International Journal of Human–Computer Interaction

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/hihc20

AVICol: Adaptive Visual Instruction for Remote Collaboration Using Mixed Reality

Lili Wang, Xiangyu Li, Jian Wu, Dong Zhou, Im Sio Kei & Voicu Popescu

To cite this article: Lili Wang, Xiangyu Li, Jian Wu, Dong Zhou, Im Sio Kei & Voicu Popescu (18 Feb 2024): AVICol: Adaptive Visual Instruction for Remote Collaboration Using Mixed Reality, International Journal of Human–Computer Interaction, DOI: 10.1080/10447318.2024.2313920

To link to this article: https://doi.org/10.1080/10447318.2024.2313920

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
https://doi.org/10.1080/10447318.2024.2313920

AVICol: Adaptive Visual Instruction for Remote Collaboration Using Mixed Reality

Lili Wang[a,b], Xiangyu Li[a], Jian Wu[a], Dong Zhou[a], Im Sio Kei[c], and Voicu Popescu[d]

[a]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China; [b]Peng Cheng Laboratory, Shengzhen, China; [c]Macao Polytechnic University, Macau, China; [d]Purdue University, West Lafayette, IN, USA

ABSTRACT

This article describes a mixed reality visual instruction approach for remote collaboration between a trainee and an expert. The expert authors the visual instructions through a virtual reality interface. The instructions are shown to the trainee overlaid onto the workspace using an augmented reality interface. The approach achieves effectiveness and efficiency by addressing three challenges. First, the expert-authored visual instructions are shown to the trainee by taking into account occlusions with the 3D workspace; Second, in addition to abstract visual instructions implemented by arrows, the expert can also author highly suggestive instructions by depicting the target state of the workspace realistically by selecting, copying, pasting, and repositioning workspace objects; Third, multiple instructions can be concatenated in sequences that the trainee executes on their own, without any additional guidance from the expert; The approach has been evaluated in a controlled user study with three experiments. The experiment verification confirms that compared to the conventional instruction, this approach achieves significantly lower error rates, shorter task completion times, and lower rotation angular errors. Moreover, the approach allows the trainee to execute the entire sequence robustly, without real-time instruction from the expert.

KEYWORDS

Object manipulation; remote collaboration; synchronous collaboration; asynchronous collaboration; trainee expert collaboration; mixed reality; augmented reality; virtual reality

1. Introduction

Mixed reality (MR) enables effective collaboration over great geographic distances by linking the worlds of the collaborating parties. Remote collaboration implies that physically isolated collaborators work together to integrate their activities in a seamless way to achieve a common goal (Marques, Teixeira, et al., 2022). MR remote collaboration is becoming commonplace and entering various physical scenarios (Ens et al., 2019; Fidalgo et al., 2023), including manufacturing (Wang et al., 2020), telemedicine, and teleducation (Wang et al., 2021). In one type of remote collaboration, collaborators play asymmetric roles (Sereno et al., 2020), a *local* trainee has to manipulate physical objects in their workspace under the guidance of a *remote* expert. A powerful way of providing guidance to the trainee is through graphical annotations of the workspace (Marques et al., 2021), which can be more effective and more efficient than verbal instructions (Fussell et al., 2003). The trainee's workspace is captured using multi-viewpoint depth and color cameras that acquire the appearance and geometry of the workspace in real time. The acquired color and depth data is then sent to the remote site where the expert can visualize it immersively with a virtual reality headset. However, for such a mixed reality remote collaboration approach to reach its potential several challenges have to be overcome.

One challenge is brought by occlusions in the 3D workspace, which complicate the visualization of the graphical annotations. The trainee and the expert see the workspace from different viewpoints, so what one sees might not be visible to the other. For example, the trainee might not see an arrow drawn by the expert to indicate the translation of an object, and solving this problem by asking the trainee to adjust their viewpoint can be difficult and time-consuming.

A second challenge is to allow the expert to author visual instructions that convey specific desired workspace configurations with high accuracy. Whereas visual instructions can be conveyed through abstract graphical annotations, such as an arrow that indicates which object has to be moved where, richer visual instructions are needed to convey, for example, an object's exact desired pose.

A third challenge is achieving good communication efficiency between the trainee and the expert. When the eloquence of the visual instruction is low, the trainee and expert have to communicate one small step at a time, which leads to long task completion times and also to learning delays, as the trainee has to focus on the communication and cannot focus on learning the task; efficient remote collaboration requires concatenating long sequences of visual instructions that can be communicated to and executed by the trainee in a single batch.

CONTACT Jian Wu lanayawj@buaa.edu.cn State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China

In this article, we describe *AVICol*, a mixed reality approach for the effective and efficient authoring and execution of object manipulation instructions in a trainee-expert remote collaboration scenario to address the three challenges enumerated above. The *AVICol* approach:

- adapts the rendering of the visual instructions automatically and in real time to avoid the occlusion challenges posed by a 3D workspace seen from different viewpoints by the expert and trainee;
- allows the expert to author visual instructions that are either abstract or realistic, as needed for conveying simple instructions efficiently, or for conveying complex instructions effectively, by describing in great detail the workspace state to be achieved;
- adapts instructions to the current state of the workspace in support of an asynchronous authoring/execution loop of multi-step instruction sequences.

We have evaluated our approach in a controlled user study ($N = 24$) with three experiments, where participants served as experts and trainees. In the first experiment, participants were asked to translate objects collaboratively, where our method achieved significantly lower error rates and shorter task completion times compared to conventional visual instructions that do not take into account occlusions. In the second experiment, participants were asked to rotate objects collaboratively, where our method achieved significantly shorter instruction authoring times, shorter instruction execution times, and lower rotation angular errors, compared to verbal instructions and compared to conventional visual instructions based on coordinate system axes visualization. In the third experiment, participants serving as experts were asked to author a sequence of 10 object translations and rotations, and then, participants serving as trainees were asked to execute all 10 instructions in the sequence without any additional help from the expert; both the instruction authoring and execution times were significantly shorter compared to the time needed when the expert authored and communicated instructions to the trainee one step at the time.

The article is organized as follows: we review the related work in Section 2; Section 3 describes our *AVICol* approach. In Section 4, we present the user study design and discuss the results. Finally, we draw conclusions and discuss the future work in Section 5.

2. Related work

In addition to the brief overview of prior work below, we also refer the reader to recent comprehensive surveys of prior work approaches for visual instruction using virtual (VR), mixed, and augmented reality (AR) (de Belen et al., 2019; Druta et al., 2021; Schäfer et al., 2021). We first discuss prior approaches for co-located (Section 2.1) and for remote (Section 2.2) collaboration between two or more users; we then discuss related approaches for single-user interaction guidance (Section 2.4), with an emphasis on methods that circumvent occluders with the help of curved selection lines, which are closely related to our work (Section 2.5).

2.1. Co-located collaboration

Prior work has investigated collaboration efficiency gains afforded by AR annotations. Several collaboration scenarios have been investigated. In one scenario, the collaborators work together toward completing a task, where one of the goals is to allow the collaborators to cross-reference the workspace effectively. One study has shown that synchronizing the visual attention of collaborators can be done with salient virtual objects, e.g., a potted plant, inserted in the workspace for common reference (Müller et al., 2016). The virtual landmarks reduced the number of deictic gestures used by the collaborators in favor of less ambiguous verbal references. Another study reveals that collaborators can find a common reference more easily with visual cues, such as lines or animated cursors pointing at the object of common interest (Chen et al., 2021). The results of the study showed a contradiction between task performance, which was higher for pointing lines, and user preference, which was higher for animated cursors. A taxonomy of spatial communication partitions cues according to two dimensions: the cue's attachment, i.e., physical (AR) or virtual (VR), and the cue's animation, i.e., local or world trajectory (Irlitti et al., 2019).

2.2. Remote collaboration

In a second collaboration scenario, which is also the scenario investigated by our work, one of the collaborators is a local trainee who has to perform tasks under the guidance of a remote expert. In this scenario the collaboration system has the additional tasks of conveying the workspace to the remote expert, and of allowing the expert to author annotations. There are five essential factors for remote collaboration: task, local user, remote user, communication, and tool/interface (Kim, Billinghurst, et al., 2020). Most researches either explicitly or implicitly center on these five essential factors. One study captures the workspace either with a still image or with a video feed that is uploaded to the mentor site (Kim et al., 2013). Then the mentor annotates the workspace with a telestrator paradigm to point to or to sketch on the workspace. The study found greater collaboration efficiency when the workspace was captured with a video as opposed to with still images, especially when quick feedback is needed. Furthermore, the study found that sketched annotations were more useful than simple pointing cues. Another study focused on textual annotations of videos which helped the collaborators use converging language, making verbal communication more efficient (Chang et al., 2017).

Researchers have also investigated upgrading the acquisition of the workspace from 2D videos to 3D geometry. One study resorts to a passive acquisition of the workspace geometry. The ambiguity of 2D annotations is resolved by having the expert draw the annotations from two viewpoints and by triangulating the 2D annotations into 3D space

(Nuernberger et al., 2016). As depth cameras evolved, researchers have begun relying on active workspace acquisition (Gauglitz et al., 2014). Hand-drawn 2D visual cues are unprojected to 3D space by leveraging the workspace geometry, or by relying on simplifying planar proxy assumptions, which allows the trainee to change viewpoint freely (Gauglitz et al., 2014). Due to limited depth acquisition accuracy, users preferred the anchoring of annotations using the planar proxy assumption.

Another challenge is to convey a stable visualization of the workspace to the expert and a stable visualization of the annotations to the trainee. The first problem is particularly acute when the workspace is acquired by the trainee with a head-mounted camera, which brings the system compactness prerequisite for deployment in austere environments, but also brings frequent and substantial view direction changes as the trainee moves their head, which distracts the mentor (Lin et al., 2020). We acquire the workspace with an array of depth and color cameras mounted on a tripod, which provides good stability of the workspace visualization at the expert. The second problem, i.e., that of stabilizing the annotations in the trainee's view, requires either asking the trainee to assume a viewpoint substantially similar to that of the acquisition viewpoint or alternatively, anchoring the annotations to the 3D workspace. One example of prior work that took the latter approach relied on pre-modeling the workspace and on fiducials to maintain projector/workspace alignment (Adcock & Gunn, 2015).

Another aspect of facilitating collaboration is whether and how much of the remote collaborator is visible to the local collaborator. This recent review summarizes the visual communication cues being used, categorizing the communication cues into explicit and implicit ones. Explicit cues are categorized into pointer/sketches and annotations/hand gesture/object models. Implicit cues, such as eye gaze, avatar could help improve collaborative experience, such as co-presence, situation awareness (Huang et al., 2022). One option is to only convey gaze (Higuch et al., 2016; Piumsomboon et al., 2019). For example, letting the trainee see where the mentor's gaze is fixated during real time collaboration provides an efficient pointer, but also has to be supplemented with hand gestures to convey more complex instructions, such as those needed for rotation manipulations (Higuch et al., 2016). Researchers have also investigated bidirectional gaze sharing where both the expert and the trainee see each other's gaze (Jing et al., 2021, 2022; Lee et al., 2017), which helped, for example, in puzzle solving (Lee et al., 2017). A second option is to convey mentor hand gestures to the trainee, which have been shown to help in the context of object selection (Kim, Jing, et al., 2020). Other studies have explored the effects of sharing either gaze or gesture cues alone, or their combination, on remote collaboration effectiveness (Bai et al., 2020). A third option is to show remote collaborators through avatars (Piumsomboon et al., 2018; Yoon et al., 2019; Yu et al., 2021). For example, realistic avatars have been preferred to cartoon-like avatars in terms of social presence scores (Yoon et al., 2019). A recent work explored how the counterpart representation affects social presence. The author compared two distinct conditions: traditional video chat (collaborators always visible) and AR annotations (collaborators never available) and found that the majority of participants preferred the AR-based condition, despite the absence of team members representation, which led to slightly lower sense of social presence, but significantly higher results for the remaining dimensions of collaboration, as well as faster task resolution (Marques et al., 2023).

Researchers have investigated hybrid approaches based on several types of cues, such as gestures, virtual ray pointing, and drawing (Kim, Lee, et al., 2020; Teo et al., 2018, 2019), to find that specific combinations of cues are beneficial for specific tasks in terms of reducing task load and improving the social quality of the collaboration. Some researchers find that compared with AR annotation, using gestures and head pointing significantly improved the collaborative experience and remote interaction (Wang et al., 2019; Zhang et al., 2022).

Virtual replicas of physical objects have been used in remote collaboration. Physical objects corresponding to virtual objects are usually separated from other objects when the workspace is captured. One study divided the local workspace into background and foreground objects to support real-time updates. Background objects were scanned and 3D reconstructed and remained stationary. Foreground objects(virtual replicas) can be relocated in the workspace during the collaboration using object tracking (Chang et al., 2023). Another way is to model the foreground objects of interest as a polygon mesh in advance and reconstruct the surrounding background objects into a point cloud (Lee et al., 2021; Tian et al., 2023). The combination of Virtual replicas and other visual cues was also discussed. One study investigated the combination of visual cues of gestures, avatar, and virtual replicas and found the combination plays a positive role in improving user experience (Wang et al., 2023). Another study focused on the combination of virtual replicas and gesture cues in the 3D video and another method of using gesture cues in the 3D video. The study found that using the former can significantly improve the performance and user experience in industrial assembly tasks. Virtual replicas were also used in a web-based extended reality collaboration system to instruct the manipulation of physical objects (Lee & Yoo, 2021).

In addition to visual cues, spatial auditory cues are another aspect of facilitating collaborator. Researchers presented an MR remote collaboration system that shares both spatial auditory and visual cues between collaborators and found that compared to non-spatialized audio, the spatialized remote expert's voice and auditory beacons enabled local workers to find small occluded objects with significantly stronger spatial perception. The work also found that integrating visual cues with the spatial auditory cues significantly improved the local worker's task performance, social presence, and spatial perception of the environment (Yang et al., 2020).

Different view types will also have different impacts on remote collaboration. Researchers have investigated and

compared two view types, dependent and independent views. One work compared different visual cues across the two view types, respectively (Kim et al., 2023). Another work investigated how different collaboration styles and the two view types affect remote collaboration. Two different collaboration styles were compared, one is the scenario mentioned at the beginning of this section and the other is a mutual collaboration where neither user has a solution but both remote and local users share ideas and discuss ways to solve the real-world task (Kim et al., 2018). The impact of the remote user's role has been further studied in a mixed presence(MP) system. MP systems incorporate both face-to-face and remote users. The research found that the role of the coordinator significantly increased the remote user's engagement with increased usage of visual communication cues (Norman et al., 2019).

Most research works, so far, have been devoted to explore and evolve the necessary technology. However, it is important to identify gaps that should inform further research. One study adopted a user-centered approach with partners from the industry sector to probe how AR could provide solutions to support their collaborative efforts and identified a set of requirements (Marques, Silva, et al., 2022).

2.3. Asynchronous collaboration

Asynchronous collaboration offers several unique advantages over synchronous collaboration, such as work parallelism, and flexible time coordination. Research on asynchronous collaboration is discussed in a related comprehensive survey of collaborative work (Pidel & Ackermann, 2020). Recording and replaying is an important way to realize asynchronous collaboration. One research presented a multimodal asynchronous VR collaboration system capable of capturing, recording, and replaying multimodal messages, including speech, body gestures, and manipulations on objects (Chow et al., 2019). Another research linked recorded speech to visual objects to provide an effective communication in asynchronous collaboration (Kim et al., 2021). Museum exhibition is a broad application of AR asynchronous collaboration. One research developed a user interface that enables asynchronous exhibit browsing for visitors participating at different times. Visitors can access the exhibits and interactive information shared by other visitors (Chen et al., 2021). Another interesting collaboration scenario is multiple people building structures together, synchronously or asynchronously, on-site or remote (Guo et al., 2019). The challenges faced during the use of asynchronous collaborative AR are also discussed (Irlitti et al., 2016). The type of recording and replaying is also applied to MR remote asynchronous collaboration. To instruct the trainee, the expert can capture his actions and send the record file to the trainee. And the trainee can see a ghost of expert demonstrating the assembly steps (Mayer et al., 2022).

2.4. Interaction guidance

One of the first applications of AR was to help a user find their way through a real world scene by providing visual guidance. In one such early work, the AR system displays the distance and direction to the target (Thomas et al., 1998). In another example, the user is provided guidance by overlaying semi-transparent paths onto the user's view of the real world (Narzt et al., 2006). Virtual roads are rendered with semi-transparent colors, allowing users to select the preferred path in the presence of obstacles. AR has also been used to assist drone pilots by rendering important points on the suggested flight path, as well as connections between these waypoints (Zollmann et al., 2014). To improve the eloquence of the guidance provided, occluded target objects are revealed by rendering them semi-transparently, and the depth of midair waypoints is conveyed with vertical lines that project them onto the ground. More recently, machine learning has been used to determine when the user is in need of navigation assistance and when not, to avoid providing unnecessary guidance that can result in visual distraction (Seeliger et al., 2022).

Another application where researchers have explored providing guidance to the user through AR is that of searching in cluttered environments. Early work has found that subtle coercive mechanisms, such as target contrast manipulations do provide assistance with search tasks, without increasing visual clutter (Lu et al., 2012). A later study investigated target-based cues, such as increasing target salience through color enhancements or blinking, and directional cues, such as indicating the target with arrows and lines (Volmer et al., 2018), and found that all cues had a positive impact on user performance. Furthermore, directional cues lead to better search performance than target-based cues for complex search tasks. A more recent study (Seeliger et al., 2021) compared multiple types of visual cues, e.g., arrows, boxes, and wedges, along multiple dimensions, i.e., in-view *vs.* out-of-view, static *vs.* dynamic, sequential *vs.* simultaneous. The findings include that cues presented simultaneously shift the user's attention away from non-target objects more effectively than cues presented sequentially, and that dynamic cues helped users locate the target object more quickly than static cues. A special case is that when the user has to search for textual labels, where displaying the labels in a circular pattern with varying circle thickness has proven to reduce search times (Zhou et al., 2021).

Occlusions are not always a challenge that has to be overcome, but they can also be a desired feature of AR systems, to enable the correct depth sorting of elements of the real world scene and of the graphical annotations that describe them. One study proposed an AR authoring tool that provides accurate, fine grain visibility sorting by acquiring the workspace with a real time depth camera (Gimeno et al., 2013). Instead of depth acquisition, the workspace geometry was also approximated with pre-modeled virtual replicas of physical objects from the local environment (Elvezio et al., 2017).

For some applications, the user's view of the real world scene is insufficient to receive assistance from the AR system. An example is rock climbing instruction where the user's head is too close to the rock to have a comprehensive view of their body and of the path to take, so the user is

provided with a third-person view of themselves and of a pre-recorded expert climber (Kosmalla et al., 2017).

AR techniques can also focus on visualizing the deviation between the current and desired poses of the objects to be manipulated, to help the user reduce this deviation. For example, one approach is to visualize the deviation direction and magnitude trough the position and number of spherical particles (Jeanne et al., 2017), which was shown in a later study (Jeanne et al., 2017) to improve manipulation trajectory accuracy over conventional illustrations of the desired movement to be used as reference. In another prior work, researchers focused on the specific problem of posing a handheld object in mid-air with six degrees of freedom (Andersen & Popescu, 2020). The object is represented with a hand-held controller tracked with six degrees of freedom. The findings indicate that visual guidance placed close to the hand-held object reduce alignment times and translation errors, while interfaces that place the visual guidance far away from the hand-held object magnify the visualization of rotational errors for more accurate poses.

One prior method for providing guidance for object manipulation relies on color-coded annotations, e.g., a yellow line to guide the user to the next object to be selected, and a red circle to indicate the end point of the translation (Liu et al., 2022). The study tested concatenating several instructions and found that the best performance is achieved for two instructions (i.e., "cue" and "precue"). In addition, the study also found that rotation instructions are most useful when split across the manipulated object and its destination.

2.5. Bending ray techniques

Selection in VR and AR is complicated by occlusions. When the user has no line of sight to the selection target, one approach is to rely on the user to translate their viewpoint to establish the line of sight. However, this could be slow, unintuitive, and, physically tiring. One approach for increasing the user's ability to select in the presence of occlusion relies on a bent selection ray that can circumvent occluders. One approach uses a flexible pointer to select an object visible to the user, i.e., from the user's viewpoint, but not visible from the user's hand from where the pointing line starts (Feiner, 2003). The Bent Pick Ray (Riege et al., 2006) is a method for providing visual feedback in situations in which objects are manipulated simultaneously by multiple users. When multiple users are manipulating the same object, the selection rays are bent based on the object direction and the pick ray direction. The visual feedback helps users understand the collaborative manipulation, taking into account the different user viewpoints. Another study used a flexible ray modeled with a quadratic Bézier curve that starts out as a straight line and then bends to snap to the closest selection candidate object, facilitating selection (Steinicke et al., 2006).

Researchers have also combined the curved ray technique with other selection techniques. One approach enhances the ray with a bubble and allows the user to select candidate objects intersecting with the bubble (Lu et al., 2020). Disambiguation between selection candidates is done by bending the ray, and empirical evaluation shows performance and preference advantages over the conventional ray casting selection paradigm.

Instead of bending the selection ray, researchers have also examined overcoming occlusions by bending the camera rays used to create the visualization presented to the user. One prior work facilitates occlusions through a two-viewpoint multiperspective visualization that chooses a secondary viewpoint to optimize the image footprint of selection candidates. The larger footprints allow for easier selection by increasing the solid angle of possible selection rays (Wang et al., 2021). In another prior work, a multiperspective visualization was designed to alleviate the viewpoint disparity between an instructor and a trainee, such that the trainee can adopt the instructor's view of the workspace, while still seeing the instructor's avatar at the correct location (Wang et al., 2020). The result is that the trainee can see any part of the workspace to which the instructor is pointing, as long as the instructor sees it while being able to naturally turn to the instructor for effective non-verbal communication.

We leverage the higher accuracy of current depth cameras to use both the dominant plane of the workspace and the detailed geometry of the workspace objects. We achieve the alignment of the physical and virtual workspace by pre-setting the initial pose of the trainee (AR HMD), that is, the relative pose of the AR HMD and the physical workspace, and applying the relative pose to the virtual camera and virtual workspace in advance. The alignment of two spaces was accurate to around 5–10 cm each time, and through manual fine-tuning, the accuracy can reach around 1-2cm. Our system acquires not only the workspace but also the trainee, whom the expert can see during live collaboration. However, our system is designed to allow for asynchronous collaboration, and our goal is to allow for annotations that are sufficient to convey instructions clearly, without a live expert.

Our work focuses on a different application, that of object manipulation, but the visual eloquence and occlusion avoidance concerns noted and addressed in the context of navigation are of concern in our context as well, as they are fundamental to the problem of providing visual guidance through AR. In our context, rendering an arrow semi-transparently as it crosses through occluding objects is both challenging, as it requires accurate geometry, and potentially confusing, as it breaks the line into several pieces. Instead, we aim to disocclude by bending the arrow to clear the occluder, which not only indicates the desired position but also hints at a collision-free trajectory of the target object for it to be placed in the desired position. Whereas in the case of a search task highlighting the object might be sufficient, in our object manipulation task finding the target is just the first step, and conveying to the user the desired object pose has to rely on graphical annotations that go beyond manipulating the appearance of the target.

The idea of creating occlusion-free arrow annotation has been discussed for a long time. Moreover, it might be more natural and versatile for an expert to convey instructions through direct drawing rather than generating occlusion-free arrow annotations. However, the occlusion-free arrow

annotation's automatic generation is necessary for implementing the batch instruction method proposed in this study. Although there have been some studies on virtual replicas, the existing studies have created replicas by pre-modeling. The pre-modeling approach provides a more complete and high-fidelity representation of physical objects, but it requires coverage of all required objects and cannot instruct objects that are missing replicas. Therefore we propose a compromise approach that enables real-time guidance when objects lack virtual replicas. Furthermore, unlike most research on MR remote collaboration, our work focuses on a different application, that of object manipulation. Our work demonstrates time savings for longer sequences of instructions, i.e., 10 instructions executed asynchronously, and strengthens the eloquence of rotation instructions by showing the object in the new pose to provide real-time, detailed visual feedback to the user as they manipulate the object.

3. Methods

We describe an approach for effective and efficient visual instruction in a remote collaboration context where a trainee has to perform object manipulation tasks in their workspace under the real-time (synchronous, on-line) or preauthored (asynchronous, offline) guidance of a remote expert. Figure 1 illustrates the strengths of *AVICol* in terms of alleviating occlusions, of providing a rich visual description of the desired state of the workspace, and of adapting instructions to the actual state of the workspace to allow for the robust asynchronous execution of a sequence of instructions as a single batch. We also refer the reader to the video accompanying our article.

We first describe the mixed reality system setup that enables the remote collaboration (Section 3.1). Then we describe our contributions of occlusion-aware visual instruction for object translation (Section 3.2), of realistic future state visual instruction for object rotation (Section 3.3), and of adaptive sequences of visual instructions for asynchronous collaboration (Section 3.4).

3.1. Mixed-reality system setup

The system setup is shown in Figure 2. The trainee stands in front of their physical workspace and manipulates objects based on visual instructions shown through an optical see-through augmented reality head mounted display (AR HMD). The geometry and color of the workspace and of the trainee is acquired in real time with a 4×2 RGBD camera acquisition rig. A point-based geometry plus color model of the workspace and trainee is transmitted to the remote expert site, where the expert visualizes it using a virtual reality head mounted display (VR HMD). The expert authors visual instructions using a virtual laser pointer paradigm to select 3D workspace points and objects, to draw arrows, or to copy, paste, and orient 3D objects (see Section 3.3). The instructions are transmitted to the local trainee site, where they are rendered for using the trainee's AR HMD. In addition, we achieve the alignment of the physical and virtual workspace by pre-setting the initial pose of the trainee(AR HMD), that is, the relative pose of the AR HMD and the physical workspace, and applying the relative pose to the virtual camera and virtual workspace in advance. The alignment of two spaces was accurate to around 5–10 cm each time, and through manual fine-tuning, the accuracy can reach around 1–2 cm.

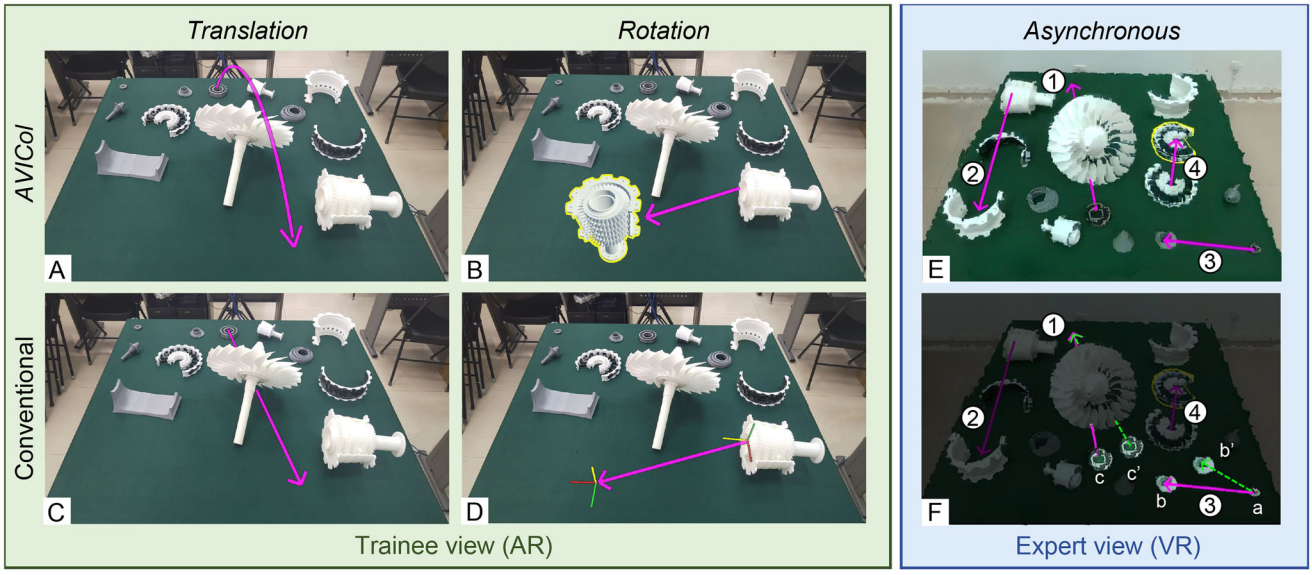


Figure 1. AR trainee interface (left panel): comparison between our *AVICol* approach and a conventional approach for providing visual instructions for object translation and object rotation. *AVICol* bends the translation arrow upwards to clear the occluding object (a), which hides the arrow with the conventional approach (C). *AVICol* accurately conveys the desired pose of the object to be rotated with a rendering of the geometric model of the object (B), as opposed to the cryptic coordinate system axes visualization of the conventional approach (D). *AVICol* instruction adaptation in support of asynchronous collaboration (right panel): sequence of four instructions authored by the expert (E), adapted in real time as they are executed asynchronously by the trainee (F, where the workspace is darkened for illustration clarity). Instruction 1 is adapted from c to reflect the actual position c′ of the object to be translated; instruction 3 asks the user to stack the object from a onto the object at b, which is updated to b′.

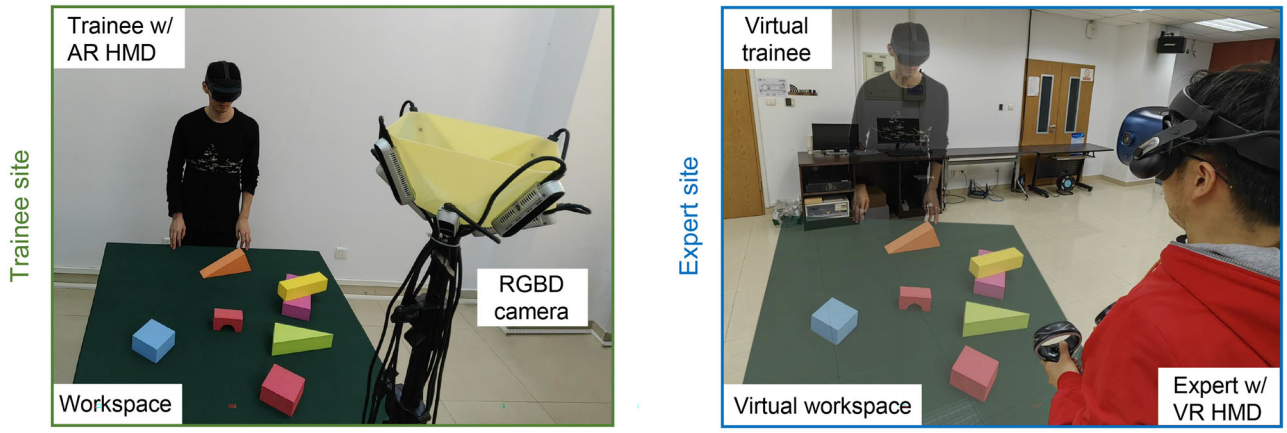


Figure 2. Mixed reality remote collaboration setup between a trainee (left) and an expert (right). The trainee manipulates objects in their workspace. The workspace is acquired with an array of RGBD cameras. The color and depth workspace data is transmitted to the remote expert. The workspace and trainee are shown to the expert through a VR HMD. The expert annotates the workspace to provide visual instruction shown to the trainee through an AR HMD.

One important design consideration is where to place the acquisition rig and where to place the expert in relation to the workspace and the trainee. Some collaboration scenarios do benefit from the expert and trainee having similar viewpoints. For example, expert surgeons will stand side by side with their trainee surgeon to illustrate complex surgical instrument manipulations in a way that avoids the left/right mirroring introduced by opposite viewpoints. In other collaboration scenarios, the expert stands to benefit more from seeing not only the workspace but also the trainee, which is easier done with an acquisition viewpoint opposite the trainee. Furthermore, in the *remote* collaboration context, providing the expert with the trainee's viewpoint is challenging because that requires placing the acquisition device at the trainee's location, which is infeasible since the acquisition device would encumber the trainee's workspace and the trainee would block the acquisition device's line of sight to the workspace.

To meet the hard constraints of avoiding workspace encumbrance and acquisition line of sight blockage, and to allow for the expert to see the trainee, we have opted for a configuration where the acquisition rig is opposite the trainee with respect to the workspace. The acquisition of the workspace in both geometry and color does support offsets between acquisition and visualization viewpoints. Therefore the remote expert can make any movement from the starting position to change his viewpoint to get an independent point of view. However, reprojecting the acquired workspace and trainee data to a visualization viewpoint that is substantially different from the acquisition viewpoint does result in a low quality visualization due to occlusions and to low depth resolution (Figure 3(B), the corresponding real scene is Figure 3(A)). We resort to showing the expert the workspace *and the trainee* from a viewpoint similar to that of the acquisition viewpoint, which provides a high visualization quality conducive to the situational awareness needed for effective collaboration (Figure 3(D), the corresponding real scene is Figure 3(C)).

3.2. Occlusion aware visual instruction

Given an object that has to be translated from a starting position P_s to an ending position P_e, simply connecting P_s and P_e with a straight-line arrow can provide confusing visual instruction when the arrow intersects other workspace objects, without correctly resolving visibility between the arrow and the workspace geometry. One option is to z-buffer the straight line arrow with the workspace geometry, which provides the correct visual cue of the arrow traversing the 3D workspace, but this comes at the cost of a partial occlusion of the arrow (Figure 1(C)). We take the approach of computing a curved arrow that steers clear of the occluding objects to provide a good indication of the suggested translation, without visibility inconsistencies (Figure 1(A)).

Algorithm 1. Unoccluded curved translation arrow generation.

Input: starting point P_s, ending point P_e, trainee viewpoint V, acquisition viewpoint A, acquired depth buffer ZB_A
Output: Unoccluded curve *finalArrow* from P_s to P_e

ZB_V = PointBasedRender(A, ZB_A, V);
$n = 20$; $step = 1$; $stepScale = 1.2$;
$C_0 = P_s$; $C_1 = (P_s + P_e)/2$; $C_2 = P_e$; $bestVisScore = -1$;
while $bestVisScore \neq n$ **and** $C_1.y \leq ||P_sP_e||$ **do**
  $candArrow = Bezier(C_0, C_1, C_2)$;
  $visScore = 0$;
  for $i = 0; i < n; i{+}{+}$ **do**
    $P_i = PointOnBezierCurve(candArrow, i/(n-1))$;
    P_{ip} = Project(P_i, ZB_V, V)
    if $P_{ip}.z < ZB_V[P_{ip}.x, P_{ip}.y]$ **then**
      $visScore{+}{+}$;
    end if
  end for
  if $visScore > bestV\ isScore$ **then**
    $finalArrow = candArrow$;
    $bestVisScore = visScore$;
  end if
  $C_1.y\ +{=}\ step$; $step\ {*}{=}\ stepScale$;
end while
return *finalArrow*;

Alg. 1 describes the construction of the curved arrow in a way that alleviates occlusions with the 3D space. The algorithm is run for every trainee frame, as occlusions change based on

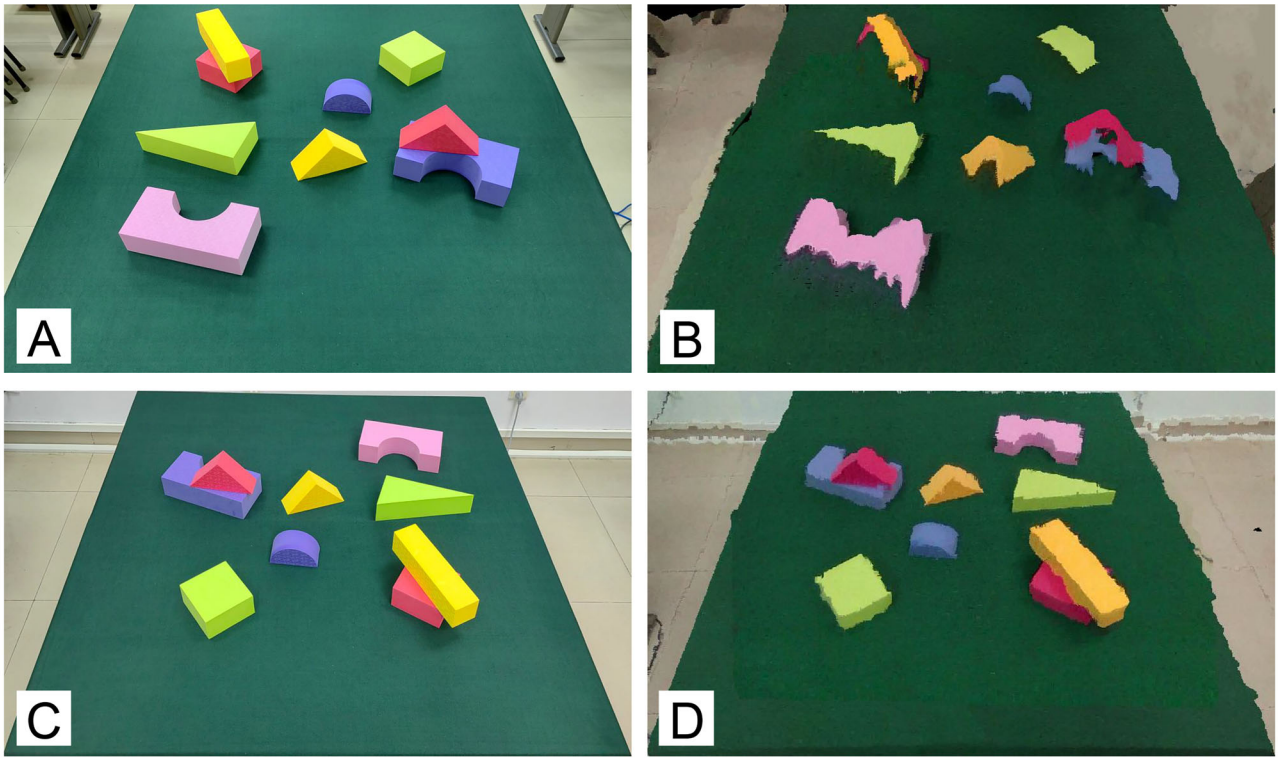


Figure 3. Ground truth visualization (photo) of workspace from trainee viewpoint (A), point-based rendering of workspace from trainee viewpoint (B), ground truth visualization (photo) of workspace from expert viewpoint (C), point-based rendering of workspace from expert viewpoint (D). Since the workspace is acquired from a viewpoint similar to that of the expert, rendering the workspace from the expert viewpoint results in a higher quality visualization (D) compared to rendering it from the trainee viewpoint (B).

the current trainee viewpoint. The input to the algorithm are the starting and ending points P_s and P_e specified by the expert, the workspace depth buffer ZB_A acquired from viewpoint A, and trainee viewpoint V for the current frame.

The algorithm starts out by rendering the workspace depth buffer from the current trainee viewpoint (Line 1). We use a point-based rendering with a constant 3D splat size to obtain the trainee view depth buffer ZB_V which is then used to solve translation arrow occlusion checks.

The algorithm has three parameters whose values are given in Line 2. Parameter n defines the number of points along the arrow where occlusion is checked. In all our experiments we have used $n = 20$ which allows estimating occlusion with sufficient acuity while keeping the number of occlusion checks low, as needed for computational efficiency. The parameter *step* defines the vertical offset by which each iteration of the algorithm raises the candidate arrow above the workspace for it to escape occlusions. *step* starts at 1 cm and then increases exponentially by a factor given by parameter *stepScale*, e.g., by 20% at each iteration.

The algorithm starts from a straight line arrow (Line 3) and then investigates iteratively arrows that have increasing out of plane curvature (Lines 4–19), see Figure 4. An arrow is modeled as a quadratic Bézier curve with three control points C_0, C_1, and C_2. The first and last control points C_0 and C_1 are always P_s and P_e. The middle control point C_1 starts out as the midpoint of line segment P_sP_e (Line 3) and then raises above the workspace. The algorithm keeps track of the best candidate arrow in terms of visibility through the variable *bestVisScore* whose value equals the number of points along the arrow that are visible, initialized to −1 (Line 3).

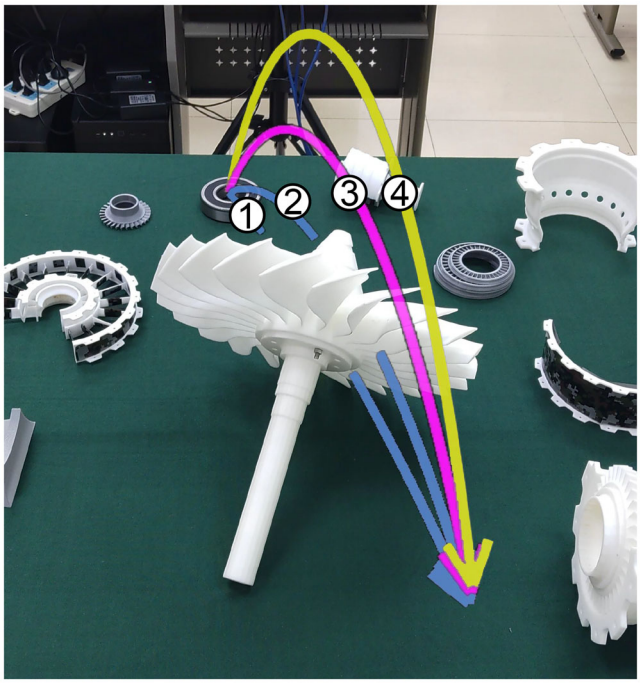


Figure 4. Illustration of candidate arrows for Alg. 1. Arrows 1 and 2 are partially occluded, arrow 3 is the first one to curve sufficiently to clear the occluding object. The algorithm returns arrow 3 (and does not even consider arrow 4).

The algorithm iterates while no arrow has been found with all of its points visible and while the vertical displacement *step* of the arrow does not exceed a maximum value (Line 4). We define the maximum value as the length of the segment P_sP_e. For example, an arrow that has to connect

starting and ending positions that are 50 cm apart should not curve more than a height of 50 cm.

The algorithm returns the best arrow it found (Line 20).

3.3. Realistic target state visual instruction

Instructions like those for the rotation of an object stand to benefit from richer visualizations. Indeed, a visualization that relies on arrows and coordinate system axes to convey such an operation is difficult to understand and execute accurately (Figure 1(D)). Once the trainee moves the object away from the original position and orientation, the trainee has to remember the axes in the initial state to be able to interpret correctly the destination state. Furthermore, abstract rotation visual instructions based on arrows and coordinate systems are not only difficult to execute but also to author.

We propose visual instructions based on a realistic depiction of the target state with the goal of supporting fast and accurate authoring and execution. The pipeline for authoring such instructions is illustrated in Figure 5. The pipeline provides two options (blue and green arrows between pipeline stage illustrations).

The first two stages (A) and (B) are the same for both options. In stage A, the expert indicates the object to be manipulated by creating a selection polygon one vertex at the time to define a loose-fitting bounding box of the object. The expert uses a virtual laser attached to the tracked handheld controller. The intersection point between the virtual laser beam and the workspace depth buffer is computed by sliding a point along the laser beam from near to far until the point becomes hidden. The expert creates a selection polygon vertex using a controller button.

In stage B, the selection is refined from the loose bounding polygon to the subset of the workspace point cloud that belongs to the object. The points that belong to the object are segmented from all points inside the bounding polygon. This is done with a filter that keeps only the bounding polygon points that are above the planar workspace table and whose color and depth is sufficiently similar to the mean value over all points, i.e., within three standard deviations. Once the object point cloud is determined, the pipeline proceeds in one of two ways.

If the geometric model of the object *is not* available, the expert relies on the object's point cloud representation to indicate the object's desired pose (C). If the geometric model of the object *is* available, the expert selects the model from a gallery of objects, manually, with the virtual laser interaction paradigm. The approximate point could representation of the object is replaced with the complete and high-fidelity representation provided by the object's geometric model. The initial position of the geometric model is the same as the average position of all points in the point, and the initial principal axis orientation is the same as the longest axis of the point cloud(including only x, y, z). The expert defines the desired pose by grabbing, translating, and orienting the object. We rely on a conventional on—handle manipulation, i.e., handheld controller trigger button to grab, and then fist translation and rotation to define the pose (Grandi et al., 2019). Once the object point cloud is replaced by its geometric model, the object is visualized at interactive rates with high-fidelity and without occlusion errors (E).

3.4. Adaptive sequences of visual instructions

To increase the efficiency of expert/trainee communication, to allow an expert to provide guidance simultaneously to multiple trainees, and to allow for asynchronous expert/trainee collaboration, we have devised a procedure for authoring and executing a sequence of multiple instructions.

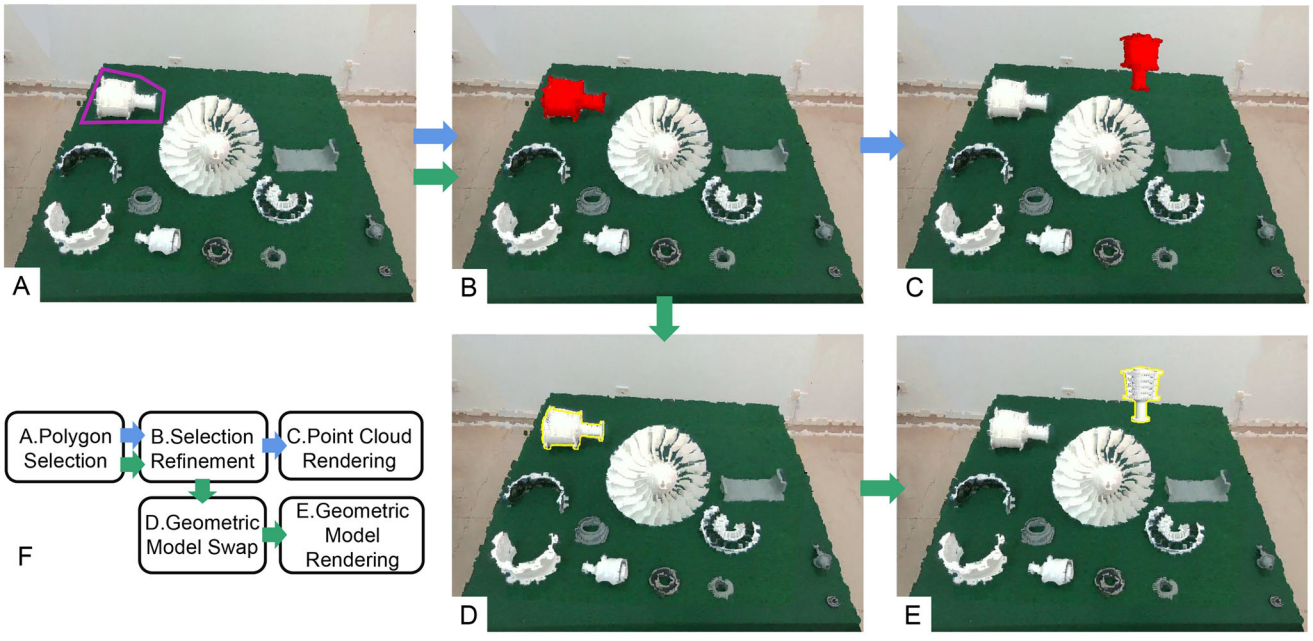


Figure 5. Pipeline for realistic target visual instruction (F), when the geometric model of the object is not available (A–C), and when the geometric model of the object is available (A,B,D,E).

The expert authors multiple instructions, one at the time, in chronological order. The instructions are authored in a pre-acquired color and depth point cloud model of the workspace. In other words, the workspace is not acquired in real time, which allows for the trainee to submit their workspace for annotation asynchronously. The trainee executes the sequence of instructions one at the time. The instructions that remain to be executed are updated based on the actual execution of earlier instructions. This maintains the accuracy of the instructions despite small variations in the execution of the earlier instructions. This robustness is particularly important in the case of long sequences of instructions where errors can accumulate making future instructions unusable if they do not adapt to the actual state of the workspace.

Figure 6 illustrates the need for adaptivity to support accurate sequences of instructions. The top row (A–C) shows the case when the starting position of a second instruction depends on the ending position of a first instruction. The first instruction prescribes moving the object to an intermediate location and then the second instruction prescribes moving it to a final position (A). If the trainee moves the object to a slightly different position, the starting point for the arrow of the second instruction is incorrect (B). With our approach, the second instruction is adapted for the arrow to start correctly at the center of the object in its actual intermediate position (C). The bottom row (D–F) shows the case when the ending position of a second instruction depends on the ending position of a first instruction. Here the two instructions (1 and 2 in D) ask the trainee to move the blue block to a new location and then to stack the pink block on top of the blue block. Since the user moves the blue block at a position slightly different from the one prescribed by the first instruction, the ending point of the arrow for the second instruction does not coincide with the actual center of the object (E). With our approach (F), the ending point of the arrow for the second instruction is adapted to track the center of the object moved by the trainee. This way the second instruction correctly indicates the desired stacking of the pink block on top of the blue block.

We adapt the current instruction k to the actual state of the workspace (1) by computing the actual position A of the object acted upon by instruction $k-1$, and (2) by adapting the arrow of instruction k to A. (1) The actual position of the object acted upon is computed by subtracting the depth buffer of the workspace after instruction $k-1$ and from the depth buffer after instruction k. The depth buffers are computed from a top view, perpendicular to the workspace table, which minimizes occlusions. To compute the difference robustly, we average 30 depth frames and we filter out differences below a threshold of 1 cm, a parameter established empirically. The samples with non-zero depth differences define region R. The center of the axis aligned bounding box of R approximates the center of the object at its actual position after instruction $k-1$. (2) Once the actual position of the object is known, the starting or ending point of the arrow of instruction k is adapted accordingly, with the updated arrow endpoints being passed to Alg. 1 for occlusion free visualization.

Our depth difference approach to analyzing the workspace provides robustness with deviations from the prescribed state of previous instructions. For example, it does not matter how far the trainee places an object from the prescribed location–*AVICol* will find the object and will use the incorrect position as the starting point for the next instruction, so the inaccuracy of the previous instruction execution does not compromise future instructions. Similarly, if the trainee moves the wrong object, *AVICol* can detect that the incorrect object was manipulated and that the correct object was left at its original location. Depth differences are ineffective if the object manipulated is placed in the occlusion shadow of another object. Such cases are rare since we use a depth camera with multiple acquisition viewpoints and since we reproject the workspace geometry to obtain top-view depth buffers. Nonetheless, if an object is moved under the overhang of another object, *AVICol* loses the object.

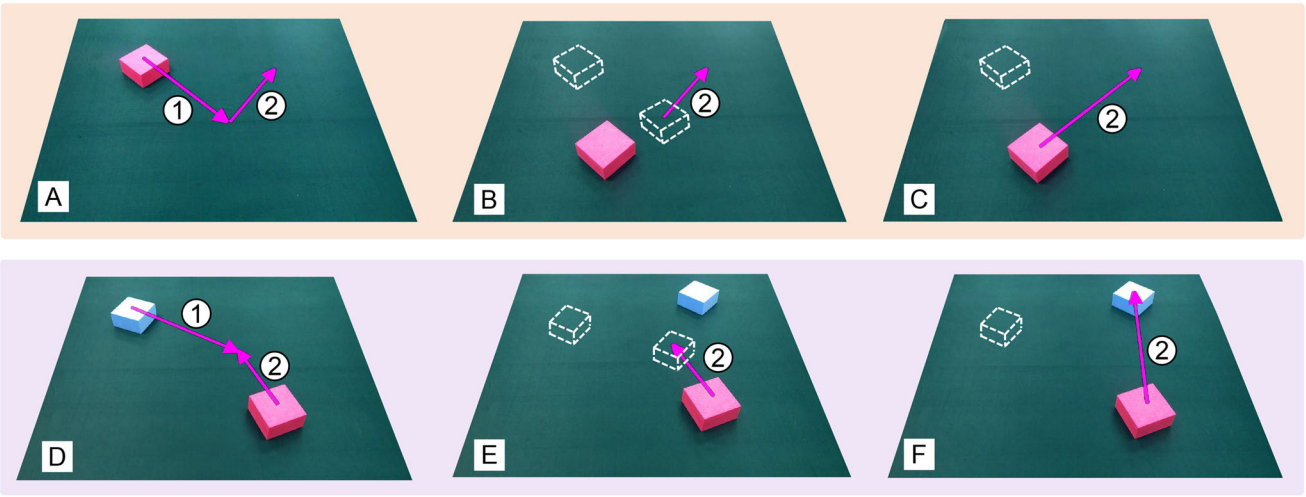


Figure 6. The need for instruction adaptivity to the actual state of the workspace. Top: the starting point of the second arrow (2) has to be modified from its initial position (A,B) to coincide with the actual position of the pink block after the execution of the first instruction (C). Bottom: the ending point of arrow 2 has to be modified from its initial position (D,E) to coincide with the actual position of the blue block (F).

4. User study

We have conducted a user study to investigate the potential advantages of our *AVICol* approach. The study is designed around the following research hypotheses:

- *Hypothesis A*. The *AVICol* occlusion aware visualization of instructions leads to more accurate translation instruction executions and faster task completion times compared to a visualization that does not take into account occlusions.
- *Hypothesis B*. The *AVICol* realistic visualization of the desired pose of a workspace object leads to faster and more accurate rotation instruction executions compared to conveying the rotation instructions verbally, and compared to a visualization based on local coordinate systems.
- *Hypothesis C*. The *AVICol* batch authoring and execution of a sequence of 10 instructions leads to faster authoring and execution times compared to a sequential authoring and execution of individual instructions, one at the time.

We investigate each of the three research hypotheses in a separate experiment. Experiment 1 investigates *Hypothesis A* in the context of object translation. Experiment 2 investigates *Hypothesis B* in the context of object rotation. Experiment 3 investigates *Hypothesis C* in the context of the asynchronous authoring and execution of a sequence of 10 instructions. Experiment 3 also serves as a summative evaluation of the occlusion awareness and realistic visualization elements of *AVICol*, put together in these longer instruction sequences.

4.1. Participants

We have recruited $N = 24$ participants from our undergraduate and graduate student population who were recruited randomly within schools and had different professional backgrounds. Participants were divided randomly into 12 pairs, with one participant serving as the trainee and one serving as the expert, assigned randomly. Any participant can serve as expert since the expert is provided a script of instructions to communicate. The average participant age is 24 years, 10 participants self-reported as women, 20 had prior experience with virtual reality and augmented reality applications, none reported suffering from balance disorder or color blindness conditions, and all had good or corrected vision. The same participants were used in all three experiments, in a single session. A session lasted ∼1 hr and participants had 5 min breaks between experiments. Participants performed experiments in the same order, as Experiment 3 subsumes Experiments 1 and 2. The user study was awarded and approved by the Biology and Medical Ethics Committee of Beihang University.

4.2. Implementation

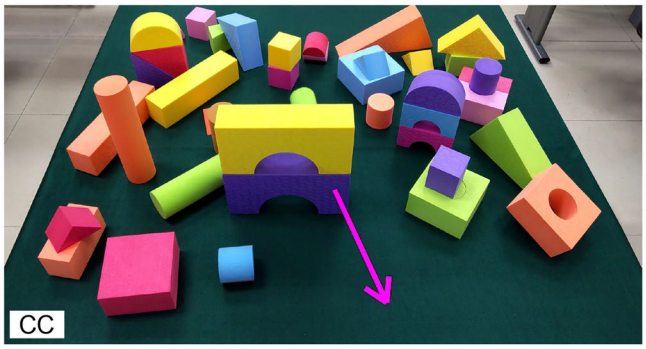
We used a Microsoft HoloLens 2[1] for the trainee's AR HMD and an HTC Vive Cosmos[2] for the expert's VR HMD. The trainee workspace was acquired with a RealSense D455 depth camera. The software implementation relied on Unity[3] Internet communication relied on the Mirror[4] Unity plugin. The average amount of color and depth point cloud data transmitted for each frame is 8×3000 KB. All data is transmitted through the same LAN connected by network cable and the transmission delay of data is ∼300 ms. The time taken to reconstruct a frame through point cloud data is about 30–40 ms. The frame rate of the expert site is about 20–30 fps, and the frame rate of the trainee site is about 90 fps. The point cloud is a set of RGBD samples, which are unprojected to 3D at the client using the known acquisition camera parameters. These are fixed external parameters preset for each camera that are used to stitch the point clouds from eight cameras into a single point cloud. The points had constant 3D size and were projected onto the user's view and splatted as 2D points with a footprint derived from the distance to the viewer. All instruction data authored by the expert will be synchronized and rendered to the trainee in real time through the Mirror plug-in.

4.3. Experiment 1: Translation tasks

4.3.1. Experimental design

Participants were divided into 12 pairs randomly, with one participant serving as the trainee and one serving as the expert. The trainee was in front of a table on which there were blocks (Figure 2) to be manipulated under the instruction of the remote expert. The trainee wore an optical see-through AR HMD. The expert was located in a different room, standing in an empty space, wearing a VR HMD. The workspace was acquired in real time with an array of depth and color cameras, connected to a workstation. The workstation sent the color and depth data to the expert over the internet. The expert's VR HMD receives the workspace data and renders it in real time. The expert authors a visual instruction for a single object translation. The visual instruction is sent to the trainee over the internet where it is rendered by the trainee's AR HMD.

The visual instruction is rendered in one of two ways. In the control condition (CC), the translation arrows are rendered as straight lines with z-buffering (Figure 7 CC). We did not use the method of alleviating occlusion by rendering directly above other objects, although this method is easier to implement. Because this method of alleviating occlusion is achieved through incorrect depth relationships, which sometimes causes further problems, such as the inability to determine the actual position of the instr. There have been many works in the past that have also made efforts to provide a correct occlusion relationship between virtual and real objects (Gimeno et al., 2013). In the experimental condition (EC), the translation arrows are rendered as curves to alleviate occlusions using our algorithm (Alg. 1), as shown in Figure 7 EC. A participant serving as a trainee performs tasks under both conditions. This within-subject design allows for greater statistical power with fewer participants. The order of conditions is counterbalanced. A participant



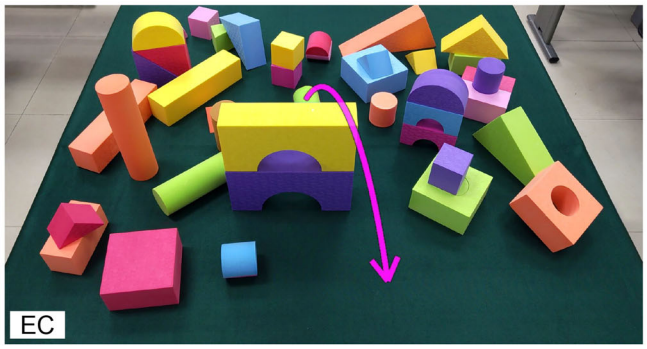


Figure 7. Control condition (CC) and experimental condition (EC) for Experiment 1. The user has to translate the green cylinder to the front, with the translation arrow occluded in CC and occlusion-free in EC.

serving as a trainee never serves as an expert, and a participant serving as an expert never serves as a trainee.

4.3.2. Tasks

For each trainee-expert pair, the expert issues eight object translation instructions, in each condition, for a total of 16 trials. For each trial, the expert was informed of which translation task to communicate to the trainee using a 3D rectangle texture mapped with a screen capture of the desired translation. The same eight translations were used for each trainee-expert pair, in the same order. Once the trainee places the object in the new location, the current trial ends, without iterative refinement of the object location. After each trial, the expert is asked whether the correct object was manipulated and whether the new object location is correct. Errors are committed when either the starting or the ending point of the arrow is occluded from the trainee. When the starting point is not visible, the incorrect object might be manipulated. When the ending point is not visible, the destination location might be off substantially. Therefore, all experts can easily make consistent decisions regarding translation correctness. The experts are not asked to estimate the translation error, e.g., in cm, and then to make the subjective call on whether the translation error is acceptable or not. Our goal is for the trainees to understand the intent of the experts. For example, the expert authorized the translation of object A to object B, but the trainee translated the object A to another object C. We told the expert the judgment criteria before the experiment, and the expert could easily determine such translation errors.

4.3.3. Data collected

For each trial, we recorded an object manipulation error when either the wrong object was manipulated, or its location was incorrect. We also recorded the total time needed for the expert to author the instruction and for the trainee to execute it.

4.3.4. Data analysis

For each participant/expert pair, we computed average selection accuracy, location accuracy, and completion time over all trials, for each condition. We checked the data normality assumption using the Shapiro-Wilk test (Shapiro & Wilk, 1965). Normally distributed data was analyzed using repeated measures ANOVA (Gelman, 2005). When the normality assumption did not hold, the analysis was performed using Wilcoxon's signed-rank test (Rey & Neuhäuser, 2011). In addition to the p-value of the statistical test, we also estimated the size of the effect using Cohen's d (Frees & Kessler, 2005). We derive a qualitative estimate of effect size using the following effect size thresholds (Cohen, 2013): *Huge* ($d > 2.0$), *Very Large* ($2.0 > d > 1.2$), *Large* ($1.2 > d > 0.8$), *Medium* ($0.8 > d > 0.5$), *Small* ($0.5 > d > 0.2$), and *Very Small* ($0.2 > d > 0.01$).

Table 1. Object translation error rates. Significant difference is marked with asterisk.

Condition	Avg ± std. dev.	(CC-EC)/CC	p	Cohen's d	Effect size
EC	0.00 ± 0.00				
CC	4.55 ± 6.01	100.0%	0.02*	1.07	Large

4.3.5. Results and discussion

The object translation error rates are shown and compared in Table 1 and in the graph in Figure 9(a). In all graphs, statistically significant differences are indicated with an asterisk. There were no errors recorded in EC, and a 4.55% error rate in CC (column 2), which corresponds to a 100% error reduction (column 3). Examples when participants are likely to make errors are shown in Figures 7 and 8. The difference in manipulation error rates is significant (column 4) and the effect size is large (column 6). Throughout this article, significance is indicated with an asterisk by the p-value. The task completion times are shown and compared in Table 2 and in the graph in Figure 9(b). *EC* has significantly shorter completion times than CC, and the effect size is large. We conclude that alleviating occlusions helps the user perform object translation tasks faster and more accurately, in support of *Hypothesis A*.

4.4. Experiment 2: Rotation tasks

4.4.1. Experimental design

We used the same experimental design as for Experiment 1. The only difference is that now the expert authored and the trainee executed instructions for object rotation (and not translation. The translation distances included in the tasks

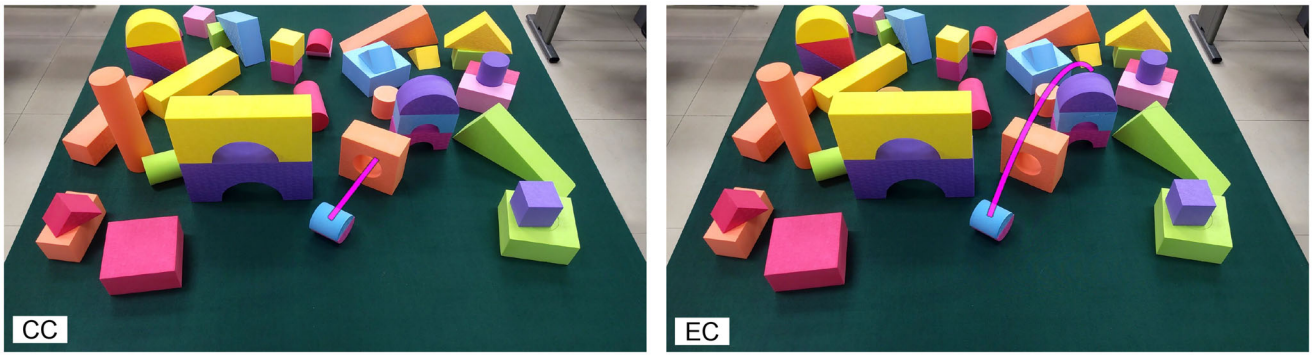


Figure 8. Control condition (CC) and experimental condition (EC) for Experiment 1. The arrow is severely occluded in CC which leads the participant to place the object at an incorrect position. Enough of the arrow is visible in EC for the participant to place the object correctly.

Table 2. Object translation authoring + execution times, in seconds. Significant difference is marked with asterisk.

Condition	Avg ± std. dev.	(CC-EC)/CC	*p*	Cohen's *d*	Effect size
EC	132.83 ± 11.52				
CC	143.76 ± 12.41	7.6%	0.003*	0.91	Large

are very small, all less than the size of objects, and consistent across different control conditions. Therefore we ignore the effect of translation on the task.), and we used two control conditions.

In one control condition (CC_1) the expert provides verbal instructions that are transmitted to the trainee in real time using an audio communication channel. In a second control condition (CC_2), the rotation instructions are provided visually by rendering the local coordinate system axes of the object to be manipulated (Figure 10 CC2). In the experimental condition (EC), the rotation instructions are implemented with our approach by showing the object in the desired pose as a point cloud (Figure 10 EC).

4.4.2. Tasks

For each trainee-expert pair, the expert issues five object rotation instructions, in each condition, for a total of 15 trials. Like for translations, the expert was informed of the rotation to communicate to the trainee using a 3D rectangle texture-mapped with a pre-recorded screen capture of the rotation. Once the trainee places the object in the new orientation, the current trial ends, without iterative refinement. We did not use objects invariant under rotation, such as cylinders. For symmetrical objects, such as a cube, any of the poses was accepted (four poses for the cube).

4.4.3. Data collected

For each trial, we recorded the object rotation error as the square root of the sum of the squares of the angular errors for each of the three local coordinate system axes of the object (i.e., Euclidian distance in the 3D rotation space). Whereas for translation, the tip of the arrow conveys the destination location accurately as long as it is visible, for rotation, the axes, even when visible, might not convey the desired pose with great accuracy. Therefore we measure the rotation error objectively, on a continuous scale, without relying on the experts. We also recorded separately the time needed for the expert to author the instruction, and the time needed for the trainee to execute it.

4.4.4. Data analysis

The data was analyzed like for Experiment 1, with the only difference being that in Experiment 2 there were two control conditions, i.e., CC_1 and CC_2, each of which was compared to the experimental condition EC.

4.4.5. Results and discussion

Table 3 and the graph in Figure 9(c), show and compare the times needed by experts to author the rotation instructions across conditions. The experimental condition is significantly faster than either of the control conditions. The difference is larger compared to verbal instruction (CC_1). Table 4 and the graphs in Figure 9(d), show and compare the times needed by trainees to execute the rotation instructions. Again, the experimental condition has a significant advantage over either of the two control conditions, and the difference is larger compared to verbal instruction (CC_1). Table 5 shows and compares the instruction execution errors across conditions. EC is significantly more accurate than either CC_1 or CC_2. This result is in agreement with prior work that has also noted the difficulty in posing objects accurately using short local coordinate system axes visualizations (Andersen & Popescu, 2020). We conclude that rotations are performed faster and more accurately with *AVICol* compared to conveying instructions verbally, and to conveying instructions through local coordinate system visualizations, in support of *Hypothesis B*.

4.5. Experiment 3: Translation and rotation sequences

4.5.1. Experimental design and tasks

We used the same general experimental design as for the other two experiments, with two differences. One difference is that the collaboration now entailed a sequence of 10 instructions for object manipulation. Out of the 10 instructions, six ask the trainee to translate an object, while four ask the trainee to translate and rotate an object. Like before, the expert was informed of the instructions to prepare for the trainee using a texture-mapped billboard. The instructions are sequenced in chronological order, whit each

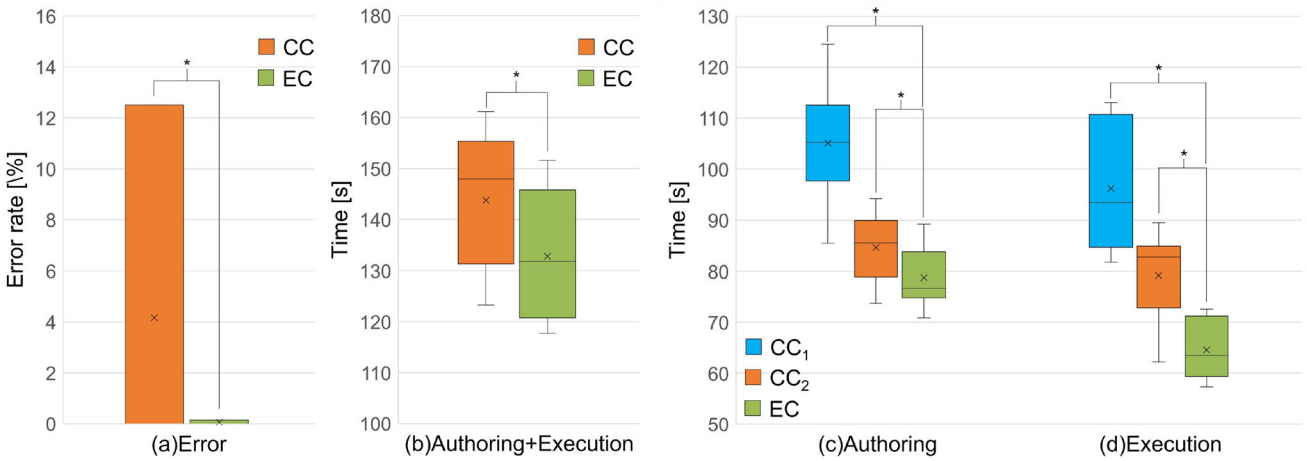


Figure 9. Left: Error rates (a) and authoring + execution times (b) box plots for object translation, for the control and experimental conditions. Right: Rotation instruction authoring (c) and execution times (d) box plots for the control and experimental conditions.

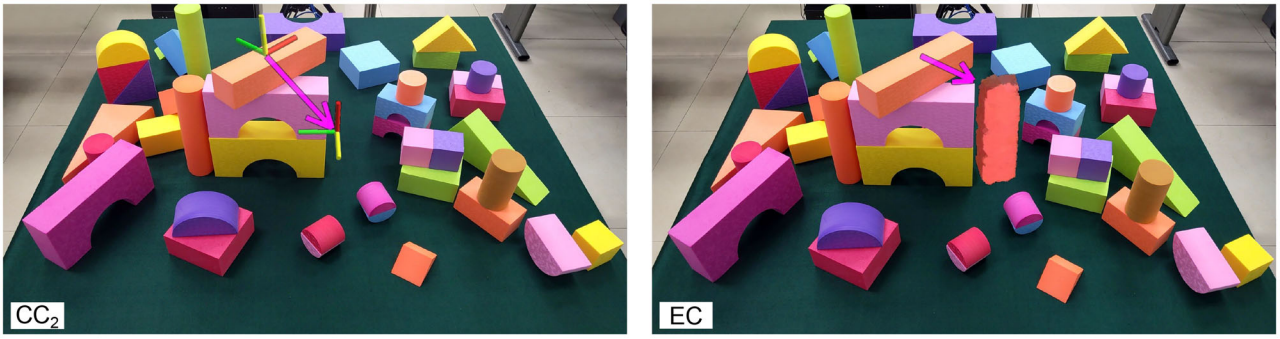


Figure 10. Second control condition (CC2) and experimental condition (EC) for Experiment 2. The participant is asked to place the long orange block vertically adjacent to the two bridge blocks, which is more clearly conveyed in EC by showing the block in the desired position, compared to CC2 where the desired position and orientation are conveyed with a coordinate system axes visualization.

instruction starting from workspace state at the end of the previous instruction (Figure 1(E)). A second difference is that we used two workspaces: the blocks workspace (Blocks) used in Experiments 1 and 2 and shown in Figures 7, 8, and 10, and the complex parts workspace (Parts) shown in Figure 1.

There were two conditions. In both conditions, the instructions were executed using our approach for translation and rotation. In one condition (SC), the instructions of a sequence were authored and executed synchronously, one at a time, with live communication between the expert and trainee. In a second condition (AC), the instructions of a sequence were authored and executed asynchronously, all at once, without any live communication between the expert and trainee. The instructions are not rendered to the trainee all at once, but only one at a time in the expert authoring order. After self-identifying completion of one, the trainee manually switches to the next one (*via* a hololens button).

Table 3. Rotation instruction authoring times (VR), in seconds. Significant difference is marked with asterisk.

Condition	Avg ± std. dev.	$(CC_i\text{-}EC)/CC_i$	p	Cohen's d	Effect size
EC	78.83 ± 5.62				
CC_1	105.11 ± 12.35	25.0%	<0.001*	2.71	Huge
CC_2	84.65 ± 6.90	6.9%	0.003*	0.91	Large

Table 4. Rotation instruction execution times (AR), in seconds. Significant difference is marked with asterisk.

Condition	Avg ± std. dev.	$(CC_i\text{-}EC)/CC_i$	p	Cohen's d	Effect size
EC	64.56 ± 5.81				
CC_1	96.20 ± 12.20	32.9%	<0.001*	3.31	Huge
CC_2	79.16 ± 9.12	18.4%	<0.001*	1.91	Very large

Table 5. Rotation instruction execution errors, in degrees. Significant difference is marked with asterisk.

Condition	Avg ± std. dev.	$(CC_i\text{-}EC)/CC_i$	p	Cohen's d	Effect size
EC	10.18 ± 5.75				
CC_1	17.46 ± 4.69	41.7%	<0.001*	1.44	Very large
CC_2	14.09 ± 5.46	27.8%	0.036*	0.72	Medium

4.5.2. Data collected

Figure 6 shows that failing to adapt instructions to the state of the workspace makes it impossible to execute subsequent instructions correctly. For example, moving the pink cube to its old destination will fail to stack up the cubes as intended. Experiment 3 focuses on quantifying the time advantage brought by the asynchronous execution of instruction sequences, which is made possible by *AVICol*'s ability to adapt instructions. For this, we record the time needed by the expert to author the sequence and the time needed for the trainee to execute the sequence. The final states of the workspace were checked for correctness, and all final states

were correct in both the asynchronous and synchronous conditions. Furthermore, we collected usability data using a prior art questionnaire (Kim et al., 2015) that has six questions with answers on a 9-point Likert scale, as well as task load data using the NASA TLX (Hertzum, 2021) questionnaires.

4.5.3. Data analysis

The data was analyzed like for Experiment 1 (two conditions, significance level threshold of $p = 0.05$).

4.5.4. Results and discussion

The authoring times across conditions and workspaces are given in Table 6 and in the graph in Figure 11(a), and the execution times are given in Table 7 and in the graph in Figure 11(b). Our method is beneficial for both the simple (Blocks) and the complex (Parts) workspaces. The times are significantly longer when the expert has to provide guidance for individual steps of the sequence. By adapting the instructions to the actual state of the workspace, our approach allows concatenating long sequences of instructions and executing them robustly, saving time, and enabling asynchronous collaboration, as needed for example to accommodate different time zones or to allow an expert to assist multiple trainees at the same time. We conclude that using *AVICol*, authoring and executing instructions asynchronously leads to faster authoring and execution times, in support of *Hypothesis C*.

Figure 11, right, graphs the NASA TLX task load scores for the two workspaces, i.e., Blocks (c) and Parts (d), and for the synchronous (SC) and the asynchronous (AC) conditions. For the trainee, there are no significant differences between the two conditions, for either of the two workspaces. This is a positive result since it indicates that the trainee can execute the sequence of 10 instructions alone, with no help from the expert, without the task become more complex. Furthermore, for the expert, the authoring task has a significantly lower load in the asynchronous condition, which is expected, as the expert can author more instructions more quickly when they do not have to also be executed by the trainee. In an absolute sense, all tasks have load values below 35, the threshold customarily used to differentiate between low and high load tasks (Hertzum, 2021).

Figure 12 gives the usability scores which are uniformly high. The only exception is for effectiveness (EF) which is significantly higher for the asynchronous condition (AC) compared to the synchronous condition (SC), confirming that the batch execution is preferred by both the trainee and the expert.

Overall, the user study confirms the advantages of the occlusion awareness and of the realistic visualization capabilities of *AVICol*, both in isolation (Experiment 1 and

Table 6. Instruction sequence authoring times (VR), in seconds. Significant difference is marked with asterisk.

Work-space	Condition	Avg ± std. dev.	(SC-AC)/SC	*p*	Cohen's *d*	Effect size
Blocks	AC	147.54 ± 14.47				
	SC	216.59 ± 21.38	31.9%	<0.001*	3.78	Huge
Parts	AC	151.75 ± 10.53				
	SC	225.42 ± 18.66	32.7%	<0.001*	4.86	Huge

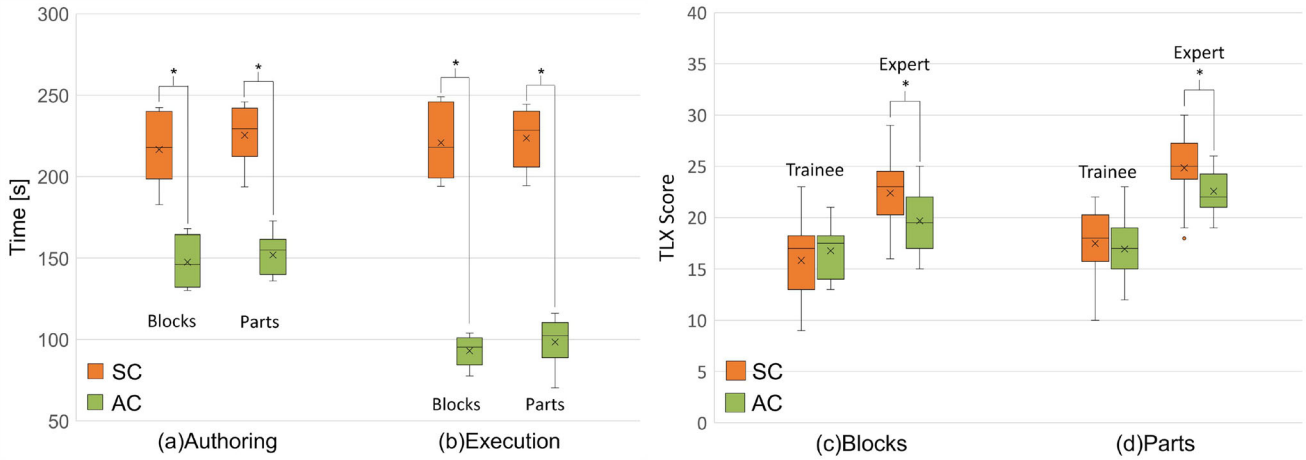


Figure 11. Left: Instruction sequence authoring (a) and execution times (b) box plots for the two conditions and the two workspaces. Right: NASA TLX task load scores for the authoring (expert) and the execution (trainee) tasks, for the two workspaces (Blocks and Parts), and for the synchronous (SC) and asynchronous (AC) conditions.

Table 7. Instruction sequence execution times (AR), in seconds. Significant difference is marked with asterisk.

Work-space	Condition	Avg ± std. dev.	(SC-AC)/SC	*p*	Cohen's *d*	Effect size
Blocks	AC	93.08 ± 7.59				
	SC	220.98 ± 22.18	57.9%	<0.001*	7.71	Huge
Parts	AC	98.41 ± 15.18				
	SC	223.61 ± 18.99	56.0%	<0.001*	7.28	Huge

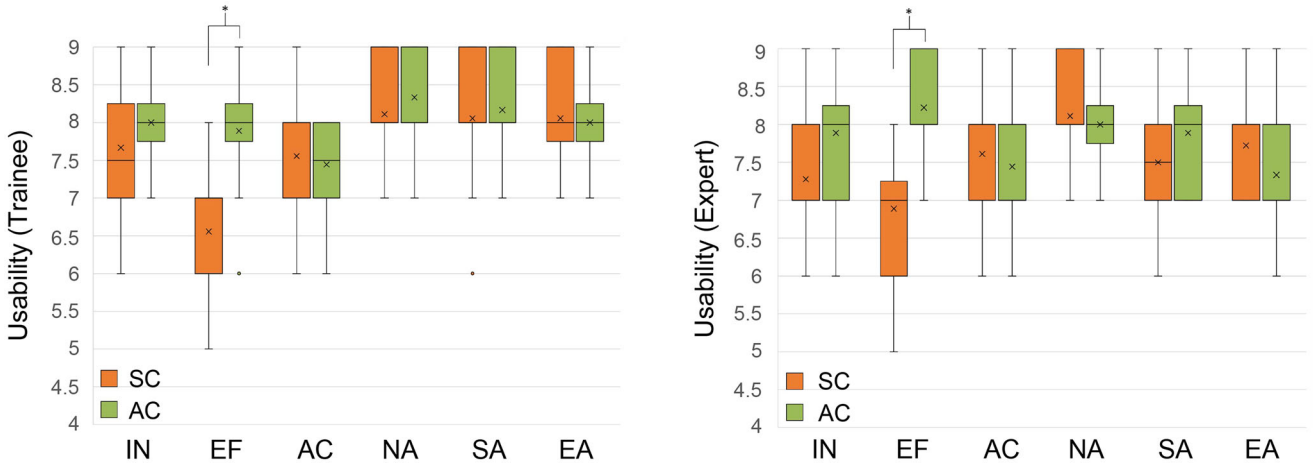


Figure 12. Scores for the six questions of the usability score for the trainee (left) and the expert (right). The questions gauge intuitiveness (in), effectiveness (EF), accuracy (AC), naturalness (NA), satisfaction (SA), and easiness (EA). The highest possible usability value is 9, and the smallest is 0.

Experiment 2), and in conjunction (Experiment 3). The workspace monitoring achieved through real-time depth acquisition enables for visual instructions to be authored and executed quickly and accurately both synchronously, one instruction at the time, and also asynchronously, in batches of 10 instructions. The asynchronous collaboration is possible by relaxing the accuracy requirement for the execution of individual instructions, which in turn is afforded by *AVICol*'s ability to adapt instructions to the current state of the workspace.

4.6. Limitations

Our experiments have several limitations. Our experiments use simple tasks in a table-scale collaboration scenario which are the basic components of more complex tasks, and future work can expand the experimental validation with tasks that involve the accurate concatenation of multiple such basic components, i.e., engine assembly training as shown in 1. And our approach can easily be applied to larger scale collaborations. Our acquisition range is not limited to the table-scale and the algorithms in 3.2 and 3.3 are also not only applicable to the table. But we always attempt to alleviate occlusions by bending the translation arrows up, whereas in more complex scenarios more directions might need to be considered. Also, the algorithm in Section 3.4 needs to acquire a larger range of depth buffers.

Our experiments focus on one-to-one scenarios. Some researches have focused on exploring the one-to-many remote scenarios, including view-sharing techniques (Lee et al., 2020; Marques et al., 2022a), workloads of remote experts (Otsuki et al., 2022; Wang et al., 2022), etc. The topics need to be addressed by the research community in one-to-many scenarios were also discussed (Marques et al., 2022b). Our system can be extended to one-to-many scenarios based on the following facts. Our acquisition rig allows more local trainees to work at the same time. And expert can move freely throughout the acquisition space and author instructions at any location. Authored instructions can also be shared among trainees. In addition, in the experiments, only one visual instruction is rendered to the trainee at a time. It is feasible to visualize multiple instructions to refer to multiple objects at the same time but requires further design. For example, the order of execution between instructions and the mutual occlusion of instructions. Finally, Our study uses one or two control conditions for each of the three experiments. Future work could compare *AVICol* to additional control conditions. For example, one could compare the *AVICol* performance against non-AR conditions, such as providing guidance to the trainee on a nearby monitor, with the trainee having to memorize the instruction and then to execute it from memory on the actual workspace. Future work could also extend the empirical validation on additional workspaces and additional tasks. The workspaces we have investigated are characterized by objects with salient features which are easy to disambiguate. We foresee that the need to alleviate occlusions and to adapt instructions is even greater for workspaces with similar objects where occlusion in approximately executed earlier instructions can lead to errors.

5. Conclusions, limitations, and future work

We have presented *AVICol* a mixed-reality approach for effective and efficient remote visual instruction. Our approach relies on a VR interface for the expert to author instructions, and on an AR interface for the trainee to execute them. Our approach makes three contributions: (1) it curves object translation arrows to alleviate occlusions; (2) it allows depicting the desired state of a rotation instruction with high fidelity by allowing the expert to place a point-cloud or geometric-model copy of the object in the desired pose; (3) it allows the trainee to execute a long sequence of instructions asynchronously, without real-time help from the expert. We have tested our approach in a user study with three experiments that confirm significant accuracy and time advantages of our approach over conveying translations with a straight-line arrow and over conveying rotations with arrow and coordinate system visualizations. Furthermore, the study shows a time advantage of asynchronous over synchronous instruction authoring and execution.

Our approach has several limitations. We always attempt to alleviate occlusions by bending the translation arrows up, whereas simpler solutions might be available by considering bending the arrows laterally. The resolution of the point cloud is limited, which means that distant or small objects are captured at low resolution which lowers the quality of the point-based rendering when a geometric is not available. Even though our depth camera has multiple acquisition viewpoints, acquisition occlusions remain, which truncates the visualization of workspace objects when the viewpoint flips across the workspace from the acquisition viewpoint to the trainee viewpoint. One future option is to arrange multiple pods. The scalability of our system could support the installation of more pods, which can be implemented in both software and hardware simply by copy and reuse. However, this approach needs to take into account cost and rate constraints, and is therefore not adopted for the time being.

Our current implementation focuses on adapting the current instruction based on the state of the workspace after the trainee completes the previous instruction, and it does not check for execution correctness. In other words, our approach focuses on leniency when instructions are not executed precisely, and it does not place a limit on the magnitude of the deviation between the prescribed and the executed instructions. One future work could examine interaction paradigms where the trainee is asked to refine the execution of the current instruction until it meets predetermined requirements.

Whereas *AVICol* tolerates large deviations between intended and achieved positions of the manipulated object, there are certain incorrect object manipulation scenarios from which our current implementation cannot recover. One such scenario is when the user moves the wrong object and then ignores the warning about the error and proceeds with moving a second object. Another scenario is when the user moves multiple objects simultaneously, e.g., one object with each hand, or all workspace objects by bumping the workspace table. Future work could aim to increase the robustness of the system by implementing longer undo sequences, by keeping track of multiple incorrectly executed instructions, and by assisting the user during the undo process. Another approach for increasing robustness is to try to handle a workspace where the positions of many or all objects have been perturbed. This requires mapping the perturbed configuration to the default configuration, thereby porting automatically the instructions to the new configuration. *AVICol* uses an overhead depth buffer analysis for workspace understanding, which objects hiding under other objects can evade. Future work could extend the analysis beyond single states to consider sequences of states that could reveal which object hides where. In addition, our depth difference approach may also be affected by depth variations produced by environmental changes. 3D object registration and tracking techniques may be able to address the above limitations. Firstly, it can provide more accurate object positions and determine the correct start and ending positions of instructions even when the environment changes. And semantics information allows checking for execution correctness, identifying more types of errors as mentioned above. However, this solution is currently limited by the maturity of model-free object tracking technology, tracking stability and other factors, and remains to be studied.

In addition to addressing these limitations, future work could investigate alternative sensor placement for the acquisition of the workspace. For example, modern AR headsets do acquire workspace color and depth, which allows, in principle, acquiring the scene from the trainee viewpoint. However, the viewpoint of the trainee is uncontrollable for the expert, which is a distraction for the expert in some cases. The point cloud information acquired in our experimental scenarios is already enough for the expert to make correct instructions. In the future, we will consider how to take into account the fusion of the two perspectives, so that the expert can obtain more information about the scene and the first AR perspective without causing confusion, give more personalized instructions, and improve the accuracy and efficiency of collaboration. Another direction of future work is to investigate alternative implementations of the AR interface, for example through a video see-through AR HMD, to remove the field of view limitation of the optical see-through AR HMD we have used, or through overhead projection, to bypass the encumbrance brought by AR HMDs whose form factor is still inadequate for comfortable extended use. Finally, future work could investigate deploying our approach in real world expert-trainee collaboration scenarios.

Notes

1. Microsoft HoloLens 2. https://www.microsoft.com/en-us/hololens.
2. Vive Cosmos. https://www.vive.com/cn/product/vive-cosmos/overview.
3. Unity 2020.3.20f. https://unity3d.com.
4. Mirror. https://mirror-networking.com/.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Adcock, M., & Gunn, C. (2015). Using projected light for mobile remote guidance. *Computer Supported Cooperative Work*, *24*(6), 591–611. https://doi.org/10.1007/s10606-015-9237-2

Andersen, D., & Popescu, V. (2020). AR interfaces for mid-air 6-dof alignment: Ergonomics-aware design and evaluation. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (pp. 289–300). https://doi.org/10.1109/ISMAR50242.2020.00055

Bai, H., Sasikumar, P., Yang, J., & Billinghurst, M. (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* (pp. 1–13). https://doi.org/10.1145/3313831.3376550

Chang, E., Lee, Y., & Yoo, B. (2023). A user study on the comparison of view interfaces for VR-AR communication in XR remote collaboration. *International Journal of Human–Computer Interaction*, 1–16. Advance online publication. https://doi.org/10.1080/10447318.2023.2241294

Chang, Y.-C., Wang, H.-C., Chu, H-K., Lin, S.-Y., & Wang, S.-P. (2017). Alpharead: Support unambiguous referencing in remote collaboration with readable object annotation. In *Proceedings of the*

2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (pp. 2246–2259).

Chen, L., Liu, Y., Li, Y., Yu, L., Gao, B., Caon, M., Yue, Y., & Liang, H.-N. (2021). Effect of visual cues on pointing tasks in co-located augmented reality collaboration. In *Symposium on Spatial User Interaction* (pp. 1–12). https://doi.org/10.1145/3485279.3485297

Chen, W., Shan, Y., Wu, Y., Yan, Z., & Li, X. (2021). Design and evaluation of a distance-driven user interface for asynchronous collaborative exhibit browsing in an augmented reality museum. *IEEE Access*, *9*(40), 73948–73962. https://doi.org/10.1109/ACCESS.2021.3080286

Chow, K., Coyiuto, C., Nguyen, C., & Yoon, D. (2019). Challenges and design considerations for multimodal asynchronous collaboration in VR. *Proceedings of the ACM on Human–Computer Interaction*, *3*(CSCW), 1–24. https://doi.org/10.1145/3359142

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic Press.

de Belen, R. A. J., Nguyen, H., Filonik, D., Del Favero, D., & Bednarz, T. (2019). A systematic review of the current state of collaborative mixed reality technologies: 2013–2018. *AIMS Electronics and Electrical Engineering*, *3*(2), 181–223. https://doi.org/10.3934/ElectrEng.2019.2.181

Druta, R., Druta, C., Negirla, P., & Silea, I. (2021). A review on methods and systems for remote collaboration. *Applied Sciences*, *11*(21), 10035. https://doi.org/10.3390/app112110035

Elvezio, C., Sukan, M., Oda, O., Feiner, S., & Tversky, B. (2017). Remote collaboration in AR and VR using virtual replicas. In *ACM SIGGRAPH 2017 VR Village* (pp. 1–2). https://doi.org/10.1145/3089269.3089281

Ens, B., Lanir, J., Tang, A., Bateman, S., Lee, G., Piumsomboon, T., & Billinghurst, M. (2019). Revisiting collaboration through mixed reality: The evolution of groupware. *International Journal of Human–Computer Studies*, *131*(SI), 81–98. https://doi.org/10.1016/j.ijhcs.2019.05.011

Feiner, A. O. S. (2003). The flexible pointer: An interaction technique for selection in augmented and virtual reality. In *Proceedings of UIST* (Vol. 3, pp. 81–82).

Fidalgo, C. G., Yan, Y., Cho, H., Sousa, M., Lindlbauer, D., & Jorge, J. (2023). A survey on remote assistance and training in mixed reality environments. *IEEE Transactions on Visualization and Computer Graphics*, *29*(5), 2291–2303. https://doi.org/10.1109/TVCG.2023.3247081

Frees, S., Kessler, G. (2005). Precise and rapid interaction through scaled manipulation in immersive virtual environments. In *IEEE Proceedings. Virtual Reality, 2005* (pp. 99–106).

Fussell, S. R., Setlock, L. D., & Kraut, R. E. (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* (pp. 513–520). https://doi.org/10.1145/642611.642701

Gauglitz, S., Nuernberger, B., Turk, M., & Höllerer, T. (2014). In touch with the remote world: Remote collaboration with augmented reality drawings and virtual navigation. In *Proceedings of the 20th ACM Symposium on Virtual Reality Software and Technology* (pp. 197–205).

Gelman, A. (2005). Analysis of variance. *Quality Control & Applied Statistics*, *20*(1), 295–300.

Gimeno, J., Morillo, P., Orduña, J. M., & Fernández, M. (2013). A new AR authoring tool using depth maps for industrial procedures. *Computers in Industry*, *64*(9), 1263–1271. https://doi.org/10.1016/j.compind.2013.06.012

Grandi, J. G., Debarba, H. G., & Maciel, A. (2019). Characterizing asymmetric collaborative interactions in virtual and augmented realities. In *IEEE Conference on Virtual Reality and 3d User Interfaces*.

Guo, A., Canberk, I., Murphy, H., Monroy-Hernández, A., & Vaish, R. (2019). Blocks: Collaborative and persistent augmented reality experiences. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *3*(3), 1–24. https://doi.org/10.1145/3351241

Hertzum, M. (2021). Reference values and subscale patterns for the task load index (TLX): A meta-analytic review. *Ergonomics*, *64*(7), 869–878. https://doi.org/10.1080/00140139.2021.1876927

Higuch, K., Yonetani, R., & Sato, Y. (2016). Can eye help you? effects of visualizing eye fixations on remote collaboration scenarios for physical tasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5180–5190).

Huang, W., Wakefield, M., Rasmussen, T. A., Kim, S., & Billinghurst, M. (2022). A review on communication cues for augmented reality based remote guidance. *Journal on Multimodal User Interfaces*, *16*(2), 239–256. https://doi.org/10.1007/s12193-022-00387-1

Irlitti, A., Piumsomboon, T., Jackson, D., & Thomas, B. H. (2019). Conveying spatial awareness cues in XR collaborations. *IEEE Transactions on Visualization and Computer Graphics*, *25*(11), 3178–3189. https://doi.org/10.1109/TVCG.2019.2932173

Irlitti, A., Smith, R. T., Von Itzstein, S., Billinghurst, M., & Thomas, B. H. (2016). Challenges for asynchronous collaboration in augmented reality. In *2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)* (pp. 31–35). https://doi.org/10.1109/ISMAR-Adjunct.2016.0032

Jeanne, F., Soullard, Y., Oker, A., & Thouvenin, I. (2017). Ebagg: Error-based assistance for gesture guidance in virtual environments. In *2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT)* (pp. 472–476). https://doi.org/10.1109/ICALT.2017.32

Jeanne, F., Thouvenin, I., & Lenglet, A. (2017). A study on improving performance in gesture training through visual guidance based on learners' errors. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology* (pp. 1–10).

Jing, A., May, K., Matthews, B., Lee, G., & Billinghurst, M. (2022). The impact of sharing gaze behaviours in collaborative mixed reality. *Proceedings of the ACM on Human–Computer Interaction*, *6*(CSCW2), 1–27. https://doi.org/10.1145/3555564

Jing, A., May, K. W., Naeem, M., Lee, G., & Billinghurst, M. (2021). Eyemr-vis: A mixed reality system to visualise bi-directional gaze behavioural cues between remote collaborators. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–4). https://doi.org/10.1145/3411763.3451545

Kim, H., Lee, G., & Billinghurst, M. (2015). A non-linear mapping technique for bare-hand interaction in large virtual environments. In *Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction* (pp. 53–61). https://doi.org/10.1145/2838739.2838774

Kim, S., Billinghurst, M., & Kim, K. (2020). *Multimodal interfaces and communication cues for remote collaboration* (Vol. 14). Springer.

Kim, S., Billinghurst, M., & Lee, G. (2018). The effect of collaboration styles and view independence on video-mediated remote collaboration. *Computer Supported Cooperative Work*, *27*(3–6), 569–607. https://doi.org/10.1007/s10606-018-9324-2

Kim, S., Huang, W., Oh, C.-M., Lee, G., Billinghurst, M., & Lee, S.-J. (2023). View types and visual communication cues for remote collaboration. *Computers, Materials & Continua*, *74*(2), 4363–4379. https://doi.org/10.32604/cmc.2023.034209

Kim, S., Jing, A., Park, H., Lee, G. A., Huang, W., & Billinghurst, M. (2020). Hand-in-air (HIA) and hand-on-target (HOT) style gesture cues for mixed reality collaboration. *IEEE Access*, *8*, 224145–224161. https://doi.org/10.1109/ACCESS.2020.3043783

Kim, S., Lee, G., Billinghurst, M., & Huang, W. (2020). The combination of visual communication cues in mixed reality remote collaboration. *Journal on Multimodal User Interfaces*, *14*(4), 321–335. https://doi.org/10.1007/s12193-020-00335-x

Kim, S., Lee, G. A., Sakata, N., Dünser, A., Vartiainen, E., & Billinghurst, M. (2013). Study of augmented gesture communication cues and view sharing in remote collaboration. In *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (pp. 261–262). https://doi.org/10.1109/ISMAR.2013.6671795

Kim, T. S., Kim, S., Choi, Y., & Kim, J. (2021). Winder: Linking speech and visual objects to support communication in asynchronous collaboration. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–17). https://doi.org/10.1145/3411764.3445686

Kosmalla, F., Daiber, F., Wiehr, F., Krüger, A. (2017). Climbvis: Investigating *in-situ* visualizations for understanding climbing movements by demonstration. In *Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces* (pp. 270–279).

Lee, G., Kang, H., Lee, J., & Han, J. (2020). A user study on view-sharing techniques for one-to-many mixed reality collaborations. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (pp. 343–352). https://doi.org/10.1109/VR46266.2020.00054

Lee, G. A., Kim, S., Lee, Y., Dey, A., Piumsomboon, T., Norman, M., & Billinghurst, M. (2017). Improving collaboration in augmented video conference using mutually shared gaze. In *ICAT-EGVE* (pp. 197–204).

Lee, Y., & Yoo, B. (2021). XR collaboration beyond virtual reality: Work in the real world. *Journal of Computational Design and Engineering*, *8*(2), 756–772. https://doi.org/10.1093/jcde/qwab012

Lee, Y., Yoo, B., & Lee, S.-H. (2021). Sharing ambient objects using real-time point cloud streaming in web-based XR remote collaboration. In *The 26th International Conference on 3D Web Technology* (pp. 1–9). https://doi.org/10.1145/3485444.3487642

Lin, C., Rojas-Munoz, E., Cabrera, M. E., Sanchez-Tamayo, N., Andersen, D., Popescu, V., Barragan Noguera, J. A., Zarzaur, B., Murphy, P., Anderson, K., Douglas, T., Griffis, C., Wachs, J. (2020). How about the mentor? Effective workspace visualization in AR telementoring. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (pp. 212–220). https://doi.org/10.1109/VR46266.2020.00040

Liu, J.-S., Tversky, B., & Feiner, S. (2022). Precueing object placement and orientation for manual tasks in augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, *28*(11), 3799–3809. https://doi.org/10.1109/TVCG.2022.3203111

Lu, W., Duh, B.-L H., & Feiner, S. (2012). Subtle cueing for visual search in augmented reality. In *2012 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (pp. 161–166). https://doi.org/10.1109/ISMAR.2012.6402553

Lu, Y., Yu, C., & Shi, Y. (2020). Investigating bubble mechanism for ray-casting to improve 3D target acquisition in virtual reality. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (pp. 35–43). https://doi.org/10.1109/VR46266.2020.00021

Marques, B., Ferreira, C., Silva, S., Dias, P., & Santos, B. S. (2023). Is social presence (alone) a general predictor for good remote collaboration? Comparing video and augmented reality guidance in maintenance procedures. *Virtual Reality*, *27*(3), 1783–1796. https://doi.org/10.1007/s10055-023-00770-7

Marques, B., Silva, S., Alves, J., Araujo, T., Dias, P., & Santos, B. S. (2021). A conceptual model and taxonomy for collaborative augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, *28*(12), 5113–5133. https://doi.org/10.1109/TVCG.2021.3101545

Marques, B., Silva, S., Alves, J., Rocha, A., Dias, P., & Santos, B. S. (2022). Remote collaboration in maintenance contexts using augmented reality: Insights from a participatory process. *International Journal on Interactive Design and Manufacturing*, *16*(1), 419–438. https://doi.org/10.1007/s12008-021-00798-6

Marques, B., Silva, S., Dias, P., & Santos, B. S. (2022a). Comparing a large-scale display and an interactive projector for one-to-many mixed reality (MR) remote collaboration. In *Proceedings of the 2022 ACM Symposium on Spatial User Interaction* (pp. 1–2). https://doi.org/10.1145/3565970.3568187

Marques, B., Silva, S., Dias, P., Santos, B. S. (2022b). One-to-many remote scenarios: The next step in collaborative extended reality (XR) research. In *Workshop on Analytics, Learning & Collaboration in Extended Reality (XR-WALC). ACM International Conference on Interactive Media Experiences (IMX 2022)* (pp. 1–6).

Marques, B., Teixeira, A., Silva, S., Alves, J., Dias, P., & Santos, B. S. (2022). A critical analysis on remote collaboration mediated by augmented reality: Making a case for improved characterization and evaluation of the collaborative process. *Computers & Graphics*, *102*, 619–633. https://doi.org/10.1016/j.cag.2021.08.006

Mayer, A., Combe, T., Chardonnet, J.-R., & Ovtcharova, J. (2022). Asynchronous manual work in mixed reality remote collaboration. In *International Conference on Extended Reality* (pp. 17–33).

Müller, J., Rädle, R., & Reiterer, H. (2016). Virtual objects as spatial cues in collaborative mixed reality environments: How they shape communication behavior and user task load. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 1245–1249).

Narzt, W., Pomberger, G., Ferscha, A., Kolb, D., Müller, R., Wieghardt, J., Hörtner, H., & Lindinger, C. (2006). Augmented reality navigation systems. *Universal Access in the Information Society*, *4*(3), 177–187. https://doi.org/10.1007/s10209-005-0017-5

Norman, M., Lee, G. A., Smith, R. T., & Billingurst, M. (2019). The impact of remote user's role in a mixed reality mixed presence system. In *Proceedings of the 17th International Conference on Virtual-Reality Continuum and Its Applications in Industry* (pp. 1–9). https://doi.org/10.1145/3359997.3365691

Nuernberger, B., Lien, K.-C., Grinta, L., Sweeney, C., Turk, M., & Höllerer, T. (2016). Multi-view gesture annotations in image-based 3D reconstructed scenes. In *Proceedings of the 22nd ACM Conference on Virtual Reality Software and Technology* (pp. 129–138). https://doi.org/10.1145/2993369.2993371

Otsuki, M., Wang, T.-Y., & Kuzuoka, H. (2022). Assessment of instructor's capacity in one-to-many AR remote instruction giving. In *Proceedings of the 28th ACM Symposium on Virtual Reality Software and Technology* (pp. 1–5). https://doi.org/10.1145/3562939.3565631

Pidel, C., & Ackermann, P. (2020). Collaboration in virtual and augmented reality: A systematic overview. In *Augmented Reality, Virtual Reality, and Computer Graphics: 7th International Conference, AVR 2020, Lecce, Italy, September 7–10, 2020, Proceedings, Part i 7* (pp. 141–156).

Piumsomboon, T., Dey, A., Ens, B., Lee, G., & Billinghurst, M. (2019). The effects of sharing awareness cues in collaborative mixed reality. *Frontiers in Robotics and AI*, *6*, 5. https://doi.org/10.3389/frobt.2019.00005

Piumsomboon, T., Lee, G. A., Hart, J. D., Ens, B., Lindeman, R. W., Thomas, B. H., Billinghurst, M. (2018). Mini-me: An adaptive avatar for mixed reality remote collaboration. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1–13).

Rey, D., & Neuhäuser, M. (2011). Wilcoxon-signed-rank test. In *International encyclopedia of statistical science*. https://doi.org/10.1002/9780471462422.eoct979

Riege, K., Holtkamper, T., Wesche, G., & Frohlich, B. (2006). The bent pick ray: An extended pointing technique for multi-user interaction. In *3D User Interfaces (3DUI'06)* (pp. 62–65).

Schäfer, A., Reis, G., & Stricker, D. (2021). A survey on synchronous augmented, virtual and mixed reality remote collaboration systems. *ACM Computing Surveys*, *55*(6), 1–27. https://doi.org/10.1145/3533376

Seeliger, A., Merz, G., Holz, C., & Feuerriegel, S. (2021). Exploring the effect of visual cues on eye gaze during AR-guided picking and assembly tasks. In *2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)* (pp. 159–164). https://doi.org/10.1109/ISMAR-Adjunct54149.2021.00041

Seeliger, A., Weibel, R. P., & Feuerriegel, S. (2022). Context-adaptive visual cues for safe navigation in augmented reality using machine learning. *International Journal of Human–Computer Interaction*, 1–21. Advance online publication. https://doi.org/10.1080/10447318.2022.2122114

Sereno, M., Wang, X., Besancon, L., McGuffin, M. J., & Isenberg, T. (2020). Collaborative work in augmented reality: A survey. *IEEE Transactions on Visualization and Computer Graphics*, *28*(6), 2530–2549. https://doi.org/10.1109/TVCG.2020.3032761

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3–4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591

Steinicke, F., Ropinski, T., & Hinrichs, K. (2006). Object selection in virtual environments using an improved virtual pointer metaphor. In *Computer Vision and Graphics: International Conference, ICCVG 2004, Warsaw, Poland, September 2004, Proceedings* (pp. 320–326).

Teo, T., Lee, G. A., Billinghurst, M., & Adcock, M. (2018). Hand gestures and visual annotation in live 360 panorama-based mixed

reality remote collaboration. In *Proceedings of the 30th Australian Conference on Computer–Human Interaction* (pp. 406–410). https://doi.org/10.1145/3292147.3292200

Teo, T., Lee, G. A., Billinghurst, M., & Adcock, M. (2019). Investigating the use of different visual cues to improve social presence within a 360 mixed reality remote collaboration. In *The 17th International Conference on Virtual-Reality Continuum and Its Applications in Industry* (pp. 1–9).

Thomas, B., Demczuk, V., Piekarski, W., Hepworth, D., & Gunther, B. (1998). A wearable computer system with augmented reality to support terrestrial navigation. In *Digest of Papers. Second International Symposium on Wearable Computers (Cat. no. 98ex215)* (pp. 168–171). https://doi.org/10.1109/ISWC.1998.729549

Tian, H., Lee, G. A., Bai, H., & Billinghurst, M. (2023). Using virtual replicas to improve mixed reality remote collaboration. *IEEE Transactions on Visualization and Computer Graphics*, *29*(5), 2785–2795. https://doi.org/10.1109/TVCG.2023.3247113

Volmer, B., Baumeister, J., Von Itzstein, S., Bornkessel-Schlesewsky, I., Schlesewsky, M., Billinghurst, M., & Thomas, B. H. (2018). A comparison of predictive spatial augmented reality cues for procedural tasks. *IEEE Transactions on Visualization and Computer Graphics*, *24*(11), 2846–2856. https://doi.org/10.1109/TVCG.2018.2868587

Wang, L., Chen, J., Ma, Q., & Popescu, V. (2021). Disocclusion headlight for selection assistance in VR. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)* (pp. 216–225). https://doi.org/10.1109/VR50410.2021.00043

Wang, L., Wu, W., Zhou, Z., & Popescu, V. (2020). View splicing for effective VR collaboration. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (pp. 509–519). https://doi.org/10.1109/ISMAR50242.2020.00079

Wang, P., Bai, X., Billinghurst, M., Zhang, S., Zhang, X., Wang, S., He, W., Yan, Y., & Ji, H. (2021). AR/MR remote collaboration on physical tasks: A review. *Robotics and Computer-Integrated Manufacturing*, *72*, 102071. https://doi.org/10.1016/j.rcim.2020.102071

Wang, P., Wang, Y., Billinghurst, M., Yang, H., Xu, P., & Li, Y. (2023). Behere: A VR/SAR remote collaboration system based on virtual replicas sharing gesture and avatar in a procedural task. *Virtual Reality*, *27*(2), 1409–1430. https://doi.org/10.1007/s10055-023-00748-5

Wang, P., Zhang, S., Bai, X., Billinghurst, M., Zhang, L., Wang, S., Han, D., Lv, H., & Yan, Y. (2019). A gesture-and head-based multimodal interaction platform for MR remote collaboration. *The International Journal of Advanced Manufacturing Technology*, *105*(7–8), 3031–3043. https://doi.org/10.1007/s00170-019-04434-2

Wang, P., Zhang, S., Billinghurst, M., Bai, X., He, W., Wang, S., Sun, M., & Zhang, X. (2020). A comprehensive survey of AR/MR-based co-design in manufacturing. *Engineering with Computers*, *36*(4), 1715–1738. https://doi.org/10.1007/s00366-019-00792-3

Wang, T.-Y., Otsuki, M., & Kuzuoka, H. (2022). Evaluating workload in one-to-many remote collaboration. In *Proceedings of the 15th International Conference on Pervasive Technologies Related to Assistive Environments* (pp. 296–297). https://doi.org/10.1145/3529190.3534717

Yang, J., Sasikumar, P., Bai, H., Barde, A., Sörös, G., & Billinghurst, M. (2020). The effects of spatial auditory and visual cues on mixed reality remote collaboration. *Journal on Multimodal User Interfaces*, *14*(4), 337–352. https://doi.org/10.1007/s12193-020-00331-1

Yoon, B., Kim, H-i., Lee, G. A., Billinghurst, M., & Woo, W. (2019). The effect of avatar appearance on social presence in an augmented reality remote collaboration. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (pp. 547–556). https://doi.org/10.1109/VR.2019.8797719

Yu, K., Gorbachev, G., Eck, U., Pankratz, F., Navab, N., & Roth, D. (2021). Avatars for teleconsultation: Effects of avatar embodiment techniques on user perception in 3d asymmetric telepresence. *IEEE Transactions on Visualization and Computer Graphics*, *27*(11), 4129–4139. https://doi.org/10.1109/TVCG.2021.3106480

Zhang, X., Bai, X., Zhang, S., He, W., Wang, P., Wang, Z., Yan, Y., & Yu, Q. (2022). Real-time 3D video-based MR remote collaboration using gesture cues and virtual replicas. *The International Journal of Advanced Manufacturing Technology*, *121*(11–12), 7697–7719. https://doi.org/10.1007/s00170-022-09654-7

Zhou, Z., Wang, L., & Popescu, V. (2021). A partially-sorted concentric layout for efficient label localization in augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, *27*(11), 4087–4096. https://doi.org/10.1109/TVCG.2021.3106492

Zollmann, S., Hoppe, C., Langlotz, T., & Reitmayr, G. (2014). Flyar: Augmented reality supported micro aerial vehicle navigation. *IEEE Transactions on Visualization and Computer Graphics*, *20*(4), 560–568. https://doi.org/10.1109/TVCG.2014.24

About the authors

Lili Wang received her PhD degree from Beihang University, Beijing, China. She is a professor with the School of Computer Science and Engineering of Beihang University and a researcher with the State Key Laboratory of Virtual Reality Technology and Systems. Her interests include virtual reality, real-time rendering and HCI.

Xiangyu Li is a master student in the School of Computer Science and Engineering of Beihang University, China. His current research focuses on virtual reality, augmented reality, and HCI.

Jian Wu is a PhD student in the School of Computer Science and Engineering of Beihang University, China. His current research focuses on virtual reality, augmented reality, visualization and HCI.

Zhou Dong is a full professor at Beihang University. He received his PhD in System Engineering at Beihang University. His research interests are multidisciplinary, including virtual and augmented reality, computer-aided design, industrial maintenance and assembly, human factor and knowledge management.

Im Sio Kei received his PhD degree from Queen Mary University of London, United Kingdom. He is a professor at the Faculty of Applied Sciences, Macau Polytechnic University. His research interests include video coding, image processing, machine learning for NLP and multimedia.

Voicu Popescu received his PhD degree in computer science from the University of North Carolina at Chapel Hill, USA in 2001. He is an associate professor with the Computer Science Department of Purdue University. His research interests lie in the areas of computer graphics, computer vision, and visualization.