Human Systems Engineering and Design (IHSED 2024), Vol. 158, 2024, 155–164
https://doi.org/10.54941/ahfe1005539

Evaluation of Human Adaptation in a Monoscopic Video See-Through Headset

Long Cheng[1,3], Valentin Holzwarth[2], and Andreas Kunz[3]

[1]RhySearch, Buchs, 9470, Switzerland
[2]Atlas VR, Schlieren, 8952, Switzerland
[3]ETH Zurich, Zurich, 8092, Switzerland

ABSTRACT

Video see-through (VST) cameras, integrated in virtual reality (VR) headsets, offer a convenient means to bridge the real world and virtual objects. However, performing tasks within a VST system can deviate from real-world experiences. In this study, we assess the technical specifications of a commercially available consumer-grade virtual reality headset and investigate user performance across various tasks within a VST environment using a pilot user study and the prism adaptation method. Tasks include object relocation, drinking, screwing, and typing on a physical tablet. Our findings reveal a decline in performance for tasks requiring close-range interaction and screen-based operations, accompanied by a user adaptability to the studied tasks. We also note mild motion sickness symptoms but find no discernible aftereffects associated with the tasks examined.

Keywords: Mixed reality, Human factor study, Human adaptation, Virtual reality

INTRODUCTION

The VR Head-Mounted Display (HMD) has emerged as an indispensable tool for representing immersive experiences, including industrial education, entertainment (Jeong, 2023), and social interactions. Thanks to the mass production of cutting-edge technologies, the HMD has made significant improvements, ensuring a user’s comfort and engagement even during extended wear. To democratize VR and make it accessible to a broader audience, innovations have been introduced to create lightweight devices (utilizing pancake lenses, lighter straps, and novel materials), more potent processing units (transitioning from a reliance on PC power to standalone applications), enhanced customization (including automatic inter-pupillary distance adjustments and face covers tailored via 3D scanning), and advanced see-through cameras (progressing from single black-and-white cameras to high-resolution, full-color cameras equipped with depth-sensing capabilities (Meta, 2023).

Traditional augmented reality (AR) experiences have been realized through optical see-through (OST) systems like the Microsoft HoloLens or VST systems on mobile phones. OST systems offer users a real-world view without any latency. However, they suffer from limitations such as a limited field of view (FOV), chromatic aberration, contrast ratio, bulkiness, high power consumption, overheating, and challenges in maintaining accurate and consistent real-world alignment. In contrast, mobile phones provide a more accessible AR experience through their digital screens. This well-established AR platform has found applications in diverse fields, including dimension measurement, interior design, and mobile gaming. Nevertheless, current mobile AR solutions still struggle with tracking accuracy, FOV constraints, battery consumption, and the need for users to continually hold or position the device. Additionally, the user's perspective of the augmented content often differs from that of the phone's camera, leading to a disconnect from the intended immersive experience.

In the context of HMDs, VST technology empowers users with a broader FOV through outward-facing camera(s). Seamlessly integrating virtual content into the real-world environment captured by these cameras, VST delivers an immersive AR experience that harmoniously blends the virtual and real world. Recent advancements in HMDs have expanded the applications of VST systems into gaming (Jeong, 2023), immersive media, and office productivity. Enhanced hand tracking capabilities enable users to interact smoothly with virtual content overlaid on the real world. Depth sensors further enhance the VST system's capabilities, enabling it to comprehend the environment and provide depth-based occlusion of virtual objects, thus narrowing the gap of HMDs in terms of AR capabilities. This technology has showcased its potential in ocular safety-critical scenarios like welding and laser experiments (Li, 2022), where users must protect themselves from hazardous light sources. Moreover, VST systems have demonstrated their effectiveness in the context of automotive interior design (Goedicke, 2022), providing designers with valuable insights for improving in-car environments.

While considerable advancements have been made in the field of VST systems, much of the research has gravitated towards applications of specific interest to researchers, rather than addressing real-world adoption rates. At the time of this study, limited attention has been devoted to investigating human performance within VST systems, particularly in the context of manual and office tasks. Furthermore, the HMDs used in such fields are typically developed as prototypes and often remain inaccessible for subsequent research purposes. Additionally, the exploration of depth perception and adaptation capabilities within VST systems remains an underexplored area of study.

In this study, we conducted experiments using a commercially available monoscopic VST system, specifically the Pico 4 Pro/Enterprise. Our investigation involved an examination of its hardware specifications and an in-depth evaluation of human performance across four distinct manual tasks, as part of a pilot user study. Our findings revealed that users demonstrated adaptability to the monoscopic VST-HMD environment, although we observed that human performance suffered from noticeable impairments in tasks requiring close-range interaction (i.e., within a distance of less than 0.5 meters) when compared to tasks performed at greater distances. Additionally, our exploration brought to light several limitations inherent to the monoscopic VST system.

RELATED WORK

Human Perception

Various studies have focused on human perception in VST systems. Due to offsets and camera quality, users do not have the exact same experience as seeing without the HMD. Visual deterioration in VST systems can be categorized into visual displacement and deterioration of viewing quality (Lee, 2023). Efforts to counter visual displacement include prototypes involving light field camera rays (Kuo, 2023) and neural perspective rendering (Xiao, 2022). However, the high computational demand in VST systems introduces challenges related to weight, heating, and chip design within the headset. To achieve a life-like VR experience, it is essential to achieve latency levels lower than 10 milliseconds (ms), along with a horizontal FOV of 162 degrees, and a vertical FOV of 135 degrees per eye (Cuervo, 2018). Meeting these specifications demands significantly more computational power than what is currently available in existing hardware. In a study measuring latency for consumer headsets, it was found that most HMDs on the market exhibited latencies exceeding 70 ms (Gruen, 2020).

Task Performance

In the realm of VST systems, (Rolland, 1995) conducted pioneering research into task performance within VST conditions and revealed a substantial 43% degradation in manual task performance when the camera was positioned 165 mm forward and 62 mm upward from the eye location. This study also observed aftereffects among the participants. Lee (2020) conducted a comprehensive assessment of human performance across various tasks and examined the aftereffects of a stereoscopic VST headset using the prism adaptation method over three days. The study tracked participants' performance before, during, and after exposed to binocular VST glasses. The findings indicated that humans can adapt to stereoscopic VST, albeit with an initial performance drop during the first exposure. Importantly, no significant after-effects were reported following this exposure. It is worth noting that the HMD used in their experiment had relatively low resolution (640 × 480 pixels) and a narrow field of view (horizontal FOV 48 degrees and vertical FOV 18 degrees), and it is not commercially available. Another study by Serefoglou (2008) found that depth perception did not play a significant role in hand-eye coordination tasks when using stereo cameras at various offset values. However, in the context of an all-in-one VST system, performance of real manual operation was not studied. In this work, we delve into the manual tasks and human factors associated with a monoscopic VST system.

METHODOLOGY

Hardware

We chose to use the Pico 4 Enterprise in our experiment, a consumer product launched in 2022. It has several notable features, including pancake lenses, automatic inter-pupillary distance adjustment (62 mm - 72 mm), and a monoscopic full-color VST camera (Sony IMX 471) with a 16-megapixel sensor, offering a horizontal FOV of 130 degrees and a vertical FOV of 115 degrees (Pico, 2023). Specific details about the photon-to-photon latency of the Pico 4 Pro were not available up to the date of publication. To address this gap in information, we conducted ten repetitions of the latency measurement process introduced in (Gruen, 2020) to calculate latency and ensure precision. Through these measurements, we calculated the photon-to-photon latency of the HMD under experimental lighting conditions, yielding a mean value of 55 ms, with a standard deviation of 3.4 ms. Additionally, we measured the axial offset, defined as the distance from the center of the eyes to the camera, to be 72 ms. Other stereoscopic all-in-one VST-HMDs such as the Meta Quest 3 are not considered. They are not yet available during the time of experiment. After tests, they were found to have large distortions for close range interactions.

Pilot Study Design

To investigate the impact of the monoscopic VST system on human perception and work performance, we devised a set of four office desk tasks carefully selected to explore the potential influence of depth information and axial offset. These tasks included relocating a glass from an arm's reach distance to a closer position, drinking water from a glass, tightening screws with screwdrivers, and a typing task on a 9.7-inch touchscreen tablet computer, specifically designed to assess the resolution and display quality of the device. The experiment overview is depicted in Figure 1.

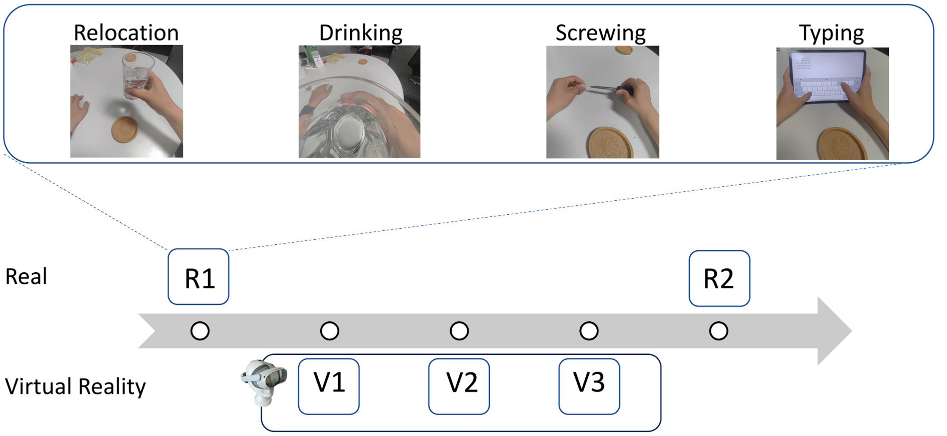

Figure 1: Experiment overview, the experiment commences with one cycle in the real world (R1), followed by three cycles in virtual reality (V1 to V3), and concludes with a second cycle in the real world (R2).

The experiment comprises five cycles denoted as R1 (Reality cycle 1), V1 (Virtual Reality cycle 1), V2 (Virtual Reality cycle 2), V3 (Virtual Reality cycle 3), and R2 (Reality cycle 2). This prism methodology is adapted from Lee (2020). Each cycle involves four identical tasks: relocation, drinking, screwing, and typing, with each complete cycle lasting approximately 90 seconds. Tasks in each cycle are described as:

- Relocation: transporting a glass of water from a designated far glass holder to a fixed nearby glass holder.
- Drinking: consuming the water from the glass placed in the first task.
- Screwing: assembling a screw and bolt, inserting the screw into the bolt, picking up a screwdriver, and tightening the screw. Timing for this task concludes when the screw completes one full rotation.
- Typing: entering a consistent line of text on the tablet computer. The text contains twenty seven characters. The participant is not required to correct typing errors.

Our study involved the participation of 8 individuals, each of whom underwent a preliminary familiarization phase with all five tasks without the use of HMDs for a single cycle (though this was not recorded). During the experiment, participants were seated and instructed to behave naturally. Upon entering the virtual environment, participants were prompted to measure their Inter Pupillary Distances (IPD) without glasses. Following the IPD measurement and adjustment, they continued the task in VR while wearing glasses or contact lenses, if needed, to correct vision. Within each cycle, participants were required to sequentially complete the relocation, drinking, screw tightening, and typing tasks. In each task, participants initiated with both hands resting on the table and concluded by placing both hands back on the table. These tasks were carefully designed to encompass mid-range, close-range, two-handed mid-range, and two-handed close-range manual operations. All experiments were conducted in indoor settings, specifically on office tables with standard lighting conditions. Throughout the experiment, an experimenter recorded task completion times and errors as quantitative measurements. Timing for all tasks commences when both hands lift off the table. Following the completion of the drinking task, the experimenter removed the glasses. Errors were defined to encompass instances where the glass touched the boundary of the holder or moved outside of it, cases where water splashed during the user study, instances of misalignment between the screwdriver and the screw, as well as instances where incorrect characters were selected on the tablet. Subsequent to the user study, the experimenter conducted informal interviews with each participant to collect qualitative data, including assessment of motion sickness level, perceived VST camera quality, and how demanding the task was.

For the quantitative data collected, we employed R for statistical analysis. This involved conducting the Shapiro-Wilk test to assess normality assumptions, and performing posthoc Tukey HSD analysis for groupwise comparisons. These analyses were applied to both, task completion times and the observed errors during the user study. The examination of both qualitative and quantitative data provides a comprehensive understanding of the impact of the monoscopic VST system on human perception and work performance.

RESULT AND DISCUSSION

All our participants were students with previous experience in all tasks researched. They did not wear glasses during automatic IPD adjustment but had normal or corrected-to-normal vision during the user study.

The task completion time of each task across all cycles conforms to a normal distribution according to the Shapiro-Wilk test, while the number of errors does not. The quantitative comparison of task completion times is depicted in Figure 2. It is shown in the chart that task completion time for all tasks exhibits improvement across VR cycles. Notably, all tasks display enhanced performance and learning effects in real cycles, denoted as task completion time of R2 < task completion time of R1. Referring to Figure 3, we observe that R1 significantly differs from V1 in the tasks of drinking and typing. In these two tasks, the VR cycles demonstrate noteworthy performance improvements (V1 > V2 > V3). As for task errors, minimal errors occurred during the first three tasks, in VR cycles or in real cycles, with the majority of mistakes concentrated in the typing task, as illustrated in Figure 4. Errors gradually decreased as participants repeated the typing task. It is noteworthy that the initial VR cycle (V1) witnessed the highest error rate, while the second Reality cycle (R2) predominantly exhibited correct task execution.

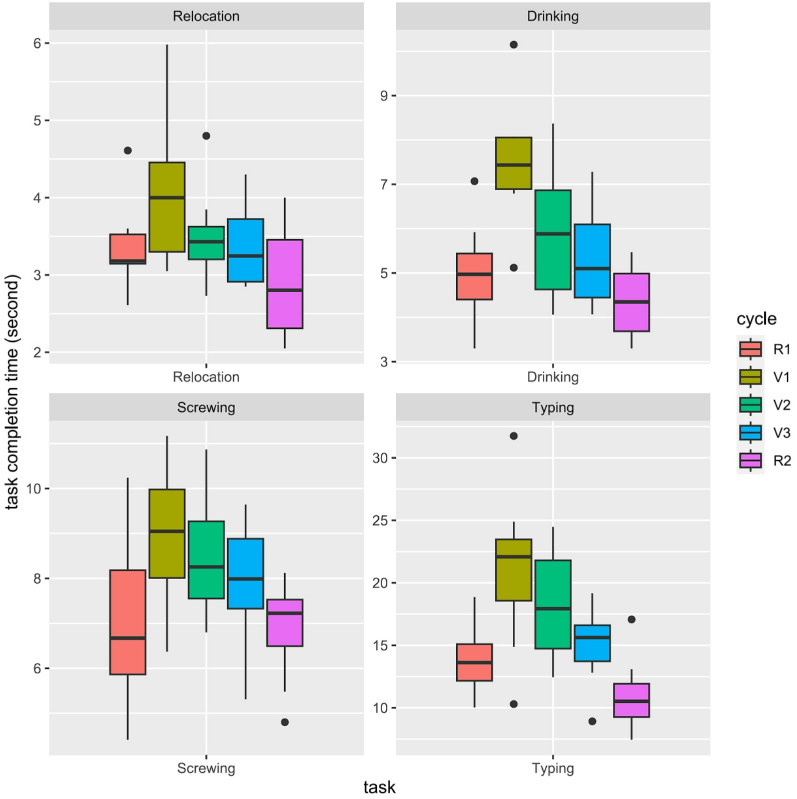

Figure 2: Quantitative comparison of task completion times.

From a qualitative user feedback, 3 out of 8 users reported experiencing a minor degree of discomfort during the VR cycles, although this discomfort dissipated after the conclusion of the user study. Participants also noted instances of scene object distortion, particularly in peripheral vision. In the context of monoscopic VST, participants gave positive feedback and expressed optimism about the system's potential benefits, assuming that camera quality will improve. However, some participants raised concerns regarding the typing task. They cited issues with touchscreen keyboard distortion and difficulties related to the display panel lighting, both of which made it challenging to type accurately and recognizing previous inputs on the screen.

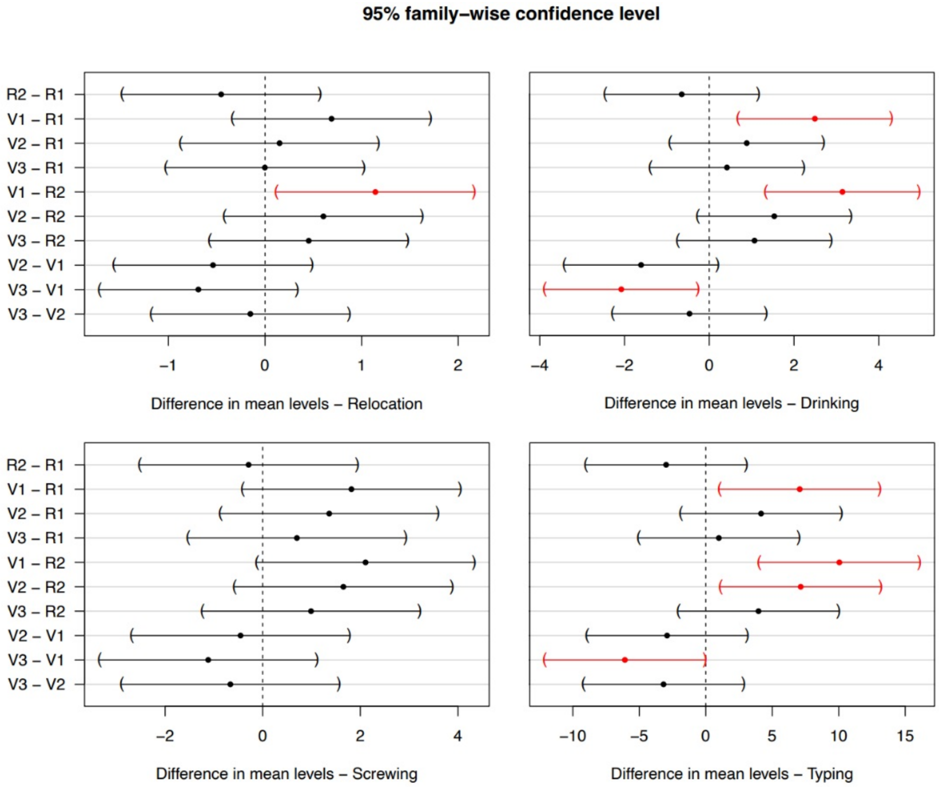

Figure 3: Tukey HSD analysis of task completion times. Significant differences are highlighted in red. Notably, no significant differences were observed in the screwing task.

The significance observed in the drinking task primarily relates to the VST intrinsic factors associated with axial offset. This value, measured to be 38 mm as detailed in section introducing hardware measurements, leads to an inaccurate perception of the mouth's location. The system can not handle vergence accommodation conflict and this effect is evident for close range interactions. It is worth noting that drinking is the only task where the streaming from the camera can be potentially misleading. In reality, individuals typically do not require visual assistance to drink, if they hold the glass. This corresponds to the "Finger-to-Nose" proprioception test conducted in psychology to assess proprioceptive impairments (Bo, 2019). In the VST environment, the visual feedback can significantly influence the user's perception. However, participants demonstrated an ability to adapt to the VST experience over time, as shown by the significance found for typing task and drinking task in V1 and V3 show adaptability across VR cycles.

The significance observed in the typing task can be attributed to default font size on the tablet as well as the camera quality. The camera used for this device has a fixed focus. In order to clearly recognize the text, the user has to move close to the tablet, which is far from focus distance of the camera. Interviews conducted with study participants revealed that most of them faced difficulties in clearly seeing the content they were typing. This experience is further compounded by the latency introduced by the HMD, as measured in section of hardware. It resulted in a distinctive difference from the real-world typing experience. Without a direct control of the camera shutter speed and frame rate, the user might observe light or screen flickering because of the unmatching frequencies. This problem applies to all camera systems including VST-HMDs, but will be influential for the human vision, that is replaced by the VST camera in the virtual environment.

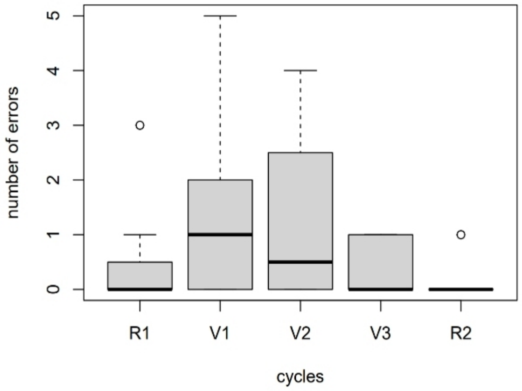

Figure 4: Number of errors in the typing task.

The lower significance found in the relocation task and the screwing task is also worth mentioning. In the relocation task, we discovered that participants could still gain depth cues from the relative size of the cup holders or shadow information. The deprivation of binocular vision might not lead to a complete loss of depth perception. For the screwing task, participants relied on tactile feedback from the screwdrivers and screws to help localize the screw position on the bolt.

CONCLUSION

Building upon the findings, akin to those in (Lee, 2020), we conclude that adaptation effects are present within the context of the VST system, accompanied by minor aftereffects associated with task performance in this VST environment. Discomfort reported by some participants can be attributed to lens focus and reduced keyboard clarity during the typing task. The statistical significance observed in the drinking and typing tasks suggests that depth perception at close distances significantly influences task performance in the monoscopic VST system.

Low-cost VST systems often exhibit vision quality inferior to that of normal human perception, as seen in phone-based VST solutions and entry-level VST-HMDs. The use of a single camera in these setups helps avoid discrepancies that can arise from the synchronization of multiple cameras. Notably, some recent mobile phones have introduced innovative technologies, such as the capability of recording spatial video using two built-in cameras. These dual-camera setups, while offering new possibilities, also present challenges related to differing parameters, alignment, and offsets. Nonetheless, monoscopic VST remains a prevalent choice for mobile phone-based AR systems and low-cost HMDs. The challenge of enabling seamless interaction with objects at varying distances under monoscopic VST persists. The introduction of new sensors, such as the Lidar or depth sensor for scene understanding, holds the promise of shaping the future of the VST industry (Meta, 2023).

Continued research is required to explore the applications of monoscopic cameras and to conduct comprehensive comparisons between monoscopic and stereo VST systems. It is essential to delve into how human perception and task performance vary across different devices, owing to the pivotal role played by outward-facing camera(s) in shaping perception. At the same time, more VST applications are required to fully unleash the potential of the VST systems.

REFERENCES

Bo, K. (2019). Finger nose proprioception test (case study). In *Clin Med (Lond)*, 19 (Suppl. 3), pp. 1–20.

Cuervo, E., Chintalapudi, K. and Kotaru, M. (2018). Creating the Perfect Illusion: What will it take to Create Life-Like Virtual Reality Headsets? In *Proceedings of the 19th International Workshop on Mobile Computing Systems & Applications, HotMobile '18*, New York, NY, USA, 12–13 February, pp. 7–12.

Goedicke, K., Bremers, A. W. D., Lee, S., Bu, F., Yasuda, H. and Ju, W. (2021). XR-OOM: Mixing virtual driving simulation with real cars and environments safely, in *13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '21 Adjunct)*, New York, NY, USA, 09–10 and 13–14 September 2021, pp. 67–70.

Gruen, R., Ofek, E., Steed, A., Gal, R., Sinclair, M. and Gonzalez-Franco, M. (2020). Measuring System Visual Latency through Cognitive Latency on Video See-Through AR Devices. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, Atlanta, GA, USA, 22–26 March, pp. 791–799.

Jeong, Y., Kim, D., Jang, H. and Ryu, J. (2023). Designing interactive space for the XR boardgame, *Immersive Learning Research - Practitioner*, 1(1), pp. 12–16.

Kuo, G., Penner, E., Moczydlowski, S., Ching, A., Lanman, D. and Matsuda, N., 2023. Perspective-Correct VR Passthrough Without Reprojection. In *ACM SIGGRAPH 2023 Conference Proceedings*, *SIGGRAPH '23*, New York, NY, USA, 6–10 August 2023, pp. 1–9.

Lee, J. H. and Park, J. H. (2020). Visuomotor adaptation to excessive visual displacement in video see-through HMDs. *Virtual Reality*, 24, pp. 211–221.

Lee, J. H., Yeom, K. and Park, J.-H. (2023). The effect of video see-through HMD on peripheral visual search performance, *IEEE Access*, 11, pp. 85184–85190.

Li, K., Choudhuri, A., Schmidt, S., Lang, T., Bacher, R., Hartl, I., Leemans, W. and Steinicke, F. (2022). Stereoscopic video see-through head-mounted displays for laser safety: An empirical evaluation at advanced optics laboratories, in *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, Singapore, Singapore, 17–21 October 2022, pp. 112–120.

Meta. (2023). Meta Quest 3 Homepage. Viewed 24 October 2023, https://www.meta.com/ch/en/quest/quest-3.

PICO Global. (2023). PICO 4E Specs. PICO Global Homepage. Viewed 24 October 2023, https://www.picoxr.com/global/products/pico4e/specs.

Rolland, J. P., Biocca, F. A., Barlow, T. and Kancherla, A. (1995). Quantification of adaptation to virtual-eye location in see-thru head-mounted displays. In *Proceedings of the Virtual Reality Annual International Symposium '95*, Research Triangle Park, NC, USA, 11–15 March, pp. 56–66.

Serefoglou, S., Schmidt, L., Radermacher, K., Schlick, C. and Luczak, H. (2008). Hand-Eye Coordination Using a Video See-Through Augmented Reality System. *The Ergonomics Open Journal*, 1, pp. 46–53.

Xiao, L., Nouri, S., Hegland, J., Garcia, A. and Lanman, D. (2022). NeuralPassthrough: Learned Real-Time View Synthesis for VR. In *ACM SIGGRAPH 2022 Conference Proceedings, SIGGRAPH '22*, New York, NY, USA, 8–11 August, 40, pp. 1–9.