the VR mirror image of remote participants to physical participants, and the 360 camera is used to present physical participants to remotes.

Lagging Hardware. Most AR-wearable headsets have limited ability to maintain a consistent AR digital position localized without noticeable drift. Additionally, reality capture systems tend to require compute heavy processing making constant capture of LIDAR depth captures over several hours challenging. Memory issues also plague headsets and handheld devices, not having the capacity to handle large scanned models of environments. Using a Level-of-Detail (LoD) mechanism for models, and selectively removing models in AR dramatically helped to support less powerful devices.

Privacy. Fundamental to camera-based AR technologies are issues of privacy. These systems need access to sensors, for example, to perform computer vision tasks such as plane detection or perform optical flow or computing occlusion from real-time depth sensors like LIDAR. In Web-based XR systems, access to these is mediated by browsers that need to walk a thin line between adequate, privacy-preserving, defaults and obtrusive permission dialogues.

We also saw many situations where people worried about potential conversation snooping between AR and VR. In real life, when people are behind a barrier, it is usually quite apparent they are there since you can subtly see or hear them. If those people are in VR, this becomes much less obvious. A radar map of nearby users can help inform when others are within earshot.

6 CONCLUSION AND FUTURE WORK

We presented a hybrid conferencing system designed to facilitate interactions between remote and in-person attendees with a shared sense of the venue. Our solution allows remote users in VR to interact with in-person users in AR in various manners. VR users navigate a 3D scanned version of the environment with avatars or floating video cubes representing AR users in the physical environment. AR users can see VR user avatars (and audio/video) as AR holograms. We experimented with several techniques like portals, 360 video bubbles, and phones/tablets that could act as windows between the two worlds. The most accessible example is using a mobile phone's rear-facing camera to capture the AR user while the screen shows a pass-through AR projection of VR content. All devices in AR are registered in the real-world.

Through several user studies, we learn that people prefer CVE-style platforms for social interactions but prefer standard 2D telepresence for more directed interactions like talks. We also see that people tend to prefer 3D video cubes instead of synthetic avatars as a representation since they provide many of the routine, often subtle, social cues while also giving a sense of location and directionality in the virtual space. We also provided some general guidelines learned from many poster sessions, lab meetings, and social gatherings that helped us understand the impact of event sizing, acoustic setup, and general content layout. Finally, we discussed several open challenges, both technical and social, that need to be overcome in order for this sort of approach to be more broadly adopted.

While our system provides a glimpse at the future of hybrid conferences, we have a long way to go before we achieve truly immersive telepresence. 3D scans of the environment and real-time scene capture are relatively low-resolution, have sparse coverage, and struggle to update at typical video frame rates. In addition, AR wearable technology needs improvements in terms of form factor, battery life, resolution, motion-to-photon latency, and field of view. In the near future, we will focus on portal hardware with high resolution 360 video and more immersive and practical displays.

REFERENCES

[1] Jeremy N. Bailenson. 2021. Nonverbal Overload: A Theoretical Argument for the Causes of Zoom Fatigue. *Technology, Mind, and Behavior* 2, 1 (feb 23 2021). https://tmb.apaopen.org/pub/nonverbal-overload.

[2] James H Clark. 1976. Hierarchical geometric models for visible surface algorithms. *Commun. ACM* 19, 10 (1976), 547–554.

[3] Mingsong Dou, Henry Fuchs, and Jan-Michael Frahm. 2013. Scanning and tracking dynamic objects with commodity depth cameras. In *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 99–106. https://doi.org/10.1109/ISMAR.2013.6671769

[4] Barrett Ens, Joel Lanir, Anthony Tang, Scott Bateman, Gun Lee, Thammathip Piumsomboon, and Mark Billinghurst. 2019. Revisiting collaboration through mixed reality: The evolution of groupware. *International Journal of Human-Computer Studies* 131 (2019), 81–98. https://doi.org/10.1016/j.ijhcs.2019.05.011 50 years of the International Journal of Human-Computer Studies. Reflections on the past, present and future of human-centred technologies.

[5] Inc. Gather Presence. 2022. Gather Website. https://www.gather.town/. Online. Accessed: May 2022.

[6] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[7] Facebook Inc. 2022. Facebook Spaces. https://www.facebook.com/spaces. Online. Accessed: May 2022.

[8] Google Inc. 2022. Google Hangouts Website. https://hangouts.google.com/. Online. Accessed: May 2022.

[9] Microsoft Inc. 2022. AltspaceVR. https://altvr.com/. Online. Accessed: May 2022.

[10] Microsoft Inc. 2022. Microsoft Teams Website. https://www.microsoft.com/en-us/microsoft-teams/group-chat-software. Online. Accessed: May 2022.

[11] Optitrack Inc. 2022. Optitrack Motion Capture Systems. https://optitrack.com/. Online. Accessed: May 2022.

[12] Vatom Inc. 2021. Vatom Virtual Spaces. https://www.vatom.com/platform/virtual-spaces/. Online. Accessed: May 2021.

[13] VRChat Inc. 2022. VRChat. https://hello.vrchat.com/. Online. Accessed: May 2022.

[14] Zoom Inc. 2022. Zoom Website. https://zoom.com/. Online. Accessed: May 2022.

[15] Brennan Jones, Yaying Zhang, Priscilla N. Y. Wong, and Sean Rintel. 2021. Belonging There: VROOM-Ing into the Uncanny Valley of XR Telepresence. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 59 (apr 2021), 31 pages. https://doi.org/10.1145/3449133

[16] Duc Anh Le, Blair MacIntyre, and Jessica Outlaw. 2020. Enhancing the Experience of Virtual Conferences in Social Virtual Environments. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. 485–494. https://doi.org/10.1109/VRW50115.2020.00101

[17] Gun A. Lee, Theophilus Teo, Seungwon Kim, and Mark Billinghurst. 2017. Mixed Reality Collaboration through Sharing a Live Panorama. In *SIGGRAPH Asia 2017 Mobile Graphics and Interactive Applications* (Bangkok, Thailand) *(SA '17)*. Association for Computing Machinery, New York, NY, USA, Article 14, 4 pages. https://doi.org/10.1145/3132787.3139203

[18] Blair MacIntyre. 2020. VR2020: Audio Design for Public Rooms. https://blairmacintyre.me/2020/04/03/vr2020-design-of-a-poster-room/. Online. Accessed: May 2022.

[19] Edwin Olson. 2011. AprilTag: A robust and flexible visual fiducial system. In *ICRA*. IEEE.

[20] Nuno Pereira, Anthony Rowe, Michael Farb, Ivan Liang, Edward Lu, and Eric Riebling. 2021. ARENA - The Augmented Reality Edge Networking Architecture. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE.

[21] Patrick Stotko, Stefan Krumpen, Matthias B Hullin, Michael Weinmann, and Reinhard Klein. 2019. SLAMCast: Large-scale, real-time 3D reconstruction and streaming for immersive multi-client live telepresence. *IEEE transactions on visualization and computer graphics* 25, 5 (2019), 2102–2112.

[22] Ivan Sutherland. 1965. The ultimate display. (1965).

[23] Theophilus Teo, Louise Lawrence, Gun A. Lee, Mark Billinghurst, and Matt Adcock. 2019. Mixed Reality Remote Collaboration Combining 360 Video and 3D Reconstruction *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300431

[24] John Wang and Edwin Olson. 2016. AprilTag 2: Efficient and robust fiducial detection. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 4193–4198.

[25] Cheng Xiao and Zhang Lifeng. 2014. Implementation of mobile augmented reality based on Vuforia and Rawajali. In *ICSESS*. IEEE.