Understanding Multi-View Collaboration between Augmented Reality and Remote Desktop Users

Arthur Fages, Cédric Fleury, Theophanis Tsandilas

Understanding Multi-View Collaboration between Augmented Reality and Remote Desktop Users

ARTHUR FAGES, Université Paris-Saclay, CNRS, Inria, LISN, France
CÉDRIC FLEURY, IMT Atlantique, Lab-STICC, UMR CNRS 6285, France
THEOPHANIS TSANDILAS, Université Paris-Saclay, CNRS, Inria, LISN, France

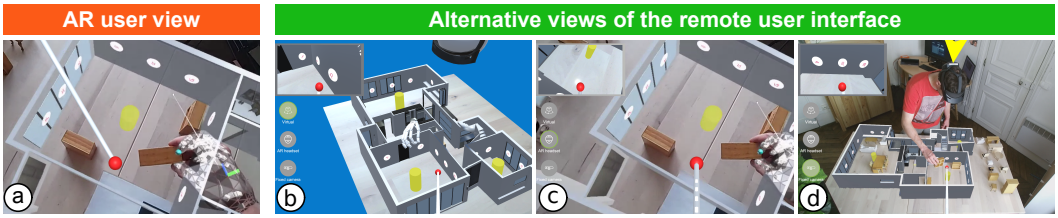

Fig. 1. An AR user and a desktop remote collaborator achieve a physical furniture arrangement task in a virtual 3D house model. (a) AR user view displayed in the headset and three alternative views of the remote collaborators available in the ARgus interface: (b) a fully virtual view, (c) a first-person view streamed from the headset, and (d) an external view streamed from a depth camera. Project page: https://argus-collab.github.io

Establishing an effective collaboration between augmented-reality (AR) and remote desktop users is a challenge because collaborators do not share a common physical space and equipment. Yet, such asymmetrical collaboration configurations are common today for many design tasks, due to the geographical distance of people or unusual circumstances such as a lockdown. We conducted a first study to investigate trade-offs of three remote representations of an AR workspace: a fully virtual representation, a first-person view, and an external view. Building on our findings, we designed ARgus, a multi-view video-mediated communication system that combines these representations through interactive tools for navigation, previewing, pointing, and annotation. We report on a second user study that observed how 12 participants used ARgus to provide remote instructions for an AR furniture arrangement task. Participants extensively used its view transition tools, while the system reduced their reliance on verbal instructions.

CCS Concepts: • **Human-centered computing** → **Computer supported cooperative work**; **Mixed / augmented reality**; **Collaborative interaction**.

Additional Key Words and Phrases: Augmented reality, remote collaboration, video-mediated communication.

ACM Reference Format:
Arthur Fages, Cédric Fleury, and Theophanis Tsandilas. 2022. Understanding Multi-View Collaboration between Augmented Reality and Remote Desktop Users. *Proc. ACM Human-Computer Interaction* 6, CSCW2, Article 549 (November 2022), 27 pages. https://doi.org/10.1145/3555607

1 INTRODUCTION

Augmented reality (AR) technologies radically change the way 3D design teams work together. AR users can move away from the screen of their computer to interact directly with the objects of a virtual scene and naturally navigate in their physical space. AR also strengthens collaboration by

Authors' addresses: Arthur Fages, Arthur.Fages@lisn.upsaclay.fr, Université Paris-Saclay, CNRS, Inria, LISN, Orsay, France; Cédric Fleury, Cedric.Fleury@lisn.upsaclay.fr, IMT Atlantique, Lab-STICC, UMR CNRS 6285, Brest, France; Theophanis Tsandilas, Theophanis.Tsandilas@inria.fr, Université Paris-Saclay, CNRS, Inria, LISN, Orsay, France.

adding virtual aids [39] while preserving traditional communication channels, such as voice, gaze and gestures. Previous work has investigated the use of AR for a diverse range of collaborative tasks, from interior design for couples [48] and science teaching [51] to industrial manufacturing [60]. Unfortunately, real-time collaboration is a challenge when users work remotely and, consequently, they do not share the same physical environment and do not all have access to AR equipment. Such situations have become commonplace during the still ongoing COVID19 pandemic [5]. Many design and research teams have found themselves to work remotely, relying on video-communication software to collaborate together [63]. Some experts predict that such situations are not temporary – they will largely persist after the pandemic [10]. HCI research thus needs to better understand how different remote workspace configurations support collaboration in these new contexts.

While screen sharing has been a valuable tool of collaboration for remote desktop users, sharing the workspace of a collaborator wearing an AR headset requires a new set of tools that considers both the physical and the virtual space of the AR user. In this direction, several AR technologies such as the Microsoft HoloLens enable AR users to video-stream their view. Yet, such views are not interactive and do not offer independent camera control to remote viewers. According to Tait and Billinghurst [52], increased view independence results in stronger collaboration performance. However, view independence requires that the physical environment of the AR user is reconstructed in real time, such that it can be smoothly integrated into the 3D virtual scene. Unfortunately, existing solutions for reconstructing independent AR views have serious limitations. For example, techniques based on multiple depth sensors [7, 9] require heavyweight instrumentation, consume large volumes of bandwidth, while the quality of their reconstructed models is still limited and largely unrealistic [25]. Other 3D reconstruction techniques [19, 36] pose significant constraints on the view possibilities of remote users.

The alternative approach that we investigate here is to offer remote users multiple view representations, where each provides a different aspect of the workspace of the local AR worker. We focus, in particular, on tasks that require access both to a virtual model and to its physical context, or to physical objects that interact with the virtual model. In this case, remote collaborators must make decisions about which representation to use to effectively complete the task. We study three complementary representations: (i) a first-person view as provided by the AR headset, (ii) an augmented third-person view as captured by a fixed camera with a depth sensor, and (iii) a fully virtual representation. The first two representations show the real-word scene but do not support view independence. The last representation, in contrast, supports full view independence but does not capture the real-world scene. However, by providing tools for switching between these representations, we expect that remote users will develop strategies that leverage their complementary roles. We frame our research questions as follows:

- **RQ_1: How do remote users perceive the trade-offs of the three representations when providing instructions to an AR worker?** Several past studies [18, 20, 51] have studied the trade-offs of first-person and third-person views, but as we discuss in this paper, their results are somehow contradictory and non-conclusive. Others [52] have studied fully independent views, but only ones that rely on the 3D reconstruction of the real scene.
- **RQ_2: If we offer remote users the possibility to switch between representations, how will they make use of them?** To explore answers to this question, we integrate the three representations into ARgus (see Fig. 1), a remote collaboration system. A key contribution of ARgus is on how its user interface merges representations through a collection of interactive tools for previewing, between- and within-view navigation, camera control, 3D pointing, and annotation.

We report on the results of two user studies, one for each question. The first study examines strengths and weaknesses of the three representations, focusing on the collaboration experience of remote users when communicating spatial instructions. The second user study investigates how 12 remote participants use ARgus to guide a local AR user to complete an AR furniture arrangement task. Our results provide a fresh perspective on the trade-offs of each representation. They also help us characterize participants' view-switching strategies, evaluate the perceived effectiveness and utility of ARgus, and understand whether and how it assists remote communication.

2 RELATED WORK

Our research builds upon a rich volume of previous HCI work on remote collaboration.

The role of viewpoint in video-mediated collaboration. When people do not share the same space, video is the most common communication medium. Its role is to bring a common ground of understanding (or *conversational grounding* [18]) and support *workspace awareness* [14] or a *"shared person space"* that includes *"facial expressions, voice, gaze and body language"* [12].

The HCI literature has long examined the role of different views in video-mediated communication, especially in the context of physical tasks that involve spatial object manipulation and construction. Back in the 90s, Kuzuoka [31]

investigates spatial workspace collaboration through SharedView, a video communication system. Kuzuoka's study requires a remote expert to explain a 3D task to a local worker in a machining center and shows that the viewpoint of the video can affect the efficiency of communication.

Gaver et al. [20] study the use of five camera views for a remote-collaboration design task.

Their task requires a participant in a local office to arrange the furniture in a dollhouse in collaboration with a remote partner.

Results show that participants largely preferred task-centered views than face-to-face communication. The authors also observe that view switching can be problematic. In particular, multiple views can interfere with establishing a common frame of reference, introduce discontinuities, and impede coordination.

Ten years later, Fussel et al. [18] compare two remote-view configurations: (i) a head-mounted camera with eye tracking, and (ii) a scene camera placed at the back of the worker, providing a wider but fixed view of the working environment. The scene camera is shown to be preferable and improve communication efficiency, while the head-camera view does not add any benefit compared to an audio-only condition. Similarly, adding a second, head-camera view to a scene-camera view seems to deteriorate rather than to improve collaboration performance. A more recent study [51] in the context AR video collaboration for 3D guidance tasks also shows that a third-person view results in better task performance and higher user satisfaction than a first-person view.

However, other studies show advantages in combining multiple alternative views. For example, Schafer and Bowman [45] study a virtual furniture arrangement task and observe that the availability of two alternative representations (virtual 3D and floor plan) *"enabled the users to investigate different aspects of the space."* Ranjan et al. [41] find that remote users complete complex lego-construction tasks faster with automatic pan-tilt-zoom camera than with a static camera. Giusti et al. [21] investigate how a local user and a remote expert configure a mobile phone and a tablet to repair a Lego model or replace a punctured bike tube. When both a phone and a tablet was available, local users tended to fix the tablet's camera view to show an overview of their workspace and sometimes their face, while they used the camera of their mobile phone when they needed to zoom in on specific parts to show details. Lanir et al. [32] investigate user performance and behavior with respect to who (the local vs. the remote user) has the camera control. Their conclusion is that the outcome depends on the situation and task at hand. Overall, results are far from conclusive but

seem to suggest that the most suitable strategy is to give users control over alternative views, each adapted to a different type of task. Our goal is to verify this hypothesis and investigate mechanisms that help users effectively control these views.

Finally, in the context of remote AR collaboration, Tait and Billinghurst [52] evaluate how varying degrees of view independence affect collaboration. They find that more independent views result in faster task-completion time, higher user confidence, and fewer verbal instructions. Unfortunately, the approach of Tait and Billinghurst [52] requires the physical environment of the local user to be reconstructed as a virtual 3D model. This model does not capture dynamic changes in the real environment, is limited in space and resolution, and does not include a natural representation of the local AR user, who is represented instead as a virtual view frustum. Furthermore, to detect the manipulation of physical objects and communicate it to remote users, the authors use a sophisticated optical tracking system and attach infrared markers to a small set of preregistered physical objects. Clearly, such configurations are extremely hard to set up and do not scale to real-world collaboration tasks. Next, we discuss the limitations of 3D reconstruction methods in more depth and present the state-of-the-art of view transition techniques.

Remote representation of an AR workspace. A key challenge for remote AR collaboration is how to communicate information about the physical space while enabling users to navigate in the scene and manipulate objects. A common solution is using virtual replicas of the physical objects. For example, Oda et al. [38] focus on remote collaboration between an expert wearing a VR headset and a local worker wearing an AR headset. Their system enables the expert to provide guidance by moving or annotating the 3D model of an existing physical part (virtual replica) in the worker's virtual space.

Unfortunately, virtual replicas provide partial only information about the physical environment of the local AR user. Feick et al. [17] combine, instead, two parallel views for a remote expert user: (i) a video feed showing the other user manipulating a physical object, and (ii) a 3D scene that allows the expert to gesture over a virtual proxy of the object. Kumaravel et al. [57] take this approach even further. They study two representations that communicate the virtual and physical workspace of a local user: (i) a 2D video stream and (ii) an *hologlyph*, a 3D representation of spatial data captured by depth cameras and rendered as a point cloud. In other mixed-reality systems, remote collaborators can switch between a 360° panorama video and a 3D reconstructed scene [56] or even navigate in a point-cloud representation of the remote workspace through multiple depth cameras that produce a real-time 3D reconstruction of the scene [9]. Other research explores techniques for communicating cues about the gaze of collaborating users [24].

Despite their technical sophistication, the above systems have serious limitations. First, they either support static 3D models or require remote users to have access to specialized and hard to set up equipment. Second, even the most compelling systems suffer form artifacts that limit the realism of the reconstructed workspace. For example, the system of Bai et al. [9] (one of the very few to support real-time scene reconstruction) can only display low-resolution 3D panoramas and simplistic avatar representations of the local user. But as Jones et al. [25] observe, the reduced quality of a full 3D reconstruction can distort collaborators' expressiveness and make them experience an *"uncanny valley of XR [extended reality] telepresence."* The authors also report that *"the more immersive an XR Telepresence system is, the more amplified technical issues such as latency, video quality, and control become"* [25].

Other very active research in AR mobile collaboration [19, 36, 49] has introduced techniques that enable remote users to interact with a reconstructed 3D representation of the remote workspace. These techniques have similar limitations. Based on KinectFusion [23], BeThere [49] requires the local user to pre-capture the 3D geometry of the workspace with a mobile depth camera and the

remote user to use a device with a depth sensor to interact with it. SLAM systems [19] provide a limited range of 3D navigation that is constrained by the image viewpoints seen by the camera of the local user. Finally, systems based on light fields [36] lack depth information, making occlusion management problematic.

Since we do not expect the above problems to be solved any time soon, we limit our scope to techniques of augmented video-mediated communication, as those require more lightweight setups, consume less bandwidth, and do not suffer from 3D reconstruction problems. Furthermore, as we study tasks that involve both virtual and physical objects, we are also interested in how streamed video can be coupled with fully virtual representations that afford free navigation.

View transition techniques. Purely virtual environments offer considerable freedom for remote collaboration through arbitrary virtual cameras and views. For example, Photoportals [30] and Spacetime [62] provide a range of imaginative techniques for viewpoint control in VR. In contrast, AR collaboration is largely constrained by the position and coordination of physical cameras in the environment of the local user. Previous work has tried to deal with this problem in different ways. Rasmussen and Huang [42] show previews from multiple cameras to remote users who can then switch between them. Sukan et al [50] enable mobile AR users to quickly switch between snapshots of their past views. Komiyama et al. [28] provide techniques for a smooth transition among the views of multiple physical cameras. Finally, Tatzgern et al [55] study how to seamlessly transition between AR and VR views. Our system design draws inspiration from all this line of work.

3 DESIGN PROBLEM

We are interested in asymmetric collaboration setups that involve a *local user* with an AR headset (e.g., a Microsoft HoloLens) and *remote collaborators* who participate from distance through a desktop application. In contrast to approaches that require users at both ends to wear an AR or a VR headset [9, 58], such setups are relatively lightweight and easy to employ, as they only require the local user to have access to AR equipment. These setups thus offer high flexibility to the *remote collaborators*, allowing them to work in many different situations, such as while traveling or in a crowded open office where physical space is limited.

Video has become the most common medium of remote collaboration and has taken a dominant role during the ongoing COVID19 pandemic [63]. Our goal is not to replace video communication but to enhance it with new visual and interaction modalities that leverage the benefits of AR systems. A major challenge is how to deal with the asymmetry in the views of remote collaborators, in particular how to enable them to easily navigate in the 3D environment of the AR user, inspect the virtual content, and provide directions that require spatial orientation and awareness.

As we already discussed, we also dismiss solutions that require the reconstruction of the physical workspace [9, 26, 52, 56], either because they cannot keep track of dynamic changes in the environment of the local user, or because they provide a largely unrealistic representation of the scene and the local user, break the collaborators' experience due to the *"uncanny valley of XR telepresence"* [25], and amplify network outage problems [8].

We restrict our design space to lightweight configurations that use a single external depth camera in addition to the camera of the AR user's headset. This external camera could be replaced by a webcam or a smartphone since more and more devices are now equipped with a depth sensor. We may even rely on standard webcams or smartphones in the near future, as a single monocular camera can be sufficient to provide depth data [34].

Focusing on the views of the *remote collaborators*, we investigate three design dimensions:

Workspace representation. It refers to the representation used by the system to help collaborators perceive each other and their shared workspace. This representation may consist

Fig. 2. Remote-view configurations tested by our first study: Headset View (left), External View (middle) and Virtual View (right). A remote participant gives oral instructions to the AR user on how to position 3D shapes on a virtual support.

of a virtual 3D scene, video, or alternatively a combination of these two. Ideally, it should provide spatial information about both virtual and physical objects in the workspace but also information about the actual AR user, such as her or his body position and gestures.

Scene viewpoint. It determines which virtual and physical objects are visible at a given moment or during the whole collaborative task and from which perspective. Previous literature often makes a distinction between a *first-person* and a *third-person* perspective (e.g., see Komiyama et al. [28]). The former refers to the perspective of the AR user. It can be captured by a head-mounted and communicated to remote collaborators. The latter refers to an out-of-body perspective as captured by external cameras.

View independence. A key problem is how to enable remote users to independently navigate in the 3D space of the AR user to obtain a convenient view, e.g., a view that helps them inspect details of the virtual model or avoids occlusions, and point to a position in space, e.g., to indicate a physical or virtual object to the local user.

Additional dimensions, such as display configuration and means of communication, can emerge from this design space. We chose dimensions that focus on the collaboration process itself rather than ones that deal with how collaboration is made possible, since many tools have already been presented for this purpose [20, 51, 56, 61]. To simplify our user studies, we also decided to focus on one-to-one collaboration. We defer the study of the more general case where multiple remote collaborators participate to our future work.

4 USER STUDY 1

We conducted a user study to investigate our first research question (RQ_1). The study examines trade-offs of different workspace representations and scene viewpoints. In particular, it observes how users provide remote instructions under three configurations:

Headset View is an augmented video from a first-person viewpoint. We capture the video directly from the AR headset to simulate the situation where the remote user sees the scene "through the eyes" of the local user (see Fig. 2-left). The video feed integrates the virtual 3D content into the physical scene of the AR user. The key strength of this configuration is that collaborators share a common frame of reference. So they do not need to mentally rotate the 3D space [47] to communicate.

External View is an augmented video from an external (third-person) viewpoint. We use a depth camera (Microsoft Kinect V2) to provide an overview of the full workspace of the local user. A key question is how to optimally position the camera. In previous studies [18, 51], in which the local worker remains seated, the external camera is positioned at the back left (or right) side of the worker. This way, the two collaborators view the scene from a similar perspective. Unfortunately, in such configurations, the face, hands, and other key parts of the worker's body may not be visible. Furthermore, if the worker freely moves around the

model of interest, his or her body may occlude parts of the workspace. For these reasons, we excluded this alternative. For optimal visibility, the camera is positioned in front of the AR user (at 2m height and approximately 2.5m away) and oriented 30° downwards. We also ensure that the board on which the local user places objects is centered in the recorded image. The video feed of this camera is augmented with the virtual 3D content visible in the AR user's workspace (see Fig. 2-middle). Compared to first-person views, external views have been shown to increase communication efficiency [18] and improve performance and satisfaction [51]. Other authors observe that users strongly prefer them for *"placing objects recommended by themselves"* [11].

Virtual View is a fully virtual representation with a free viewpoint. Remote collaborators see a virtual representation of the 3D scene. A simplified avatar shows the head and hands of the AR user (see Fig. 2-right). Remote users can freely navigate in the 3D scene and choose their preferred viewpoint. This approach follows the naive metaphor of birds that can choose the most convenient position to observe the AR user. Previous results [52] suggest that this additional freedom in the choice of views can improve both the performance and the confidence of remote collaborators.

The study took place during the COVID19 pandemic. To eliminate risks of contamination, the experimenter (first author) acted as the local user wearing the AR headset for all sessions of the study. Participants acted as remote collaborators and completed the study tasks from their home or office environment. The experimental protocols of our studies were approved by a local ethical committee.

4.1 Participants

24 volunteers (11 women and 13 men) participated. They were 21 to 41 years old (Median = 26.5 years). All were frequent or occasional users of at least one video-communication tool, such as Skype or Zoom. Seven participants frequently or occasionally used an AR or a VR headset. 11 participants were frequent or occasional users of 3D games, game engines, or 3D modeling environments. Participants were recruited by word of mouth and responses to a recruitment email sent to our lab's mailing lists. No compensation was given.

4.2 Apparatus

The experimenter set up the workspace in his home environment and interacted with the scene through a Microsoft HoloLens 2. For the calibration, the experimenter defined the HoloLens origin by manually positioning a 3D object on an AprilTag [59] marker. The Kinect camera was automatically calibrated by detecting this marker using the ViSP library [35]. Communication between the participants and the experimenter was established through commercial video-communication software (Skype or Discord). The Headset View and External View were presented to participants through screen sharing. For the Headset View, we used the Microsoft HoloLens 2 video-sharing application [2] to stream live video from the headset. For the External View, our implementation considered potential occlusions between virtual and real objects as "seen" by the Kinect camera. For each pixel, a shader chose to display either the streamed video or the virtual object by respecting their depth information from the camera. For the Virtual View, participants downloaded and executed a client application, which rendered an interactive 3D scene synchronized with the HoloLens application via a remote server. This architecture is implemented with Unity 2019.4 and used the Unet library [6] for network communication. Participants could pan, zoom, and rotate the 3D scene using their mouse and keyboard. Finally, we used a website to guide participants in

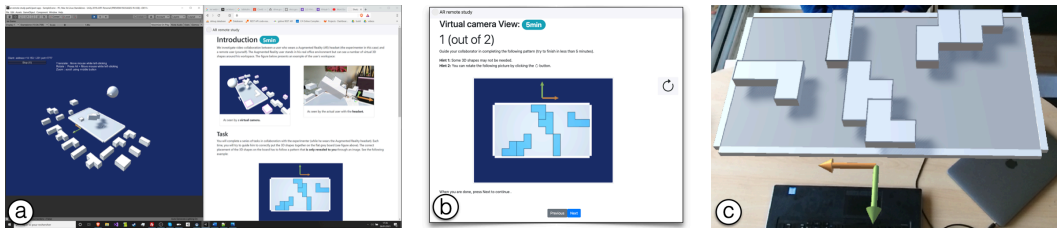

Fig. 3. (a) Remote participant interface used for our first study: tested view configuration on the left (Virtual View in this example) and website used to give instructions on the right. (b) Close-up of the website showing the target pattern: the UI widget on the right allows participants to rotate the pattern image. (c) Zoom-in on the AR user workspace showing the virtual board with the finalized task: colored axes help participants make the correspondence between the pattern on the image and the virtual board shown on the view.

the course of the experiment (see Fig. 3-a). This website provided information and instructions regarding the configurations and the task and linked to our online questionnaires.

4.3 Task

Participants were asked to place 3D pieces of nine different shapes on a virtual board by giving oral instructions to the experimenter who acted as a surrogate (see Fig. 2). The experimenter used close and distant manipulation tools provided by the Microsoft HoloLens 2: its direct manipulation gestures, its hand-ray tool and air-tapping for selection.

The solution to the task was a 2D top-view pattern that described how to position pieces in any order. The pattern was randomly generated to contain eight pieces out of 18 pieces available in the workspace. It was presented to participants as an image on the website and was unknown to the experimenter (see Fig. 3-b). Its default orientation shown to participants reflected the experimenter's perspective. The pattern was thus inverted with respect to the External View. To help participants adapt the orientation of the 2D pattern as they would do with a piece of paper, we included UI widgets for rotating the pattern. We also added colored axes both in the views and pattern images to make correspondence clear. The virtual board was composed of a 9×5 grid of squares with side length 10 cm (see Fig. 3-c). When a 3D piece was placed on the board, it was snapped to the grid. Pieces had a maximum length of 30 cm (i.e., three grid squares).

As Kuhlen and Brennan [29] discuss, using a confederate in studies that involve conversations between humans is a common research method, but its practice *"might be hazardous"* to collected data. In particular, if confederates have an active, uncontrolled participation in the dialog and are aware of the hypotheses of the study, they can bias the results. To reduce the risk of bias, we established a minimalistic communication protocol for the experimenter. The experimenter followed the participant's instructions and only verbally intervened: (i) to ask the participant to repeat an instruction if the instruction was not understood; (ii) to request confirmation for a planned action; and (iii) to request confirmation for a completed action. The experimenter could also answer questions concerning the user interface or the task, but we tried to respond to such questions as much as possible during training. In contrast, the task required participants to take the initiative as speakers, as Kuhlen and Brennan [29] also recommend.

4.4 Design

To keep sessions short, we simplified the experimental design by dividing the user study into two independent parts. Focusing on the viewpoint (first-person vs. third-person) of AR video, Part I compared the Headset View with the External View. Focusing on the workspace representation (virtual vs. AR) and the type of navigation control (remote user vs. local user control), Part II

compared the Headset View with the Virtual View. We divided our participants into two groups of 12 participants, one for each part, trying to balance gender. We followed a within-participant design, where all 12 participants tested both configurations. Half of them were first exposed to the Headset View, and the other half starts with the second condition. For each configuration, participants completed two main tasks, preceded by a training task with a simplified pattern with three only pieces.

4.5 Procedure

After signing a consent form, participants completed an online demographic questionnaire. Participants went through a short tutorial that explained the two communication configurations. They were then introduced to the training and two main tasks of each configuration. Participants evaluated the configurations and the task through a set of questions divided into multiple short questionnaires. Each participant answered seven questionnaires in total: one after each task (2 tasks $\times$ 2 configurations), one after each configuration (2 configurations), and one after the full session. The full procedure lasted approximately 50-70 minutes.

4.6 Data Collection and Measures

We collected: (i) participants' answers to the online questionnaires, (ii) recordings of the participants' voice during the tasks, and (iii) logs of low-level software events (view positions, trajectories, and time stamps). As we discussed above, the presence and collaboration role of the experimenter adds bias in the way tasks are completed. As a result, task performance measures such as task-completion time and errors are not reliable, and we do not consider them here. We focus instead on how participants perceived difficulty for different components of the task. We also report on the participants' preferences and their feedback about trade-offs of the compared conditions. Finally, we examine the strategies that participants followed to complete the tasks. Consider that our analyses are exploratory and should be interpreted as such.

4.7 Results

We present our main results. Anonymized data from this study and the R code of our analyses are available as supplementary material at https://osf.io/3nqrg.

Perceived task difficulty. Participants rated the difficulty for each sub-task through 5-point Likert items (1 = very difficult, 5 = very easy). We miss the answers of one participant for these questions in Part I. The analysis of ordinal data with metric models is generally problematic [33]. We therefore use state-of-the-art *cumulative probit* regression models [13, 33] that enable us to map ordinal scales to a latent (i.e., not observable) continuous variable and then express estimates of differences between conditions as standardized effect sizes. For an extensive justification of this method and a comprehensive tutorial, we refer the interested reader to Bürkner and Vuorre [13]. The method is based on a Bayesian statistics [27] framework, but we emphasize that we do not use informative priors here. Figure 4 presents the results of our analysis, where we compare the perceived difficulty of our configurations through estimates of mean standardized differences expressed as 95% *credible intervals*[1]. Those are differences over a continuous (rather than ordinal) physiological variable of difficulty and are expressed in standard deviation (SD) units. In contrast to common non-parametric significance tests that rely on rank transformations, the approach enables us to estimate the magnitude of the observed effects by means of probabilistic interval estimates and effect sizes and thus better evaluate the statistical evidence about these effects.

[1]A credible interval is the Bayesian analog of a confidence interval. Unlike a 95% confidence interval, which is often misinterpreted, a 95% credible interval expresses a range in which the parameter of interest lies with 95% probability [27].

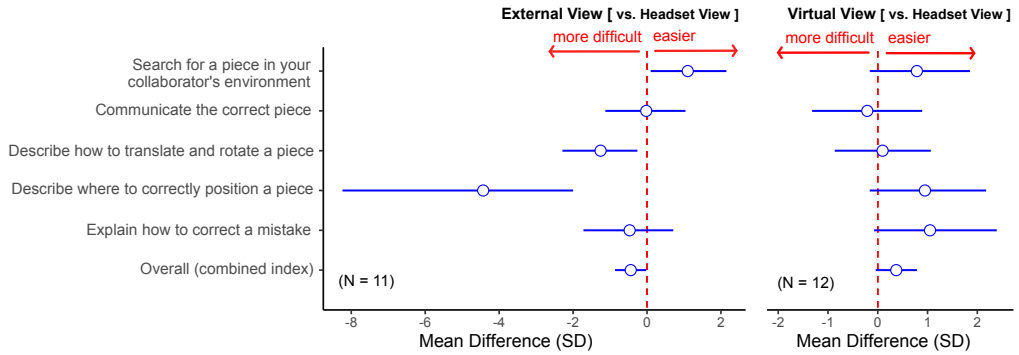

Fig. 4. Comparing the perceived difficulty of different subtasks between configurations. For our analysis, we use Bayesian ordinal (cumulative probit) models [13], which map the original ordinal scale of Likert items to a *latent* continuous variable. The bars in the graph represent 95% *credible intervals* of mean differences over this continuous variable and can be treated as estimates of standardized effect sizes. Note that the unit of these differences is the standard deviation (SD) of the distribution of the latent variable.

The results indicate that participants perceived that the External View was easier than the Headset View for searching pieces in their collaborator's environment. In contrast, the External View was more difficult for describing how to translate or rotate a piece and how to correctly position a piece. This latter effect is especially pronounced. When exposed to the External View, several participants struggled to correctly map their image of the pattern to the workspace of the AR user. Because of the position of the external camera, the participants had to mentally perform a rotation transformation to give the correct instructions. We further discuss this problem below. For the other subtasks (communicate the correct piece and explain how to correct a mistake), we do not observe any clear difference between the two configurations.

Differences between the Virtual View and the Headset View are more uncertain. There is a trend that the Virtual View was perceived as easier for searching pieces in their collaborator's environment, for describing how to correctly position a piece, and for explaining how to correct a mistake. However, the low size of the sample does not let us draw clear conclusions.

Preferences. We also asked participants to compare the configurations that they tested on six different aspects of the collaboration task. Figure 5 summarizes our results. We observe that participants see different benefits in each configuration. They appreciated the ability of the External View to provide awareness about the remote environment and help them search and locate pieces effectively. However, most participants expressed an overall preference for the Headset View, as it helped them perceive their collaborator's actions, facilitated communication, and helped them complete the task more effectively. The Virtual View, in turn, was especially appreciated for helping participants search and locate pieces effectively but also complete the task more effectively than the Headset View. Overall preferences between the Virtual View and the Headset View were equally split.

Trade-offs. Open-ended questions in the questionnaires asked participants to elaborate on the strengths and weakness of each configuration. All 12 participants of Part I reported that providing a global view of the workspace was the main strength of the External View. *"The strongest aspect was being able to see the overview of the scene and the entire puzzle we are building as a whole"* (P1).

"The fixed camera implies that all the items always stay in view of the distance person, easier if the collaborator cooperates less" (P11). As a comparison, in the Headset View *"the environment is reduced, and it takes more time to find your way around and locate all the items"* (P3). *"I do not have an autonomy of my vision angle, I only see what he sees" (P5).*

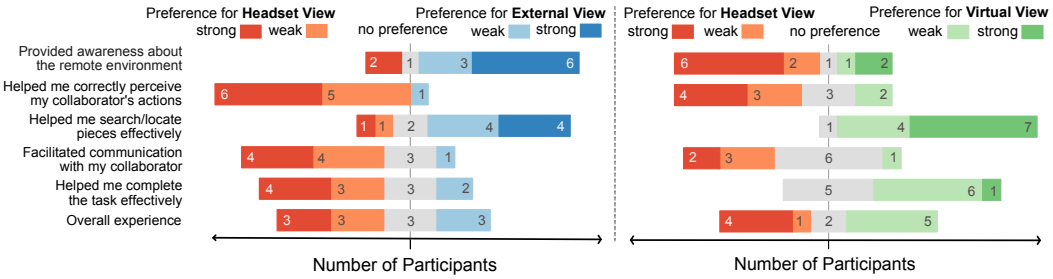

Fig. 5. Distribution of participants' preferences: Headset View vs. External View (left) and Headset View vs. Virtual View (right).

However, most participants evaluated this very same property of the Headset View as its strongest aspect: *"giving directions is much easier because I can just tell the partner to what I am doing!"* (P5). According to P12, *"you see through the eyes of [your partner], so you could exactly guide his gestures like a puppet."* In contrast, eight participants explicitly mentioned the inversion of left and right as a major problem of the External View: *"You are located on the opposite side so everything is going to be the reverse to explain."* (P12). Even though we allowed users to rotate the reference image with the solution pattern (see Fig. 3-b), only half of them used this function, and even this strategy did not seem to solve the problem for them.

As additional limitations of the External View, participants complained about distance distortions (P10), difficulties in correctly perceiving depth (P1), a sense of *"distantiation"* (P12), and a weaker sense of participation (P3).

The responses of the participants of Part II focused on the same qualities and drawbacks of the Headset View but raised additional concerns that the camera can be *"shaky"* (P19) and can *"induce motion sickness"* (P13). Concerning the Virtual View, participants especially appreciated its navigation capabilities: *"The user may navigate independently of the operator, make it possible to change point of view, or see things out of the operator's sight"* (P13); *"you are totally autonomous on the vision of the environment"* (P21). However, participants also identified several weaknesses: *"I do not really know where my collaborator is looking at"* (P15); *"lack of information about the real environment of the other user"* (P18); *"less points of reference than the previous configuration"* (P22); *"users need to be used to 3D applications in order to place [their] view correctly"* (P20).

Communication strategies. All participants frequently referred to their partner's "left" and "right" to communicate orientation. A common approach for indicating specific objects was to verbally describe their shape, e.g., by means of a letter of a similar shape ("Z", short "L", long "L", etc). A small number of participants (four in total) responded that they sometimes or frequently made use of physical objects in the experimenter's space as reference for the two AR views. To provide directions about how to rotate objects, strategies were more diverse. Several participants described the angle (90 or 180 degrees) of the rotation and its direction (clockwise/anticlockwise or left/right), while two participants acknowledged difficulties in finding an efficient strategy. For translations, most participants used the edges and corners of the virtual table for reference, but for higher precision, they also referred to the borders of other pieces on the table. In the Virtual View, participants' dominant approach was to place the virtual camera above the head of the avatar of their partner to obtain a similar viewpoint. According to our logs, four participants moved around the board to discover a better viewpoint but also ended up placing the camera at this position.

User feedback. Two participants proposed to place the camera of the External View slightly behind (P5) or above the head (P12) of the AR user, while P2 proposed to approach the camera closer

to the table. P14 and P21, instead, wondered about the possibility to increase the field of view of the Headset View, e.g., by adding extra cameras, while three participants (P12, P13, P15) proposed to combine multiple views together. Finally, several participants made suggestions about pointing techniques: a cursor for *"indicating locations"* (P17), a laser to *"target specific parts"* (P18), clicking with the mouse to *"illuminate a piece"* (P22) or to *"ping"* at a certain position as the Headset View moves (P16), and *"add a vocabulary to easier describe pieces"* (P9).

4.8 Discussion

The task required the AR user to manipulate virtual only objects. This choice was made to ensure that participants could complete the task under all three configurations. Clearly, it overrates the utility of the Virtual View, which lacks support for physical objects. Furthermore, we notice that some participants expressed strong preference for the Headset View over the External View. The External View was also rated as more difficult for certain subtasks. This finding is somehow at odds with results of past studies [18, 51], suggesting that the specificities of the task and the camera viewpoint may have an important influence to the success of a representation. In particular, in the external views that those two studies compared, the camera was conveniently located to the back left of the worker. As Shepard and Metzler [47] have shown, the time needed to perform a *mental rotation* in 3D space linearly increases with the angular offset of a viewer's viewpoint. This mental-rotation model predicts longer reaction times for our 180° camera configuration and implies a greater mental effort. An 180° offset also requires collaborators to reverse their wording, e.g., to replace every egocentric "right" with a "left" [46].[2]

Despite the above shortcomings, the External View presents several benefits over the Headset View. First, the view provided global awareness about the remote environment. Second, most participants felt that it helped them search for and locate pieces with less effort (see Fig. 4-5). The External View is also the only configuration that allows remote users to see the face and real full body of their collaborators. Although the role of such information was not directly evaluated with our task, it can be essential for supporting empathy [53] between participants and establishing communication awareness [14].

5 *AR*GUS: A MULTI-VIEW COLLABORATION SYSTEM

The results of our first study show that each view configuration has unique qualities that are difficult to substitute by the other two. The External View supports global awareness about the physical environment of the local worker and helps the remote user search for objects that are spread around the workspace. The Virtual View supports independent navigation, helping the remote user to provide instructions (e.g., about how to correct mistakes) from a convenient but also stable point of view. Finally, the Headset View is especially effective for perceiving the actions of the AR user and communicating egocentric instructions. Our research efforts thus focus on how to combine them and how to give remote desktop collaborators direct control over their use. To this end, we developed ARgus, a multiview collaboration system for 3D modeling (see Fig. 1). ARgus's implementation reflects three design goals:

DG$_1$. Communicate both real and virtual representations but without requiring the 3D reconstruction of the local workspace. We rely instead on video for capturing the physical environment of the local AR user and his or her real body. As we discussed in previous sections, this approach avoids problems associated with the 3D reconstruction of a physical workspace.

[2] A mirror configuration would transfer the problem to rotational directions, e.g., a "clockwise" direction should become "anticlockwise." Given their complexity, we suspect that the mental effort of such transformations would be even greater.

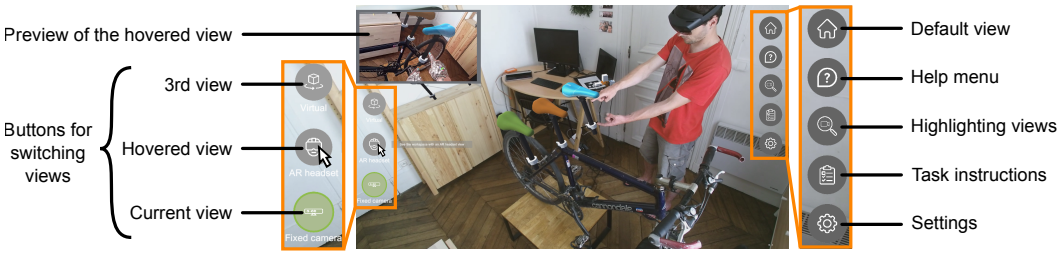

Fig. 6. Desktop interface of ARgus used by a remote collaborator for the redesign of a bicycle saddle.

DG$_2$. Support both first-person and third-person views of varying levels of view independence. This goal is consistent with the results of our formative study and recommendations of several older studies [20, 26, 40, 43, 52]. A challenge for ARgus was how to design effective and consistent mechanisms for switching and navigating between and within views.

DG$_3$. Provide tools that minimize communication effort and facilitate coordination. According to Schober [46], speakers try to minimize the mental effort of their addresses and their own by replacing *speaker-centered descriptions* (e.g., at "my left" or "your right") by *neutral descriptions*. ARgus provides aids for neutral descriptions via direct-pointing and spatial-annotation tools.

Below, we present the main features of ARgus. Although the system supports bidirectional communication, we focus in this paper on its design for remote desktop collaborators.

5.1 Combining Multiple Views

ARgus receives the augmented video streams from both an AR headset and an external depth camera located in the AR user's physical space. Furthermore, it maintains a synchronized version of the virtual 3D scene and can generate virtual views from any workspace location. Remote users can seamlessly switch between virtual and augmented video representations, as well as freely navigate to any viewpoint on the 3D scene.

ARgus also offers the possibility to display live previews of all three views (Headset View, External View, and Virtual View). These previews are video thumbnails of alternative views displayed in a small embedded window on top of the user's current view. They allow users to take a quick look at a different view, e.g., to inspect details of the physical environment that are not visible in the current view or to decide whether it is worth switching views. This mechanism aims to prevent the short bursts of switching between views observed by Gaver et al. [20] and facilitates coordination when users ask their collaborator to temporarily switch to their viewpoint to approve the veracity of their discovery [40].

5.2 Supporting Navigation

We provide several solutions for displaying previews, switching between views, and navigating in the 3D scene.

Main user interface. The main window of ARgus' user interface displays three circular buttons for selecting views and getting feedback about the active view (see Fig. 6). When users hover over a button, a live video preview is displayed on the top-left corner of the window. Clicking on the button activates the view. We use a trajectory and field-of-view interpolation based on Cinemachine [1] to animate the virtual camera in the 3D scene.

This solution ensures visual consistency among views, helps users understand the location of distant viewpoints, and avoids disorientation. We also use a blur effect to smooth out transitions

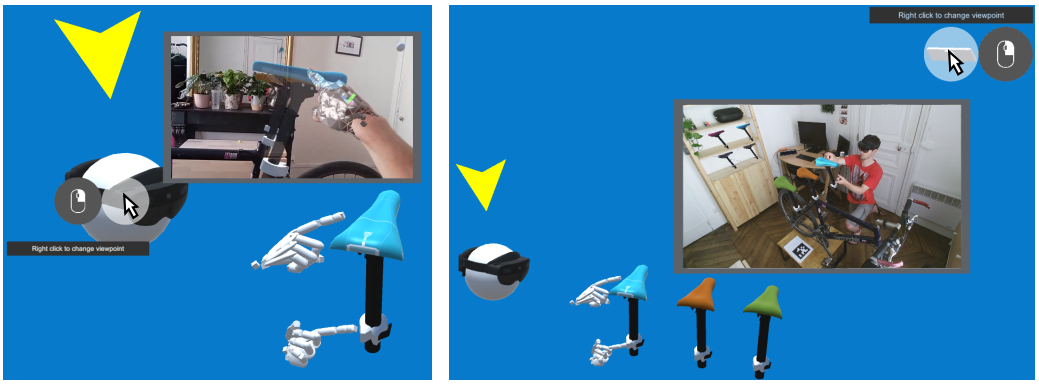

Fig. 7. The remote user hovers the mouse over the headset of the 3D avatar (left) and Kinect 3D model (right) to display the preview of the Headset View and the External View respectively.

between augmented video and virtual representations. We let users customize the duration of view transitions.

Interacting with the 3D scene. The 3D scene of ARgus' Virtual View serves as the basis for 3D navigation. It also offers an alternative solution for switching between views through interactive virtual camera representations. In the Virtual View, users can use the mouse to rotate their viewpoint around the center of the 3D scene and translate it (pressing alt). The same navigation capabilities are available in the two augmented-video representations, the Headset View and the External View. However, since remote users do not have direct control of the position of the two physical cameras (i.e., the external and the head-mounted camera), navigation actions within these views immediately cause the view representation to turn to virtual. This design approach ensures that interaction is consistent across all views.

The 3D scene includes virtual representations of the physical cameras themselves. Users can interact with them to preview or activate their corresponding views. For example, Fig. 7 shows the active Virtual View of a desktop user who remotely collaborates for the redesign of a bicycle saddle. The virtual view does provides no information about the real scene. Therefore, the remote user hovers the mouse over the headset of the 3D avatar to better understand what her partner sees (Fig. 7-left). She then hovers over the model of the Kinect camera (Fig. 7-right) to compare how the three saddle designs look together with the bicycle's real frame. Users may also decide to click the mouse to switch to this view. Finally, the 3D scene includes guides (arrows and highlighting effects) that help users locate the cameras and orient themselves in the 3D space.

Navigating with spherical views. Using base 3D rotation and translation interactions to closely inspect specific parts of a 3D model can be tedious and time consuming. To facilitate such tasks, we adapt *Navidget* interaction technique [22] and integrate it into ARgus' user interface as a Spherical View tool. Activated with a mouse right-click within either the External View or the Virtual View, the tool visualizes a sphere centered on the selected point. Users can move a virtual camera on the surface of the sphere, and a camera preview is shown (see Fig. 8-left). The sphere radius can be adjusted with the mouse wheel, causing the virtual camera to zoom in or out. Users can release the mouse to switch to a desired view or press esc to keep the current viewpoint.

Viewpoint recording. Following the approach of Sukan et al. [50], we allow users to record viewpoint locations (pressing a key) when they spot interesting views that they want to later reuse. Viewpoint recordings are represented as virtual cameras. As all other cameras (see above), they

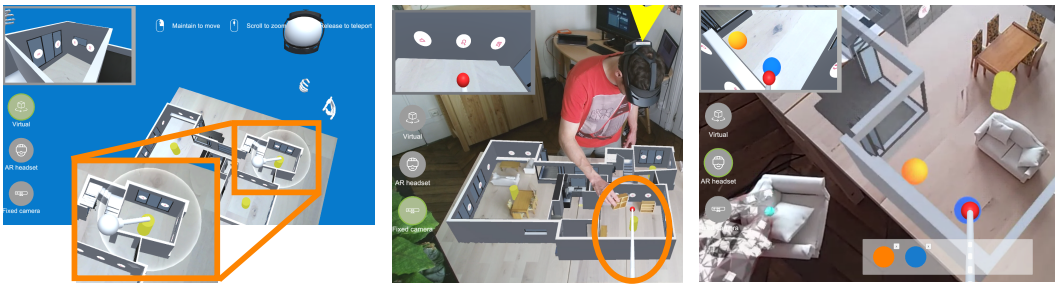

Fig. 8. Tools available in ARgus: Spherical View (left), Virtual Stick (middle) and annotations (right).

have a visual representation in the 3D scene, and users can interact with them to preview or switch to their views.

5.3 Facilitating Communication

Other tools in ARgus focus on how to facilitate the communication of users (DG_3).

AR user representation. The Virtual View includes a synchronized representation of the AR user with a simplified avatar composed of a sphere wearing the 3D model of a Microsoft HoloLens 2 and virtual hands (see Fig. 7). Each hand is represented by 24 joints, connected by canonical shapes, such as cylinders and squares. Both hands and head positions are retrieved from the MRTK libraries [4]. In the External View, a vertical arrow on top of the real head of the AR user communicates an interaction point for previewing and selecting the Headset View.

Pointing stick. As several participants of our formative study proposed, it is often useful to directly point in the remote scene, e.g., to indicate an object or provide instructions about where to place it. ARgus provides such functionality through a Virtual Stick (see Fig. 8-middle). The stick starts from the viewpoint's origin. Its direction is controlled with the mouse, while its length can be adjusted with the mouse wheel. A small sphere represents its tip, which is red if colliding with a 3D element and grey otherwise. A dotted line indicates its pointing direction, starting from its tip and projected until its collision with a 3D model in the scene. We considered results by Brown et al. [11], who report that users express a strong preference for surface-constrained pointing under all circumstances. A virtual camera is attached to the tip of the stick, and a preview of this camera is displayed on the top-left corner of the main window, helping users perceive depth and understand where the Virtual Stick is pointing at. In the Headset View, the view is frozen from the time users activate the Virtual Stick until they stop using it. Like in TransceiVR [58], freezing the moving view allows users to focus on an interesting viewpoint and achieve more accurate pointing.

Annotations. Overlaying information in an AR workspace in a spatially meaningful way can improve human performance and decrease mental workload [54]. Likewise, using shared virtual landmark increase user experience and facilitate spatial referencing in collaboration [37, 46]. In all views of ARgus, remote users can use the Virtual Stick to add annotations represented as colored spheres. In Fig. 8-right, for example, the remote user has added a yellow and a blue annotation to suggest target locations for placing furniture. The user interface shows a list of all activate annotations (up to five in our evaluation study), allowing users to quickly review and remove them.

5.4 Architecture and Implementation

ARgus was developed in Unity 2019.4. Its architecture relies on a client-server model connecting a remote desktop user and a local AR headset to a local server (see Fig. 9). The server keeps a

[3]The figure includes icons made by Freepik and Good Ware from www.flaticon.com.

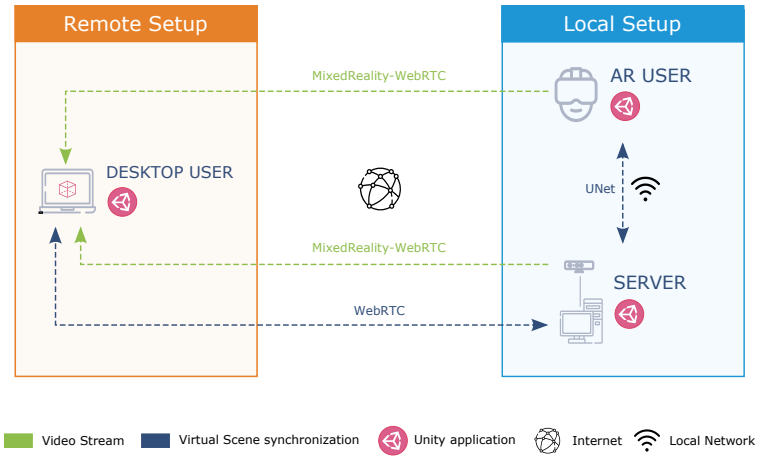

Fig. 9. System's architecture and implementation.[3]

synchronized version of the 3D scene and records the AR user's physical workspace with the external depth camera. It generates the External View by augmenting the camera video feed with the objects of the 3D scene. Occlusions between the virtual objects and the real objects are managed through the depth map of the external camera: for each pixel, a shader displays either the streamed video or the virtual object according to their depth information.

The AR headset is connected to the server as a client using the Unet library. It maintains a synchronized version of the 3D scene, which is used both to render AR user's augmented view in the headset and to generate the video feed of the Headset View. It also transmits the AR user's head and hands positions and orientations. To calibrate the AR headset reference frame and the depth camera reference frame, the virtual space origin is defined (i) manually by the AR user who needs to position a 3D object on an AprilTag [59] marker and (ii) automatically by the depth camera, which detects this marker using the ViSP library [35].

The application of the remote user is also connected to the server as a client using WebRTC. We built a custom protocol based on this technology to synchronize 3D object states (position and rotation) and software events (tools, logs, etc.). The application can thus render a synchronized version of the 3D scene to create the Virtual View. In addition, it receives the video feeds from the AR headset and the server based on the Mixed-Reality WebRTC libraries [3] to display the External View and the Headset View.

6 USER STUDY 2

We conducted a second user study that investigates our second research question (RQ_2). The study examines how remote collaborators use ARgus to provide instructions to a local AR designer.

As for our first study, we opted for an experimental design that avoids contamination risks due to the COVID19 pandemic. The experimenter (first author) acted as the local user wearing the AR headset, while participants acted as remote collaborators and completed the tasks from their home or office. A preregistration [16] of the study is available at https://osf.io/6dhzn.

6.1 Participants

12 volunteers (4 women and 8 men) participated in the study with an age ranging from 24 to 29 years old (Median = 27.5 years). All were frequent or occasional users of at least one video-communication

application. Two participants frequently or occasionally used an AR or a VR headset, while five participants had no previous experience with AR/VR technologies. Eight participants were frequent or occasional users of 3D games, game engines, or 3D modeling environments. Before starting the tasks, we verified that all participants had a stable internet connection (we replaced four initial participants who could not continue due to connection problems). We followed the same recruitment process as for our first study.

6.2 Apparatus and Conditions

As for the first study, the experimenter interacted with a Microsoft HoloLens 2 in a workspace created in his home environment. We evaluated a simplified version of the ARgus (written here as ARgus) to help participants quickly master the key features of the interface. More specifically, we deactivated its support for viewpoint recording since it was not useful in our experimental task. We also used pre-selected positions for the spherical view, suitable for the 3D model used in this study. To activate the tool, participants had to right-click on a yellow cylinder located at three relevant positions of the model (one for each room of a house model). The cylinder then became the rotation center of the Spherical View. As we observed in our first study, finding a good placement for the external depth camera is not trivial and largely depends on the task. We decided to use the same configuration as for the first study: we positioned the camera at 2m height, 30° downwards to face the experimenter and to capture his moving body and his augmented workspace, minimizing occlusions. We used the Headset View as control condition. As in our first study, this condition did not provide any interaction capabilities.

Participants downloaded and executed a single Unity application for both conditions on their personal computer. The user interface had a fixed-size window with a 1920×1080 resolution. A step-by-step tutorial about the system functionality and the tasks was directly embedded in the system. For verbal communication between the participants and the experimenter, we used a commercial application (Skype or Discord).

6.3 Task

ARgus' functionalities can support remote mixed-reality participatory design in a range of domains, such as furniture arrangement [11] and urban planning [15, 44]. We decided to focus on a furniture arrangement task because it was used in the past by other related studies [20, 26, 45]. As in our formative study, this task requires participants to search for 3D pieces in the workspace of the experimenter, find a target location for them, and instruct the experimenter to place them correctly. In contrast, we now looked for tasks that would involve both physical and virtual objects in a scene. We considered two alternatives: (i) the AR user manipulates virtual pieces within a larger physical frame of reference (e.g., as in Fig. 6); or (ii) the AR user manipulates physical pieces (miniature furniture) within the virtual model of a house. We opted for the second alternative (see Fig. 1), as it provides richer opportunities for virtual navigation and better captures the trade-offs of different representations. The task simulates the situation where a remote buyer communicates with a furniture designer (or seller). The furniture designer follows instructions to try miniature models of his or her collection in a virtual model of the buyer's house.

We introduced several constraints to create various arrangement tasks unknown to the experimenter. Zodiac symbols were randomly displayed on pre-defined positions on the virtual house model's walls. We chose these symbols as they are easy to identify but hard to verbally describe. This way, we forced participants to rely on intrinsic landmarks of the model for communicating positions, rather than artifacts that are absent in real-world tasks. Two symbols were randomly assigned to each participant. In each room, these two symbols were located on perpendicular walls

and defined a cross-shaped forbidden area: the line in front of each symbol was not available to place furniture.

Participants were asked to arrange furniture for three *thematic spots* randomly chosen among nine. The functional aspect of these spots was described textually. For example, a "living spot" was described as "a place where people can meet and spend some time together". To perform this task, participants could choose miniature furniture among six storage cabinets, four tables and ten chairs (see Fig. 1-d). To complicate the task, we required each miniature chair to be appropriately oriented so that sitting people can see a virtual window without moving their head too much. To be valid, a spot had to include at least two pieces of furniture, meet the placement constraints and represent an harmonious layout (according to the participant's preferences). The symbols, constraints and spot description were communicated to participants at the beginning of each task and made available at any time in a specific panel of the interface (see Fig. 6). This information was unknown to the AR local user.

As for Study 1, we tried to reduce the experimenter's influence [29] by constraining his verbal interventions to only ones required for the completion of the task, i.e., asking the participant to repeat instructions, and asking for confirmation of planned or completed actions.

6.4 Design and Procedure

We followed a within-participant design, where all 12 participants tested both user interface configurations. Half of them were first exposed to the Headset View. The other half were first exposed to ARgus. After signing a consent form, participants completed an online demographic questionnaire. They were then introduced to the two configurations. For ARgus, participants went through a tutorial presenting each tool step-by-step. For each configuration, participants completed a practice and main task. The practice task required the arrangement of one thematic spot.

At the end, participants completed a questionnaire that evaluated their experience with the two configurations that they tested. The full procedure lasted approximately 70 to 90 minutes.

6.5 Data Collection and Measures

We collected participants' answers to a pre- and a post-questionnaire. The post-questionnaire evaluated the efficiency of each user interface configuration with a Likert scale of four items with seven levels (*1 = Inefficient*, *7 = Efficient*). It also assessed the importance of verbal communication for each configuration with a Likert scale of four items with five levels (*1 = Not important*, *5 = Very important*). The questionnaire further evaluated the utility of the views and interactive tools of ARgus configuration and collected participants feedback about their use. We also collected logs of low-level events that describe the use of interactive tools and view transitions during the task. Due to technical problems, logs were not collected for one participant (P5).

Finally, we recorded and manually transcribed participants' voice during the tasks. We then distinguished between phrases that provide remote instructions and other non-instructional content, such as transitional ("ok", "now") and thinking-aloud sentences. Instructions were further classified into three subtask categories: identifying & reaching an object, manipulating an object, and moving in the scene. These categories cover the full set of instructions that we identified and do not overlap. We started with a finer-grained coding scheme. In particular, we initially tried to differentiate between instructions on identifying and reaching objects or locations, and between instructions that concerned different types of manipulation actions. However, these categories were often fused, which made their coding uncertain and unreliable. We thus finally opted for larger categories.

The first and second author decided together on how to segment the transcripts and code the segments by inspecting the data of the first participant. They independently coded the transcripts of three additional participants. They then discussed and finalized the segmentation and coding scheme.

As a last step, the first author re-coded all the transcripts, while the second author independently coded the transcripts of the last two participants. We calculated inter-coder reliability at the word level both for distinguishing between instructions and non-instructions (Krippendorff's $\alpha = .98$, 95% CI [.97, .99]) and for the overall classification that also considers the type of instruction (Krippendorff's $\alpha = .97$, 95% CI [.96, .98]). Inter-coder reliability scores are high, so we count and analyze the words in participants' transcripts for all the above categories.

6.6 Questions and Hypotheses

We expected that participants might develop diverse strategies to complete the tasks. Our goal was to observe and understand these strategies. We were particularly interested in two questions:

Q_1: Will participants find the three views of ARgus useful, and how will they make use of them?

Q_2: Will participants find the user interface tools useful, and how will they make use of them?

Furthermore, we wanted the participants to reflect about how they completed tasks with the two configurations and report on their trade-offs. Tait and Billinghurst [52] found that increased view independence reduces the number of verbal instructions between collaborators. Likewise, we expected that ARgus would reduce reliance on verbal communication, because it gives more viewing freedom to remote users and provides opportunities for completing the task more efficiently. More formally, we tested the following three hypotheses:

H_1: The mean perceived efficiency will be higher for ARgus.

H_2: The mean perceived importance of verbal communication will be lower for ARgus.

H_3: The mean number of words for communicating instructions will be lower for ARgus.

Like Tait and Billinghurst [52], we are interested in the link between view independence and communication performance. However, our studies are distinct from each other. First, since we do not reconstruct the model of the real scene, we investigate view independence through complementary views with different levels of navigation control. Therefore, we also try to identify the view-control strategies that participants develop to carry out the task. Second, our system includes an external view, which also shows the real body of the local user. Note that Tait and Billinghurst [52] recognize the potential benefits of an external view and identify it as a promising configuration for future studies. Third, Tait and Billinghurst [52] test the positioning of physical objects on a physical table. We study instead a more complex task that requires collaborators to position physical pieces within a larger virtual model. In our case, collaborators need to deal with occlusions in the AR scene, thus both physical and virtual navigation are essential for completing the task. Finally, annotations in their system are virtual replicas of a small collection of physical objects, which are conveniently placed on the surface of a table. Our annotation mechanism is simpler but more generic, as it lets remote participants mark any virtual or physical object and location in the 3D workspace with little manipulation effort.

6.7 Results

Anonymized data from this study and the R code of our analyses are available as supplementary material at https://osf.io/3nqrg. Here, we summarize our results.

Use of tools and view representations. We first summarize the strategies that participants used to complete the task under the ARgus condition. For each participant (except for P5), Figure 10 visualizes the active views during the task and the use of previews, the pointing stick, and the spherical view. We emphasize that we did not encourage participants to be fast, and the time range that we show does not always reflect active collaboration time. Some participants (e.g., P1) spent

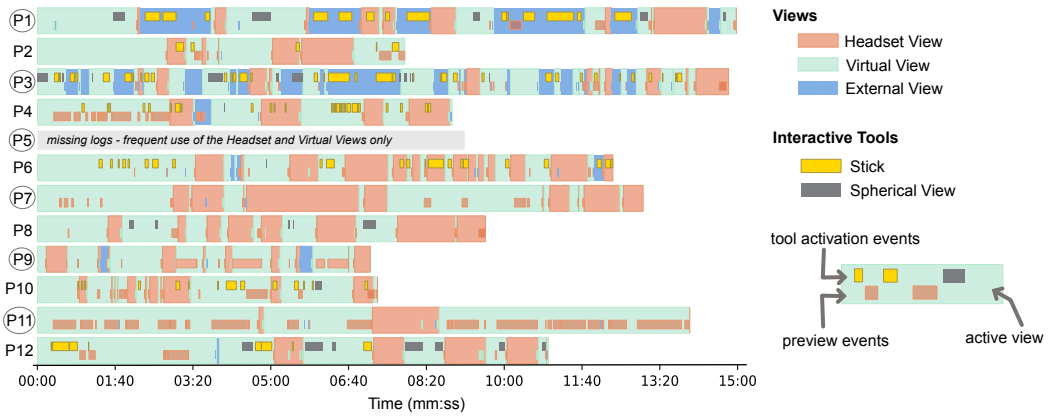

Fig. 10. Use of the three view representations, the pointing stick, and the spherical view by the participants of the evaluation study for the main task under ARgus. Circled participants were exposed to ARgus first.

initial time to think about the constraints of the task and further explore the available tools. It is not a surprise that the slowest participants in Figure 10 were exposed to ARgus first (in circle).

Overall, all participants frequently transitioned between views during the task, which demonstrates the utility of our approach. However, we observe that the Virtual View and the Headset View dominated the participants' choices. The External View was heavily used by P1 and P3 and sparingly by three other participants. Participants' questionnaire responses are consistent with these patterns.

Three only participants found the External View to be useful (P3) or very useful (P1, P9).

P2 explained that he did not *"feel the need"* to use it but *"in a bigger environment it could have been useful to guide the partner quickly from one point to another."*

The three view representation were used in two different ways: (i) as main active views or (ii) through the preview window. Figure 10 shows that several participants (P2, P4, P7, P9, P10, and P11) extensively used the Headset View in preview mode from the Virtual View. According to P4, *"the headset view caused dizziness [...] I stayed in the virtual view and watched the headset view from the window."* P2 agrees that *"having the headset view showing in the corner while navigating and pointing in virtual view was the ideal setup."*

The stick was activated in all three representations either as a pointing or as an annotation tool. For example, P1 and P3 regularly used it from the External View to indicate furniture pieces. P4, P6, and P12 used in combination with the Virtual View to indicate target positions. Other participants did not feel the need to use it: *"I did not use the stick as the rooms had enough identifiable elements to allow my partner to understand my instructions"* (P5). Finally, a smaller group of participants made use of the spherical view. According to P1, it is *"the best to manage the constraints"* but other participants did not agree: *"I was comfortable enough with virtual navigation not to feel the need to resort to the spherical view"* (P2); *"I tried to use the spherical view but I am not enough comfortable with in comparison with rotate and translate so I abandoned."* (P6); *"I would have liked a 2D mapping"* (P5). The spherical mapping that we used is generic but may not be the most appropriate for the specific task. Alternative mappings that better adapt to the geometry of the virtual model might indeed improve the usability of the tool.

Perceived efficiency. We compare the efficiency of the two user interface configurations as perceived by our participants. We use again Bayesian cumulative probit models [13] for our analysis (see Section 4). Figure 11-left summarizes our results. Overall, participants rated ARgus as

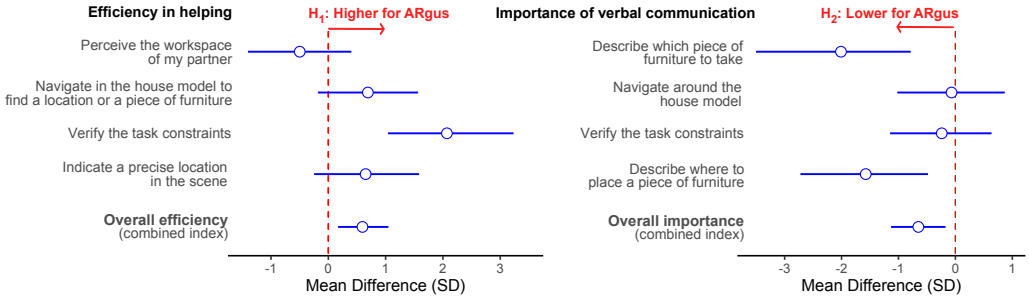

Fig. 11. Comparing the perceived efficiency of the two user interface configurations and the importance of verbal communication for each of them ($N = 12$). We use again Bayesian ordinal (cumulative probit) models [13]. The bars in the graph represent 95% credible intervals of mean differences over a latent continuous variable and can be treated as estimates of standardized effect sizes.

more efficient (see Hypothesis H_1). This was especially the case for verifying the constraints in the scene. For this task, free navigation through the virtual view seemed to be crucial. According to P6, the Headset View causes *"seasickness"*, while P9 commented that its resolution *"was not so effective to perceive accurately the symbols on the walls when having a wide point of view."* In contrast, seven participants rated the Headset View as more efficient for helping them to perceive the workspace of their partner despite the fact that the ARgus configuration provided a richer set of views and options for observing the remote space. The added complexity of this interface can explain this result: *"Having only one solution forces to rely on it and in the case of the headset, forces to establish an efficient communication with the partner, that can be lacking when overwhelmed by all the possibilities of the different views and the difficulty to master them all"* (P3).

Reliance on verbal instructions. Figure 11-right compares the mean difference between configurations in participants' perception about the importance of verbal communication. Overall, verbal communication was perceived as less important for ARgus (see Hypothesis H_2), particularly for describing which pieces of furniture to take and where to place them. P10 explained that verbal communication is more important for the Headset View *"because you cannot point with as much precision as with the stick and you cannot see equally well symbols and distances."*

Our transcript analysis provides additional information about how participants verbally communicated instructions. Figure 12 summarizes our results. Overall, the ARgus user interface reduced the number of words that belonged to instructions by 151.8, 95% CI $[25.7, 278.0]$, $t(11) = 2.65$, $p = .023$ (see Hypothesis H_3). To put this number in perspective, participants pronounced on average 834.5 words with the Headset View, where 435.6 of these words were instructions. We observe that clear differences between conditions only concern instructions that ask the experimenter to move around the model. Surprisingly, there is no clear difference in the number of words used by participants to guide the experimenter on how to identify, reach, and manipulate (e.g., translate or rotate) objects. A possible explanation of this result is the fact that five participants did not at all use the stick (see Fig. 10) and relied on verbal instructions for these subtasks. Indeed, a post hoc analysis shows a strong correlation between the use of the stick (binary variable) and the difference of words used for these subtasks (*Point-biserial correlation* = .79, 95% CI [.40, .94]). Seven participants who used the stick pronounced 156.9 fewer words (95% CI [41.8, 271.9]) with ARgus when they provided instructions for these subtasks. This result, however, must be treated with caution because uncontrolled ordering effects may exaggerate the difference.

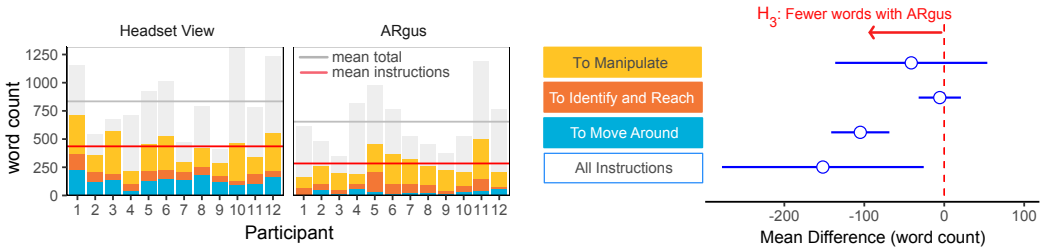

Fig. 12. Results of transcript analysis. We compare the number of words pronounced by the 12 participants to provide instructions. The grey boxes at the left show the total number of words with non-instructions. The error bars at the right represent 95% confidence intervals derived from the t-distribution.

7 DISCUSSION

Overall, our results confirm that remote desktop collaborators can benefit from the multiple views of ARgus, since each view is best adapted to a different aspect of the task. The Virtual View makes navigation in the virtual model easier and independent of the position and visual focus of the local AR user. The External View provides a static overview of the workspace, showing both virtual and physical objects. Finally, the Headset View allows remote users to directly observe the view and actions of their local partner and provide direct instructions. Our participants demonstrated various strategies on how to combine these views with the tools of ARgus. Given previous results [18, 51], we expected a more extensive use of the External View. However, using all three views can be complex, increasing cognitive costs. So many participants judged that the Virtual View and the Headset View were enough for completing the task. Nevertheless, mastering all combinations of views and previews, as well as developing strategies to use them effectively in various steps of the collaboration, may require a long learning process that we did not assess in our studies. Finding a good viewpoint for an external camera also remains a problem. A solution may be to reposition the external camera on the fly depending on the collaborative situation, as explored in Giusti et al. [21]. The nature of the task may also explain why most participants largely relied on the Virtual View to complete the task. It is reasonable to expect that if key objects and landmarks in the scene were mostly physical rather than virtual, the Virtual View might be less appropriate, while the two other views might be more frequently used. Clearly, there are trade-offs in the choice of each view that largely depend on where the task falls in the continuum between virtual and physical.

The results support our three hypotheses. Participants perceived on average that ARgus was more efficient than the control Headset View condition (H_1) and lessened the importance of verbal communication (H_2). We also found that ARgus reduced the average number of words of remote instructions (H_3), which corroborates previous evidence [52] that increased view independence reduces the prevalence of verbal instructions.

We acknowledge that our experimental method and setup present several limitations. The experimenter took the role of the local collaborator in all experimental sessions, which inevitably limits the external validity of our results. The variable quality of the internet connection and the limited resolution of the HoloLens frontal camera may have had an effect as well. Furthermore, we studied one only part of the bilateral collaboration, neglecting how the local AR user perceives and interprets instructions given by the remote collaborator through multiple complementary views. Future user studies should thus examine the collaboration strategies (verbal communication, physical navigation, and gestural interaction) of local users, and their need for awareness of remote user actions.

Another interesting problem is how to extend ARgus to support multiple remote users and enable them to collaboratively interact with but also edit a shared AR scene. This problem poses

significant challenges for the user interface of both local and remote users, since users will now have to coordinate and follow an increased number of viewpoints. Finally, we are interested in enriching ARgus' pointing, annotation, and hybrid navigation tools and evaluate their collaboration effectiveness with more specialized experimental tasks.

8 CONCLUSION

We studied how different views can help a remote desktop user to collaborate with a local user wearing an AR headset on design tasks that may require manipulation of virtual and physical objects. We presented a user study that compared three view representations: (i) a Headset View, augmented video from a first-person viewpoint, (ii) an External View, augmented video from an external third-person viewpoint, and (iii) a Virtual View, a virtual representation with a free viewpoint. Structured as two independent sub-studies with 12 participants each, the study confirmed that each view presents different benefits, targeting a different aspect of a collaboration task. Based on these insights, we developed ARgus, a multi-view collaboration system that provides tools for effectively switching between views, virtually navigating in the remote AR workspace, pointing, and annotating the model. We then ran a second user study to evaluate how 12 remote participants used ARgus to provide instructions to a local user wearing an AR headset for a furniture arrangement task. We observed that participants frequently switched between views or concurrently used them through ARgus' preview functionality. Our results also suggest that the added flexibility of ARgus' multi-view interface allows remote users to verify spatial constraints more efficiently and reduces their reliance on verbal instructions. Future work needs to understand the role of such as a multi-view system from the perspective of the local AR user and extend its scope to multiple remote users.

ACKNOWLEDGMENTS

This work was supported by French government funding managed by the National Research Agency under the Investments for the Future program (PIA) grant ANR-21- ESRE-0030 (CONTINUUM). We thank Olivier Gladin for his precious help during the development of ARgus.

REFERENCES

[1] 2021. *Cinemachine.* Suite of modules for operating the Unity camera https://unity.com/fr/unity/features/editor/art-and-design/cinemachine. Retrieved July 28th 2021.

[2] 2021. *Microsoft HoloLens Application.* Microsoft Hololens 2 Desktop Application https://www.microsoft.com/en-us/p/microsoft-hololens/9nblggh4qwnx. Retrieved July 26th 2021.

[3] 2021. *MixedReality-WebRTC.* Microsoft Mixed Reality WebRTC libraries. https://github.com/microsoft/MixedReality-WebRTC. Retrieved July 26th 2021.

[4] 2021. *MRTK.* Mixed Reality Toolkit for Unity https://docs.microsoft.com/en-us/windows/mixed-reality/mrtk-unity/?view=mrtkunity-2021-05. Retrieved July 28th 2021.

[5] 2021. *UN/DESA Policy Brief 92: Leveraging digital technologies for social inclusion.* United Nation, Department of Economic and Social Affairs, Economic Analysis. https://www.un.org/development/desa/dpad/publication/un-desa-policy-brief-92-leveraging-digital-technologies-for-social-inclusion/. Retrieved November 26th 2021.

[6] 2021. *UNet.* Unity Multiplayer and Networking. https://docs.unity3d.com/Manual/UNet.html. Retrieved January 22th 2021.

[7] Matt Adcock, Stuart Anderson, and Bruce Thomas. 2013. RemoteFusion: Real Time Depth Camera Fusion for Remote Collaboration on Physical Tasks. In *Proceedings of the 12th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry* (Hong Kong, Hong Kong) *(VRCAI '13)*. Association for Computing Machinery, New York, NY, USA, 235–242. https://doi.org/10.1145/2534329.2534331

[8] Tooba Ahsen, Zi Yi Lim, Aaron L. Gardony, Holly A. Taylor, Jan P de Ruiter, and Fahad Dogar. 2021. The Effects of Network Outages on User Experience in Augmented Reality Based Remote Collaboration - An Empirical Study. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 313 (oct 2021), 27 pages. https://doi.org/10.1145/3476054

[9] Huidong Bai, Prasanth Sasikumar, Jing Yang, and Mark Billinghurst. 2020. 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*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376550

[10] Nickolas Bloom. 2020. Working from Home and the Future of U.S. Economic Growth under COVID. https://www.youtube.com/watch?v=jtdFIZx3hyk.

[11] Gordon Brown and Michael Prilla. 2019. Evaluating Pointing Modes and Frames of Reference for Remotely Supporting an Augmented Reality User in a Collaborative (Virtual) Environment: Evaluation within the Scope of a Remote Consultation Session. In *Proceedings of Mensch Und Computer 2019* (Hamburg, Germany) *(MuC'19)*. Association for Computing Machinery, New York, NY, USA, 713–717. https://doi.org/10.1145/3340764.3344896

[12] William A. S. Buxton. 1992. Telepresence: Integrating Shared Task and Person Spaces. In *Proceedings of the Conference on Graphics Interface '92* (Vancouver, British Columbia, Canada). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 123–129.

[13] Paul-Christian Bürkner and Matti Vuorre. 2019. Ordinal Regression Models in Psychology: A Tutorial. *Advances in Methods and Practices in Psychological Science* 2, 1 (2019), 77–101. https://doi.org/10.1177/2515245918823199 arXiv:https://doi.org/10.1177/2515245918823199

[14] Gutwin Carl and Greenberg Saul. 2002. A Descriptive Framework of Workspace Awareness for Real-Time Groupware. *Computer Supported Cooperative Work (CSCW)* 11, 3 (Sept. 2002), 411–446. https://doi.org/10.1023/A:1021271517844

[15] T. Chassin, J. Ingensand, M. Lotfian, O. Ertz, and F. Joerin. 2019. Challenges in creating a 3D participatory platform for urban development. *Advances in Cartography and GIScience of the ICA* 1 (2019), 3. https://doi.org/10.5194/ica-adv-1-3-2019

[16] Andy Cockburn, Carl Gutwin, and Alan Dix. 2018. HARK No More: On the Preregistration of CHI Experiments. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173715

[17] Martin Feick, Anthony Tang, and Scott Bateman. 2018. Mixed-Reality for Object-Focused Remote Collaboration. In *The 31st Annual ACM Symposium on User Interface Software and Technology Adjunct Proceedings* (Berlin, Germany) *(UIST '18 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 63–65. https://doi.org/10.1145/3266037.3266102

[18] Susan R. Fussell, Leslie D. Setlock, and Robert E. Kraut. 2003. Effects of Head-Mounted and Scene-Oriented Video Systems on Remote Collaboration on Physical Tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 513–520. https://doi.org/10.1145/642611.642701

[19] Steffen Gauglitz, Benjamin Nuernberger, Matthew Turk, and Tobias Höllerer. 2014. World-Stabilized Annotations and Virtual Scene Navigation for Remote Collaboration. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology* (Honolulu, Hawaii, USA) *(UIST '14)*. Association for Computing Machinery, New York, NY, USA, 449–459. https://doi.org/10.1145/2642918.2647372

[20] William W. Gaver, Abigail Sellen, Christian Heath, and Paul Luff. 1993. One is Not Enough: Multiple Views in a Media Space. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems* (Amsterdam, The Netherlands) *(CHI '93)*. Association for Computing Machinery, New York, NY, USA, 335–341. https://doi.org/10.1145/169059.169268

[21] Leonardo Giusti, Kotval Xerxes, Amelia Schladow, Nicholas Wallen, Francis Zane, and Federico Casalegno. 2012. Workspace Configurations: Setting the Stage for Remote Collaboration on Physical Tasks. In *Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design* (Copenhagen, Denmark) *(NordiCHI '12)*. Association for Computing Machinery, New York, NY, USA, 351–360. https://doi.org/10.1145/2399016.2399071

[22] Martin Hachet, Fabrice Decle, Sebastian Knoedel, and Pascal Guitton. 2008. Navidget for Easy 3D Camera Positioning from 2D Inputs. In *Proceedings of the IEEE Symposium on 3D User Interfaces (3DUI)*. United States, 83–88. https://hal.archives-ouvertes.fr/hal-00308251

[23] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, and Andrew Fitzgibbon. 2011. KinectFusion: Real-Time 3D Reconstruction and Interaction Using a Moving Depth Camera. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology* (Santa Barbara, California, USA) *(UIST '11)*. Association for Computing Machinery, New York, NY, USA, 559–568. https://doi.org/10.1145/2047196.2047270

[24] 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*. Association for Computing Machinery, New York, NY, USA, Article 283, 7 pages. https://doi.org/10.1145/3411763.3451844

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

https://doi.org/10.1145/3449133

[26] Nicolas Kahrl, Michael Prilla, and Oliver Blunk. 2020. Show Me Your Living Room: Investigating the Role of Representing User Environments in AR Remote Consultations. In *Proceedings of the Conference on Mensch Und Computer* (Magdeburg, Germany) *(MuC '20)*. Association for Computing Machinery, New York, NY, USA, 267–277. https://doi.org/10.1145/3404983.3405520

[27] Matthew Kay, Gregory L. Nelson, and Eric B. Hekler. 2016. Researcher-Centered Design of Statistics: Why Bayesian Statistics Better Fit the Culture and Incentives of HCI *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 4521–4532. https://doi.org/10.1145/2858036.2858465

[28] Ryohei Komiyama, Takashi Miyaki, and Jun Rekimoto. 2017. JackIn Space: Designing a Seamless Transition between First and Third Person View for Effective Telepresence Collaborations. In *Proceedings of the 8th Augmented Human International Conference* (Silicon Valley, California, USA) *(AH '17)*. Association for Computing Machinery, New York, NY, USA, Article 14, 9 pages. https://doi.org/10.1145/3041164.3041183

[29] Anna K. Kuhlen and Susan E. Brennan. 2013. Language in dialogue: when confederates might be hazardous to your data. *Psychonomic Bulletin & Review* 20, 1 (2013), 54–72. https://doi.org/10.3758/s13423-012-0341-8

[30] André Kunert, Alexander Kulik, Stephan Beck, and Bernd Froehlich. 2014. Photoportals: Shared References in Space and Time. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Baltimore, Maryland, USA) *(CSCW '14)*. Association for Computing Machinery, New York, NY, USA, 1388–1399. https://doi.org/10.1145/2531602.2531727

[31] Hideaki Kuzuoka. 1992. Spatial Workspace Collaboration: A SharedView Video Support System for Remote Collaboration Capability. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Monterey, California, USA) *(CHI '92)*. Association for Computing Machinery, New York, NY, USA, 533–540. https://doi.org/10.1145/142750.142980

[32] Joel Lanir, Ran Stone, Benjamin Cohen, and Pavel Gurevich. 2013. Ownership and Control of Point of View in Remote Assistance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 2243–2252. https://doi.org/10.1145/2470654.2481309

[33] Torrin M. Liddell and John K. Kruschke. 2018. Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology* 79 (2018), 328–348. https://doi.org/10.1016/j.jesp.2018.08.009

[34] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. 2020. Consistent Video Depth Estimation. *ACM Trans. Graph.* 39, 4, Article 71 (jul 2020), 13 pages. https://doi.org/10.1145/3386569.3392377

[35] E. Marchand, F. Spindler, and F. Chaumette. 2005. ViSP for visual servoing: a generic software platform with a wide class of robot control skills. *IEEE Robotics and Automation Magazine* 12, 4 (December 2005), 40–52.

[36] Peter Mohr, Shohei Mori, Tobias Langlotz, Bruce H. Thomas, Dieter Schmalstieg, and Denis Kalkofen. 2020. Mixed Reality Light Fields for Interactive Remote Assistance. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376289

[37] Jens Müller, Roman Rädle, and Harald Reiterer. 2017. Remote Collaboration With Mixed Reality Displays: How Shared Virtual Landmarks Facilitate Spatial Referencing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 6481–6486. https://doi.org/10.1145/3025453.3025717

[38] Ohan Oda, Carmine Elvezio, Mengu Sukan, Steven Feiner, and Barbara Tversky. 2015. Virtual Replicas for Remote Assistance in Virtual and Augmented Reality. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software and Technology* (Charlotte, NC, USA) *(UIST '15)*. Association for Computing Machinery, New York, NY, USA, 405–415. https://doi.org/10.1145/2807442.2807497

[39] Niklas Osmers and Michael Prilla. 2020. Getting out of Out of Sight: Evaluation of AR Mechanisms for Awareness and Orientation Support in Occluded Multi-Room Settings. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3313831.3376742

[40] Kyoung S. Park, Abhinav Kapoor, and Jason Leigh. 2000. Lessons Learned from Employing Multiple Perspectives in a Collaborative Virtual Environment for Visualizing Scientific Data. In *Proceedings of the Third International Conference on Collaborative Virtual Environments* (San Francisco, California, USA) *(CVE '00)*. Association for Computing Machinery, New York, NY, USA, 73–82. https://doi.org/10.1145/351006.351015

[41] Abhishek Ranjan, Jeremy P. Birnholtz, and Ravin Balakrishnan. 2007. Dynamic Shared Visual Spaces: Experimenting with Automatic Camera Control in a Remote Repair Task. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1177–1186. https://doi.org/10.1145/1240624.1240802

[42] Troels A. Rasmussen and Weidong Huang. 2019. SceneCam: Using AR to Improve Multi-Camera Remote Collaboration. In *SIGGRAPH Asia 2019 XR* (Brisbane, QLD, Australia) *(SA '19)*. Association for Computing Machinery, New York, NY,

USA, 36–37. https://doi.org/10.1145/3355355.3361892

[43] Patrick Salamin, Daniel Thalmann, and Frédéric Vexo. 2006. The Benefits of Third-Person Perspective in Virtual and Augmented Reality?. In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology* (Limassol, Cyprus) *(VRST '06)*. Association for Computing Machinery, New York, NY, USA, 27–30. https://doi.org/10.1145/1180495.1180502

[44] Sheree May Saßmannshausen, Jörg Radtke, Nino Bohn, Hassan Hussein, Dave Randall, and Volkmar Pipek. 2021. *Citizen-Centered Design in Urban Planning: How Augmented Reality Can Be Used in Citizen Participation Processes*. Association for Computing Machinery, New York, NY, USA, 250–265. https://doi.org/10.1145/3461778.3462130

[45] Wendy A. Schafer and Doug A. Bowman. 2005. Integrating 2D and 3D Views for Spatial Collaboration. In *Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work* (Sanibel Island, Florida, USA) *(GROUP '05)*. Association for Computing Machinery, New York, NY, USA, 41–50. https://doi.org/10.1145/1099203.1099210

[46] Michael F. Schober. 1995. Speakers, addressees, and frames of reference: Whose effort is minimized in conversations about locations? *Discourse Processes* 20, 2 (1995), 219–247. https://doi.org/10.1080/01638539509544939 arXiv:https://www.tandfonline.com/doi/pdf/10.1080/01638539509544939

[47] R. N. Shepard and J. Metzler. 1971. Mental Rotation of Three-Dimensional Objects. *Science* 171, 3972 (Feb. 1971), 701–703. https://doi.org/10.1126/science.171.3972.701

[48] Joon Gi Shin, Gary Ng, and Daniel Saakes. 2018. Couples Designing Their Living Room Together: A Study with Collaborative Handheld Augmented Reality. In *Proceedings of the 9th Augmented Human International Conference* (Seoul, Republic of Korea) *(AH '18)*. Association for Computing Machinery, New York, NY, USA, Article 3, 9 pages. https://doi.org/10.1145/3174910.3174930

[49] Rajinder S. Sodhi, Brett R. Jones, David Forsyth, Brian P. Bailey, and Giuliano Maciocci. 2013. BeThere: 3D Mobile Collaboration with Spatial Input. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 179–188. https://doi.org/10.1145/2470654.2470679

[50] Mengu Sukan, Steven Feiner, Barbara Tversky, and Semih Energin. 2012. Quick Viewpoint Switching for Manipulating Virtual Objects in Hand-Held Augmented Reality Using Stored Snapshots. In *Proceedings of the 2012 IEEE International Symposium on Mixed and Augmented Reality (ISMAR) (ISMAR '12)*. IEEE Computer Society, USA, 217–226. https://doi.org/10.1109/ISMAR.2012.6402560

[51] Hongling Sun, Yue Liu, Zhenliang Zhang, Xiaoxu Liu, and Yongtian Wang. 2018. Employing Different Viewpoints for Remote Guidance in a Collaborative Augmented Environment. In *Proceedings of the Sixth International Symposium of Chinese CHI* (Montreal, QC, Canada) *(ChineseCHI '18)*. Association for Computing Machinery, New York, NY, USA, 64–70. https://doi.org/10.1145/3202667.3202676

[52] Matthew Tait and Mark Billinghurst. 2015. The Effect of View Independence in a Collaborative AR System. *Comput. Supported Coop. Work* 24, 6 (dec 2015), 563–589. https://doi.org/10.1007/s10606-015-9231-8

[53] Chiew Seng Sean Tan, Kris Luyten, Jan Van Den Bergh, Johannes Schöning, and Karin Coninx. 2014. The Role of Physiological Cues during Remote Collaboration. *Presence: Teleoperators and Virtual Environments* 23, 1 (02 2014), 90–107. https://doi.org/10.1162/PRES_a_00168 arXiv:https://direct.mit.edu/pvar/article-pdf/23/1/90/1625375/pres_a_00168.pdf

[54] Arthur Tang, Charles Owen, Frank Biocca, and Weimin Mou. 2002. Experimental Evaluation of Augmented Reality in Object Assembly Task. In *Proceedings of the 1st International Symposium on Mixed and Augmented Reality (ISMAR '02)*. IEEE Computer Society, USA, 265.

[55] Markus Tatzgern, Raphael Grasset, Denis Kalkofen, and Dieter Schmalstieg. 2014. Transitional Augmented Reality navigation for live captured scenes. In *2014 IEEE Virtual Reality (VR)*. 21–26. https://doi.org/10.1109/VR.2014.6802045

[56] 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* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300431

[57] 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* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 161–174. https://doi.org/10.1145/3332165.3347872

[58] Balasaravanan Thoravi Kumaravel, Cuong Nguyen, Stephen DiVerdi, and Bjoern Hartmann. 2020. TransceiVR: Bridging Asymmetrical Communication Between VR Users and External Collaborators. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 182–195. https://doi.org/10.1145/3379337.3415827

[59] John Wang and Edwin Olson. 2016. AprilTag 2: Efficient and robust fiducial detection. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

[60] Peng Wang, Shusheng Zhang, Mark Billinghurst, Xiaoliang Bai, Weiping He, Shuxia Wang, Mengmeng Sun, and Xu Zhang. 2020. A comprehensive survey of AR/MR-based co-design in manufacturing. *Engineering with Computers* 36 (2020), 1715–1738. Issue 4. https://doi.org/10.1007/s00366-019-00792-3

[61] Michael Wittkämper, Irma Lindt, Wolfgang Broll, Jan Ohlenburg, Jan Herling, and Sabiha Ghellal. 2007. Exploring Augmented Live Video Streams for Remote Participation. In *CHI '07 Extended Abstracts on Human Factors in Computing Systems* (San Jose, CA, USA) *(CHI EA '07)*. Association for Computing Machinery, New York, NY, USA, 1881–1886. https://doi.org/10.1145/1240866.1240915

[62] Haijun Xia, Sebastian Herscher, Ken Perlin, and Daniel Wigdor. 2018. Spacetime: Enabling Fluid Individual and Collaborative Editing in Virtual Reality *(UIST '18)*. Association for Computing Machinery, New York, NY, USA, 853–866. https://doi.org/10.1145/3242587.3242597

[63] Longqi Yang, David Holtz, Sonia Jaffe, Siddharth Suri, Shilpi Sinha, Jeffrey Weston, Connor Joyce, Neha Shah, Kevin Sherman, Brent Hecht, and Jaime Teevan. 2021. The effects of remote work on collaboration among information workers. *Nature Human Behaviour* (2021). https://doi.org/10.1038/s41562-021-01196-4

Received January 2022; revised April 2022; accepted August 2022