Volumetric Mixed Reality Telepresence for Real-time Cross Modality Collaboration

Andrew Irlitti*
The University of Melbourne
Melbourne, Australia

Mesut Latifoglu
The University of Melbourne
Melbourne, Australia

Qiushi Zhou
The University of Melbourne
Melbourne, Australia

Martin Reinoso
The University of Melbourne
Melbourne, Australia

Thuong Hoang
Deakin University
Geelong, Australia

Eduardo Velloso
The University of Melbourne
Melbourne, Australia

Frank Vetere*
The University of Melbourne
Melbourne, Australia

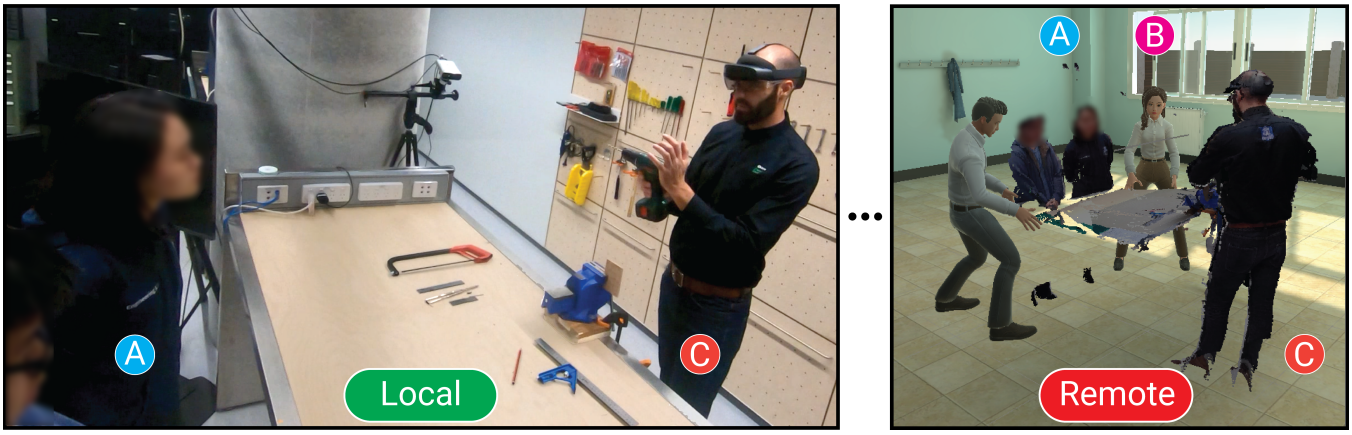


Figure 1: A multi-user, cross-modality, mixed reality telepresence communication platform, supporting face-to-face, AR, and VR collaboration. A volumetric capture of the local environment is transmitted to several remote sites, where the environment is reconstructed with all local users represented as volumetric hologram avatars standing inside the physical environment (A, C). At each remote site, the reconstruction is presented inside a virtual environment, with remote users represented by an avatar (B). Users wearing head-mounted displays (B, C) can see both local and remote users co-existing inside the same space. All users can hear one another, allowing for communication to occur as if everyone were present together in the same space.

ABSTRACT

Mixed-reality telepresence allows local and remote users feel as if they are present together in the same space. In this paper we report on a mixed-reality volumetric telepresence system that is adaptable, multi-user and cross-modal, i.e. combining augmented and virtual reality technologies with face-to-face interactions. The system extends state-of-art by creating full-body and environmental volumetric renderings in real-time over local enterprise networks. We report findings of an evaluation in a training scenario which was adapted for remote delivery and led by an industry professional. Analysis of interviews and observed behaviours identify varying attitudes towards virtually mediated full-body experiences and highlight the impact of volumetric mixed-reality telepresence to facilitate personal experiences of co-presence and to ground communication with interlocutors.

* ✉ andrew.irlitti@unimelb.edu.au — ✉ f.vetere@unimelb.edu.au

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04…$15.00
https://doi.org/10.1145/3544548.3581277

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **Mixed / augmented reality**; **Empirical studies in HCI**.

KEYWORDS

Extended reality, augmented reality, virtual reality, volumetric capture, telepresence, avatars, collaboration, conversational grounding

ACM Reference Format:
Andrew Irlitti, Mesut Latifoglu, Qiushi Zhou, Martin Reinoso, Thuong Hoang, Eduardo Velloso, and Frank Vetere. 2023. Volumetric Mixed Reality Telepresence for Real-time Cross Modality Collaboration. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3544548.3581277

1 INTRODUCTION

In this paper, we describe the design, implementation and evaluation of *Virtual Co-Presence* (VCP), a volumetric capture telepresence system. VCP uses mixed reality to support immersive and grounded communication, where collaborators are able to coordinate and share a mutual understanding of their practices and contexts in multi-user environments (Figure 1). Motivated by Holoportation [44], the implementation builds on this seminal example by producing real-time telepresence on readily available consumer hardware, affording deployment in dynamically configured environments. The network protocol supports real-time delivery to multiple simultaneous clients connected over an existing local enterprise network. The capture and calibration environment does not require any additional setup, such as green screening, and allows for cameras to be placed in a variety of configurations to achieve the required coverage. The platform captures people and environments as volumetric objects, incorporating simultaneous multi-user volumetric and full-body 3D avatars, offering a platform to explore the human experience of virtual telepresence.

This experience is investigated in a training scenario involving five users, in a cross-modal environment across face-to-face, augmented, and virtual realities, spanning the entire Mixed Reality continuum [43]. The volumetric avatars and volumetric environments generate a rich context for communication that supports shared workspace awareness, co-presence, and multi-modal proxemics. The study aims to provide valuable insights into the human experiences and attitudes of virtually mediated full-body interactions. The study involves a face-to-face training task for on-boarding novice users in a fabrication lab. This real-world context offers unique insights into better understanding the potential of mixed reality telepresence in unstructured environments. The findings of the study contribute to understanding user experiences in multi-user mixed reality environments and inform technical improvements to mixed reality telepresence systems.

The core contributions of this paper are:

- technical details about the design and implementation of a mixed reality telepresence system that supports the capture and visualisation of full-body, real-time collaborative teams across physical and virtual spaces in dynamically configured environments.
- insights about the human experience of mixed reality telepresence in multi-user, multi-modal environment, for supporting co-presence, workspace awareness, and communication.

2 RELATED WORK

2.1 Mixed Reality

The *Mixed Reality continuum* [43] describes a shifting user experience where the physical environment is gradually replaced with a virtual computer-generated world. On one end of the continuum, Augmented Reality (AR) presents a view of the physical world where visual and auditory senses are augmented with virtual information that enhances the user experience [2]. On the other end, Virtual Reality (VR) replaces the user's view of the physical world with a fully synthetic representation, affording unique experiences that are unattainable within the limits of their physical form [9]. Through mixing realities, new opportunities for collaborative human experiences are available by leveraging the technology to support remote mediation [8, 41, 60, 62]. Capitalizing on the characteristic of supporting viewpoint independence [10], mixed reality affords mutual grounding of communication, where collaborators establish a common understanding of their partner's actions and expectations by working together [13].

Telepresence extends this idea by incorporating the human body as part of the virtual experience. At its core, telepresence is typically mediated through some form of mixed reality [10]. In telepresent experiences people are rendered as virtual copies and broadcast to a remote location, ideally affording remote embodied collaborative experiences [44].

2.2 Volumetric Environments

Multiview camera environments have attracted notable attention in their ability to capture and create three dimensional content, emphasizing the interactive opportunities to update and control viewpoints within the scene [33]. This digitization is achieved through a process called volumetric capture, where an array of cameras capture and reconstruct the content in a volume form [16, 17, 32]. Seminal examples [31, 34] captured remote environments with large arrays of RGB cameras using stereo reconstruction. Collet and colleagues [14] combined 53 pairs of RGB and IR cameras in a green-screen environment to reconstruct recorded actions within the capture space. The high-quality results were post-processed down into network streamable video. *Motion2fusion* [16] further refined the quality of recorded movements using their scalable approach for single or multiple camera environments in arbitrary challenging scenes. The *Office of the Future* [51] proposed the illusion of connecting remote environments through wall projections, affording users the ability to view into each other's office environment with correct perspective as if looking through a glass window. With the introduction of commodity depth sensors, such as the Kinect, a variety of real-time multi viewpoint examples emerged using a small number of cameras, demonstrating their effectiveness of reconstructing environments to support interactivity [7, 19, 32, 57, 65]. *RoomAlive* proposed an interactive projection mapping system combining depth cameras alongside projectors providing real-time updating of environmental surfaces for dynamic projection mapping scenarios [30, 64]. Fender and colleagues further explored the use of the *RoomAlive* toolkit to produce an omni-directional display system to support collaborative face-to-face group meetings, turning an environment into display surfaces to share content [20].

The *Proximity toolkit* [40], based on proxemic concepts of how people utilize interpersonal space to communicate with other people [24] and other interactive devices [5], uses a sensor enabled environment to track objects and people, inferring proxemic interactions based on orientation, distance, identity, motion, and

location. The *Creepy Tracker* toolkit [57] incorporates an array of depth sensors to provide multi-user interactivity in a context-aware environment. The toolkit supported numerous co-located and remote experiences, demonstrating the flexibility of using RGB, depth, and skeleton data from RGBD sensors to create interactive experiences. To support the increasing mechanisms to spatially monitor an environment, the *Society of Devices toolkit* [56] abstracted sensor data to allow a plug and play experience, fusing multiple sensor sources for multi-user, multi-device ubiquitous deployments. The *Velt* framework supports the integration of multiple RGBD and skeleton streams in a camera network, offering a flexible and modular platform to control, inspect, and playback camera feeds for environmental interactions [19].

The range or various toolkits demonstrate that environmental and contextual understanding is important in mixed reality experiences. As people and environments are captured and rendered, the underlying physical relationships are also captured and rendered, offering unique opportunities for experiences between people and environments. Proxemic and temporal relationships between users and objects, are essential considerations for realizing mixed reality telepresence.

2.3 Telepresence

The notion of telepresence has been widely evoked in science fictions, where people are transported to a remote environment [25, 63]. Maimone, Fuchs and colleagues demonstrated early telepresence experiences through the combination of multiple depth sensors. Their work allowed users to co-exist within the capture space with correct occlusion, reconstructing the physical environment in real-time [37–39]. More recently, *LiveScan3D* [32] allowed for the real-time capture and reconstruction of users surrounded by an array of depth sensors. Multiple commercial solutions have also become available to deliver volumetric capture installations [42, 54].

The *Beaming* telepresence system [58] "beamed" a single participant into a remote group of users. Local users were presented with an avatar representation of the beamed user on a spherical display. *Room2Room* [46] used depth sensors in two remote locations to capture and transmit a user's pose and actions to a colleague in the other room. The system used projectors to produce view-dependent renderings on the wall and floor in front of each user, augmenting the remote collaborator into their local space. Beck and colleagues supported eight simultaneous users across two sites to support remote experiences [7]. Using their stereoscopic projection environment, the system used depth cameras to capture and recreate meshes from each group, projecting the results on a projection screen at each site, with individual shutter glasses being used to display view dependent renderings for each user. *Remixed Reality* [35] offered a unique perspective on asynchronous telepresence, using a volumetric capture space as an environment to experience temporal augmentations in virtual reality. As an alternative to head-mounted displays, Fairchild et. al. [18] designed a multimodal remote collaborative system to support non-verbal communication in a collaborative space operation. The *withyou* telepresence system incorporated a single display, a wall display, a 2-sided CAVE, and a desktop computer across two European countries with 3D avatars used to represent users not using a volumetric capture site. Further examples have demonstrated telepresence without users needing to wear head-mounted displays, with *ImmerseBoard* [26] allowing users the experience of writing side-by-side with a remote collaborator on a whiteboard, and 3D-board [65], where the remote user is visualized inside and alongside the interactive content.

Holoportation [44] established a real-time end-to-end telepresence process between two physical-twin environments. Each environment combined 8 nodes containing depth, near infrared, and color cameras to produce a high-fidelity mesh of the user and their local environment. The environment data from one site is transmitted across a 10GbE network and rendered using a Microsoft Hololens with correct occlusion inside another site. While the system did not use any environmental mediating technology, such as green-screening, the high visual fidelity and performance required substantial custom hardware and costs.

2.4 Visual Fidelity

Fidelity of the virtual application is often highly sought after in virtual user experiences [17, 23, 44]. However, high visual fidelity is not always needed for effective user experience of telepresence. Regenbrecht and colleagues [52] showed that users can hold feelings of presence in a volumetric rendering with limited graphical resolution. Additionally, too much photorealistic fidelity can be detrimental to user experiences. Cho et al. [12] compared 2D video, a prescanned avatar, and live volumetric capture from a single RGBD camera during two collaborative tasks. Their results indicated that in tasks motivated by communication between participants, prescanned photo-realistic avatars suffered from the uncanny valley effect, while volumetric capture increased perceived social presence. In *Loki* [61], researchers explored the use of volumetric capture for remote instruction of physical tasks in mixed reality, enabling a learner and instructor to engage through the combination of virtual avatar and volumetric capture using mixed reality displays. While *Loki* only incorporated a brightly coloured head to represent the remote learner and instructor inside a raw point cloud reconstruction, users found *Loki* to be supportive to collaboration, acknowledging feelings of being connected with their peer and the impact of social dynamics in local and remote spaces.

2.5 Avatar Representation

Avatar representation is an important consideration when interacting with others in mixed reality social environments [45]. Avatar representation is commonly abstract or cartoon-like in design [4, 36]. The virtual doppelganger is an alternative approach which produces a digital lookalike avatar of our physical selves [3, 22]. Virtual doppelgangers are created using a set of digital photographs and head modeling software. Bailenson and Segovia [3], posit that virtual doppelgangers can have a significant influence on our behaviour and memory due to the self-identification process, where a user can recognise and accept ownership of the avatar, even if its creation is from a set of static images. Volumetric capture is the logical extension of the virtual doppelganger, by presenting real-time, full-body captures of the user [16, 17, 44]. Users can be represented by an avatar from their own digital representation, in much the same way as traditional video recording but embracing a three dimensional structure. Volumetric capture technology can

enable a sense of co-presence [44, 52] through the holistic virtualization of users and their surrounding environment [13]. Anjos et al. considered the effect of volumetric avatar sizing on communication [15], rendering a remote avatar as an adequate size to allow existing communication techniques to be applied through the wearing of Head Mounted Displays (HMDs). Piumsomboon and colleagues evaluated the effects on social presence and engagement with adaptive avatars [49], while also considering interaction techniques to support the effective communication with a tangible giant-miniature collaborator [48]. In both examples, users found the avatar provided social grounding, and assisted in conveying spatial awareness cues in the shared environment.

2.6 Research Themes

This review of the research, concepts and examples of Mixed-Reality Telepresence has identified several key themes.

- **Cross-modal designs.** The review shows examples of telepresence systems that leverage technologies across Milgram's Mixed Reality continuum [43]. Some examples focus on telepresence between physical environments and augmented reality environments [39, 44, 46], and others between physical environments and virtual environments [35, 52, 61]. However, to the best of our knowledge, none have explored the possibilities of a telepresence that incorporates the entire range of the continuum; integrating the physical, augmented and virtual realities simultaneously to create a cross-modal design.
- **Multi-user experiences.** While examples of telepresence systems have significantly evolved over recent years, they are mostly designed to support only two users simultaneously [44, 46, 61]. Human experiences of social interactions typically involve more than one other person. A rich exploration of telepresence should allow interactions with multiple people.
- **Impact of fidelity on user experience.** Telepresence systems typically aim for high-fidelity and high-resolution rendering of avatars and environments. Yet, findings by Regenbrecht and colleagues [52] suggest that lower fidelity telepresence systems may not be detrimental to user experiences. The impact of fidelity of virtual and augmented representations of avatars and environments in multi-user telepresence contexts is unclear.

This study investigates these themes through the Virtual Co-Presence (VCP) system. To the best of our knowledge, it is the first work to investigate the user experience of a telepresence system where there are more that than two users simultaneously interacting across multiple modes (physical, augmented and virtual) in contexts that prioritise the human activity rather than fidelity of the representation.

3 VCP: VIRTUAL CO-PRESENCE

Virtual Co-Presence (VCP) is a volumetric capture platform that enables multi-user remote collaborative instruction using mixed reality telepresence. The system incorporates a configurable number of off-the-shelf Azure Kinect depth cameras, leveraging the rendering and interaction properties of both virtual and augmented reality to explore the social behaviours in cross-modality collaborations. Users can be represented by either a real-time volumetric reproduction of their physical appearance or embody a rigged 3D avatar model to represent their actions. The following section provides a description of the system and its contribution beyond prior work.

3.1 Technical Platform

VCP is developed using Unity 2020.3 LTS on the Windows platform. The capture processing pipeline is achieved on an off-the-shelf Alienware Aurora R13 Gaming Desktop PC, with an Intel Core i7-12700KF, and a NVIDIA GeForce RTX 3080 GPU. The PC is connected to a cabled network using a $1000BaseT$ connection over an existing enterprise network. To support real-time transmission of volumetric recordings to multiple XR clients, the process employs optimizations to reduce the data throughput, including the dynamic masking of depth and colour images using a combination of body index information and environmentally registered bounding volumes. We use publicly available dynamic link libraries for the sensor SDK and bodytracking SDK from Microsoft alongside the own internal C# library to manage calibration, memory management, rendering, transmission, and temporal storage.

For Both AR and VR modules, the client rendering PC receives synchronized frames from the volumetric host PC, and rendered as described in Section 3.3. Client rendering is achieved using an Alienware 17 R5 Gaming laptop, with an Intel Core i9-8950HK, and a Nvidia Geforce GTX 1080 GPU. The AR module is deployed using a Microsoft Hololens 2. We utilize Microsoft's Mixed Reality Toolkit (MRTK) and the Holographic Remoting procedure, a process which offloads the cost of natively rendering computationally expensive graphics on the device to a connected computer. The VR module is deployed using a Meta Quest 2 head mounted display alongside Oculus Integration and Open XR, supporting both controller and gesture based interactions as part of the delivery of communication. To increase performance, we use use Oculus Air-link to offload rendering to a dedicated PC.

Volumetric renderings are created from the combination of color and depth streams from an array of three[1] Azure Kinect RGB-D sensors[2]. The position and orientation of the cameras will be directly reflected on the intended capture information. Once calibrated, the cameras produce a real-time volumetric rendering of the scene. The calibration process does not require any specialized environmental modifications such as green screening, and can be easily extended to 360° capture. The number and positioning of cameras dictates the overall spread, redundancy, and coverage of the environment. The calibration process does not require cameras to be facing the same direction, as is required for calibration procedures that use 2D visual markers. This offers greater flexibility in camera positioning.

3.2 Calibration

To calibrate the array of cameras into a single coordinate space, the algorithm performs an extrinsic alignment of each sensor to every other sensor in the array matching synchronized key frames between captures. In this pose estimation calibration, for each camera

[1]While only three cameras are used in this study, dependent on available USB controllers, four devices can be connected directly to a motherboard controller resulting in stable performance on a single PC

[2]https://docs.microsoft.com/en-us/azure/kinect-dk/hardware-specification

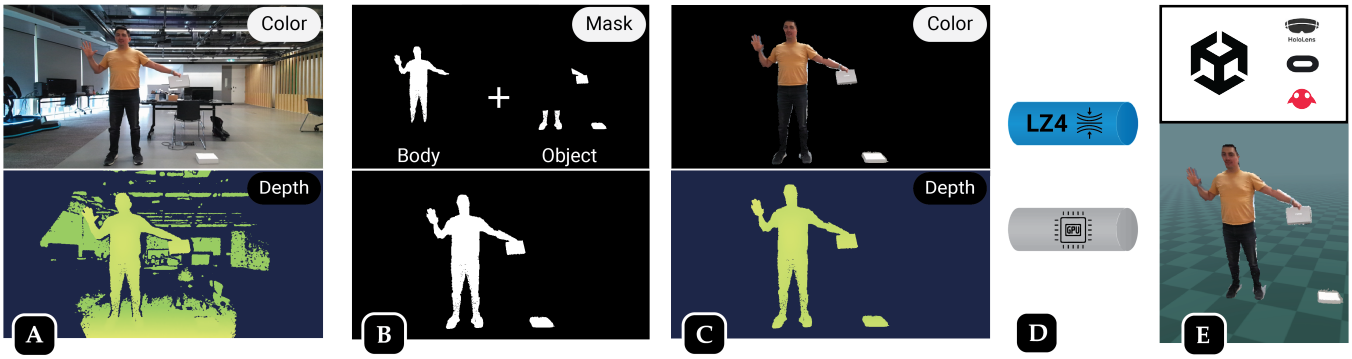


Figure 2: (A) Color and depth images are captured from each sensor and intrinsically aligned for further processing for each camera in the array, (B) Areas of interest are masked from the depth image using a combination of body and object techniques, (C) Masks are applied to color and depth images, exchanging unnecessary information with black pixels. (D) LZ4 compression is applied to each color and depth frame and passed onto the broadcast server to send to each connected client. If rendering is done locally, the frames are passed directly to the GPU for further processing, (E) The arriving depth and color images are transferred onto the GPU and rendered to an XR device such as the Hololens 2, Meta Quest 2, or Magic Leap One using Unity's visual effects graph.

in the array the pose of an optical rigid body marker is matched between frames using OpenCV in similar manner to Kaufmann et al. [47]; model fitting of the marker in each camera's frame, internal pose estimation, then finally projecting the pose into a primary sensor's coordinate space. During this process, the algorithm shows immediate results once an initial alignment is calculated but accuracy is refined over time as more frames are captured. Once extrinsic alignment is complete, the optical rigid body marker can be used to reorient the calibrated array to a known ground origin, with provided up and forward vectors. This final step allows the resulting volumetric information to have a consistent frame of reference regardless of the underlying positioning of the sensors between sessions. The whole calibration process is done prior to running the telepresence system, and can be completed within 30 seconds.

3.3 Volumetric Rendering and Recording

Baseline data is captured from each camera at 30 frames per second, with a depth resolution of 640x576 and color resolution of 1280x720 (Figure 2A). All cameras are synchronized via cable to reduce depth sensor interference. The depth resolutions can be dynamically altered to meet real-time rendering or recording requirements.

For rendering, on the server-side, once frames have been requested, the associated memory is transferred to Unity's Visual Effects (VFX) Graph for rendering purposes using Unity's Universal Rendering Pipeline[3] (Figure 2D). Client renderers receive packets from the network, transfer and inject into a VFX graph on the GPU, with all processing and masking done on the server side. While we utilize PC-hosted rendering techniques to increase performance, this design supports native rendering on non-windows based clients without requiring additional libraries, such as the Meta Quest 2, Magic Leap One, and Hololens 2 (Figure 2E).

3.4 Network Transmission

Once data is transferred to the GPU, we deploy optimizations to reduce the visual features rendered locally, and reduce the bandwidth necessary to send real-time data to connected clients. We use the body index map from incoming body-tracking frame data to isolate each user's body from the background. The mask sometimes misses key details, as seen in (Figure 2B) where the user's left hand (holding the item) and feet are not captured by the mask. We supplement this process with the shape culling technique which uses spheres and boxes to isolate areas of the combined depth images. To aid in compression of frame data over the network, we mark regions not of interest to black, and combine with the bounding volumes of the valid pixels (Figure 2C). The resulting depth and colour maps are sent to the compression and transmission pipeline using LZ4 compression and broadcast over eNet[4], a reliable UDP prototcol, which offers additional bandwidth by reducing protocol handshaking (Figure 2D). Using a single large environment shape culling approach, the network payload is approximately 40mbps of data per camera while capturing a body alone (@ 15% coverage), the payload can be reduced to a manageable 15mbps per camera. The positioning and number of culling volumes will directly impact the underlying size of each frame generated by the LZ4 compression process. There is a necessary trade-off in quality (higher-resolution imagery, more content, more cameras) or reduced bandwidth (lower-resolution imagery, less content, less cameras), which can be adapted to each platform deployment.

3.5 Avatar Representation

The VCP platform supports both real-time volumetric reproduction of users and rigged 3D avatar model representations. This flexibility offers the opportunity for users to remotely connect to learning environments using only an AR/VR headset. The addition of the volumetric camera array can be used to replace 3D avatar model

[3]https://unity.com/srp/universal-render-pipeline

[4]http://enet.bespin.org

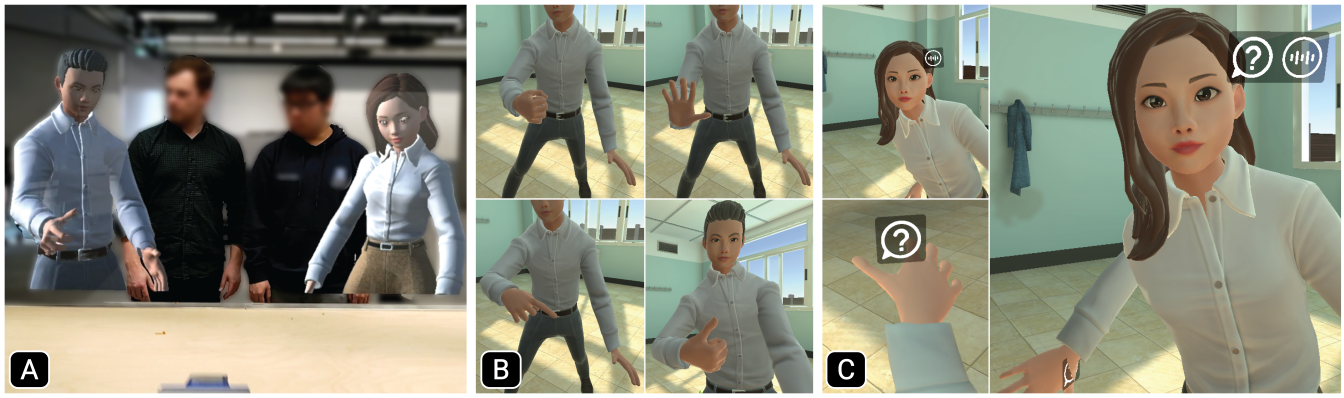


Figure 3: (A) Example remote avatars co-existing with local learners captured through the Hololens 2. Avatar movement is achieved using inverse kinematics from head and hand tracking. (B) Hand gestures are provided to support deictic and non-verbal communication. (C) A question bubble appears to assist in drawing attention to the avatar's intention to ask a question and a speech bubble appears when the avatar is talking. A question notation appears on the avatar's hand to provide visual feedback to the current question state.

representations, bringing components of the physical environment into the virtual environment.

Virtual avatars are provided by *ReadyPlayerMe*[5] (Figure 3). Users are provided a full-body avatar however can only see their avatar's hands, arms and legs from a first-person perspective in VR. Avatars are animated from the room-scale tracking of the Meta Quest 2 HMD and its two handheld controllers or integrated hand tracking. Hand gesture control is supported however the use of controllers reduces inexperienced participants needing to learn any gestural input to activate operations. Hand gestures using controllers are supported via press and hold button combinations on the controller including pointing, open palm, closed fist, and thumbs up (Figure 3B). Final IK[6] provides the inverse kinematics to animate the arms and legs as the avatar moves through the environment. Each user is calibrated with their avatar, matching the avatar's virtual height with that of the physical user by measuring the distance of the headset from the floor using internal tracking. This provides a consistent and expected look and feel between physical and virtual environments for the user.

3.6 Voice Communication

A network provides synchronized voice communication between all the locations through Dissonance VoIP[7]. The voice chat communication is configured to transmit and receive positional audio from the connected clients. This results in an spatial audio experience, allowing the participant to hear other participants' audio coming from their location inside the virtual world. In physical spaces, we incorporate a single desk speakerphone to transmit and receive audio from connected remote clients, requiring the disabling of spatial audio for this location. To aid in the visual communication between instructor and remote participants, visual cues appear at opportune moments to provide additional context. When a participant speaks, a *speech bubble* automatically appears next to their head (Figure 3C). When asking a question, a participant can press a button on the controller to toggle a *question mark* to appear alongside their head. As the user cannot see this cue, an additional private cue is also positioned on the user's hand when the question prompt is activated.

4 USER STUDY

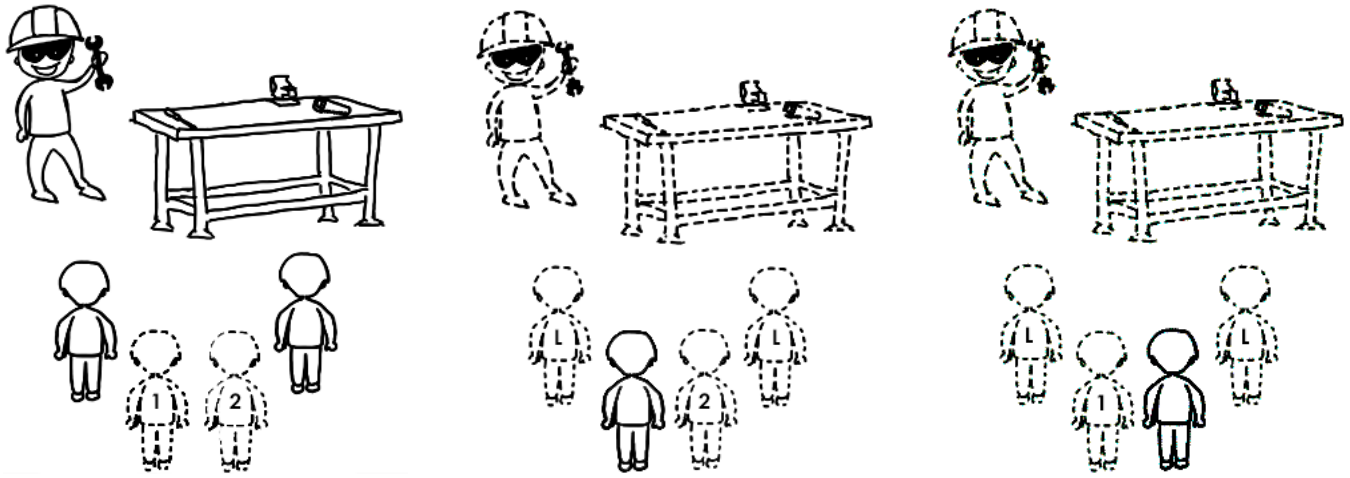
A study was conducted to explore telepresence in a mixed reality collaborative environment. The study aimed to provide insights to the human experiences of volumetric representations in telepresence systems. It seeks to understand expressions of co-presence, and to explore the role of user avatars on real-time remote interactions in a cross modality collaborative scenario involving in-person, augmented, and virtual reality participation. The study address three user experiences questions arising from telepresent collaboration driven from a cross-modal, multi-user design:

(1) How does the presence of head-worn mixed reality displays modify the way participants communicate during a multi-user telepresence experience?
(2) How does an instructor's communication with local learners differ to their communication with remote learners through a multi-user telepresence experience?
(3) Is there a difference in perception and communication between different avatar representations (e.g. full-body 3D, 180° volumetric and 360° volumetric)?

4.1 Instructional Design for Hybrid Volumetric Mixed Reality

The instructional design for this study was adapted from an established vestibule training currently offered by our university's fabrication lab. The fabrication lab instruction consisted of training in three tasks: measuring and scribing, sawing, and drilling. The training introduced the tools and techniques required to safely complete each task using two types of materials, wood and metal.

[5]https://readyplayer.me
[6]http://root-motion.com
[7]https://placeholder-software.github.io/Dissonance/



Local View: The instructor wears an AR HMD and sees four leaners. Two learners are in person. The other two learners are in remote sites 1 and 2 (shown as dashed lines) and are rendered as cartoon-like avatars.

Remote Site 1 View: The remote learner in site 1 wears a VR HMD and sees the instructor, the workbench, the tools, and the local learners (L) as volumetric holograms. The remote learner in site 2 is seen as a cartoon-like avatar.

Remote Site 2 View: The remote learner in site 2 wears a VR HMD and sees the instructor, the workbench, the tools, and the local learners (L) as volumetric holograms. The remote learner in site 1 is seen as a cartoon-like avatar.

Figure 4: Multi-user environments in a cross-modal training scenario

The instruction was focused on the (a) physical features of each tool, (b) operation of the tools, emphasising safety, (c) use of correct tools for each material, (d) sensory feedback from each material to illustrate safety concerns or near-task-completion (such as drilling and/or sawing). At the conclusion of the verbal instruction, students practiced the lesson on their own set of materials and tools. Once complete, students returned to a group, where the instructor explained the next task.

In preparation for the study, the authors participated in the on-site instruction. The authors were afforded direct insight into the experiences of the learners. We observed how learners interacted with the trainer and each other, when communication was undertaken and how it was delivered. We were provided with the script used by the fabrication lab instructors. The script was adapted for a hybrid volumetric mixed reality system. The revised training maintained an instructor led experience, but redesigned its delivery to support remote participation through mixed reality. The following changes were made to the in-person instruction: (1) the instructor was asked to demonstrate each task on both materials, rather than delay the second instruction once students had left the group; (2) students did not attempt to complete the tasks individually, thereby creating a similar learning experience for local and remote students; (3) greater onus was placed on the instructor to prompt student participation and gauge their understanding. The adapted training procedure was designed to be undertaken by four learners (two local and two remote) and an instructor. Instruction for each material (wood and metal) required approximately 20 minutes to deliver, and the entire training taking approximately 45 minutes.

4.2 Training Sites

The hybrid classroom evaluation was conducted across three sites (Figure 4). The local site was a simulated makerspace environment with workbench and tools. The two remote sites were two separate rooms in a nearby usability laboratory. The three sites were connected via the university cabled network. The training space at the local site was captured by three cameras, creating a volume of a 3m x 2.5m x 2.5m. The volume encompasses the instructor, the local students, and the training work area around the workbench. The physical layout did not require any special equipment (i.e, green screens), and was set-up using a mix of tripods and lighting truss stands (Figure 5A).

The instructor operated in an area of 1m x 0.5m, with a blue workbench clamp positioned in the horizontal centre of the capture volume. The two local students stood on the opposite side of the bench, and were free to take up any position (Figure 5B). The training involved instructions about using a Japanese pull saw, a cordless drill, a hacksaw, a combination square, and a scribe on the demonstrated materials.

The remote students, wearing a Meta Quest 2, were placed in two separate rooms. There were no obstacles in the room. The virtual experience of the remote learners was visualised as a virtual classroom environment[8] with the volumetric rendering of the physical environment appearing in the middle of the room. Each remote learner was initialized as standing in front of the volumetric workbench and looking toward the instructor. Remote learners were free to move around a physical space of 1.5m^2. The remote learners were free to move in the virtual environment, including

[8]https://assetstore.unity.com/packages/3d/props/interior/school-classroom-60343

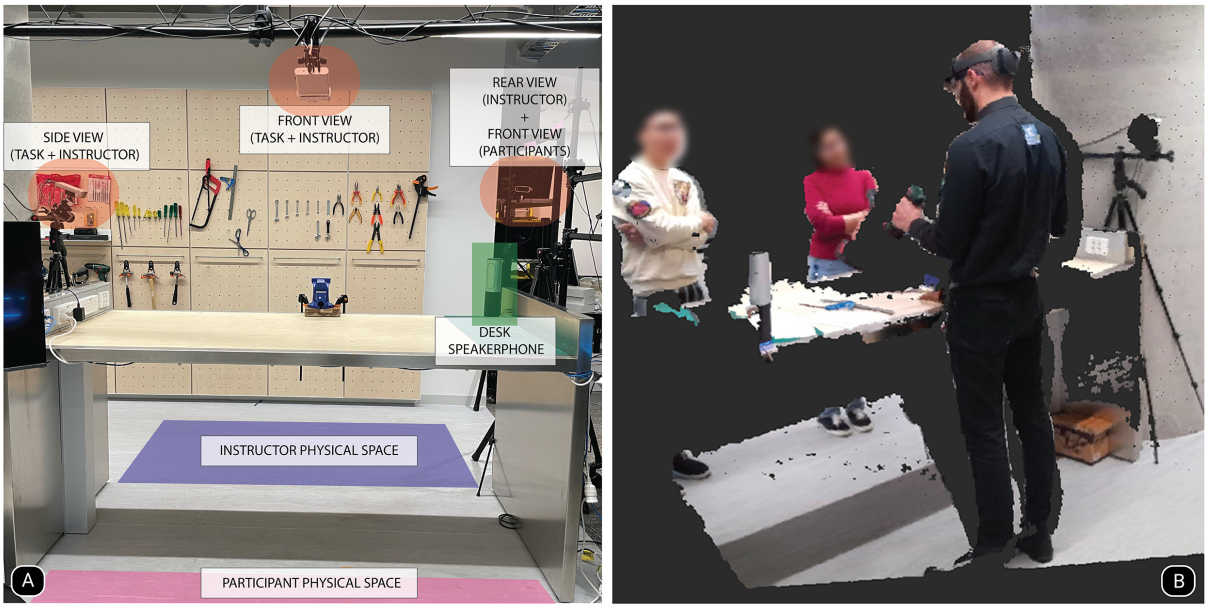


Figure 5: (A) The physical environment used in the evaluation incorporated a workbench, 1m in height, with a blue vice attached. The annotations show the location of the 3x Azure Kinect cameras, the desk speakerphone used for audio communication, and the areas where the instructor and local learners were positioned throughout the experience. (B) An example frame reconstruction from the study.

around, inside and behind the volumetric environment, volumetric instructor and volumetric learners in the local site.

The hand and power tools used in the study were borrowed from the fabrication lab, with no special considerations for their appearance or effect inside the volumetric capture volume.

4.3 Avatar Representations

Two different styles of avatar representation were presented; a full body virtual cartoon-like avatar for remote participants, and a full-body volumetric representation for local participants. Two further variations of volumetric reconstruction were provided; the instructor was presented as a 360° volumetric capture digital reconstruction, while the two local learners were presented as front facing 180° volumetric capture. Remote learners were represented with a gender-matched full bodied virtual avatar, with hand and arm control through hand-held controllers and inverse kinematics to control associated body joint movement. The instructor was physically present with two local learners, while two remote learners joined a volumetric reconstruction of the physical classroom in VR. The instructor was provided an optical see-through HMD to maintain visual awareness of both local and remote learners (Figure 1). The local classroom and each of the remote virtual sites were connected over a university network.

4.4 Participants

We recruited 8 learners and one instructor to participate in the study (Table 1). Participants self-rated their prior experience in three areas: (a) extended reality devices, (b) hand tools/power tools, and (c) maker training. The learners (P1-P8) were aged between 22 and 44 (5 male, 3 female). The instructor (T) was a full-time employee and instructor in the fabrication lab. He conducted the training in the configured mixed reality environment. He had no prior experience with augmented reality or virtual reality. Prior to the study, the instructor was given training on the use of VCP. The training included use of safety glasses with the HMD (Hololens 2) and eye-calibration.

4.5 Study Procedure

The study involved two identical training sessions. Learners P1-P4 participated in the first learning session (group 1) and P5-P8 in the second session (group 2). The instructor (T) lead both sessions. The two groups were created to have a relatively even distribution of the knowledge across tools, mixed reality and prior training.

For each training session, we divided the four participants into two pairs. Each pair contained an expert or a participant who had received prior training and a novice in the domain of tools.

The instructor delivered the training from the workbench to the two local learners and the two remote learners. After the training in the first material was completed, participants completed a questionnaire about their experience, and then swapped roles. The same procedure was followed for the second training session with participants asked to complete the same questionnaire again on their alternate role.

For the duration of the study, two researchers observed and recorded participants' behaviour, communication, and body language. The audio and video of the training experience at all three sites was recorded for later analysis by roof mounted recording

Table 1: Self-Reported Participant Information

ID	Tools Experience	XR Experience	Prior Training	Gender	Age
P1	Novice	Novice	No	F	23
P2	Novice	Novice	No	M	26
P3	Expert	Familiar	Yes	M	44
P4	Novice	Passing	No	F	27
P5	Passing	Passing	Yes	M	24
P6	Passing	Familiar	No	F	25
P7	Novice	Passing	No	M	22
P8	Familiar	Familiar	Yes	M	25
T	Expert	None	Yes	M	27

equipment. At the conclusion of the training, all participants (learners and instructors) were invited to a semi-structured group interview. The entire training and interview lasted approximately two hours for each group. Participants were given a $50 gift card compensation for their efforts. The study received ethics approval from our University's Office of Research Ethics and Integrity.

5 FINDINGS

Our data included a total of 5.5 hours of video and audio recordings from interviews and training studies, 8 pages of written responses, and researcher observation notes from each session. To understand the experience of our users in the varying roles in our telepresence experience, we followed a deductive thematic analysis approach outlined by Braun and Clarke [11]. To support the data familiarization process, the lead author digitized all written notes, and transcribed all audio from the recordings. The lead author open coded the first interview transcript, creating a set or preliminary codes. The remainder of the transcripts and written data were then coded using these preliminary codes, with further added as they emerged from the data. The research team met to discuss the generated codes, and through discussion defined a set of themes.

The instructor, who had no prior training with augmented reality headsets, was able to easily transfer their existing delivery of training content into our telepresence environment. We observed patterns of behaviour between different roles throughout the training. The following sections present our findings on: the user experience of volumetric visualisation, the group experience, communication between different modalities, and proxemics.

5.1 Volumetric Experience

When participants were interviewed about their experiences, there were a variety of responses which show the potential but also limitations of our approach. P1 expressed how easy it was to recognise their colleagues when represented as a volumetric rendering in the virtual space, however they also shared their desire in improvements in visual fidelity, due to *"the pixelated appearance."* P5 agreed that you could follow the instructors body movements, including *"the teacher's interactions, the outline of the drill and how you hold it and stuff like that."* Participants also openly proposed the suitability for the technology to be used for other techniques such as yoga, *"where capturing the body and hips are important (P1)"*, where body movements are of significant interest. While our participants shared their appreciation to the visual appearance of users, this was not shared for the finer details. P8 reflected that while it was cool to see rendered volumetric imagery, *"you couldn't see clearly the little details."* This view was shared by multiple participants raising concerns such as when the instructor was cutting through steel, *"you couldn't actually see the blade (P7)"*, and the presence of the Hololens on the instructor not rendering correctly, *"you'd see the inside of the back of his skull (P3)."* In the training, the metal surfaces of the tools produced unexpected behaviour from the time-of-flight cameras. Figure 6 shows examples of these errors. The top row of the figure illustrates missing details around the instructor's face, the bottom row illustrates time-of-flight depth miscalculations, resulting in straight surfaces appearing curved.

5.2 Group Experiences in Cross Modal Collaborative Tasks

Our study design produced asymmetrical visual representations of our learners. Remote learners viewed local learners as volumetric renderings, while being unseen by local learners. Audio communication was symmetrical so both sides could hear one another. Both roles expressed the impact of the experience on their feelings of being co-present with their colleagues. For remote learners, due to the camera layout, local learners were sometimes occluded by the instructor, while for local learners, they had no visual feedback from the system. P1 reflected on their amazement when they took the remote condition after the local condition:

> *"Initially, I thought that these guys (the other group) couldn't see us. I went to the other side (remote), that's when I realised that half of us were like avatars and the other were like, sort of real representations."*

Contrasting this with the learner's sense of co-presence when discussing their local experience:

> *"If they (remote learners) hadn't made some sounds, I wouldn't have known, it was just like the three of us."*

This view appeared to be shared by all participants as local learners, with the lack of any visual feedback impacting their sense of being co-present with remote colleagues, reduced them to only a sound, *"ah yeah that's right, the speakers talking" (P4).* While this was viewed as a limitation, P2 voiced the similarities with existing videoconferencing tools *"if you are using Zoom, it is the same, you are people on the outside and cannot (visually) recognise who asks questions."* This emphasised the disconnection between the local and the remote learners, after their repeated realisation that there were other participants present, even after experiencing the remote condition first. To counteract this scenario, P3 elaborated as a local learner on their desire to wear a Hololens to trigger their awareness of being co-present with virtual learners, *"I could have seen what was going on in front of me, but if I looked around, I could have seen the other people as well."* This is an interesting contrast, as it highlights the differences in perception of the experience between the two learner roles. For remote learners, they desired higher fidelity and clearer content, both of their collaborators, and the physical environment, in volumetric forms. However for local learners, even

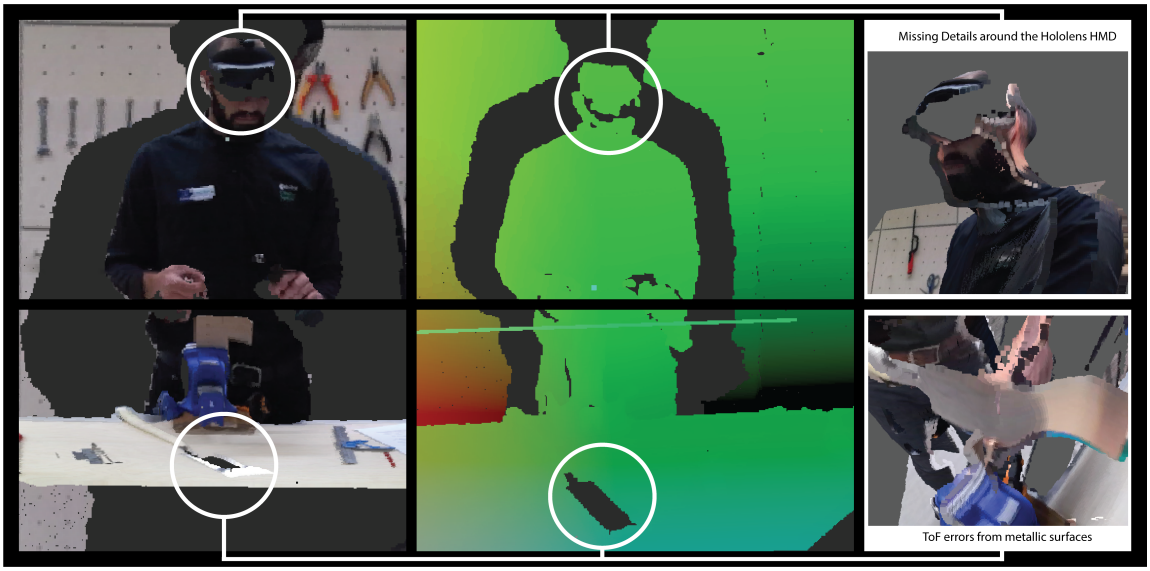


Figure 6: (Top) While volumetric renderings of user's appear clear and recognisable, some missing details were conveyed due to reflective properties from the Hololens 2 on the instructor. (Bottom) Time-of-flight depth error estimation due to reflective surfaces. The missing details can be seen in the depth image. When the item is rendered whilst in use, the straight edged saw appears to be curvy as depth calculation errors are passed into the rendering pipeline.

knowing wearing technology would diminish their remote appearance, they were more motivated in their desire to just *see* other remote learners.

5.3 Communicating between Modalities

The instructor acted as an intermediary between local and remote participants and could see and hear both groups of learners simultaneously. The instructor demonstrated traditional conversational behaviour towards both groups of learners, focusing their visual attention towards both local and remote participants when responding to questions. When reflecting on their experience, they were fully aware of all learners, both remote and local *"you look up, look at them, whether that's a real person or avatar and then speak."* However, awareness did not translate to effective communication, with numerous instances creating difficulties interpreting the remote avatar's body language. The instructor emphasized the importance of body language in following student engagement, highlighting while physical people don't have an auditory *ding* to ask questions, they kind of shuffle, giving an anticipation that a question is incoming. For VR avatars however, body language was not present, resulting in constant verbal verification by the instructor *"I had to keep asking the same thing, do you understand this?"* In these situations, the instructor was observed using diectic expressions, which were acknowledged from the remote learners to confirm their understanding, giving support to the value of volumetric representations in generating environments where conversations are supported without requiring any additional apparatus to convey body driven communication. On reflection of their role as a mediator between learners, the instructor highlighted their conscious efforts to include all participants. The instructor was conscious that addressing an 'unseen' remote participant might have been distracting for local learners *"I'm not paying attention to them (local learners), even though I'm saying stuff that is important for them."*

While visual representations are an integral component for increasing the feelings of co-presence [12, 48], the mutual awareness of remote learners' existence in the session is also important. Remote learners were 'spatially' respectful by not virtually occupying the same space as the local learners despite knowing that they could not be seen (more discussion in Section 5.4). It was the acknowledgement of the instructor they were experimenting with and seeking after *"there weren't any implications (to my actions), they didn't say,* 'hey, can you stop moving?'" In a live setting, participant's recognised that while you could move around, it could be really distracting for the rest of the class, but in the remote condition, those feelings did not translate, resulting in the experience of *"I felt (more) like an observer."* One participant commented on being self-conscious that the instructor was acknowledging them: *"I knew they were looking at me so I was about to just go around and look like right behind them, but that stopped me, I would have been embarrassed (if I was caught) (P7)."*

Facial expressions carry essential non-verbal cues that help support mutual understanding. While it was acknowledged that the capture of the instructor's face due to the presence of the Hololens detracted from the same level of interaction as face-to-face, there was a consensus that *"(in VR) you could still talk to them, still use the same methods, stand and listen and ask questions as appropriate (P3)."* When discussing their experience wearing a head-mounted display during face-to-face instruction, the instructor noted the ease in communication, *"I saw the physical people fine. I was just wearing another set of safety glasses, (but) a section was darker than the rest of my vision."* Inside the display, there is a small panel which conveys the augmented imagery to its wearer. This darkening was

also noticeable by the local learners, who expressed their desire to maintain eye contact during instruction, but its presence resulted in a loss of eye contact, with learners needing to rely only on head gaze. Due to the cross modality design of the study, this simple directional cue created confusion when the same 'physical' space was shared between multiple participants, *"(The instructor) was looking at me, so I was anticipating some sort of communication from him towards me because he looked into my eyes (P1) (but they weren't talking to me)."*

5.4 Mixed-Modality Proxemics

Due to the asymmetrical nature of the collaboration, there were almost two planes of existence, one virtual (remote) and one physical (local). VCP combines these planes into a telepresence collaborative environment, and alters the proxemics of the 'combined' space [5, 24].

Initially, the remote and local learners arranged themselves at social distances in a typical classroom layout, standing opposite the teacher (see Figure 1). As learners re-positioned themselves, the social distances collapsed into intimate distances, such as when a local learner stepped out from inside a remote learner's body, *"You don't have someone just jump out of your skin. Usually, it's watching a horror movie (P3)."* P3 (remote) and P1 (local) occupied the same physical/virtual space. This was not on purpose, as the remote learner commented *"only because I guess I got to a spot where I thought I could claim early on, (if I moved) I was going to be further away from the action, I didn't really want to do that."* The rule of proxemics regarding personal space was challenged [24], as the virtual plane forcefully overrides physical territoriality [53].

The struggle with physicality was felt most strongly by the remote participants, making them hesitant to move due to the intermittent appearances of bodies appearing from within their bodies, *"I looked around the room like, where did that come from? I couldn't work out where it was, I didn't know where to go."* Taking interpersonal distances into account, while the local learners had no visual feedback of the remote learner's location, the remote learner reflected *"it was worse for them than what it was for me!"*, validating their internal acceptance of occupying a space which could create unwanted feelings towards peers.

There were mixed feelings about maintaining social relationship distances. Some learners validated the underlying theory of proxemics, *"Even though it's a virtual environment, it still gives me the feeling that the other person is standing near me, so I don't want to overlay them (P6)"*, and *"I don't want to step on someone. I don't want to touch someone; I don't want to go inside someone's (body) (P1)."* Alternatively, the flexibility of the remote virtual condition offers unique experiences, such as embodying a local learner by standing within their body and *"wearing someone's skin (P4)"* to see their first person perspective.

The underlying theory of proxemics favoured the local learners with their lack of visual awareness of the virtual environment. Remote learners enjoyed the freedom to virtually position themselves against proxemics rules with little effect on the local learners experience. If local learners also had a headset, remote learners *"might be cautious of how close they should be, or where they stand might impact your experience (P4)."* Remote learners embraced this transitory freedom, exploring the training from different angles and locations, even getting really close to the table and behind the instructor. Compared to video-mediated communication, the free viewpoint was seen as an important feature for conveying task instruction, a feature highlighted in previous literature [44]. An additional benefit raised from this freedom was in the context of safety, *"In real life, everything is real, so you don't come too close to the danger. In VR, I'm free to move very close without any threat (P8)."* This emphasizes an appealing use case for collaborative telepresence in tasks where there is an inherent danger, which could increase with a greater number of physical individuals.

6 DISCUSSION & FUTURE WORK

6.1 Volumetric avatars – virtually *being there*

Our unique study considered the role of volumetric representations in a mixed modality telepresence hybrid classroom involving an instructor and two groups of four learners. Between the locations, three types of avatar representations were provided, (1) a 180° volumetric capture of local learners, (2) a 360° volumetric capture of the instructor and classroom environment, and (3) a full bodied avatar of remote learners. For both local and remote locations, participants expressed feeling higher levels of social presence with colleagues within their own location. For local learners this was because it was *"just a face-to-face live setting (P8)"*, while for remote learners, the presence of *"full avatars"* walking around resulted in more engagement, in comparison to only *"a vague outline (of local learners) (P8)."* Remote participants acknowledged that in-person participants were easily recognisable in their 180° avatar representation, and respected their presence in the virtual environment, however they also mentioned that they felt "ghostly" and "passively observing". In contrast, when they were asked about the instructor, these views were not shared. The 360° avatar was more engaging, "I was focusing on the instructor a lot more." We postulate these feelings are attributed to different relationships to the space, with an inherent ownership of the space belonging to the instructor. The participant comments demonstrate the successful creation of a *shared workspace* [21], with remote participants feeling present and capable of *"being able to ask questions where necessary (P7)."* Future work will need to consider how to increase this sense of a shared workspace to support richer communication and interaction experiences between remote and local learners.

Another part of the implementation that impacted volumetric co-presence relates to the camera coverage. Remote learners were sometimes only visually aware of a single local learner alongside the instructor. This was due to the local learner on the left being partially occluded from the instructor's body. As previously highlighted, they also appeared *ghostly*, due to the back half of their body not being captured and rendered. The approach continued to render front-facing content from behind to maintain interpersonal awareness, however this approach resulted in moments where a volumetric local learner can visually appear like they are looking both forwards and backwards at the same time. While these issues can be resolved by rendering techniques or additional camera coverage, the consideration of how to handle partial renderings is an important factor in dynamic, unstructured capture volumes.

While the asymmetrical study design allowed for the exploration of avatar representations in a training scenario, the AR instructor was not provided with a volumetric representation of remote learners to draw any insights but their comments point to the potential strengths of a volumetric rendering approach compared to embodied virtual avatars. The ability of the volumetric capture process to reconstruct users in their exact spatio-temporal form affords the opportunity to capture subtle cues to further support non-verbal communication. Future work will explore the bi-directional support of volumetric avatars for both virtual and augmented reality participants.

6.2 Exploration and Awareness

Freedom to explore a space was highlighted as an important characteristic of remote virtual telepresence, in line with prior literature [44, 61]. Free from the limitations of the physical world, participants were able to manoeuvre their viewpoint to locations that were deemed to be beneficial. Initially, remote participant's respected their co-learners' personal space, wanting to ensure that they didn't accidentally touch someone. As the training continued however, remote participants started to explore the visual opportunities afforded to them in the virtual environment. At one point, P2 stepped into the body of a local learner as they tried to understand the experience from their first-person perspective. Every other remote participant moved to locations behind, next to, and in-front of, the instructor to increase their visual awareness. These interactions show the great potential for telepresence to generate shared physical workspaces. While some remote learner's expressed interest in *"moving through (in-person) participants"*, all remote learners respected the instructors personal space, with some even noting they would feel *"embarrassed"* if they got caught getting too close, demonstrating the importance of a participant's visual awareness and its impact on supporting cross-modality proxemics.

6.3 Grounding conversation through spatial awareness

This study assumed that body representations, in volumetric or avatar forms, would be sufficient to create common and grounded communication [13]. When remote learners remained in the physical area in-front of the collaborative area (Figure 3A), this assumption was somewhat true. The instructor could look up, and visually acknowledge the group of four learners together, communicating where necessary to either modality. The assumption of shared space for grounded communication was challenged when remote learners moved their position in the virtual environment from standing next to the local learners to a new location. As there were no awareness cues provided to the in-person participants to indicate the remote learners new positioning, when the instructor would then look up, they were commonly presented with a subset of learners. As the training sessions progressed, the instructor intuitively adapted their behaviour to look around the physical environment to find all participants (as discussed in section 5.3). The drawback of this interaction is the inherent reduction in the HMD's field of view as a ratio of the total interaction volume. Placing information outside an optical-see-through HMD's field of view can introduce increased cognitive load [6]. Once a remote user has moved outside the AR user's field of view, they are no longer able to rely on visual communication nor the built in visual aids, as P2 reflected *"I also used that button (to ask a question), but I was behind you so you could not really recognise me"*. Furthermore, when looking down towards the workbench, due to the limited field of view of the AR headset, the instructor is essentially blind to the virtual world. This further impacts the avenue of non-verbal body language within the telepresence collaboration.

This lack of spatial awareness resulted in remote participants feeling like they were *"not being seen"*, and *"feeling like a ghost"*. Two approaches to increase spatial awareness using the visual bandwidth in the telepresence system are to provide indicators through the AR display [28, 29, 50, 59], or in the environment [1, 27, 55]. Using the periphery of the display to provide directional cues in locating users outside the current field of view is a viable solution within the constraints of the current system design. However, this is only limited to the instructor with an AR headset, such that remote learners could still not be seen by local learners, as reflected by the participants. An alternative approach could be through the use of projectors or an external display, as shown in [1, 27]. Providing spatial feedback for local learners could also facilitate proxemic interactions with their remote counterparts [5].

6.4 Communicating space through other modalities

Spatial awareness is also facilitated through auditory channels. However, spatial audio was only supported for the remote participants through their VR Meta Quest HMD. In the remote virtual environment, each user would hear communication from their remote colleague dependent on their physical relationship. The study design kept the local learner experience as close to face-to-face as possible, allowing participants to move around the space without needing to be encumbered with additional technology. This design choice resulted in not providing headphones to each user, instead using a single desk speakerphone to replicate a traditional conference room set-up. This design choice however, impacted the sense of co-presence and spatial awareness for both modalities.

First, in VR, there was a difficulty in pinning a voice to a location, with all in-person voices originating from a position somewhere in the desk, *"I didn't know where the audio is coming from. It's like everywhere"*. When a person is positioned to the left, it is important for the audio to originate from the left. The conflicting nature of it working for one cohort of participants and not for the other had a detrimental effect on the overall experience. To facilitate the expected outcome, each user should be provided with their own microphone, and be tracked inside the interaction space. This is achievable using the current pipeline, using the generated volumetric data to produce an appropriate location and view direction for each user as they speak. Second, from the perspective of in-person participants, remote users could only be heard but could not be seen, with little interaction occurring between these modalities, *"I had almost zero engagement with virtual participants, because they're they're just a voice coming out of the speaker"*. This issue can be alleviated through the use of headphones, but this will negatively impact the face-to-face experience. Alternatively, speakers could be

strategically positioned to create a sense of volume, and combined with visual aids, could provide a suitable medium.

7 CONCLUSION

In this paper, we presented *Virtual Co-Presence*, a volumetric capture system which supports cross-modal, multi-user, remote telepresence experiences. The platform combines real-time delivery of volumetric constructed users and their environment, alongside full-body 3D avatars, to remote colleagues in real time using off-the-shelf depth cameras and extended reality devices over a local enterprise network. We conducted a qualitative study which explored its use in a multi-user training adapted for remote delivery led by a training professional, involving simultaneous users in the physical, augmented, and virtual environments.

The volumetric avatars generated a rich source of communication that supported shared workspace awareness, co-presence, and multi-modal proxemics. The training task was conducted in an unstructured environment, demonstrating the potential of telepresence using commodity hardware for its significant conveyance of individual non-verbal characteristics. The instructor easily used the system, communicating naturally with local and remote participants alike, as if everyone was just *there*. Remote participants were immediately immersed in the experience and were able to easily communicate with their peers. This study adds important insights to the use of volumetric renderings in multi-user environments, involving dynamic, configurable, remote telepresence experiences.

ACKNOWLEDGMENTS

Thuong Hoang is the recipient of an Australian Research Council Discovery Early Career Researcher Award (DE200100898) funded by the Australian Government. We would like to thank the staff from The University of Melbourne's Telstra Creator Space, for their generous time and contributions to this project.

REFERENCES

[1] Matt Adcock, David Feng, and Bruce Thomas. 2013. Visualization of Off-Surface 3D Viewpoint Locations in Spatial Augmented Reality. In *Proceedings of the 1st Symposium on Spatial User Interaction (SUI '13)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/2491367.2491378

[2] R Azuma, Y Baillot, R Behringer, S Feiner, S Julier, and B MacIntyre. 2001. Recent advances in augmented reality. *IEEE Computer Graphics and Applications* 21, 6 (2001), 34–47. https://doi.org/10.1109/38.963459

[3] Jeremy N Bailenson and Kathryn Y Segovia. 2010. Virtual Doppelgangers: Psychological Effects of Avatars Who Ignore Their Owners. In *Online Worlds: Convergence of the Real and the Virtual*, William Sims Bainbridge (Ed.). Springer London, London, 175–186. https://doi.org/10.1007/978-1-84882-825-4_14

[4] Steven Baker, Jenny Waycott, Romina Carrasco, Ryan M Kelly, Anthony John Jones, Jack Lilley, Briony Dow, Frances Batchelor, Thuong Hoang, and Frank Vetere. 2021. *Avatar-Mediated Communication in Social VR: An In-Depth Exploration of Older Adult Interaction in an Emerging Communication Platform*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445752

[5] Till Ballendat, Nicolai Marquardt, and Saul Greenberg. 2010. Proxemic Interaction: Designing for a Proximity and Orientation-Aware Environment. In *ACM International Conference on Interactive Tabletops and Surfaces (ITS '10)*. Association for Computing Machinery, New York, NY, USA, 121–130. https://doi.org/10.1145/1936652.1936676

[6] J Baumeister, S Y Ssin, N A M ElSayed, J Dorrian, D P Webb, J A Walsh, T M Simon, A Irlitti, R T Smith, M Kohler, and B H Thomas. 2017. Cognitive Cost of Using Augmented Reality Displays. *IEEE Transactions on Visualization and Computer Graphics* 23, 11 (nov 2017), 2378–2388. https://doi.org/10.1109/TVCG.2017.2735098

[7] S Beck, A Kunert, A Kulik, and B Froehlich. 2013. Immersive Group-to-Group Telepresence. *IEEE Transactions on Visualization and Computer Graphics* 19, 4 (2013), 616–625. https://doi.org/10.1109/TVCG.2013.33

[8] Mark Billinghurst and Hirokazu Kato. 2002. Collaborative Augmented Reality. *Commun. ACM* 45, 7 (jul 2002), 64–70. https://doi.org/10.1145/514236.514265

[9] Frank Biocca. 1992. Virtual Reality Technology: A Tutorial. *Journal of Communication* 42, 4 (dec 1992), 23–72. https://doi.org/10.1111/j.1460-2466.1992.tb00811.x

[10] Frank Biocca and Mark R Levy. 2013. *Communication in the age of virtual reality*. Routledge.

[11] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[12] S Cho, S w. Kim, J Lee, J Ahn, and J Han. 2020. Effects of volumetric capture avatars on social presence in immersive virtual environments. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 26–34. https://doi.org/10.1109/VR46266.2020.00020

[13] Herbert H Clark and Susan E Brennan. 1991. Grounding in communication. In *Perspectives on socially shared cognition*. American Psychological Association, Washington, DC, US, 127–149. https://doi.org/10.1037/10096-006

[14] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. 2015. High-Quality Streamable Free-Viewpoint Video. *ACM Trans. Graph.* 34, 4 (jul 2015). https://doi.org/10.1145/2766945

[15] Rafael Kuffner dos Anjos, Maurício Sousa, Daniel Mendes, Daniel Medeiros, Mark Billinghurst, Craig Anslow, and Joaquim Jorge. 2019. Adventures in hologram space: exploring the design space of eye-to-eye volumetric telepresence. In *25th ACM Symposium on Virtual Reality Software and Technology*. 1–5.

[16] Mingsong Dou, Philip Davidson, Sean Ryan Fanello, Sameh Khamis, Adarsh Kowdle, Christoph Rhemann, Vladimir Tankovich, and Shahram Izadi. 2017. Motion2fusion: Real-Time Volumetric Performance Capture. *ACM Trans. Graph.* 36, 6 (nov 2017). https://doi.org/10.1145/3130800.3130801

[17] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, Pushmeet Kohli, Vladimir Tankovich, and Shahram Izadi. 2016. Fusion4D: Real-Time Performance Capture of Challenging Scenes. *ACM Trans. Graph.* 35, 4 (jul 2016). https://doi.org/10.1145/2897824.2925969

[18] A J Fairchild, S P Campion, A S García, R Wolff, T Fernando, and D J Roberts. 2017. A Mixed Reality Telepresence System for Collaborative Space Operation. *IEEE Transactions on Circuits and Systems for Video Technology* 27, 4 (2017), 814–827. https://doi.org/10.1109/TCSVT.2016.2580425

[19] Andreas Fender and Jörg Müller. 2018. Velt: A Framework for Multi RGB-D Camera Systems. In *Proceedings of the 2018 ACM International Conference on Interactive Surfaces and Spaces*. ACM, New York, NY, USA, 73–83. https://doi.org/10.1145/3279778.3279794

[20] Andreas Rene Fender, Hrvoje Benko, and Andy Wilson. 2017. MeetAlive: Room-scale omni-directional display system for multi-user content and control sharing. *Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces, ISS 2017* (oct 2017), 106–115. https://doi.org/10.1145/3132272.3134117

[21] Darren Gergle, Robert E Kraut, and Susan R Fussell. 2004. Action as Language in a Shared Visual Space. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work (CSCW '04)*. Association for Computing Machinery, New York, NY, USA, 487–496. https://doi.org/10.1145/1031607.1031687

[22] Geoffrey Gorisse, Olivier Christmann, Samory Houzangbe, and Simon Richir. 2019. From Robot to Virtual Doppelganger: Impact of Visual Fidelity of Avatars Controlled in Third-Person Perspective on Embodiment and Behavior in Immersive Virtual Environments . , 8 pages. https://www.frontiersin.org/article/10.3389/frobt.2019.00008

[23] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, Danhang Tang, Anastasia Tkach, Adarsh Kowdle, Emily Cooper, Mingsong Dou, Sean Fanello, Graham Fyffe, Christoph Rhemann, Jonathan Taylor, Paul Debevec, and Shahram Izadi. 2019. The Relightables: Volumetric Performance Capture of Humans with Realistic Relighting. *ACM Trans. Graph.* 38, 6 (nov 2019). https://doi.org/10.1145/3355089.3356571

[24] Edmund T Hall and Edward Twitchell Hall. 1966. *The hidden dimension*. Vol. 609. Anchor.

[25] Robert A. Heinlein. 1999. Waldo (1942). In *The Fantasies of Robert A. Heinlein* (1st ed.). Tor Books, New York, NY, USA, Chapter 4, 125—-212.

[26] Keita Higuchi, Yinpeng Chen, Philip A Chou, Zhengyou Zhang, and Zicheng Liu. 2015. Immerseboard: Immersive telepresence experience using a digital whiteboard. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2383–2392.

[27] Andrew Irlitti, Thammathip Piumsomboon, Daniel Jackson, and Bruce H Thomas. 2019. Conveying spatial awareness cues in xR collaborations. *IEEE Transactions on Visualization and Computer Graphics* 25, 11 (nov 2019), 3178–3189. https://doi.org/10.1109/TVCG.2019.2932173

[28] Yoshio Ishiguro and Jun Rekimoto. 2011. Peripheral Vision Annotation: Noninterference Information Presentation Method for Mobile Augmented Reality. In *Proceedings of the 2nd Augmented Human International Conference (AH '11)*.

Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/1959826.1959834

[29] Allison Jing, Kieran William May, Mahnoor Naeem, Gun Lee, and Mark Billinghurst. 2021. EyemR-Vis: Using Bi-Directional Gaze Behavioural Cues to Improve Mixed Reality Remote Collaboration. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (CHI EA '21)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411763.3451844

[30] Brett Jones, Rajinder Sodhi, Michael Murdock, Ravish Mehra, Hrvoje Benko, Andrew Wilson, Eyal Ofek, Blair MacIntyre, Nikunj Raghuvanshi, and Lior Shapira. 2014. RoomAlive: Magical Experiences Enabled by Scalable, Adaptive Projector-Camera Units. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology (UIST '14)*. Association for Computing Machinery, New York, NY, USA, 637–644. https://doi.org/10.1145/2642918.2647383

[31] T Kanade, P Rander, and P J Narayanan. 1997. Virtualized reality: constructing virtual worlds from real scenes. *IEEE MultiMedia* 4, 1 (1997), 34–47. https://doi.org/10.1109/93.580394

[32] Marek Kowalski, Jacek Naruniec, and Michal Daniluk. 2015. Livescan3d: A fast and inexpensive 3d data acquisition system for multiple kinect v2 sensors. In *2015 international conference on 3D vision*. IEEE, 318–325.

[33] Akira Kubota, Aljoscha Smolic, Marcus Magnor, Masayuki Tanimoto, Tsuhan Chen, and Cha Zhang. 2007. Multiview imaging and 3DTV. *IEEE signal processing magazine* 24, 6 (2007), 10–21.

[34] Gregorij Kurillo, Ruzena Bajcsy, Klara Nahrsted, and Oliver Kreylos. 2008. Immersive 3D Environment for Remote Collaboration and Training of Physical Activities. In *2008 IEEE Virtual Reality Conference*. IEEE, 269–270. https://doi.org/10.1109/VR.2008.4480795

[35] David Lindlbauer and Andy D Wilson. 2018. Remixed reality: Manipulating space and time in augmented reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[36] J L. Lugrin, M Ertl, P Krop, R Klüpfel, S Stierstorfer, B Weisz, M Rück, J Schmitt, N Schmidt, and M E Latoschik. 2018. Any "Body" There? Avatar Visibility Effects in a Virtual Reality Game. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 17–24. https://doi.org/10.1109/VR.2018.8446229

[37] Andrew Maimone, Jonathan Bidwell, Kun Peng, and Henry Fuchs. 2012. Enhanced personal autostereoscopic telepresence system using commodity depth cameras. *Computers & Graphics* 36, 7 (2012), 791–807. https://doi.org/10.1016/j.cag.2012.04.011

[38] A Maimone and H Fuchs. 2011. Encumbrance-free telepresence system with real-time 3D capture and display using commodity depth cameras. In *2011 10th IEEE International Symposium on Mixed and Augmented Reality*. 137–146. https://doi.org/10.1109/ISMAR.2011.6092379

[39] A Maimone and H Fuchs. 2012. Real-time volumetric 3D capture of room-sized scenes for telepresence. In *2012 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*. 1–4. https://doi.org/10.1109/3DTV.2012.6365430

[40] Nicolai Marquardt, Robert Diaz-Marino, Sebastian Boring, and Saul Greenberg. 2011. The Proximity Toolkit: Prototyping Proxemic Interactions in Ubiquitous Computing Ecologies. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology (UIST '11)*. Association for Computing Machinery, New York, NY, USA, 315–326. https://doi.org/10.1145/2047196.2047238

[41] B Marques, S Silva, J Alves, T Araújo, P Dias, and B S Santos. 2022. A Conceptual Model and Taxonomy for Collaborative Augmented Reality. *IEEE Transactions on Visualization and Computer Graphics* 28, 12 (2022), 5113–5133. https://doi.org/10.1109/TVCG.2021.3101545

[42] Microsoft. 2022. Microsoft Mixed Reality Capture Labs. https://www.microsoft.com/en-us/mixed-reality/capture-studios

[43] Paul Milgram and Fumio Kishino. 1994. Taxonomy of mixed reality visual displays. *IEICE Transactions on Information and Systems* E77-D, 12 (1994), 1321–1329.

[44] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L Davidson, Sameh Khamis, Mingsong Dou, Vladimir Tankovich, Charles Loop, Qin Cai, Philip A Chou, Sarah Mennicken, Julien Valentin, Vivek Pradeep, Shenlong Wang, Sing Bing Kang, Pushmeet Kohli, Yuliya Lutchyn, Cem Keskin, and Shahram Izadi. 2016. Holoportation: Virtual 3D Teleportation in Real-Time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST '16)*. Association for Computing Machinery, New York, NY, USA, 741–754. https://doi.org/10.1145/2984511.2984517

[45] Minna Pakanen, Paula Alavesa, Niels van Berkel, Timo Koskela, and Timo Ojala. 2022. "Nice to see you virtually": Thoughtful design and evaluation of virtual avatar of the other user in AR and VR based telexistence systems. *Entertainment Computing* 40 (jan 2022), 100457. https://doi.org/10.1016/j.entcom.2021.100457

[46] Tomislav Pejsa, Julian Kantor, Hrvoje Benko, Eyal Ofek, and Andrew Wilson. 2016. Room2Room: Enabling Life-Size Telepresence in a Projected Augmented Reality Environment. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 1716–1725. https://doi.org/10.1145/2818048.2819965

[47] Thomas Pintaric and Hannes Kaufmann. 2007. Affordable infrared-optical pose-tracking for virtual and augmented reality. In *Proceedings of Trends and Issues in Tracking for Virtual Environments Workshop, IEEE VR*. 44–51.

[48] Thammathip Piumsomboon, Arindam Dey, Barrett Ens, Gun Lee, and Mark Billinghurst. 2019. The Effects of Sharing Awareness Cues in Collaborative Mixed Reality. https://www.frontiersin.org/articles/10.3389/frobt.2019.00005

[49] Thammathip Piumsomboon, Gun A Lee, Jonathon D Hart, Barrett Ens, Robert W Lindeman, Bruce H Thomas, and Mark Billinghurst. 2018. *Mini-Me: An Adaptive Avatar for Mixed Reality Remote Collaboration*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173620

[50] Thammathip Piumsomboon, Youngho Lee, Gun Lee, and Mark Billinghurst. 2017. CoVAR: A Collaborative Virtual and Augmented Reality System for Remote Collaboration. In *SIGGRAPH Asia 2017 Emerging Technologies (SA '17)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3132818.3132822

[51] Ramesh Raskar, Greg Welch, Matt Cutts, Adam Lake, Lev Stesin, and Henry Fuchs. 1998. The Office of the Future: A Unified Approach to Image-Based Modeling and Spatially Immersive Displays. In *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '98)*. Association for Computing Machinery, New York, NY, USA, 179–188. https://doi.org/10.1145/280814.280861

[52] Holger Regenbrecht, Katrin Meng, Arne Reepen, Stephan Beck, and Tobias Langlotz. 2017. Mixed voxel reality: Presence and embodiment in low fidelity, visually coherent, mixed reality environments. In *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 90–99.

[53] Robert D Sack. 1983. Human territoriality: a theory. *Annals of the association of American geographers* 73, 1 (1983), 55–74.

[54] Scatter. 2022. DepthKit. https://www.depthkit.tv/

[55] Susanne Schmidt, Frank Steinicke, Andrew Irlitti, and Bruce H Thomas. 2018. Floor-Projected Guidance Cues for Collaborative Exploration of Spatial Augmented Reality Setups. In *Proceedings of the 2018 ACM International Conference on Interactive Surfaces and Spaces (ISS '18)*. ACM, New York, NY, USA, 279–289. https://doi.org/10.1145/3279778.3279806

[56] Teddy Seyed, Alaa Azazi, Edwin Chan, Yuxi Wang, and Frank Maurer. 2015. SoD-toolkit: A toolkit for interactively prototyping and developing multi-sensor, multi-device environments. *Proceedings of the 2015 ACM International Conference on Interactive Tabletops and Surfaces, ITS 2015* (nov 2015), 171–180. https://doi.org/10.1145/2817721.2817750

[57] Maurício Sousa, Daniel Mendes, Rafael Kuffner Dos Anjos, Daniel Medeiros, Alfredo Ferreira, Alberto Raposo, João Madeiras Pereira, and Joaquim Jorge. 2017. Creepy tracker toolkit for context-aware interfaces. *Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces, ISS 2017* 17 (oct 2017), 191–200. https://doi.org/10.1145/3132272.3134113

[58] A Steed, W Steptoe, W Oyekoya, F Pece, T Weyrich, J Kautz, D Friedman, A Peer, M Solazzi, F Tecchia, M Bergamasco, and M Slater. 2012. Beaming: An Asymmetric Telepresence System. *IEEE Computer Graphics and Applications* 32, 6 (2012), 10–17. https://doi.org/10.1109/MCG.2012.110

[59] Francesco Strada, Edoardo Battegazzorre, Enrico Ameglio, Simone Turello, and Andrea Bottino. 2022. Assessing Visual Cues for Improving Awareness in Collaborative Augmented Reality BT - Extended Reality, Lucio Tommaso De Paolis, Pasquale Arpaia, and Marco Sacco (Eds.). Springer International Publishing, Cham, 200–218.

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

[61] Balasaravanan Thoravi Kumaravel, Fraser Anderson, George Fitzmaurice, Bjoern Hartmann, and Tovi Grossman. 2019. Loki: Facilitating Remote Instruction of Physical Tasks Using Bi-Directional Mixed-Reality Telepresence. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology (UIST '19)*. Association for Computing Machinery, New York, NY, USA, 161–174. https://doi.org/10.1145/3332165.3347872

[62] Peng Wang, Xiaoliang Bai, Mark Billinghurst, Shusheng Zhang, Xiangyu Zhang, Shuxia Wang, Weiping He, Yuxiang Yan, and Hongyu Ji. 2021. AR/MR Remote Collaboration on Physical Tasks: A Review. *Robotics and Computer-Integrated Manufacturing* 72 (2021), 102071. https://doi.org/10.1016/j.rcim.2020.102071

[63] William Gibson. 1984. *Neuromancer*. Ace Books, New York, NY, USA. 271 pages.

[64] Andrew D. Wilson and Hrvoje Benko. 2016. Projected augmented reality with the RoomAlive Toolkit. *Proceedings of the 2016 ACM International Conference on Interactive Surfaces and Spaces: Nature Meets Interactive Surfaces, ISS 2016* (nov 2016), 517–520. https://doi.org/10.1145/2992154.2996362

[65] Jakob Zillner, Christoph Rhemann, Shahram Izadi, and Michael Haller. 2014. 3D-board: a whole-body remote collaborative whiteboard. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*. 471–479.