Science Arts & Métiers (SAM)

is an open access repository that collects the work of Arts et Métiers Institute of Technology researchers and makes it freely available over the web where possible.

This is an author-deposited version published in: https://sam.ensam.eu
Handle ID: .http://hdl.handle.net/10985/26441

To cite this version :

Bruno CABY, Guillaume BATAILLE, Florence DANGLADE, Jean-Rémy CHARDONNET - Environment Spatial Restitution for Remote Physical AR Collaboration - IEEE Transactions on Visualization and Computer Graphics - Vol. 31, n°5, p.3067-3076 - 2025

Any correspondence concerning this service should be sent to the repository Administrator : scienceouverte@ensam.eu

Arts et Métiers Institute of Technology

Environment Spatial Restitution for Remote Physical AR Collaboration

Bruno Caby, Guillaume Bataille, Florence Danglade, and Jean-Rémy Chardonnet

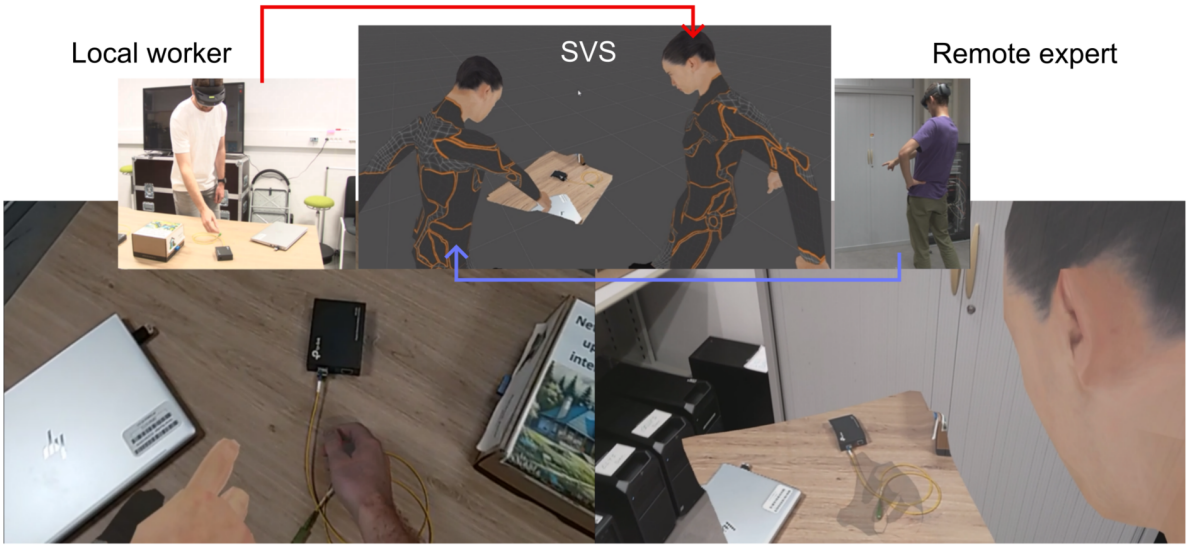


Fig. 1: Our AR remote collaborative system. On the left, the local worker has to perform an assembly. On the right, the expert helps the local worker thanks to the Shared Virtual Space.

Abstract— The emergence of spatial immersive technologies allows new ways to collaborate remotely. However, they still need to be studied and enhanced in order to improve their effectiveness and usability for collaborators. Remote Physical Collaborative Extended Reality (RPC-XR) consists in solving augmented physical tasks with the help of remote collaborators. This paper presents our RPC-AR system and a user study evaluating this system during a network hardware assembly task. Our system offers verbal and non-verbal interpersonal communication functionalities. Users embody avatars and interact with their remote collaborators thanks to hand, head and eye tracking, and voice. Our system also captures an environment spatially, in real-time and renders it in a shared virtual space. We designed it to be lightweight and to avoid instrumenting collaborative environments and preliminary steps. It performs capture, transmission and remote rendering of real environments in less than 250ms. We ran a cascading user study to compare our system with a commercial 2D video collaborative application. We measured mutual awareness, task load, usability and task performance. We present an adapted Uncanny Valley questionnaire to compare the perception of remote environments between systems. We found that our application resulted in better empathy between collaborators, a higher cognitive load and a lower level of usability, remaining acceptable, to the remote user. We did not observe any significant difference in performance. These results are encouraging, as participants' observations provide insights to further improve the performance and usability of RPC-AR.

Index Terms—Collaborative interactions, augmented-reality, spatial environment rendering.

1 Introduction

The rise of immersive systems allows new means of spatial remote collaboration. Industrials increasingly need their employees to collaborate remotely. This is due to team fragmentation and remote work adoption. Remote Physical Collaborative (RPC) Extended Reality (XR) consists in creating a Shared Virtual Space (SVS) [3, 5, 17] and rendering a remote user's environment capture in the SVS [28] (see Section 2). In spatial RPC systems, collaborators embody avatars in the SVS, providing copresence and shared spatial knowledge [3]. Thus, the SVS contains cues for remote collaboration such as spatial replicas of physical task environments, avatars, and interactions between collaborators [34]. These cues are essential to allow remote collaborators to help local workers [17] (see Section 2). Physical world rendering within an SVS is still little studied [9]. Several studies use 360° cameras to capture real environments [21, 32] and project resulting videos onto spheres surrounding remote users. Other studies use depth sensors to capture real environments spatially and render them as meshes [43] or point clouds [12]. Spatial capture, reconstruction, and transmission are restrictive because they require important computing power and the potential use of additional equipment [10]. This requires positioning various devices in the environment, or capturing it beforehand. In most cases, the remote collaborator is immersed in Virtual Reality

- *Bruno Caby is with Orange Labs and Arts et Métiers Institute of Technology, LISPEN. E-mail: bruno.caby@ensam.eu.*
- *Guillaume Bataille is with Orange Labs. E-mail: guillaume2.bataille@orange.com.*
- *Florence Danglade is with Arts et Métiers Institute of Technology, LISPEN. Research. E-mail: florence.danglade@ensam.eu.*
- *Jean-Rémy Chardonnet is with Arts et Métiers Institute of Technology, LISPEN. Research. E-mail: jean-remy.chardonnet@ensam.eu.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

(VR) [12, 22, 27, 32].

Alternatively, the use of Augmented Reality (AR) by all collaborators keeps them aware of their own physical environment [3]. Indeed, during industrial use cases, workers often cannot isolate themselves in safe and empty areas enabling safe VR experiences. In addition, AR helps enhance empathetic collaboration [9]. Also, Optical See-Through Head Mounted Displays (OST-HMD) minimise cybersickness [18, 41]. AR does not require any initial virtual environment to immerse users, unlike VR. For these reasons, we aim to study and enhance RPC-AR systems that require no additional equipment other than OST-HMDs. In this manner, collaborators remain aware of their real environment, without any prior installation or capture.

Contributions: compared to prior works, our main contributions in this paper are :

- a novel **collaborative system** where AR-HMDs are used by all collaborators. Our system shares remote environments where embodied users collaborate. It captures, transmits, and remotely renders a physical environment spatially in less than 250ms. Our RPC-AR system is fully embedded in Microsoft HoloLens 2[1] headsets. It does not require any prior environmental instrumentation, capture, or preprocessing, enhancing its usability in industrial use cases (see Section 3);
- a **cascading user study** (n = 31) comparing our system with a first-person 2D video solution *TeamViewer Assist AR*[2]. We compared task performance, task load, mutual awareness, and usability (see Section 4);
- we also present an **adapted questionnaire** to evaluate environment restitution quality based on the Uncanny Valley measure (see Section 4.5).

2 Related Work

Designing an RPC-XR application involves creating a shared virtual space: a space that enables interpersonal communication including verbal and non-verbal communication, and contextual elements such as remote environment rendering [4, 31].

Remote Environment Rendering

Asymmetric remote environment rendering involves rendering only one of the collaborators' environments. This method is much studied as it corresponds to the most common use cases. For example, when a worker needs help or supervision during a task [9, 27]. In this section, we focus on these use cases. We distinguish two categories of immersive environment restitution: 360° videos and spatial reconstructions.

360° video is commonly used in the literature. This method requires to use a head-mounted 360° camera. The rendering is then carried out by projecting this video onto a sphere whose center is the remote collaborator immersed in VR [21, 38]. A major difference with classic video is the ability to make the remote collaborator's view independent from the local collaborator's view, i.e., the remote collaborator has a three-degree-of-freedom (3 DoF) vision in rotation [22, 39]. Fig. 2 shows the principle of independent vision: dependent vision makes it easier to understand the remote collaborator's viewpoint, while an independent vision allows greater freedom in collaboration. It offers the possibility to find a better angle of vision and is preferred by users [22]. In other systems, the camera is placed on tripods [32, 42], thereby enabling the remote collaborator to maintain a 3 DoF vision from an alternative viewpoint.

Spatial reconstruction affords the remote collaborator the capacity to freely navigate into the space, thereby conferring upon him a six-degree-of-freedom (6 DoF) view [38, 43]. This reconstruction can be rendered on a screen, allowing the remote user to navigate freely in the remote space [37]. However, when combined with immersive technologies, it also enhances the user's spatial presence in the environment thanks to stereovision [39]. A point cloud results from an environment capture and rendering technique consisting in capturing and representing real objects as virtual sets of points in space [12]. Adding colors to each point, provided that the quality and the density of points are sufficient, allows users to identify remote objects. Conversely, dense reconstruction by mesh requires meshing points to display the surface of an object [43] and project a texture onto it. The emergence of novel view synthesis techniques, such as Gaussian Splatting [19] and Neural Radiance Fields [26], has facilitated environment restitution. This is achieved by storing an understanding of the scene derived from photogrammetric algorithms and rasterizing a novel view for each frame, thereby creating a comprehensive representation of the environment.

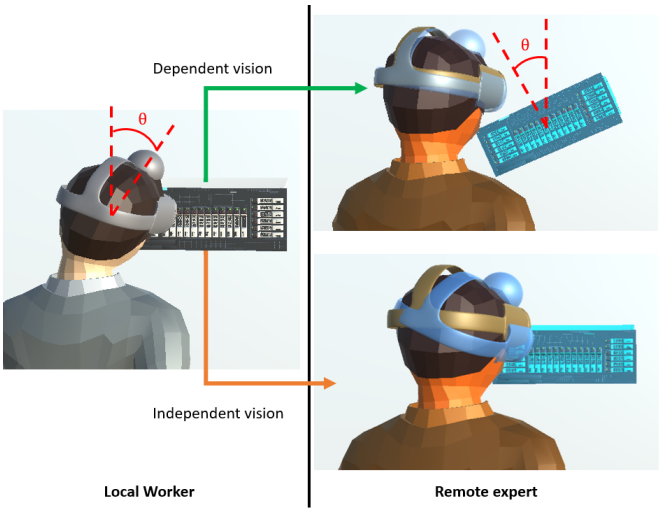


Fig. 2: Vision dependency. Elements in blue on the remote side are virtual restitution of the local collaborator's environment. In the case of dependent vision, remote users cannot control their viewing angle. In the case of independent vision, the restitution allows remote users to freely seek their viewpoint.

Cited spatial reconstruction systems are static by design. Indeed, the environment capture and processing are conducted prior to use. Rendering this reconstruction is possible, but it remains no longer updated and does not provide real-time feedback to the remote collaborator [12]. Teo et al. [39] presented a system that combines a static mesh with 360° video to texture it, enabling a semi-static reconstruction, as the spatial data is not updated while its associated texture is. Latest research employs semi-dynamic reconstructions, whereby a capture and preliminary processing are conducted, followed by real-time updates part-by-part [27, 40]. This capture mode seems to us to be a better solution in terms of quality as it allows the use of high-resolution cameras, neural networks, and pre-processing of the prior capture. However, preliminary configuration and offline processing make the solution less flexible and can potentially be tedious [40]. Fully dynamic reconstruction could help avoid this part. In 2020, Bai et al. presented a system that enabled dynamic spatial reconstruction of a remote environment. This system required a cluster of depth cameras, and data transmission was performed using a 10-Gbps Ethernet connection [1].

The choice of a prior reconstruction or wired data transmission methods is also guided by the use of VR by remote collaborators, considering that VR requires an initial environment in which users are immersed. Then, using AR for the remote user represents an optimal use case for a fully dynamic reconstruction transmitted with wireless IP transmission. Thus, Remote Physical Collaborative Augmented Reality (RPC-AR) offers a more convenient and seamless experience for all local and remote users. Local users benefit from avoiding any prior configuration, while remote users stay aware of their spatial context [3], and avoid cybersickness [18, 41].

[1]Microsoft Hololens 2: https://www.microsoft.com/hololens
[2]TeamViewer Germany GmbH: https://www.teamviewer.com/

Table 1: Classification of RPC-XR studies cited in this article. Items in bold are similar to our system. Technology acronyms are Asymmetric (Asym.), Symmetric (Sym.), Virtual Reality Head Mounted Display (VR-HMD), and Augmented Reality Head Mounted Display (AR-HMD). Environment Restitution acronyms are Panoramas (Pano.), Dynamic (Dyn.) and Static (Stat.). Communication Cues: Eye Tracking (ET), Hand Tracking (HT), and Field of View frustum (FoV). Conditions acronyms are non-verbal communication (NVC), with (w.), and without (w.o.). Measures acronyms are Task Completion Time (TcT), Iterations (Iter.), Preferences (Pref.), Observations (O.), Co-presence (Co-p.), Spatial-presence (Spa-p.), Social-presence (So-p.), Cognitive Load (CL.) Cybersickness (CS.), Empathy (Emp.), Usability (U.), and Quality (Qual.). The homogeneity of different presence measures is based on definitions provided by Lombard et al. [23].

Year	Reference	Technology		Environment Restitution		Communication Cues	Task symmetry	Compared Conditions	Measures	
		Symmetry	Type	Method	Dynamism				Objective	Subjective
2015	Tait et al. [37]	Asym.	Screen AR-HMD	**Point Cloud** + Tracking	Stat. + Tracking	Annotation, 3D Model, FoV	**Asym.**	**First-person Video** ± Freeze, Video, Independent Vision	**TcT**, Accuracy	Pref., **U., O.**
2016	Gupta et al. [13]	Asym.	Screen AR-HMD	Video	**Dyn.**	**ET**, Pointer	**Asym.**	Without NVC, ET, Pointer, Both	**TcT**	Pref., **O.**, Co-p, **Emp.**
2016	Le Chénéchal et al. [7]	Asym.	VR-HMD AR-HMD, Screen	Video, Mesh	**Dyn.**	**HT**, Draw	**Asym.**	Video, Mesh	**TcT**	Pref.
2017	Piumsomboon et al. [28, 29]	Asym.	VR-HMD AR-HMD	Mesh	Stat.	**ET**, Gestures, FoV	**Asym.**.	w. NVC, w.o NVC	Iter.	**O.**
2017	Gao et al. [12]	**Sym.**	VR-HMD	**Point Cloud**	Stat., **Dyn.**	**HT**, FoV	**Asym.**	Stat., **Dyn. Point Cloud**	**TcT**	Pref., **CL., O.**
2017	Lee et al. [21]	Asym.	VR-HMD AR-HMD	Video 360°	**Dyn.**	FoV + Arrows, **HT**	**Asym.**	No cue, FoV, FoV + Arrows		Pref.
2018	Lee et al. [22]	Asym.	VR-HMD AR-HMD	Video 360°	**Dyn.**	FoV + Arrows, **HT** + Halo	**Asym.**	Dependent, Independent Vision	**TcT**	Pref., **O.**, Co-p, Spa-p., **CS., CL.**
2018	Zillner et al. [43]	Asym..	VR-HMD AR-HMD	Mesh	Stat.	Replicas, Draw	**Asym.**			
2019	Teo et al. [38]	Asym.	VR-HMD AR-HMD	Video 360°, Mesh	**Dyn.**, Stat.	FoV + Arrows, **HT** + Halo, Pointer, Annotation	**Asym.**	Video 360°, Mesh, Both	**TcT**	Pref., Spa-p., Co-p, **CS.**
2020	Rhee et al. [32]	Asym.	VR-HMD AR-HMD	Video 360°	**Dyn.**	Pointer, **HT**	**Asym.**	Local user, Remote user		Pref., Spa-p., Co-p
2020	Teo et al. [39]	Asym.	VR-HMD AR-HMD	Video + 360° Pano., Mesh	Semi-Stat., **Dyn.**	FoV, Pointer	**Asym.**	Projected, 360°	TcT	U., O., Pref., So-p., Spa-p., CL., CS.
2020	Bai et al. [1]	Asym.	VR-HMD AR-HMD	**Point Cloud**	**Dyn.**	**ET, HT**	**Asym.**	w.o NVC, HT, ET, HT + ET	**TcT**	**U., O.**, So-p., Spa-p., **CL.**
2022	Niedermayr et al. [27]	Asym.	VR-HMD AR-HMD	**Point Cloud** + Video	Semi-Dyn.	**HT**, Pointer	**Asym.**	Local user, Remote user		**CL.**
2023	Tian et al. [40]	Asym.	VR-HMD AR-HMD	**Point Cloud**	Semi-Dyn.	**HT**, Body, Draw, Virtual Replicas	**Asym.**	Draw, Virtual Replicas	**TcT**	**CL.**, Pref., So-p, Spa-p, **U.**
2023	Zaman et al. [42]	Asym.	VR-HMD AR-HMD, Screen	Video 360°, Video	**Dyn.**	**HT**, Annotation, Pointer	**Asym.**	Video 2D, 360° w.o NVC, 360° w. NVC	**TcT**	Pref., So-p, Spa-p, **U.**
	Our approach	Sym.	AR-HMD	Point Cloud	Dyn.	HT, ET	Asym.	First-person Video, Dyn. Point Cloud	TcT	CL., CS., U., O., Qual., Emp.

Communication Cues

Non-verbal communication is essential during collaborative tasks. It fosters mutual understanding between interlocutors. It leads to a more empathetic collaboration whose effects are remarkable both in performance and user's feelings like co-presence and usability [8, 25]. We distinguish three categories of communication cues: intention, attention, and emotion.

The main system for reproducing **intentions** is the use of hand tracking to transmit either replication of hands or pre-recorded gestures in the SVS [22, 28, 29]. This replication may be replaced or combined with deictic support as the pointer metaphor [13, 35]. Annotations and diagrams may also contribute to share intentions [43].

The restitution of **attention** consists in indicating the focal point of attention for each interlocutor within the SVS. In the case of dependent vision, this indicator is less useful, given that collaborators perceive the same visual input [37]. However, it is more useful when the field of view of one of the collaborators is reduced [22]. The use of eye tracking is widespread and has the advantage of reinforcing the feeling of co-presence and the collaboration quality, enabling individuals to engage with the gaze of their interlocutors [13, 28, 33]. In the context of asymmetric technology systems (e.g., AR-VR collaboration) where the field of view differs between headsets, rendering a frustum representing remote collaborators' field of view enhances empathetic collaboration [12, 22].

Technologies capturing **emotions** are rarely present natively on devices [17]. Fairchild et al. have created a technology that enables the generation of realistic avatars based on real-time video capture of users. This technology allows for the virtual transcription of facial expressions [11].

In addition to these cues, we mention that Lee et al. [22] employed directional indicators, such as arrows or a light halo to assist collaborators in locating cues not visible.

In an RPC system, the physical environment is central to collaboration. We therefore wanted collaborators to be able to share their intentions and attention in relation to this environment. The use of hand, head and eye tracking enables us to update the behavior of avatars[3] embodied by collaborators.

[3]Avatars made on MakeHuman: http://www.makehumancommunity.org/

Classification of XR collaborative studies

Cited studies on collaborative XR presenting a means of rendering a remote environment in a shared space have been classified in Table 1. The following table provides a summary of experimental technologies, environment restitution methods, task symmetry, studied conditions, and used metrics. Most redundant elements are the asymmetry of the task and the technology. In particular, the use of VR for the remote user and AR for the local user. To the best of our knowledge, no previous collaborative study presenting an environment rendering method has used AR-HMD symmetrically. Environment rendering methods are relatively evenly distributed between those employing spatial and two-dimensional restitution. Studies generally employ the point cloud method for spatial restitution. According to this table, it appears to be the most appropriate method for obtaining spatial restitution that can be updated during collaboration. In most studies, task performance is evaluated with task completion time. Subjective measures used are co-presence and spatial presence in order to evaluate the collaboration, and cognitive load to compare effects on users. Conditions studied in the literature are non-verbal communication cues, vision independence, rendering methods, and roles. The most frequently employed communication cues are hand tracking, field of view, and pointers.

3 System Overview

Our approach involves equipping all collaborators with HoloLens 2 OST-HMDs, enabling them to remain aware of their real environment. Our use case consists of a local worker requiring assistance and sharing his environment with a remote expert. Fig. 1 shows how collaborators interact within the SVS. They embody avatars and share a virtual space in which the local user's environment is rendered in three dimensions for the remote user. The local collaborator sees and interacts with the remote expert's avatar in his real environment. The remote expert is immersed in the reconstructed environment of the local worker and interacts with the local user's avatar.

In this section, we present the architecture of our RPC-AR system. We implemented our HMD client with Unity[4] 2022.3.14 and used MRTK[5] 2.8.3.0 for AR management. Avatars were managed by inverse kinematics provided by the VRIK plugin from RootMotion[6]. Our constraint when designing this system was to make our application as adaptable as possible for other use cases. Thus, we wanted to not require additional devices for the two HoloLens 2 and to render the environment reliably while limiting latency, using wireless IP transmission. Users will then be able to identify objects with ease, and latency will not affect communication.

3.1 Environment Reconstruction

Spatial data is captured by the HoloLens 2's built-in depth sensor. This time-of-flight sensor can be used in two different modes: (1) *short-throw mode* for near-field depth sensing with a high frame rate (45 fps), (2) *long-throw mode* with a low frame rate (5 fps) for spatial mapping. Unfortunately, the range of the short-throw mode is too small to capture an environment, so we use the long-throw mode. The maximum number of points returned per frame using this mode is 12,827, based on data from 633 captures. The HoloLens' RGB sensor completes the rendering. We use a resolution of 896x504 pixels at 30 fps.

Thus, data needed to reconstruct environments remotely are captured and transmitted at two different frequencies (see Section 3.2). Spatial correlation between spatial and color data is ensured by a specific texture's transformation matrix from world space to image space. This matrix is the result of multiplying the intrinsic matrix, which depends on camera parameters, and the extrinsic matrix, which depends on the camera position relative to the depth sensor coordinate system. We consider the temporal offset between captures to be small enough to neglect aberrations.

[4]Unity Technologies: https://unity.com/

[5]Microsoft Mixed Reality Toolkit: https://github.com/microsoft/MixedRealityToolkit-Unity

[6]Root Motion: http://root-motion.com/

The **point cloud** method was selected for the reconstruction process. This method involves displaying virtual points at positions returned by the depth sensor, which are then colored according to the RGB camera data. It has the advantage of not requiring any additional internal processing (e.g., mesh reconstruction). Since the resolution of the Hololens 2's depth sensor does not provide sufficient point density to recognize objects by displaying dots $D_p(Position_p, Color_p)$ for each point p [12], we chose to display tiles for each point. Fig. 3 shows the principle behind these quads. On the left-hand side, the local user is wearing the HMD that captures his environment. Tiles are represented by black squares and are not visible to the user. Their size is calculated to match the sensor resolution at the distance between the captured point and the sensor. On the other side, the remote user sees the virtual restitution of these tiles and the texture projected onto them. Several work use built-in frameworks for point cloud rendering [1, 40], Niedermayr et al. employed a similar particle method in their system [27]. The creation of tiles and the texture projection using the projection matrix is entirely managed by the GPU through DirectX shaders to maintain acceptable performance.

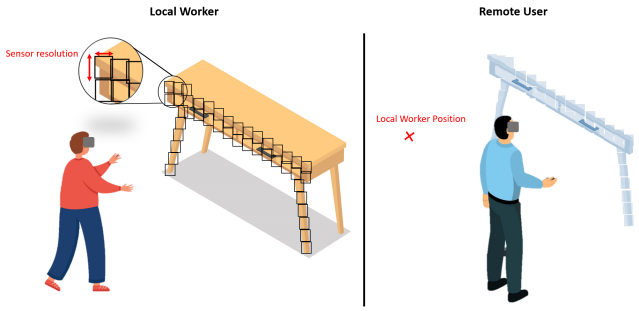


Fig. 3: Tile principles. On the left, quads represent tiles that will be displayed remotely; the local user does not actually see them. The size of tiles is determined by the sensor's resolution at the distance from the point. On the right, the remote user sees textured tiles, recognizing elements from the local user's environment.

To complete the reconstruction configuration, it is necessary to define **tiles orientation**. To illustrate this case, we define β as the angle between the local sensors, a captured point, and the remote collaborator's viewpoint. Fig. 4 shows a rendering of the environment when tiles are oriented towards the capture point (a), or the remote collaborator's viewpoint (b). If $\beta < 90°$, (b) offers a better quality for each viewpoint, while for (a) the quality of perception deteriorates from 30°. For $\beta > 90°$, (b) continues to offer acceptable quality, while the environment rendering is not visible in (a). We chose to orient these tiles towards the viewer to maintain good quality even at $\beta > 30°$. We set their orientation by aligning their normal with the observer's camera position, avoiding the need to calculate β at runtime. To the best of our knowledge, we are the first to include this aspect in our rendering method.

The **spatial data coordination** is achieved by expressing tracking and sensor data in the scene's global coordinate system. When this data is transmitted, it is then utilized in the remote scene's global coordinate system. Subsequently, on the remote side, both local and remote data are expressed in the same coordinate system. In order to adapt the scene to their surrounding environment, remote collaborators are able to use a relocation button to position themselves to the right of their interlocutor, thereby modifying the orientation of the scene's coordinate system.

3.2 Data Transmission

To render the environment remotely, we implemented a system for transmitting points' positions composing the point cloud, the RGB texture, and its associated projection matrix. Fig. 5 shows the architecture of the transmission and computation flow. We split the computational load between headsets, with the "local" headset capturing, serializing, and transmitting the data. The "remote" headset receives data, deserializes

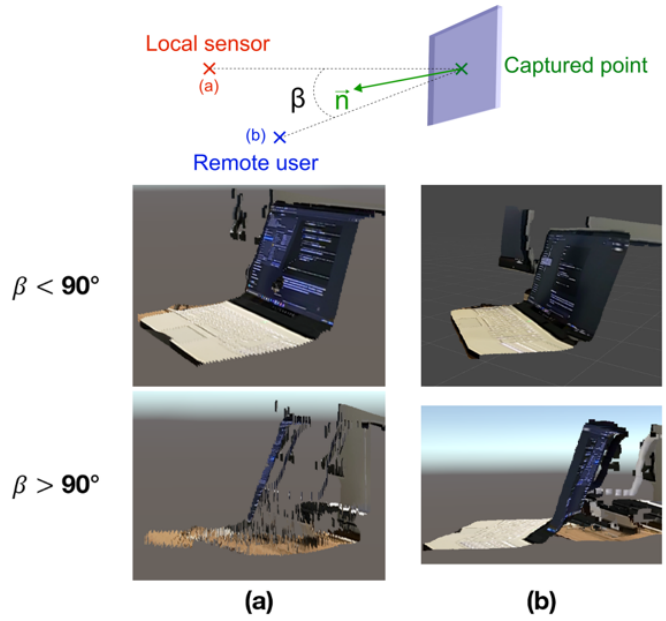


Fig. 4: Tiles Orientation. (a) represents the remote user's point of view if tiles are oriented towards the sensor; (b) represents this point of view if tiles are oriented towards him; β is the angle between the remote collaborator and the local sensor: if $\beta < 90°$, (b) offers better quality for each viewpoint; if $\beta > 90°$, (b) continues to offer acceptable quality.

it, and performs projection computations on the GPU, augmenting the remote collaborator's environment with the task environment perception. As the Hololens' computing capacity is limited, we decided not to use compression algorithms and to restrict textures' size. Interactions management, such as avatar's positions and voice chat are managed via a Photon PUN[7] 2.39 server.

The maximum amount of data to be transmitted is about 154 Kb/frame for the spatial data and about 5.4 Mb/frame for the texture. To achieve the best possible latency, we use different threads and transmission ports for texture and point cloud. We display the last frame received by each thread on the remote side. Data is transmitted using unicast UDP, so for a local Wi-Fi 4 transmission, we obtain an average latency between 2 frames for the remote user of 196 ms ($\approx$ 800 Kb/s) for the point cloud (sensor: 5 fps), and 114 ms ($\approx$ 50 Mb/s) for the texture (sensor: 30 fps), and a total latency of less than 250 ms between the capture and the remote side's feedback. We therefore obtain a low latency compared to similar works, such as Niedermayr et al. who achieve up to one-second latency [27]. Bai et al. [1] obtained a latency of approximately 300 ms for the transmission of an entire environment, using a bandwidth of up to 10 Gbps.

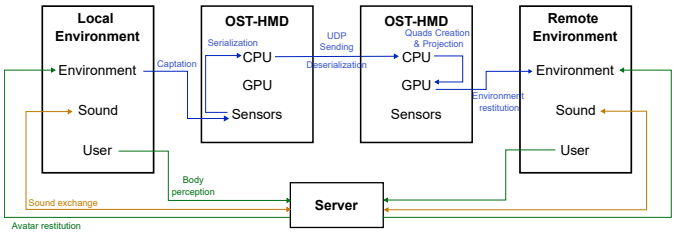


Fig. 5: Data Flow Architecture: Blue arrows represent environmental data exchanges; green ones represent tracking data exchanges; and yellow ones represent sound exchanges.

[7]Exit Games GmbH: https://www.photonengine.com/

4 User Study

The aim of this study is to compare our solution with those currently used, to evaluate the effect of a shared virtual space in three dimensions compared to a 2D video streaming solution. Indeed, remote collaboration techniques employed by industrial companies rely on real-time two-dimensional video technologies. These include the use of smartphones or tablets for video capture and the participation of people in different locations through their laptops. Then, the research question of this study is: *"How does the use of a shared virtual space in which a physical environment is spatially rendered in real-time affect collaboration compared to video capture?"*.

4.1 Conditions

Our two conditions under study are **Our RPC-AR** (**O**) system and the **TeamViewer Assist AR** (**TW**) application. We distinguish two different roles for the collaborative task. The first role is the **local technician** ($\mathbf{O_L}$, $\mathbf{TW_L}$), who shares his environment and must carry out the task. The second is the **remote expert** ($\mathbf{O_R}$, $\mathbf{TW_R}$), who must help the local technician with his task. The distribution of these roles is explained in Section 4.4.

All participants experimenting with the *O* condition wear a HoloLens 2 OST-HMD. They hear and see each other through avatars. The remote user sees a 3D rendering of his collaborator's real-world environment. The TeamViewer Assist AR application is running on an OST-HMD HoloLens 2 for the local user. The remote user in the *TW* condition uses a laptop to view the video stream captured by the local headset's camera. Collaborators communicate verbally and non verbally by placing 3D markers when they click in the local user's space to help him with his task, no other communication cue is allowed by this application. The use of an OST-HMD for the local user in both conditions allows us to keep similar modes of interaction with applications and devices. Thus, we focus on differences related to environment reconstruction.

4.2 Task

Our study focuses on network technician use cases, with a local technician and a remote expert configuration. The use case of this study consists of connecting a remote room to the home Internet network. A laptop represents the device to connect in the remote room, and the home Internet network is represented by an Internet router. The connection is made using a mock-up of a fibre optic network panel and two fibre-to-Ethernet converters. Fig. 6 shows the final assembly, consisting of the box (d), a converter (a), the panel (b), a second converter (a), and the laptop (c) connected in series.

This design enables us to evaluate the relevance of a shared virtual space where communication and task spaces are merged [24, 35], and collaborators using symmetric technologies [9]. This use case corresponds to the use of a shared space and a lightweight environment spatial restitution for a simple collaborative task, and corresponds to industrial needs. Some parts (SFP modules and dual fibre cables, see Section 4.3) of this assembly task require precise gestures, and capturing fine components such as fibre cables requires attention.

4.3 Apparatus

The local user wears a HoloLens 2 in both conditions, running our application and the TeamViewer Assist AR v15.28 standard client. In the *O* condition, the remote user also wears a HoloLens 2, while in the *TW* condition, he uses a laptop to run the standard TeamViewer v15.53 desktop application. This technological difference is because a two-dimensional rendering does not require immersive technologies. However, differences between these two technologies may have an impact on the collaboration. We discuss this impact in relation to results obtained and observations made by participants in Section 5.

All devices are connected in Wi-Fi 4 to a Wi-Fi router connected with an optical fibre connection. In the *O* condition this local Wi-Fi connection is used to transmit the environment's data. This network is independent of the network to be connected during the task. We created a network panel mock-up to represent a unidirectional fibre panel, so collaborators have to connect one fibre for sending and one for receiving, using dual fibre cables. We also needed a router, two

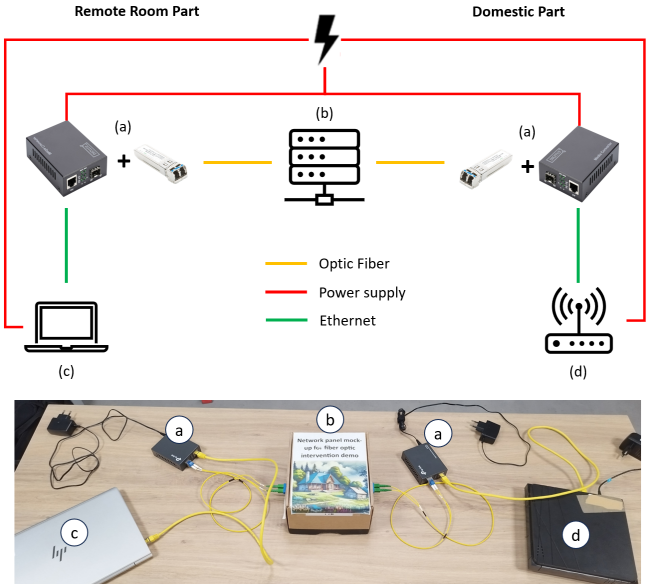


Fig. 6: Assembly diagram: (a) Media Converter and SFP module assembly; (b) Network panel mock-up; (c) The laptop symbolising devices to be connected in the remote room; (d) Home network router.

TP-Link Ethernet Media Converters, and two TP-Link SFP modules: a module that plugs into Ethernet Media Converters and allows them to be connected to optical fibres.

4.4 Procedure

Participants were randomly divided into two groups, corresponding to the two conditions. Each participant experimented with the roles of local technician and remote expert for the same condition. To consider a participant as an expert, we decided that each participant would start as a technician with his predecessor as an expert. The participant will then be the expert, helping his successor in the technician role. In this way, the expert, having already completed the assembly as a technician, would be able to guide a new participant with the help of a diagram and instructions summarising each stage of the assembly.

We welcomed each participant by presenting the use case and roles they would have to play. We then asked them to fill in information and consent forms. Then, they learned how to use the OST-HMD to launch and configure applications. After the launch, the remote expert joined the call and we verified that everything was working for them. These preliminary steps lasted about 10 minutes. After verifications, participants were asked to start the experiment and perform the task described in Section 4.2. Each session took approximately 15 minutes and was considered finished when a participant said "the assembly is complete" to give them all the time they needed to verify their assembly. Between each role, participants had to complete questionnaires cited in Section 4.5 and wait for the next participant to be ready for about 10 minutes, and start a new session in the expert's role. The experiment lasted about 1 hour per participant. We have protected personal data and anonymized participants regarding european RGPD rules. No ethics committee approval was available and required by the institution where the participants were recruited and where the experiment took place.

4.5 Metrics

We used metrics to compare our prototype with a production application (TeamViewer). We chose them to evaluate participants' ability to understand and collaborate on the environment.

Task completion time is measured as an objective metric. This measure allows us to compare participants' performance across conditions. In particular, this measure should reflect how easy it is to understand the environment restitution to give correct and efficient instructions.

We evaluate empathy between collaborators with the **Collaboration and Mutual Awareness** questionnaire [25]. This questionnaire consists of seven questions regarding participants' empathic impressions. Specifically, it evaluates their ability to understand each other's emotions.

The **NASA Task Load Index** [14] assists us in comparing the cognitive load required by both conditions. The aim is to make a comparative analysis of variables affected by the quality of the environment rendering, such as mental demand, overall effort, feelings of success, and frustration.

To compare the effect of 6 DoF vision with first-person vision, we assessed **Cybersickness** after each session using the Fast Motion Sickness Questionnaire [20]. We will not discuss further the results as they did not indicate any discernible patterns, nor did they suggest that applications induced cybersickness.

We also compare usability using the **System Usability Scale** [6] and collect **users' feedback** through an open-ended question.

In addition to these questionnaires, we adapted a questionnaire to evaluate the **environment restitution quality**. We based this questionnaire on previous Uncanny Valley Questionnaires [15, 16], initially used to assess the realism of avatar representations. The three categories from previous work were designed to measure attractiveness, humanness, and eeriness [15]. According to Ho et al. [15], eeriness and the lack of attractiveness are components leading to a mechanism of avoidance, whereas humanness corresponds to a subjective perception of photorealism of a human-like character. We chose to keep attractiveness and eeriness as scales to qualify an environment, but we adapted "humanness" to "realism" as it better corresponds to our case and preserves the initial objective of this scale. For each category, we selected a scale from those proposed by Ho et al. [16] that could be easily adapted to evaluate an evironment restitution: (1: Attractiveness) Messy - Sleek, (2: Humanness $\rightarrow$ Realism) Synthetic - Real, (3: Eeriness) Uncanny - Bland. We propose these terms as boundaries of a 7-point Likert scale with the same introduction *"The environment restitution was:"*.

4.6 Hypotheses

Our objective is to compare a first-person video collaboration solution *TW*, and our spatial RPC-AR solution *O*. Enabling users to share the same virtual space, and using symmetrical technologies. While other work has explored vision independence [22], 3D image projection [39], or interpersonal communication [1], we developed this solution to better address industrial collaborative maintenance needs by using lightweight environment reconstruction and AR for all collaborators. Motivated by prior research, we make the following hypotheses:

H1 *The empathetic aspect of the collaboration will be accentuated in the O condition.*
The use of independent vision, along with the interpersonal communication tools enabled by an SVS, should result in greater social presence and improved understanding of the other collaborator's feelings [1, 22, 31].

H2 *Environment rendering will be more uncanny for the O than the TW condition.*
Since the notion of strangeness is linked to the user's habits, the restitution of the environment should be more uncanny for the *O* than for the *TW* condition. Moreover, spatial environment rendering often creates artefacts [10] as well as image stretching and folding impact the rendering quality [39].

H3 *The task completion time will be higher in the O condition than in the TW condition.*
As prior works did not show significant differences for whole environment rendering solutions [1, 39], we think that partial environment rendering will be less performant.

H4 *The cognitive load will be higher in the O condition, in particular for the remote role.*
As for [H2], we think that the quality of rendering and the habits of the participants will be in favor of the video solution.

Table 2: NASA-TLX results. Yellow cells represent a significant difference in favor of the first condition ($m_1 > m_2$). Blue cells represent the opposite inequality ($m_2 > m_1$)

	Means ± SD			
	Local O	Remote O	Local TW	Remote TW
Mental	5.44 ± 3.95	10.6 ± 4.34	3.15 ± 2.15	6 ± 2.77
Physic	3.00 ± 1.97	4.64 ± 3.71	2.31 ± 3.55	2.23 ± 2.74
Temporal Demand	6.44 ± 4.49	7.00 ± 4.47	3.39 ± 3.20	4.62 ± 4.43
Success	18.3 ± 2.11	14.0 ± 4.46	18.6 ± 1.50	17.6 ± 1.71
Frustration	3.88 ± 3.52	8.07 ± 4.46	2.46 ± 2.93	3.31 ± 2.25
Effort	9.44 ± 5.79	12.0 ± 5.38	7.23 ± 4.66	10.8 ± 5.37
	p-values			
	Local O - TW	Remote O - TW	O Remote - Local	TW Remote - Local
Mental	.156	.004**	.014*	.061
Physic	.132	.076	.136	.375
Temporal Demand	.028*	.169	.649	.313
Success	.846	.012*	.008**	.156
Frustration	.288	.002**	.027*	.844
Effort	.276	.558	.164	.164

4.7 Subjects

31 subjects participated in this experiment, 7 females (22.6%) and 24 males (77.4%). Participants were aged between 20 and 59 years ($m = 39.4 \pm 13.8$). They were recruited at our facility, all of them worked in IT in a variety of fields (XR, AI, design, and telecoms). We asked them to rate their familiarity with XR and network hardware on a 7-point Likert scale. They were familiar with network hardware ($m = 4.4 \pm 1.5$) and XR ($m = 3.3 \pm 1.7$). Because of uncollected data due to questionnaire issues, we removed 2 outliers from the O_R condition. We also considered that the expert should not make a mistake in the assembly. Two experts made a mistake in the TW condition, they warned us before the experiment they were not familiar with network hardware at all. We decided not to include their session from our final dataset, keeping them for the technician role and their pairs for the expert role, and explaining all the assembly to them again. Finally, we analyzed results of 16 participants in the O_L condition, 14 in the O_R condition, 13 in the TW_L condition and 13 in the TW_R condition.

4.8 Results

For all data analyses, we first conducted Shapiro-Wilk's test to verify normality and Levene's test to assess the homogeneity of variances (results noted with $\neq$ if variances were significantly unequal). In most cases, we tested the equality of means using a two-tailed Student's t-test when normality was confirmed, or a two-tailed Mann-Whitney test when it was not. For paired non-normal samples, we conducted the Wilcoxon Signed-Rank test. In tables and figures, (***) denotes a significant difference with $p < .001$, (**) indicates $.001 < p < .01$, and (*) signifies $.01 < p < .05$ ($\alpha = .05$).

4.8.1 Completion Time

We recorded task completion times for participants assigned the role of expert during this session ($M_O = 602 \pm 163s$; $M_{TW} = 501 \pm 89s$; Fig.7a). A Student's t-test indicated no significant difference between the two conditions ($t(21) = 2,\ p = .058,\ \neq$).

4.8.2 Uncanny Valley

We collected results of questionnaires relative to the environment restitution's quality after the participants experienced the remote role. Mann-Whitney's tests revealed significant differences in favor of the TW condition for all three categories: the first one, dedicated to Attractiveness ($M_O = 2.71 \pm 1.44$; $M_{TW} = 5.54 \pm 1.39$; $U_1 = 15.5,\ p_1 = 6.67 * 10^{-5}$; Fig.7c), the second category for Realism ($M_O = 3.43 \pm 1.34$; $M_{TW} = 6.08 \pm .95$; $U_2 = 9.5,\ p_2 = 9.67 * 10^{-6}$; Fig.7d) and the third for Eeriness ($M_O = 2.86 \pm 1.41$; $M_{TW} = 5.77 \pm 1.48$; $U_3 = 17.5,\ p_3 = 1.18 * 10^{-4}$; Fig.7e).

4.8.3 Task Load Index

We chose to keep the results separated by different categories to better identify sources of task load in our application. Table 2 summarizes results for all conditions. We found significant differences in favor of the TW condition in the remote role for mental load ($U = 33.5,\ p = 3.95 * 10^{-3}$) and success feeling ($t(17) = 2.82,\ p = .012,\ \neq$), while the O condition led to higher frustration ($t(25) = 3.46,\ p = 1.95 * 10^{-3}$). Additionally, we found that O_R led to higher frustration than the O_L condition ($t(10) = 2.60,\ p = .027$), higher mental load ($T = 4,\ p = .013$) and lower success feeling ($T = 1.5,\ p = 7.81 * 10^{-3}$). On the local side, O_L users felt a significantly greater temporal demand than TW_L ($U = 54.5,\ p = .028$).

4.8.4 Collaboration and mutual awareness

Table 3 summarizes results for the seven questions of the collaboration and mutual awareness questionnaire. We did not find any significant difference for Q1, Q2, Q3 and Q7. For Q4, a significant effect of the O condition on communication feeling was observed only in the remote role ($U = 28.5,\ p = .002$), showing that our application may affect communication feeling. Regarding Q5, the O condition showed a significant improvement compared to the TW condition in the remote role ($U = 48.5,\ p = .038$). Additionally, for Q6 in the O condition, we found a significant difference between roles ($t(10) = -2.96,\ p = .014$), favoring the local user.

4.8.5 Usability

We compared usability results between conditions for the same role, Fig.7b shows the results of the SUS questionnaire. A Student's t-test did not reveal significant differences between local roles ($M_{OL} = 72 \pm 13$; $M_{TWL} = 74 \pm 16$; $t(25) = .22,\ p = .83$), however we found that usability was better in the TW_R condition than in the O_R condition ($M_{OR} = 57 \pm 12$; $M_{TWR} = 72 \pm 12$; $t(23) = 2.9,\ p = 8.1 * 10^{-3}$).

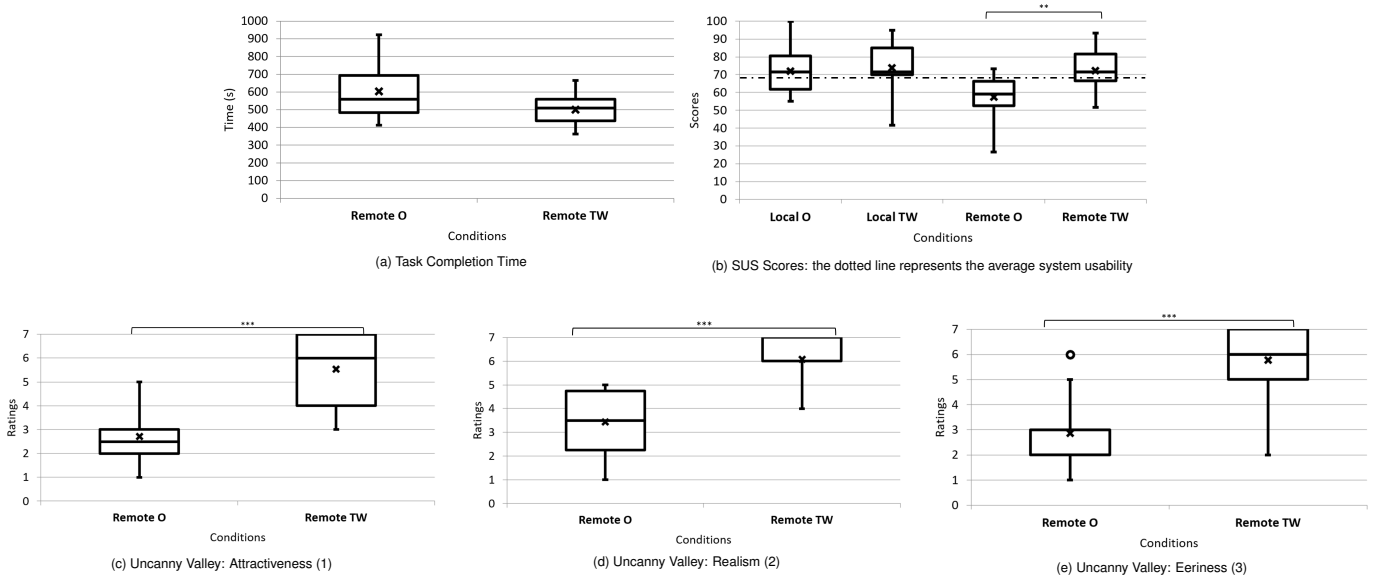
5 Discussion

We hypothesised that our environment rendering would be more uncanny than the video rendering [H2]. Results of the Uncanny Valley questionnaire show a clear difference between both conditions on the three criteria. These results allow us to validate our hypothesis and are consistent with the results obtained by prior works [7, 39] on visual comfort. The analysis of users' observations shows that 5 of the 14 participants (35.7%) in the O_R condition reported that the quality of the environment restitution was insufficient: *"the image sharpness is very low", "the image had a fairly low resolution"*. However, we note that some of O_R users did not use the relocation button and did not look for a better viewpoint during the collaboration, leading to a deterioration of their perception.

Contrary to our hypothesis H3, we did not find a significant difference in task completion time between the two conditions. In the

Table 3: Collaboration and Mutual Awareness results. Yellow cells represent a significant difference in favor of the first condition ($m_1 > m_2$). Blue cells represent the opposite inequality ($m_2 > m_1$)

	Means ± SD			
	Local O	Remote O	Local TW	Remote TW
Q1 My partner and I worked well together	6.25 ± 0.93	5.93 ± 1.00	6.31 ± 0.86	6.23 ± 0.60
Q2 It was easy to know what my partner was doing	4.25 ± 1.73	4.29 ± 1.27	3.77 ± 1.96	5.00 ± 1.53
Q3 I felt connected with my partner	5.44 ± 1.41	5.14 ± 1.17	5.69 ± 1.44	5.54 ± 1.05
Q4 My partner and I communicated together well	5.94 ± 1.18	4.50 ± 1.16	6.15 ± 1.14	6.00 ± 0.91
Q5 I understood how my partner was feeling	4.31 ± 1.54	5.64 ± 0.75	4.31 ± 1.84	4.62 ± 1.19
Q6 My partner understood how I was feeling	4.50 ± 1.79	4.00 ± 1.04	4.77 ± 1.59	4.15 ± 1.35
Q7 I am satisfied of the result	6.31 ± 1.01	5.57 ± 1.51	6.69 ± 0.48	6.39 ± 0.77
	p-values			
	Local O - TW	Remote O - TW	O Local - Remote	TW Local - Remote
Q1 My partner and I worked well together	.948	.519	.469	.625
Q2 It was easy to know what my partner was doing	.490	.197	.572	.219
Q3 I felt connected with my partner	.475	.402	1.00	.563
Q4 My partner and I communicated together well	.589	.002**	.734	.813
Q5 I understood how my partner was feeling	.914	.038*	.461	.625
Q6 My partner understood how I was feeling	.675	.741	.014*	.347
Q7 I am satisfied of the result	.351	.169	.219	.750



(a) Task Completion Time

(b) SUS Scores: the dotted line represents the average system usability

(c) Uncanny Valley: Attractiveness (1)

(d) Uncanny Valley: Realism (2)

(e) Uncanny Valley: Eeriness (3)

Fig. 7: Task completion time, usability and uncanny valley scores. Result for task completion time (a) show no significant difference. Results for usability (b) and uncanny valley (c), (d), and (e) show significant differences for participants in the remote role (b : $p < .01$; c, d, e : $p < .001$)

.

literature, results on remote collaboration comparing 2D and immersive technologies vary depending on the task and participants [7,42]. Our results show that the O condition led to more dispersed times than the TW condition. In particular, larger maxima were observed for the O condition. These differences may be due to the fact that some participants did not search for a better viewpoint, but also to differences in technology adoption or ability to interpret three-dimensional data.

We confirm our hypothesis H4 on task load. Indeed, participants in the O_R condition experienced higher mental demands and frustration and a lower sense of success than in the TW_R condition. With independent vision, users move to find their own viewpoint, while remaining aware of their partner [39], potentially resulting in a higher mental load. Moreover, these differences are also observed between roles in the O condition, as O_R users experienced a higher load than O_L users, in line with prior results [27,39]. This result is not surprising, as it is the remote collaborator who directs the collaboration and his cues depend on the environment restitution. On the local side, results showed no differences except for temporal demand, which was higher in the O_L condition than in the TW_L condition. This may be due to the higher task load experienced by the remote collaborator, which may be perceived by the local user.

Regarding mutual awareness, we confirm our hypothesis H1. The remote collaborator had a better understanding of the local collaborator's feelings in the O condition compared to the TW conditioPrior work demonstrated that using interpersonnal communication tools [1,17,31] as well as vision independence [22], contribute to improved empathy and social presence [1,25]. However, the perception of effective communication is lower in the O_R condition compared to the TW_R condition (CA - Q3). This may be due to the additional need to communicate about the other participant's position.

SUS results showed no significant differences between conditions in the local role. However, we found that users on the remote side ranked our application significantly lower than the TeamViewer application. All conditions except the O_R one received an average score of around 72, highlighting a "good" acceptance [2]. Our application received an average score of 57 for the remote role. According to Bangor et al. [2], this score shows that our application can be categorized as "fair", meaning it is usable but needs to be improved. Participants highlighted

the need to improve the environment restitution's quality, although flexibility allowed by shared virtual space and independent vision was appreciated*: "A lot of pixelation and very partial images. But it's still fun to use"*. These results are still promising, considering that the application concurently used is a market-ready application, while ours is a prototype. Future improvements to our system (improved quality, extended capture field, see Section 6), along with advancements in underlying technologies, could therefore lead to higher usability.

An open-ended question was used to gather participants' impressions, and comments about hardware issues were present in both conditions. In the O_R condition, five participants (35.7%) noted that the environment was quite limited due to the small aperture of the depth sensor *"I could only see a small part of what my partner was seeing"*. Similarly, two of them (14.3%) complained that the sensor failed to detect objects in the local user's hands: *"as soon as the person has the objects in their hand, they will no longer be visible."*. This issue arises from the use of the long-throw mode, which increases the minimum detection distance. However, its impact was limited during the experiment, as participants were instructed to keep objects on the table as much as possible. In addition, participants mainly mentioned the headset's limited field of view and the fact that sensors were aimed too high. Eight out of the 43 participants who were equipped with an HMD in all conditions (18.6%) felt that the restricted field of view hindered collaboration. They reported that it prevented them from simultaneously viewing both the non-verbal cues of their interlocutor and the environment *"the field of view is a little small, so it's sometimes difficult to see the expert or what he's showing us"*. In the TW_R condition, four participants (30.7%) reported being disturbed by camera movements *"following a video of someone watching is unbearable"*, supporting our belief that 6 degrees of freedom vision control is ideal for long collaboration sessions.

Finally, we evaluated the interest of RPC-AR compared to a 2D video rendering collaboration tool. We chose to use a laptop for the remote collaborator in the TW condition, as two-dimensional rendering does not require immersive technologies. Feedback mainly highlighted hardware issues related to the headset's design, such as field of view and sensor placement. We therefore consider that the results obtained may be biased in favor of the TW condition, especially with regard to the field of view, as the same sensors were used in both conditions. Thus, our conclusions regarding usability, performance, and mutual awareness remain valid. Indeed, improving sensor positioning and expanding the field of view could minimize the need for users to adjust their head positioning during collaboration, thereby reducing task load and enhancing usability and performance.

6 Conclusion

In this article, we presented our remote physical collaborative AR system. This system is designed to comply with industrial maintenance needs of non-controlled environments with current technologies, by using flexible acquisition methods, wireless connectivity, and limited computing power. It provides a shared virtual space and interpersonal communication tools for collaboration. These tools include lightweight remote environment rendering, verbal and non-verbal communication. Collaborators are represented by avatars in the virtual space, sharing their intentions through hand and finger tracking, and their attention through eye tracking.

We conducted a cascading user study comparing our system with a remote collaboration application that used 2D video and pointer cues for non-verbal communication. Our task design allowed us to evaluate the relevance of a shared virtual space on a simple task that tends to correspond to industrial use cases. We evaluated task load, mutual awareness, usability, quality of environment rendering through an adapted questionnaire inspired by the Uncanny Valley Questionnaire [16], and measured performance through task completion time. Although our application resulted in a higher cognitive load and a reduced usability for the remote user, we did not observe any significant difference in performance. Remote collaborators who used our application also had a better understanding of their interlocutors' emotions, making the collaboration more empathetic.

Limitations and Future Work

In a context where one of the employees is carrying sensors needed to reconstruct an environment remotely, **collaborators' placement** has a major influence on the quality of the environment feedback [39]. This placement is crucial for both local users, wearing the sensor, and remote users, who must find the optimal viewpoint. In this context, unconstrained incentives and nudges could provide a partial solution to this problem, while still keeping the experience engaging for users.

Furthermore, in our study, conditions were optimal to get the best possible **transmission performance**, as headsets were connected to the same network. However, for industrial applications, it will be important to evaluate the degradation of this performance when headsets are connected to remote networks. In such cases, the time required to compress and decompress data might be balanced by the time saved on transmission.

Participants' comments also highlighted several **technical limitations** of the headset (see Section 5). Specifically, the aperture and accuracy of the depth sensor, as well as its range and position, seem to be a hindrance : *"coarse resolution does not (yet) allow for effective assistance"*. Additionally, devices' computing power influenced the design choices we made. It would be beneficial to anticipate future hardware improvements by incorporating more powerful and better-positioned sensors [27]. Similarly, transferring computations to more powerful devices could significantly enhance our system's quality and swiftness. Despite the end of the Microsoft HoloLens program, our solution can be used with any headset allowing access to depth and RGB sensors. New video pass-through headsets such as the Apple Vision Pro or the Meta Quest 3 are interesting alternatives, although raw data sensors' access is still limited. We are also looking for the next generation of optical see-through headsets and glasses.

Accessing sensor data and transmitting it over the network also raises **privacy issues** [30, 36]. Many headsets still restrict their sensors' access for these reasons. This work does not address concerns about cybersecurity or the consent of users or other individuals or organizations that may be affected by the collection and transmission of this data. However, these issues need to be addressed before these technologies are deployed.

We highlighted several technical challenges with our task design. We focused on a network use case, taking into account constraints such as the need to capture and render very thin cables or deal with the depth sensor's minimum range when manipulating objects. Evaluating this application with **more complex tasks or in different domains** will help us identify and address additional challenges [9, 36].

We adapted the **Uncanny Valley questionnaire** [15, 16] to qualify environment renderings. This questionnaire gave consistent results in our study. However, it has not yet been tested extensively across a wide range of environment renderings. We encourage future works to use this questionnaire to complete the conclusions we draw from it.

Finally, two other comparisons seem interesting to us to complete knowledge about spatial XR collaboration. First, a comparison with our system using a laptop for the remote collaborator would allow us to study the **contribution of immersive systems to collaboration**. Second, comparing our system to a 360° video-based collaborative system could help us to evaluate **the contribution of a shared virtual space and interpersonal communication** to collaboration.

References

[1] H. Bai, P. Sasikumar, J. Yang, and M. Billinghurst. A user study on mixed reality remote collaboration with eye gaze and hand gesture sharing. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pp. 1–13, 2020. 2, 3, 4, 5, 6, 8

[2] A. Bangor, P. Kortum, and J. Miller. Determining what individual sus scores mean: Adding an adjective rating scale. *Journal of usability studies*, 4(3):114–123, 2009. 8

[3] G. Bataille, A. Lammini, and J.-R. Chardonnet. Arpuzzle: Evaluating the effectiveness of collaborative augmented reality. In *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 642–651. IEEE, 2023. 1, 2

[4] M. Billinghurst and H. Kato. Collaborative Mixed Reality. In Y. Ohta and H. Tamura, eds., *Mixed Reality*, pp. 261–284. Springer Berlin Heidelberg, Berlin, Heidelberg, 1999. doi: 10.1007/978-3-642-87512-0_15 2

[5] M. Billinghurst, I. Poupyrev, H. Kato, and R. May. Mixing realities in shared space: an augmented reality interface for collaborative computing. In *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No.00TH8532)*, vol. 3, pp. 1641–1644 vol.3, 2000. doi: 10.1109/ICME.2000.871085 1

[6] J. Brooke. Sus: a "quick and dirty" usability scale. *Usability evaluation in industry*, 189(3):189–194, 1996. 6

[7] M. L. Chenechal, T. Duval, V. Gouranton, J. Royan, and B. Arnaldi. Vishnu: virtual immersive support for HelpiNg users an interaction paradigm for collaborative remote guiding in mixed reality. In *2016 IEEE Third VR International Workshop on Collaborative Virtual Environments (3DCVE)*, pp. 9–12, Mar. 2016. doi: 10.1109/3DCVE.2016.7563559 3, 7, 8

[8] U. X. Eligio, S. E. Ainsworth, and C. K. Crook. Emotion understanding and performance during computer-supported collaboration. *Computers in Human Behavior*, 28(6):2046–2054, 2012. 3

[9] B. Ens, J. Lanir, A. Tang, S. Bateman, G. Lee, T. Piumsomboon, and M. Billinghurst. Revisiting collaboration through mixed reality: The evolution of groupware. *International Journal of Human-Computer Studies*, 131:81–98, Nov. 2019. doi: 10.1016/j.ijhcs.2019.05.011 1, 2, 5, 9

[10] A. Fages, C. Fleury, and T. Tsandilas. Understanding multi-view collaboration between augmented reality and remote desktop users. *Proceedings of the ACM on Human-Computer Interaction*, 6:1–27, 2022. 1, 6

[11] A. J. Fairchild, S. P. Campion, A. S. Garcia, R. Wolff, T. Fernando, and D. J. Roberts. A Mixed Reality Telepresence System for Collaborative Space Operation. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(4):814–827, Apr. 2017. doi: 10.1109/TCSVT.2016.2580425 3

[12] L. Gao, H. Bai, R. Lindeman, and M. Billinghurst. Static local environment capturing and sharing for MR remote collaboration. In *SIGGRAPH Asia 2017 Mobile Graphics & Interactive Applications on - SA '17*, pp. 1–6. ACM Press, Bangkok, Thailand, 2017. doi: 10.1145/3132787.3139204 1, 2, 3, 4

[13] K. Gupta, G. A. Lee, and M. Billinghurst. Do You See What I See? The Effect of Gaze Tracking on Task Space Remote Collaboration. *IEEE Transactions on Visualization and Computer Graphics*, 22(11):2413–2422, Nov. 2016. doi: 10.1109/TVCG.2016.2593778 3

[14] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, vol. 52, pp. 139–183. Elsevier, 1988. 6

[15] C.-C. Ho and K. F. MacDorman. Revisiting the uncanny valley theory: Developing and validating an alternative to the godspeed indices. *Computers in Human Behavior*, 26(6):1508–1518, 2010. 6, 9

[16] C.-C. Ho and K. F. MacDorman. Measuring the uncanny valley effect: Refinements to indices for perceived humanness, attractiveness, and eeriness. *International Journal of Social Robotics*, 9:129–139, 2017. 6, 9

[17] A. Jing, M. Frederick, M. Sewell, A. Karlson, B. Simpson, and M. Smith. How visualising emotions affects interpersonal trust and task collaboration in a shared virtual space. In *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 849–858. IEEE, 2023. 1, 3, 8

[18] A. Kemeny, J.-R. Chardonnet, and F. Colombet. Getting rid of cybersickness. *Virtual Reality, Augmented Reality, and Simulators*, 2020. 2

[19] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023. 2

[20] B. Keshavarz and H. Hecht. Validating an efficient method to quantify motion sickness. *Human factors*, 53(4):415–426, 2011. 6

[21] G. A. Lee, T. Teo, S. Kim, and M. Billinghurst. Mixed reality collaboration through sharing a live panorama. In *SIGGRAPH Asia 2017 Mobile Graphics & Interactive Applications*, pp. 1–4. 2017. 1, 2, 3

[22] G. A. Lee, T. Teo, S. Kim, and M. Billinghurst. A User Study on MR Remote Collaboration Using Live 360 Video. In *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 153–164. IEEE, Munich, Germany, Oct. 2018. doi: 10.1109/ISMAR.2018.00051 2, 3, 6, 8

[23] M. Lombard and M. T. Jones. Defining presence. *Immersed in media: Telepresence theory, measurement & technology*, pp. 13–34, 2015. 3

[24] S. Lukosch, M. Billinghurst, L. Alem, and K. Kiyokawa. Collaboration in augmented reality. *Computer Supported Cooperative Work (CSCW)*, 24:515–525, 2015. 5

[25] K. Masai, K. Kunze, M. Sugimoto, and M. Billinghurst. Empathy glasses. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 1257–1263, 2016. 3, 6, 8

[26] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[27] D. Niedermayr, J. Wolfartsberger, M. Borac, R. Brandl, M. Huber, and P. Josipovic. Analyzing the potential of remote collaboration in industrial mixed and virtual reality environments. In *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 66–73, 2022. doi: 10.1109/ISMAR-Adjunct57072.2022.00023 2, 3, 4, 5, 8, 9

[28] T. Piumsomboon, A. Day, B. Ens, Y. Lee, G. Lee, and M. Billinghurst. Exploring enhancements for remote mixed reality collaboration. In *SIGGRAPH Asia 2017 Mobile Graphics & Interactive Applications on - SA '17*, pp. 1–5. ACM Press, Bangkok, Thailand, 2017. doi: 10.1145/3132787.3139200 1, 3

[29] T. Piumsomboon, Y. Lee, G. Lee, and M. Billinghurst. CoVAR: a collaborative virtual and augmented reality system for remote collaboration. In *SIGGRAPH Asia 2017 Emerging Technologies*, pp. 1–2. ACM, Bangkok Thailand, Nov. 2017. doi: 10.1145/3132818.3132822 3

[30] V. Pooryousef, M. Cordeil, L. Besançon, R. Bassed, and T. Dwyer. Collaborative forensic autopsy documentation and supervised report generation using a hybrid mixed-reality environment and generative ai. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 9

[31] O. G. Pérez, B. Sayis, and N. P. Burguès. Analysis of interpersonal communication in a Mixed Reality full-body interaction experience to foster social initiation in children with Autism. 2020. 2, 6, 8

[32] T. Rhee, S. Thompson, D. Medeiros, R. dos Anjos, and A. Chalmers. Augmented Virtual Teleportation for High-Fidelity Telecollaboration. *IEEE Transactions on Visualization and Computer Graphics*, 26(5):1923–1933, May 2020. doi: 10.1109/TVCG.2020.2973065 1, 2, 3

[33] B. Schneider and R. Pea. Real-time mutual gaze perception enhances collaborative learning and collaboration quality. *Educational Media and Technology Yearbook: Volume 40*, pp. 99–125, 2017. 3

[34] A. Schäfer, G. Reis, and D. Stricker. A Survey on Synchronous Augmented, Virtual, andMixed Reality Remote Collaboration Systems. *ACM Computing Surveys*, 55(6):1–27, July 2023. doi: 10.1145/3533376 1

[35] M. Sereno, L. Besançon, and T. Isenberg. Point specification in collaborative visualization for 3d scalar fields using augmented reality. *Virtual Reality*, 26(4):1317–1334, 2022. 3, 5

[36] M. Sereno, X. Wang, L. Besançon, M. J. McGuffin, and T. Isenberg. Collaborative Work in Augmented Reality: A Survey. *IEEE Transactions on Visualization and Computer Graphics*, 28(6):2530–2549, June 2022. Conference Name: IEEE Transactions on Visualization and Computer Graphics. doi: 10.1109/TVCG.2020.3032761 9

[37] M. Tait and M. Billinghurst. The effect of view independence in a collaborative ar system. *Computer Supported Cooperative Work (CSCW)*, 24:563–589, 2015. 2, 3

[38] T. Teo, L. Lawrence, G. A. Lee, M. Billinghurst, and M. Adcock. Mixed Reality Remote Collaboration Combining 360 Video and 3D Reconstruction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–14. ACM, Glasgow Scotland Uk, May 2019. doi: 10.1145/3290605.3300431 2, 3

[39] T. Teo, M. Norman, G. A. Lee, M. Billinghurst, and M. Adcock. Exploring interaction techniques for 360 panoramas inside a 3d reconstructed scene for mixed reality remote collaboration. *Journal on Multimodal User Interfaces*, 14:373–385, 2020. 2, 3, 6, 7, 8, 9

[40] H. Tian, G. A. Lee, H. Bai, and M. Billinghurst. Using virtual replicas to improve mixed reality remote collaboration. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2785–2795, 2023. 2, 3, 4

[41] A. Vovk, F. Wild, W. Guest, and T. Kuula. Simulator sickness in augmented reality training using the microsoft hololens. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pp. 1–9, 2018. 2

[42] F. Zaman, C. Anslow, A. Chalmers, and T. Rhee. Mrmac: Mixed reality multi-user asymmetric collaboration. In *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 591–600. IEEE, 2023. 2, 3, 8

[43] J. Zillner, E. Mendez, and D. Wagner. Augmented Reality Remote Collaboration with Dense Reconstruction. In *2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 38–39, Oct. 2018. doi: 10.1109/ISMAR-Adjunct.2018.00028 1, 2, 3