Full body video-based self-avatars for mixed reality: from E2E system to user study

Diego Gonzalez Morin · Ester Gonzalez-Sosa · Pablo Perez · Alvaro Villegas

Received: 4 November 2022 / Accepted: 3 March 2023 / Published online: 21 April 2023
© The Author(s) 2023

Abstract
In this work, we explore the creation of self-avatars through video pass-through in mixed reality (MR) applications. We present our end-to-end system, including custom MR video pass-through implementation on a commercial head-mounted display (HMD), our deep learning-based real-time egocentric body segmentation algorithm, and our optimized offloading architecture, to communicate the segmentation server with the HMD. To validate this technology, we designed an immersive VR experience where the user has to walk through a narrow tile path over an active volcano crater. The study was performed under three-body representation conditions: virtual hands, video pass-through with color-based full-body segmentation, and video pass-through with deep learning full-body segmentation. This immersive experience was carried out by 30 women and 28 men. To the best of our knowledge, this is the first user study focused on evaluating video-based self-avatars to represent the user in a MR scene. Results showed no significant differences between the different body representations in terms of presence, with moderate improvements in some Embodiment components between the virtual hands and full-body representations. Visual Quality results showed better results from the deep-learning algorithms in terms of the whole body perception and overall segmentation quality. In this study, we provide some discussion regarding the use of video-based self-avatars and some reflections on the evaluation methodology. The proposed E2E solution is in the boundary of the state-of-the-art, so there is still room for improvement before it reaches maturity. However, this solution serves as a crucial starting point for MR applications where users can feel immersed and interact with their own bodies.

Keywords Mixed reality · Video-based self-avatars · Egocentric vision · Embodiment · Offloading

1 Introduction
The use of self-avatars, the virtual representation of the user’s body from their own perspective, is starting to become ubiquitous in virtual reality (VR) and mixed reality (MR). The possibility of seeing your own body while immersed brings many benefits. First, it increases the users’ sense of presence (SoP), as it shifts the user from being an observer to really experiencing the virtual environment (VE) (Lok et al. 2003; Slater and Usoh 1993). Besides, it enhances the spatial perception and distance estimation (McManus et al. 2011; Ebrahimi et al. 2018) while bringing a positive impact on cognitive load (Steed et al. 2016), trust, and collaboration (Pan and Steed 2017). Apart from increasing the SoP, self-avatars also increase the sense of embodiment (SoE) (Argelaguet et al. 2016; Fribourg et al. 2020).

Figure 1 shows that an avatar can be described based on its visibility, fidelity, and dynamics. The minimal configuration of an avatar’s visibility just relies on users’ hands (Dodds et al. 2011); however, the current trend is to go beyond hands and model the partial body (Gruosso et al. 2021) or even the entire human body (Ogawa et al. 2020b). Avatar fidelity accounts for shape and texture realism related to skin and clothes. Existing works for hand or full body user representations range from abstract or minimal representation (Argelaguet et al. 2016), such as the stickman (Fribourg et al. 2020), to more realistic ones with gender-matched, skin-matched limbs (Dewez et al. 2019), or even...
with realistic human models.\(^1\) The latest tendency relies on models that preserve the user’s body shape such as using a skinned multi-person linear body (SMPL) model (Loper et al. 2015), or 3D scanning systems such as point-cloud representation for fully personalized avatars (Thaler et al. 2018).

Avatar dynamics involves the capture of the user’s movements that enables or facilitates interaction tasks in MR (Pan and Steed 2019). While head and hand-tracking are provided by almost all standards HMDs or through the use of external sensors such as Leap Motion, less work has been done on full body tracking. Some works proposed using two VIVE trackers attached to the feet for allowing foot tracking and visibility (Pan and Steed 2019; Bonfert et al. 2022; Bozgeyikli and Bozgeyikli 2022). Although traditional approaches for full body capture are based on inertial measurements units (IMUs) sensors placed on a suit that the user wears (Xu et al. 2019; Waltemate et al. 2018; Jayaraj et al. 2017), there are emerging solutions based on computer vision (Bazarevsky et al. 2020; Tome et al. 2020) or depth information such as Azure Kinect v2 (Gonzalez-Franco et al. 2020b).

An alternative approach to computer-generated avatars is the use of video-based self-avatars, whose aim is to create user representation by segmenting the egocentric view of the user. The VR community has explored this idea over the last decade. Approaches based on color segmentation (Fiore and Interrante 2012; Bruder et al. 2009; Perez et al. 2019; Günther et al. 2015) can be deployed in real-time but they do not work well just with uncontrolled scenarios or different skin colors (Fiore and Interrante 2012). Alternatively, segmentation approaches based on depth information (Rauter et al. 2019; Lee et al. 2016; Alaee et al. 2018) are effective for some situations, but still lack the precision to provide a generic, realistic, and real-time immersive experience. In our previous work (Gonzalez-Sosa et al. 2022), we proposed a real-time egocentric body segmentation algorithm. This algorithm, inspired by Thundernet’s architecture (Xiang et al. 2019), achieves a frame rate of 66 fps for an input resolution of 640 \(\times\) 480, thanks to its shallow design. To train it, we created a 10,000 images dataset with high variability in terms of users, scenarios, illumination, and gender, among others. Therefore, this paper extends our previous preliminary work (Gonzalez-Morin et al. 2022a) and further investigates how to integrate the real-time egocentric user body segmentation algorithm. Our main contributions can be summarized as:

- A detailed description of our full E2E system composed of three main modules: video pass-through capable VR device, real-time egocentric body segmentation algorithm, and optimized offloading architecture.
- A subjective evaluation of the video-based self-avatar technology integrated into a gamified immersive experience, conducted by 58 users. To the best of our knowledge, it is the first work that includes a user study using full-body video-based self-avatars.

The rest of this article is structured as follows: Sect. 2 covers relevant state-of-the-art regarding the use of segmentation algorithms in augmented virtuality (AV). Section 3 describes the implementation details of our egocentric body segmentation E2E system. Section 4 gives details about the gamified immersive experience designed to validate the video-based self-avatar technology. Section 5 presents the subjective results obtained with a set of 58 different subjects and describes the between-subjects and within-subjects experiments conducted. Finally, Sect. 6 further elaborates on the benefits and drawbacks of using video-based avatars in place of CGI avatars. Section 7 concludes the paper.

2 Related works

Augmented Virtuality is a MR subcategory of the reality continuum that aims to merge portions of the reality surrounding the user with a virtual environment. In particular, introducing the user’s own body has attracted a lot of interest in the scientific community. For instance, Bruder et al. proposed a skin segmentation algorithm to incorporate users’ hands handling different skin colors (Bruder et al. 2009). Conversely, a floor subtraction approach was developed to incorporate users’ legs and feet in the VE. Assuming uniformity in floor appearance, the body was retained by simply filtering out all pixels not belonging to the floor (Bruder et al. 2009).

Chen et al., in the context of 360° video cinematic experiences, explored depth-keying techniques to incorporate all objects below a predefined distance threshold (Chen et al. 2017). This distance threshold could be changed

---

\(^1\) See for instance the Rocketbox library (Gonzalez-Franco et al. 2020c).
dynamically to control how much of the real world was shown in the VE. The user could also control the transitions between VE and the real world through head shaking and hand gestures. Some of the limitations that the authors pointed out were related to the reduced field of view of the depth sensor.

Pigny et Dominjon Pigny and Dominjon (2019) were the first to propose a deep learning algorithm to segment egocentric bodies. Their architecture was based on U-NET and it was trained using a hybrid dataset composed of images from the COCO dataset belonging to persons and a 1500-image custom dataset created following the automatic labeling procedure reported in Gonzalez-Sosa et al. (2020). The segmentation network has two decoders: one for segmenting egocentric users, and the other for exocentric users. However, their segmentation execution speed could not currently totally keep up with the camera’s update rate while maintaining a decent quality. They reported 16 ms of inference time for 256 × 256 images, and observed problems of false positives that downgrade the AV experience.

Later, in Gruosso et al. (2021), the authors developed a system able to segment the user’s hand and arms using a dataset composed of 43K images from TeGO (Lee and Kacorri 2019), EDSH (Li and Kitani 2013), and a custom EgoCam dataset, using semantic segmentation deep learning network based on DeepLabv3+ (Chen et al. 2018). For the experiment, they used an HTC Vive HMD, a monocular RGB camera, and a Leap Motion controller to provide a free-hands interaction camera. Inference time takes around 20 ms for a 360 × 360 RGB video in a workstation provided with Nvidia Titan Xp GPU (12GB memory). They manage to run VR rendering and CNN segmentation in the same workstation.

3 System design and implementation

The generation of real-time, accurate and effective video-based self-avatars for MR requires the following solutions to be integrated as a single MR end-to-end system:

- **Video pass-through capable VR device** captures the real scenario using a frontal stereo camera, accurately aligns it to the user’s view and head movement, and renders it, in real-time, within the immersive scene.
- **Real-time egocentric body segmentation algorithm** identifies the pixels corresponding to the user’s body from the frames captured by the frontal camera.
- **Optimized offloading architecture** egocentric body segmentation algorithms require high-end hardware not available in the current state-of-the-art HMDs. Consequently, we need a communication architecture that allows fast data exchange between the immersive client and the server running the segmentation algorithm.

3.1 Custom video pass-through implementation

There are several VR devices with video pass-through capabilities that are already commercially available, such as the Varjo XR-3\(^2\) or the HTC VIVE Pro.\(^3\) These devices are still wired, constraining the range of possible use cases of the proposed system. Consequently, we decided to build our own video pass-through solution, thoroughly described in (Morin et al. 2022), for the Meta Quest 2,\(^4\) as it is a well-known commercially successful standalone VR device. For this purpose, we had to build our own hardware along with the necessary software to effectively integrate the captured video into the VR scene, ensuring a perfect alignment between the captured feed and head movement.

To comply with many MR video pass-through use cases,\(^5\) we chose the ELP 960P HD OV9750 stereo camera\(^6\) which provides a maximum resolution of 2560 × 960 at a 60Hz update rate and a field of view of 90°. Most of the commercially available video pass-through MR devices incorporate stereo cameras that are aligned with the user’s eyes. On the other hand, we realized, after some initial tests, that for our particular use case tilting the stereo cameras 25° towards the ground provided a better experience and ergonomics for the user. We designed a custom 3D printed attachment to fix the stereo camera to the VR device with the described offset as in Fig. 2.

Our system, built using Unity 2019.3, has been designed to work both in wired, using Unity’s in-built C# API to access the stereo camera feed as a webcam, and wireless modes, using a custom native Android plugin based on the UVC library\(^7\) to access the camera feed. While our goal is to implement a fully wireless system, the wired mode is necessary for development purposes and to have controlled network conditions during our subjective evaluation experiments.

The rendered stereo camera’s feed and the user’s head movement must be perfectly aligned to avoid any user discomfort or VR sickness. This misalignment can come from the three non-exclusive sources: lens distortion effects, inaccurate placement of the planes where the camera feed is...
rendered, and motion-to-photon delays.\textsuperscript{8} We need to apply the necessary calibration steps and adjustments to reduce the effect of the aforementioned misalignment sources.

The captured video frames from the stereo camera are rendered into two virtual planes (in blue in Fig. 3), which we refer to as rendering planes, one for each eye. We need first to estimate the camera intrinsic parameters and distortion coefficients, using a well-known camera calibration technique (Zhang 2000), to accurately correct the camera’s lens distortion and properly scale these rendering planes.

In the second step, we aim to accurately place the rendering planes relative to the VR device’s coordinate frame, as shown in Fig. 3. While this is generally achieved using a classic camera-IMU (inertial measurement unit) calibration method (Mirzaei and Roumeliotis 2008), commercial devices don’t provide access to raw sensor data with accurate time-stamps. Consequently, we decided to design our own workaround calibration method which used the VR device’s controllers and the Aruco (Garrido-Jurado et al. 2014) library to estimate the geometric relationship between the device and rendering planes. For more details refer to Morín et al. (2022).

Finally, we need to introduce a method to remove the misalignment produced by the motion to photon latency. This latency produces an unnatural and uncomfortable decoupling between the stereo feed, the head movement, and, consequently, the virtual scene. To overcome this issue, we add a delay alignment buffer, which stores the position and orientation of the stereo camera’s coordinate frame. The size of the buffer ($t_c$) corresponds to the empirically estimated motion to photon delay. We assume $t_c$ to be constant for the same stereo camera and VR device models. By placing and rotating the stereo camera feed according to the first pose stored in the buffer, we achieve an accurate head movement to stereo feed alignment.

For more details regarding our custom video pass-through implementation please refer to Morín et al. (2022).

### 3.2 Egocentric segmentation

This is a crucial step in our setup: we need an accurate real-time egocentric body segmentation algorithm which allows identifying which pixels, from the stereo camera feed, correspond to the user’s body. Our algorithm is based on deep learning techniques. Particularly, it is based on the use of semantic segmentation networks (Guo et al. 2018). In this case, the segmentation is performed from the first point of view (egocentric vision).

The designed algorithm must meet two requirements: (1) real-time performance achieving an update rate above 60Hz, and (2) high-quality segmentation in uncontrolled conditions. Concerning the first requirement, we designed our architecture inspired by Thudernet (Xiang et al. 2019), a very shallow semantic segmentation algorithm composed of: (1) an encoding subnetwork; (2) a pyramid pooling module (PPM) that enables to extract features at different

\textsuperscript{8} Motion-to-photon delay refer to the time between the user’s head movement and the display of the corresponding changes in the VR scene.
scales; and (3) a decoding subnetwork (see Fig. 4). Several modifications were applied to the baseline architecture: (1) larger pooling factors of 6, 12, 18, 24 were used in the PPM module to better extract features of larger input images, (2) three additional long skip connections were added between encoding and decoding subnetworks for refining object boundaries, and (3) weighted cross entropy loss function was used to handle class imbalance between the human body (foreground) and background.

To achieve high-quality segmentation, a data-centric approach was followed, putting a strong emphasis on the variability of the training data. We created a dataset of almost 10,000 images composed of three datasets: (1) Ego Human dataset, a semi synthetic dataset composed of egocentric body parts (arms, lower limbs, torsos) merged with realistic backgrounds that facilitates the process of groundtruth generation; (2) a subset of THU-READ dataset (Tang et al. 2018), originally created for action recognition, whose segmentation groundtruth was created manually using Amazon Mechanical Turk; and (3) EgoOffices: an egocentric dataset that was manually captured using a portable custom. The dataset was captured by more than 25 users and multiple realistic scenarios, such as different apartments or office rooms. As the hardware setup involves the use of a stereo camera for providing egocentric stereo vision, to the user, we re-trained the Thundernet architecture with images composed of two copies of the same images resulting in a $640 \times 480$ image, to replicate stereo vision. For more details, please refer to Gonzalez-Sosa et al. (2022).

3.3 ZMQ-based offloading architecture

Our egocentric body segmentation solution can’t fulfill the real-time requirements using mobile hardware, such as smartphones or wireless VR devices. Consequently, we decided to wirelessly offload the segmentation algorithm to a nearby server. The offloading architecture must be specifically designed to allow real-time processing, providing the highest throughput and lowest latency possible. For this reason, we used ZeroMQ, a TCP-based architecture that has shown outstanding performance in most throughput and latency benchmarks (Sommer et al. 2018). We have developed our own ZMQ-based offloading architecture, which is already validated in realistic MR offloading use cases (González Morín et al. 2022b). The architecture has been specifically designed to ensure high reliability, throughput, and low latency data transmissions.

The architecture has been designed following a publisher-subscriber scheme in which one or multiple nodes advertise services to which other nodes can publish or subscribe to. These nodes can be defined as individual agents within the architecture, which can subscribe or publish to an available data channel or topic.

Each software module runs a communication library named Alga. Alga implements a PUB-SUB ZMQ messaging pattern, where each frame is encapsulated in a multipart ZMQ message, composed of three sub-messages: topic, header, and data (payload). This architecture provides the core functionality required for distributed computing: semantic routing (using topics), custom metadata (in the header), payload framing with arbitrary packet size, reliable transmission, and queue size limit (using ZMQ high water mark functionality) to control latency. ZMQ also provides the possibility to run the functionality either in the same or different nodes without changing the code, supporting in-memory, inter-process (Unix socket), and TCP transport protocols. Even though using TCP instead of UDP transport imposes an overhead in terms of throughput and latency, we have found that the achieved performance is enough for the purposes of our application. For future versions of the architecture, we are already working on a version of Alga that supports the same functionality using UDP transport.

The architecture implements several types of publishers/subscribers according to the kind of data to transmit. For our particular application, we use the Picture and Unsigned 8-bit Picture packets. The first one is specifically designed to transmit individual JPEG-coded color frames. In our case,
this packet type is used to transmit the stereo camera information. The second type is designed to transmit 8-bit single channel frames, specifically designed for segmentation masks. Consequently, this packet type is used to transmit the resulting segmentation mask back to the VR client.

The architecture is built both in C++ and Python. On the server side, we use the Python version of the architecture to align with the egocentric body segmentation algorithm. As the VR client is built in Unity, we created a C++ plugin, allowing to use of non-blocking asynchronous transmission and reception. This is necessary to avoid extra latencies coming from Unity’s update rate not being synchronized with the camera capture or segmentation algorithm.

3.4 End-to-end system

The final End-to-End (E2E) setup is depicted in Fig. 5. The principal data flow of the E2E architecture is as follows:

1. **Stereo camera frame capture** As we have already described, the VR client runs Unity and is in charge of obtaining individual frames from the stereo camera video feed. The individual frames are moved to the underlying C++ plugin running our offloading architecture.

2. **Transmit stereo color frame to server** The offloading architecture, via Alga, is in charge of packing each frame and transmitting them via TCP through an arbitrary topic. The segmentation server runs another instance of Alga and receives the frames through the same topic.

3. **Egocentric body segmentation** The server then infers the segmentation mask for each received frame.

4. **Transmit result back to VR client** The mask is transmitted back to the VR client via Alga, through another arbitrary topic.

5. **Shader application on rendering frames** Finally, the client applies the received masks to the custom shader attached to the rendering frames. The shader is in charge of exclusively rendering the stereo camera pixels corresponding to the user’s body according to the received mask.

3.4.1 Hardware setup

The egocentric segmentation server was running on a PC with an Intel Xeon ES-2620 V4 @ 2.1Ghz with 32 GB of RAM and powered with 2 GPU GTX-1080 Ti with 12GB RAM each. On the VR client side, we considered two modalities:

- **Wireless Mode**: The immersive scene is running directly on the Meta Quest 2 HMD, to which the stereo camera is directly connected.
- **Wired Mode**: The immersive scene runs on a workstation to which a Meta Quest 2 is connected via an USB-C cable using the "Link" mode. The stereo camera is connected to the workstation. In this case, the client workstation includes an Intel Core i7-11800 H and a Nvidia Geforce RTX 3070 GPU. The workstation is connected to the server via an Ethernet cable, with a 1 Gbps interface.

We used Unity version 2019.3 on the client side, where the stereo frames are captured at a resolution of $1280 \times 480$. For the validation experiments, we used a Netgear R6400 router, providing symmetric wireless 200 Mbps when a single user is connected (González Morín et al. 2022b).

3.5 Performance evaluation

The proposed offloading architecture has been thoroughly benchmarked in terms of throughput and latency in multiple relevant scenarios and wireless technologies (González Morín et al. 2022b). We decided to extend the benchmark by testing the architecture on the final E2E system. To obtain the following results we used the exact hardware setup described in Sect. 3.4.1.

3.5.1 Offloading architecture and video pass-through

We first aim to evaluate the performance of our offloading architecture alone in both hardware setups, wired and wireless.

In this first round of experiments, we removed the segmentation algorithm from the server, directly transmitting back to the client a single channel 8-bit mask of the
incoming frame. This mask is obtained using a chroma-key algorithm which adds a negligible overhead on the server side. For each scenario, we repeated a set of four experiments each transmitting 1000 frames. No other users are connected to the router. We store the times that each frame takes from the moment it is loaded in memory on the client side to when the final mask is received from the server on the client side. The initial results are shown in the first two columns of Table 1. Our E2E setup provides mean round trip times of 7.37 and 10.30 milliseconds on wired and wireless scenarios, respectively, with confidence intervals below 20 ms in both cases.

### 3.5.2 Segmentation server

In this case, the goal is to measure the total time consumed by the server in performing all the necessary steps for a frame: reception, necessary transformations such as scaling, and inference. Similarly to the previous set of experiments, we carried out four experiments each processing 1000 frames. We obtained a mean time of 16.7 ms, as shown in Table 1, which guarantees an update rate of 60 Hz, complying with the MR applications’ real-time requirements.

### 3.5.3 End-to-end system

Finally, by aggregating the results obtained in the previous two experiments we obtained E2E mean round trip times (RTT) of 29.67 ms and 32.39 ms. This latency is low enough to allow the use of delay buffers to ensure a proper alignment of the color frame and masks, as described in Sect. 3.6.

### 3.6 Delay buffers

As the obtained RTTs values are higher than the device’s update period of 16 ms, the hand pixels appear to be misaligned with respect to the segmentation mask. The misalignment comes from the fact that the segmentation mask arrives at the VR client on an average of 32.29 ms later than its corresponding stereo pixels are loaded in memory and rendered. The feeling is as if the mask follows the actual hand. This artifact, shown in Fig. 6, affects the overall experience and must be removed.

### 4 Subjective evaluation

Our study aims to investigate the role of full-body video-based self-avatars in Mixed Reality and how our E2E implementation affects the overall experience in terms of presence, embodiment, visual quality, and acceptability. We compare our video-based avatar approach with a state-of-the-art hand-tracking solution, as well as with our previous implementation of video-based avatars using chroma-key segmentation (Perez et al. 2019).

#### 4.1 Research questions

The following research questions were established:

- **RQ1**: The inclusion of a stereo camera that allows video pass-through increases the sense of presence and embodiment with respect to a virtual representation of your hands. Seeing your full body, including your feet, is relevant for the experience.
- **RQ2**: Deep learning based segmentation is a better solution for self-avatars in uncontrolled conditions than color-based segmentation.

#### 4.2 Immersive experience: design and considerations

To overcome this issue, we decided to add a frame buffer of a size equivalent to the measured mean RTT (32.29 ms) according to the camera’s update rate. By doing this we overcome the artifacts generated by the system’s added delays, ensuring the mask and its corresponding stereo pixels are rendered simultaneously. Notice that the size in time of this delay buffer must be added to the alignment buffer described in Sect. 3.1.

![left: Mask-color pixels alignment observed when no delay buffer is used. right: alignment achieved using the delay buffer](image)

To overcome this issue, we decided to add a frame buffer of a size equivalent to the measured mean RTT (32.29 ms) according to the camera’s update rate. By doing this we overcome the artifacts generated by the system’s added delays, ensuring the mask and its corresponding stereo pixels are rendered simultaneously. Notice that the size in time of this delay buffer must be added to the alignment buffer described in Sect. 3.1.

### Table 1 Performance evaluation results for the wired and wireless scenarios along with the server processing times

|                  | Wired (ms) | Wireless (ms) | Server (ms) |
|------------------|------------|---------------|-------------|
| Mean             | 7.37       | 10.39         | 16.7        |
| Variance         | 14.37      | 19.38         | 0.006       |
| 95-th Perc.      | 14.0       | 18.0          | 31.2        |

![image]
the crater of the volcano, forcing the users to pay attention to their lower limbs while walking. To provide passive haptics, to increase the feeling of being there, we incorporated 1 cm height steps, as can be seen in Fig. 7. The main purpose of the steps is to elicit visuotactile stimuli. To align the real and virtual steps, we implemented a C++ plugin running a standard Aruco tracking library Garrido-Jurado et al. (2014), and the camera calibration parameters estimated in Sect. 3.1. We placed an Aruco marker as shown in Fig. 7.

The whole path of tiles is not visible to the user from the beginning. On the contrary, the tiles are activated, in order, when users match their hands with a neon-light representation of the hands shown on the floor. The neon-light hands are placed on the ground/tiles level so that the users need to touch with their own hands the actual surface, increasing the haptic sensations. The hands are tracked using Meta Quest 2 hand-tracking solution. Finally, at the end of the tiles path, we created a virtual portal that, once reached by the users, would teleport them to a friendly scenario. The goal of this final friendly scene is to allow the users to explore the virtual environment as well as their own body representation without any stress or pressure. The representation of the entire setup and the described details are shown in Fig. 8.

The total area occupied by the setup is 4 m × 6 m, using 16 steps in total. The distribution used is shown in Fig. 7. Our E2E system has been specifically designed for wireless operation. However, is out of the scope of this work to evaluate and study how the network quality of service affects the immersive experience. To achieve a fair comparison between the different algorithms and our E2E system we need to have controlled network conditions. For this reason, we decided to use the wired setup in which the Meta Quest 2 is connected to a workstation running the scene. Furthermore, the wired

---

10 Hand-tracking was provided by Meta Quest 2.
version considerably reduces the battery consumption and increases the length of the experiment runs.

4.3 Conditions

The main goal of the experiment is to evaluate the different modalities available to create the self-avatar user representations (see Fig. 9). The following three conditions were explored:

- **Virtual Reality** the user is represented with the virtual hand models provided by the Meta Quest 2. We use it as a control condition, to assure that our video-based self-avatar brings some benefits that can not be achieved with just virtual hands.
- **Chroma** users are represented as the output of a simple algorithm that masks everything in the range of skin color, similar to Perez et al. (2019). This algorithm is directly implemented inside a shader in Unity, thus running locally on the client side, and taking less than 3 ms per frame. To ensure that users could see their feet as a reference for walking, we asked them to wear some pink booties, which were perfectly detected by the color-based filter, as can be seen in Fig. 9.
- **Deep learning** the user is represented by the output of our semantic segmentation algorithm reported in Sect. 3.2 using the described E2E architecture.

4.4 Protocol

In the beginning, participants were briefed about the experiment and its purpose, and were asked to sign a consent form. Three modalities were set up: A, B, and C. To counterbalance the order effect, users were evenly distributed among modalities. On each modality, users needed to perform the experiment under two conditions, as depicted in Fig. 10. After each condition, users answered the questionnaire described in Sect. 4.5.

Experiments took place in Madrid for three consecutive days in March 2022 at a local university. On the first and third days, room A was used, while on the second day, room B was used (see Fig. 7).

It is worth noting that, for safety reasons, users under modality A did not use the 1 cm steps described above, since the Virtual Reality condition did not allow them to see their feet.

4.5 Questionnaire

After providing some demographic information about age, gender, and previous experience with VR, users were asked to fill out a 30-item subjective questionnaire, where 12 were used to measure embodiment, 7 for presence, 6 for visual
quality, and 4 for acceptability It is an adaptation of the questionnaire used in our previous work (Villegas et al. 2020), with a stronger focus on embodiment compared to presence. As in Villegas et al. (2020), it is based on well-established tools from the state-of-the-art (Gonzalez-Franco and Peck 2018; Witmer and Singer 1998) and ITU-T Recommendations, to favor the reproducibility of the results and its comparison with other works.

**Embodiment** We have used the embodiment questionnaire (EQ) proposed by Gonzalez-Franco and Peck (2018). We chose to include a subset of 13 items, namely: Q1, Q2, and Q3 for ownership, Q6, Q7, Q8, and Q9 for agency, Q14 for location; Q17, Q19, and Q20 for external appearance, and Q25 for external stimuli (modifying it to “I had the feeling that if I fell off the bridge I was going to hurt myself”). Unlike in the original questionnaire, Q20 (“I felt like I was wearing different clothes from when I came to the laboratory”) has been reversed, since now the users should be able to see their own clothes. Items from the tactile sensation factor were not included as they were not addressed in our study. A total score has also been computed following (Gonzalez-Franco and Peck 2018), as the weighted average of all items, where items Q1-Q14, which belong to the main embodiment components (ownership, agency, and location) have double the weight than the rest of the items (Q17-Q25).

**Presence** We have used Witmer and Singer’s presence questionnaire version 3 (PQv3) (Witmer et al. 2005). PQv3 contains 32 items to explain the sense of presence using the four different sub-dimensions: involvement, interface quality, adaptation/immersion, and sensory fidelity. We used a subsampling with seven items, as it shows a good correlation with the full questionnaire in the overall score (computed as the mean of all items), as well as in the aforementioned sub-dimensions (Pérez et al. 2021).

**Visual Quality** To complete the evaluation of the experience, we have included additional questions to assess the subjective perception of the visual quality of the different elements (Perez et al. 2019). The visual quality of the virtual Environment, the virtual objects, and the user’s own Body, Arms, and Legs were assessed using ITU-T P.913 absolute category rating scale (ACR). Since some of the elements may not appear under some conditions, an additional category was included indicating that the element was not visible at all, which was rated as 0. Additionally, the annoyance of elements from the local Background that were misclassified as foreground by the segmentation algorithm (false positives) was rated using ITU-T P.913 degradation category rating (DCR). A Total score has also been computed using the average of all scores (including both ACR and DCR ratings).

**Acceptability** Four additional questions were included to explore people’s acceptability. Cybersickness was assessed using the single-item vertigo score rating (VSR) scale, as recommended by ITU-T P.919. We use the ratio of users with severe symptoms (VSR ≤ 2) as the evaluation score. Global quality of experience (QoE) was asked using the ITU-T ACR scale, and then the proportion of good-or-better (GoB) ratings was extracted (Hoßfeld et al. 2016). Finally, the Net Promoter Score was extracted with two computations: the traditional one, based on the self-reported probability of recommending the game to a friend (“NPS-R”) (Reichheld 2003), and a variant based on the self-reported willingness to pay (“NPS-P”) (Villegas et al. 2020).

**Qualitative evaluation** Users had two moments to provide free-form feedback, one answering why would you recommend the experience, and a final comment with free-form text about the overall experience.

### 4.6 Statistical analysis

The design of the experiment allows for two types of analysis (see Fig. 10): a between-subject analysis between the first execution of the three conditions, aimed at comparing the Virtual Reality baseline solution with both video-based avatars (RQ1), and a within-subject analysis between Chroma and Deep Learning implementations, to address RQ2.

Presence, Embodiment, and Video Quality measures can be modeled as normal variables, and therefore we will analyze them using parametric statistics (Narwaria et al. 2018). One-way ANOVA will be used for the between-subject analysis, with Bonferroni-corrected pairwise Student T-tests as
post-hoc analysis. One-way ANOVA with repeated measures will be used for the between-subject analysis. No post-hoc test is needed, as there are only two conditions.

Acceptability scores do not use a (weighted) average of the different ratings to provide a final score, but an analysis of the distribution of the ratings. Consequently, to test for the significance of the differences in the acceptability measures, Mann–Whitney U-tests will be applied to the distributions of ratings, with Bonferroni correction for the within-subject analysis.

A significance level of \( p < 0.05 \) is established for all the tests.

### 4.7 Participants

The game was evaluated by 58 volunteers (30 female and 28 male). The participant’s ages ranged from 18 to 63 (\( M = 22.16, SD = 8.62 \)), most of them were students from a local university and some of them were staff, with no implication in the development of the project at all. All users were Caucasian but one woman, who has Black ethnicity. Table 2 shows a fair homogeneous distribution between modalities and gender distribution between modalities. 22 of the 58 performed the experiment in Room B (see Fig. 8 right), while the remainder 36 performed the experiment in Room A (see Fig. 8 left). One woman assigned to the C modality could only accomplish the first condition, as she reported extreme fear of conducting the experience one more time.

### 5 Results

Figure 11 shows the main quantitative results for Embodiment, Presence, and Visual Quality.

#### 5.1 Embodiment

The quantitative results of the different sub-scales of the Embodiment Questionnaire, as well as a total value aggregating all items, are shown in the top row in Fig. 11, for both between- and within-subject analysis. Results are shown on a 1–7 Likert scale. All scores are high, indicating that the sense of embodiment is high under all conditions.

Between-subject analysis shows a significant effect of the video-based algorithms in Location \( (F_{2,55} = 4.56, p = .015, \eta^2 = .14) \), Appearance \( (F_{2,55} = 6.48, p = .003, \eta^2 = .19) \), and Response \( (F_{2,55} = 3.26, p = .046, \eta^2 = .11) \) scales. Pairwise T-Tests show significant differences between Virtual reality and both Chroma and Deep Learning for Appearance and Location; and between Chroma and Deep Learning for Response.

Within-subject analysis shows a significant effect between Chroma and Deep Learning algorithms in Ownership \( (F_{1,38} = 4.37, p = .043, \eta^2 = .02) \) and Appearance \( (F_{1,38} = 6.13, p = .018, \eta^2 = .03) \).

In summary, we can see moderate improvements in some Embodiment components when replacing the conventional virtual hands with a full-body video-based avatar, especially in terms of Location and Appearance. We observe small, but significant, improvements when we compare both video-based avatar solutions.

#### 5.2 Presence

The quantitative results of the different sub-scales of the Presence Questionnaire, as well as a total value aggregating all items, are shown in the middle row in Fig. 11, for both between- and within-subject analysis. Results are shown on a 1–7 Likert scale.

Presence mean values are high for all sub-scales: 5.1 for Involvement, 5.2 for Haptic sensory fidelity, and 6.1 for Adaptation. However, no significant differences are found between conditions in either between- or within-subject analysis.

#### 5.3 Visual quality

Visual Quality scores are shown in the bottom row in Fig. 11, for both between- and within-subject analysis. A Total (average) score is also shown.

Environment and Objects are common for all the conditions, and they show a MOS value of about 4.0 (Good), without significant differences. The perception of Body and Arms has a moderately lower MOS, between 3.0 (Fair) and 3.5, and the differences between algorithms are not significant either.

There is a significant difference in the perception of the Legs. This is obvious for Virtual Reality, where no legs were shown in the virtual scene, but also between Chroma and Deep Learning algorithms, in both between- and within-subject comparisons. In the within-subject comparison, the MOS for Chroma and Deep Learning is 2.38 and 3.61, respectively, which means an increase of more than one quality level.

The same difference exists in the perception of the Background false positives. In this case, the highest quality is obtained in the Virtual Reality scene, where no false positives exist. The within-subject comparison between Chroma
Fig. 11  Results for Embodiment (top), Presence (middle), and Visual Quality (bottom) questionnaires, both for between-subject (left) and within-subject (right) analysis. Bars show Mean values for each of the measures, and error bars show the standard error of the mean (SEM)
Table 3: Acceptability results summary

|               | Between |           | Within |           |
|---------------|---------|-----------|--------|-----------|
|               | VR      | Chr       | DL     | Chr       | DL       |
| Cybersickness | 0       | 5% (1)    | 5% (1) | 2.5% (1)  | 0        |
| Good-or-Better| 79%     | 95%       | 90%    | 87%       | 92%      |
| NPS-R         | 47      | 70        | 53     | 51        | 61       |
| NPS-P         | 68      | 45        | 37     | 44        | 44       |

and Deep Learning shows significant MOS values differences, with results of 2.90 and 3.77, respectively.

Consequently, there is a significant difference in the Total score both for between-subject and within-subject comparisons.

5.4 Acceptability

Table 3 shows the acceptability results. Only two participants reported severe cybersickness effect, both in the first execution of video-based avatar conditions: one in Chroma, and one in Deep Learning. The latter did not want to do the second condition, and therefore, her results have been excluded from the within-subject analysis altogether.

Quality of Experience ratings were extremely high, with above 80% of the participants reporting Good or Excellent QoE. Net Promoter Scores were also high, mostly in the range of 40–70%.

When comparing the acceptability of the three conditions, no strong differences can be observed, as different acceptability indicators show different results. Furthermore, the U-test analysis of the underlying distribution of the scores does not show any significant difference. Therefore, we can conclude that all three conditions have similar (and high) levels of acceptability.

5.5 Qualitative evaluation

Several insights could be also inferred from analyzing free text comments given by users. People, in general, were grateful to experience this new technology, stating in most cases that it was a very interesting/incredible/pleasant experience, reporting having fun while doing the experiment, and envisioning it as the future.

In all conditions, most users reported feeling teleported to the place "I felt like I was really there", and also experiencing the threat of the experience: "It's interesting to see how your body reacts unconsciously even though you know nothing will happen to you" or "Even though it is a game it gives you the feeling that you are going to fall down [...] partially tricks your brain [...] there was a moment when I had to hold on because I felt like I was really going to fall". People, in general, reported good perception of the distance, especially related to touching the tiles, but show some uncertainties with respect to their height perception. From the perspective of the game itself, people felt that the game is too short and that adding more functionalities could make the game more fun. Some people wearing prescription glasses, reported a more comfortable experience when they took them off.

In the VR scenario, some users were very happy to see their own hands without using controllers, while, at the same time, reported that it could be even better if more body parts were perceivable. Indeed, another reported a "disconcerting experience of not knowing exactly where to walk". With respect to the perception of false positives, users perceive less of them in the deep learning condition, vs the chroma condition: "the false positives were very low compared to the previous experience and the avatar looked much better than in the previous experience, it was quite good", or this comment after chroma conditions: "the second time I tried the game it looked worse, I saw many more parts of the class, which made me feel more out of the game environment".

In general, in all deep learning responses, people show high acceptance to see their own body, reporting that it makes them feel more immersed, perceiving the experience as more realistic. Besides, users also appreciated being able to see their accessories such as watches, jewelry, painted nails, etc. Especially in modalities A and B, users reported a high difference with respect to the previous conditions, where only certain body parts could be seen. For instance: "It has been really impressive to see the change and be able to see my own body including my clothes and many details. It honestly blew me away. I’ve never had a VR experience where I could actually see my body and everything" or "It looks like you’re the one going through the catwalks and, moreover, the avatar looks exactly like you". Moreover, some people reported the benefits of seeing their own legs: "Seeing my legs made me feel better in space".

In the chroma scenario, the feedback could be different depending on whether users were wearing clothes in the same range as the skin color "The avatar was very well done", or not "the credibility is less because the body is not completely represented (I could not see my legs)".

Users also reported that there is still room for improvement concerning seeing your entire body with deep learning. One thing is the fact that segmentation boundaries are not perfect, and tend to be slightly bigger (thick edge) than the actual body. "The fact of showing one’s own body helps a lot to the perception of space and movement of oneself; however, the noise, that is produced around the body distracts me a lot from the scene, diverts the attention too much and took me out of the experience" or "I think it is a unique experience, however, I think the quality of the avatar image can be
improved”. In this case, users perceived that the segmentation performed with color-based is more precise, although it can only segment skin color parts, e.g. “in this experience, the hands looked more realistic because the cropping was better, however, I was not able to see my own clothes and I saw most of the furniture in the classroom”. In general, tracking was perceived as very stable in all conditions, although the presence of long hair can block the device’s sensors.

6 Discussion

6.1 Implementation

We have presented our fully functional video-based self-avatar E2E system for MR. Our implementation allows us to run the proposed segmentation algorithm in real-time on a commercial VR device. We achieved E2E latencies below 11 ms if no algorithm is running on the server. Consequently, the system allows using other computationally demanding solutions on standalone VR devices such as object tracking or camera-based full-body tracking, extending the current boundaries of MR solutions and applications.

The system still has some limitations and we had to assume a set of trade-offs in terms of segmentation accuracy, frame rate, resolution, etc. Consequently, the presented solution cannot be fairly compared to much more mature solutions such as the virtual hand solution based on hand-tracking provided by Meta or other vendors. However, we must remark that the presented technology is in the boundary of the state-of-the-art, which leaves room for improvement and it can be a starting point for future distributed and novel MR solutions.

6.2 Evaluation

Regarding the first research question established in Sect. 4.1, our qualitative analysis shows that video-based self-avatars provide marginal improvement only in some components of the sense of embodiment with respect to the Virtual Reality condition. No significant difference is observed in the presence factor.

Regarding the second research question, the Deep Learning algorithm offers better visual quality than the Chroma in uncontrolled environments, both for the legs and for the background perception. These differences result in small improvements in Ownership and Appearance factors of the sense of embodiment. No significant difference is observed in presence either.

We can extract several interpretations from this. First, all the conditions show high levels of sense of presence and embodiment. Since in all cases users are immersed in a virtual environment with which they interact with their own hands (either “real” or virtual), the degree of embodiment and presence achieved by the baseline condition is high. Therefore, any observed improvement must be necessarily small. Second, the maturity of the virtual reality technology is much higher than the video-based avatars and, consequently, the implementation of VR looks more like an actual finished product than the video avatars. Somehow the benefits of seeing their own body may be limited by the limitations on the implementation (e.g. false positives). For example, VR only shows the hand (compared to the full arms of other solutions), but it offers better resolution in the hand image. Both facts counterbalance and, as a result, the video quality for the Arms is similar in all conditions.

A similar situation happens between Chroma and Deep Learning: the former is only able to detect the hands and some colors of clothes (red or yellow); however, these elements are segmented with better pixel precision than the Deep Learning algorithm. Besides, differences regarding the number of body parts segmented between Chroma and Deep Learning might be reduced when the subjects are wearing clothes with similar colors as the skin. This was the case for 18 out of the 58 considered subjects.

These findings are also supported by qualitative evaluation given by users: they appreciated the possibility of seeing your own body completely and, at the same time, reported room for improvement. Finally, we also experienced that some people tend to forget whether they have seen their own body limbs or not, which increases some uncertainty about the results.

Finally, it is also possible that the main benefits of using video-based avatars cannot be measured in terms of presence or embodiment. Other aspects of the overall Quality of Experience should be explored.

6.3 Methodology

First of all, we are evaluating presence, embodiment, and video quality using already existing questionnaires and methodologies, but none of them were designed for video-based avatars. As a consequence, there are some limitations.

The Presence Questionnaire (PQ) was developed for VR, and the Embodiment Questionnaire (EQ) was designed for computer-generated and animated avatars. When used in a different scenario, some items may not completely apply, or they could be even misunderstood by the users. For instance, we have seen that Q8 (I felt as if the movements of the virtual body were influencing my own movements) shows a range of
values completely different from other items, which might show that it is not being well understood by the users. Other items such as Q2, Q15, Q17, or Q20 might create some confusion for the users due to the particular case of this technology showing the user’s appearance as it is.

We have also observed some limitations of some questions related to visual quality, as we are using a single question to capture, at the same time, quality perception related to how much are you seeing from your body and how sharp, accurate and stable this vision is. These two aspects of the quality should be probably rated separately. In what concerns NPS-P, people reported that they will be willing to pay more if the game itself were longer.

6.4 Comparison

Previous works have also observed that a realistic virtual hand representation elicits a stronger sense of ownership (Argelaguet et al. 2016). Later, Fribourg et al. (2020) underlined a lower popularity of the appearance factor compared to the control and point of view, when assessing in combination (preference of choosing first the degree of control, before the degree of appearance). This latter conclusion might not be easily extrapolated to the case of a video-based self-avatar, as there is no need for avatar control when moving or interacting with real objects (only with virtual ones).

Improvement in presence and embodiment perceived in this work is in line with our previous work where virtual hands (using controllers) and real hands conditions were compared (Villegas et al. 2020), showing a higher difference between the chroma-based and pure VR conditions. This is due to the fact that, in the present study, all conditions share the same hand-tracking algorithm, while in Villegas et al. (2020), the tracking in the VR condition was done through the VR controllers.

Our results are not in line with the conclusions extracted in Gisbergen et al. (2020), where users needed to walk on a 60 m high broken pathway under two different conditions: (1) high resemblance shoes, or (2) generic shoes. Authors stated that under extreme situations which trigger psychological arousal such as stress, high avatar-owner resemblance is not a requirement.

6.5 Benefits

The use of video-based self-avatar, provides some important benefits, when compared with traditional avatars. First, there is no need to scale the avatar to match the user’s body dimensions, as segmentation is done using a 1:1 scale, allowing the user to have an accurate spatial perception. Also, there are no problems related to misalignment between the virtual and real body (Ogawa et al. 2020a), uncanny valley effect (Bhargava et al. 2021), or self-avatar follower effect (Gonzalez-Franco et al. 2020a). Extending egocentric segmentation to other items beyond human body parts, would allow the user to easily interact with real objects while immersed. Last but not least, users recognize the avatars with their own bodies with high acceptance values.

6.6 Drawbacks

There are still certain factors that can be improved for a better integration of video-based self-avatars in MR applications. One clear issue is the discrepancy in illumination conditions between the video-based self-avatar and the immersive environment where it is integrated (e.g. 360 video, VR environment). It would be also desirable to further investigate how to discard false positives from the segmentation algorithm, so that they do not downgrade the user experience, e.g. by using distance information.

6.7 Applications

For evaluation purposes, the technology of video-based self-avatar has been presented in this work by being integrated into a gamified immersive experience. However, we believe it can play an important role in many emerging use cases, namely:

- Seeing your own body and others peers with high fidelity (Joachimczak et al. 2022) on immersive telepresence systems (Arora et al. 2022; Izumihara et al. 2019) or other types of immersive communication systems (Perez et al. 2022) can further foster SoP, co-presence, and communication skills.
- Seeing your own body and being able to seamlessly interact with real objects for immersive training purposes (Ipsita et al. 2022) or for other purposes such as seeing your notebook or smartphone while attending an immersive conference.

6.8 Limitations

One of the limitations of this study is related to the lack of comparison with a full-body avatar condition due to the lack of mature full-body commercial tracking solutions that do not require additional hardware. Also, the conclusions extracted from our user study are extracted from questionnaires answered by a population subset, which, although balanced in terms of gender, lacks other diversity factors such as height, age, and skin color.
7 Conclusions

In this work, we proposed an E2E system that provides full-body video-based self-avatar user representation for MR applications. This E2E system can be divided into three different modules: the video pass-through solution for VR goggles, a real-time egocentric body segmentation algorithm, and an optimized offloading architecture. The E2E system has been evaluated using an arbitrary hardware setup. The achieved results showed the implementation's capability to fulfill the real-time requirements, providing update rates above 60 Hz. Finally, we described our delay-correction buffers to overcome the round trip latencies above the update rate of 60 Hz ensuring an accurate alignment between the color and received mask frames.

To the best of our knowledge, this is the first user study that conducts a subjective evaluation of a video-based self-avatar through a gamified experienced performed on a set of 58 participants. The goal was to evaluate and compare three different user representation solutions: virtual hands, chroma-based full-body segmentation, and deep-learning-based full-body segmentation. The overall results allowed us to reach two main conclusions: (1) video-avatars provided marginal improvement only in certain components of the perceived embodiment compared to the virtual hand's solution, (2) deep-learning solution offers better visual quality and perception of ownership and appearance than the chroma solution. The observed low differences between the conditions might be since, in all three conditions, the same robust and accurate hand-tracking algorithm is provided by the Meta Quest 2. We also observed that the scores obtained for presence or embodiment are high for all conditions, making it difficult to discriminate among them. Besides, the proposed E2E video-based solution is at the edge of the state-of-the-art, and consequently, it lacks the maturity provided by the commercial virtual-hands solution. However, the presented E2E system is already useful in its current state and can be considered the starting point for future MR distributed implementations and potential improvements. Finally, we discussed some advantages and disadvantages of video-based self-avatar solutions and potential applications. We believe this solution can be relevant in applications such as telepresence, training, and education.

Appendix A: Questionnaire

Users were asked to fill out a 30-item subjective questionnaire, where 12 were used to measure embodiment, 7 for presence, 6 for visual quality, and 4 for acceptability.

Embodiment questions (#12)

- Q1—I felt as if the virtual body was my own body
- Q2—I felt as if the virtual avatar I saw was someone else
- Q3—It seemed as if I might have more than one body
- Q6—It felt like I could control the virtual avatar as if it was my own body
- Q7—The movements of the virtual avatar were caused by my movements
- Q8—I felt as if the movements of the virtual avatar were influencing my own movements
- Q9—I felt as if the virtual avatar was moving by itself
- Q14—I felt as if my body was located where I saw the virtual avatar
- Q17—It felt as if my real body were turning into an avatar body
- Q19—At some point it felt that the virtual avatar resembled my own real body, in terms of shape, skin tone, or other visual features.
- Q20—I felt like I was wearing different clothes from when I came to the laboratory
- Q25—I had the feeling that if I fell off the bridge I was going to hurt myself

Presence questions (#7)

- Involvement—How natural did your interactions with the environment seem?
- Involvement—How compelling was your sense of objects moving through space?
- Involvement—How much did your experiences in the virtual environment seem consistent with your real-world experiences?
- Haptic—How well could you move or manipulate objects in the virtual environment?
- Adaptation—How quickly did you adjust to the virtual environment experience?
- Adaptation—How proficient in moving and interacting with the virtual environment did you feel at the end of the experience?
- Adaptation—How well could you concentrate on the assigned tasks rather than on the mechanisms used to perform those tasks?

Visual quality (#6)

- Please rate the perceived quality of the environment
- Please rate the perceived quality of the objects
- Please rate the perceived quality of yourself
• Please rate the perceived quality of your arms
• Please rate the perceived quality of your legs
• Please rate how annoying were the false positives (part of the scene beyond your own body that are wrongly segmented)

Acceptability (#4)

• Sickness—Did you feel any sickness or discomfort during the experience? Please rate it.
• Global QoE—How would you rate the quality of the experience globally?
• NPS-R—How likely is it that you would recommend this experience to a friend or colleague?
• NPS-P—How much would you pay to play this game?

Funding The funding was provide by Marie Skłodowska-Curie ETN TeamUp5G; Grant No. 813391

Data Availability The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Alae G, Deasi AP, Pena-Castillo L et al. (2018) A user study on augmented virtuality using depth sensing cameras for near-range awareness in immersive vr. In: IEEE VR's 4th workshop on everyday virtual reality (WEVR 2018), p 3
Argelaguet F, Hoyet L, Trico M et al. (2016) The role of interaction in virtual embodiment: effects of the virtual hand representation. In: Proceedings of IEEE VR, pp 3–10
Arora N, Suomalainen M, Poulke M et al. (2022) Augmenting immersive telepresence experience with a virtual body. arXiv:2202.00900
Bazarevsky V, Grishchenko I, Raveendran K et al. (2020) Blazepose: on-device real-time body pose tracking. arXiv:2006.10204
Bhargava A, Venkatakrishnan R, Venkatakrishnan R et al. (2021) Did I hit the door effects of self-avatars and calibration in a person-plus-virtual-object system on perceived frontal passability in vr. IEEE Trans Vis Comput Graph 28:4198–4210
Bonfert M, Lemke S, Porzel R et al. (2022) Kicking in virtual reality: the influence of foot visibility on the shooting experience and accuracy. In: 2022 IEEE conference on virtual reality and 3D user interfaces (VR). IEEE, pp 711–718
Bozgeylikli LL, Bozgeylikli E (2022) Tangiball: foot-enabled embodied tangible interaction with a ball in virtual reality. In: 2022 IEEE conference on virtual reality and 3D user interfaces (VR). IEEE, pp 812–820
Bruder G, Steinicke F, Rothaus K et al. (2009) Enhancing presence in head-mounted display environments by visual body feedback using head-mounted cameras. In: Proceedings of international conference on CW, pp 43–50
Chen GJ, Lee BM, Lindeman RW et al. (2017) The effect of user embodiment in AV cinematic experience. In: Proceedings of ICAT-EGVE
Chen LC, Zhu Y, Papandreou G et al. (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV), pp 801–818
Dewez D, Fribourg R, Argelaguet F et al. (2019) Influence of personality traits and body awareness on the sense of embodiment in virtual reality. In: 2019 IEEE international symposium on mixed and augmented reality (ISMAR), pp 123–134
Dodds TJ, Mohler BJ, Bülthoff HH (2011) Talk to the virtual hands: self-animated avatars improve communication in head-mounted display virtual environments. PLoS ONE 6(10):e25759
Ebrahimi E, Hartman LS, Robb A et al. (2018) Investigating the effects of anthropomorphic fidelity of self-avatars on near field depth perception in immersive virtual environments. In: 2018 IEEE conference on virtual reality and 3D user interfaces (VR). IEEE, pp 1–8
Fiore LP, Interrante V (2012) Towards achieving robust video selfavatars under flexible environment conditions. Int J VR 11(3):33–41
Fribourg R, Argelaguet F, Lécuyer A et al. (2020) Avatar and sense of embodiment: studying the relative preference between appearance, control and point of view. IEEE Trans Vis Comput Graph 26(5):2062–2072
Garrido-Jurado S et al. (2014) Automatic generation and detection of highly reliable fiducial markers under occlusion. Pattern Recogn 47(6):2280–2292
Gisbergen MSv, Sensagir I, Relouw J (2020) How real do you see yourself in VR? The effect of user-avatar resemblance on virtual reality experiences and behaviour. In: Augmented reality and virtual reality. Springer, pp 401–409
Gonzalez-Franco M, Peck TC (2018) Avatar embodiment. Towards a standardized questionnaire. Front Robot AI 5:74
Gonzalez-Franco M, Cohn B, Ofek E et al. (2020a) The self-avatar follower effect in virtual reality. In: 2020 IEEE conference on virtual reality and 3D user interfaces (VR). IEEE, pp 18–25
Gonzalez-Franco M, Egan Z, Peachey M et al. (2020b) Movebox: democratizing mocap for the microsoft rocketbox avatar library. In: 2020 IEEE international conference on artificial intelligence and virtual reality (AIVR). IEEE, pp 91–98
Gonzalez-Franco M, Ofek E, Pan Y et al. (2020c) The rocketbox library and the utility of freely available rigged avatars. Front Virtual Real 1:20
Gonzalez-Morin D, Gonzalez-Sosa E, Perez-Garcia P et al. (2022a) Bringing real body as self-avatar into mixed reality: a gamified volcano experience. In: 2022 IEEE conference on virtual reality and 3D user interfaces abstracts and workshops (VRW), pp 794–795
González Morín D, López Morales MJ, Pérez P et al. (2022b) Tcp-based distributed offloading architecture for the future of untethered immersive experiences in wireless networks. In: ACM international conference on interactive media experiences. Association for Computing Machinery, New York, NY, USA, IMX ’22, pp 121–132

Gonzalez-Sosa E, Perez P, Tolosana R et al. (2020) Enhanced self-perception in mixed reality: egocentric arm segmentation and database with automatic labeling. IEEE Access 8:146887–146900

González-Sosa E, Gagic A, Gonzalez-Morín D, et al. (2022) Real-time egocentric segmentation for video-self avatar in mixed reality. arXiv:2207.01296

Gruosso M, Capece N, Erra U (2021) Exploring upper limb segmentation with deep learning for augmented virtuality

Günther T, Franke IS, Groh R (2015) Aughanded virtuality—the hands in the virtual environment. In: Proceedings of IEEE 3DUI, pp 157–158

Guo Y, Liu Y, Georgiou T et al. (2018) A review of semantic segmentation using deep neural networks. Int J Multimed Inf Retr 7(2):87–93

Hollfeldt T, Heegaard PE, Varela M et al. (2016) QoE beyond the MOS: an in-depth look at QoQ via better metrics and their relation to MOS. Qual User Exp 1(1):1–23

Ipsita A, Erickson L, Dong Y et al. (2020) Sensing the weight of objects in virtual and augmented reality. In: Proceedings of the 2020 CHI conference on human factors in computing systems, pp 1–12

Izumiha A, Uria D, Hiyama A et al. (2019) Exleap: minimal and highly available telepresence system creating leaping experience. In: 2019 IEEE conference on virtual reality and 3D user interfaces (VR). IEEE, pp 1321–1322

Jayarat L, Wood J, Gibson M (2017) Improving the immersion in virtual reality with real-time avatar and haptic feedback in a cricket simulation. In: 2017 IEEE international symposium on mixed and augmented reality (ISMAR-adjunct). IEEE, pp 310–314

Joachimczak M, Liu J, Ando H (2022) Creating 3D personal avatars with high quality facial expressions for telecommunication and telepresence. In: 2022 IEEE Conference on virtual reality and 3D user interfaces abstracts and workshops (VRW). IEEE, pp 856–857

Lee K, Kacorri H (2019) Hands holding clues for object recognition in teachable machines. In: Proceedings of the 2019 CHI conference on human factors in computing systems, pp 1–12

Lee GA, Chen J, Billinghurst M et al. (2016) Enhancing immersive cinematic experience with augmented virtuality. In: IEEE international symposium on mixed and augmented reality, pp 115–116

Li C, Kitani KM (2013) Pixel-level hand detection in ego-centric videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3570–3577

Lok B, Naik S, Whitton M et al. (2003) Effects of handling real objects and self-avatar fidelity on cognitive task performance and sense of presence in virtual environments. Presence 12(6):615–628

Loper M, Mahmood N, Romero J et al. (2015) SMPL: a skinned multiperson linear model. ACM Trans Graph (TOG) 34(6):1–16

McManus E, Bodenheimer B, Streuber S et al. (2011) The influence of avatar (self and character) animations on distance estimation, object interaction and locomotion in immersive virtual environments. In: Proceedings of the ACM SIGGRAPH SAP, pp 37–44

Mürraiz FM, Roumeliotis SI (2008) A Kalman filter-based algorithm for IMU-camera calibration: observability analysis and performance evaluation. IEEE Trans Rob 24(5):1143–1156

Morín DG, Pereira F, Gonzalez-Sosa E et al. (2022) Democratic video pass-through for commercial virtual reality devices. In: 2022 IEEE conference on virtual reality and 3D user interfaces abstracts and workshops (VRW). IEEE, pp 790–791

Narwaria M, Krasula L, Le Callet P (2018) Data analysis in multimedia quality assessment: revisiting the statistical tests. IEEE Trans Multim 20(8):2063–2072

Ogawa N, Narumi T, Hirose M (2020a) Effect of avatar appearance on detection thresholds for remapped hand movements. IEEE Trans VCG 27:3182–3197

Ogawa N, Narumi T, Kuzuoka H et al. (2020b) Do you feel like passing through walls? Effect of self-avatar appearance on facilitating realistic behavior in virtual environments. In: Proceedings of the 2020 CHI conference on human factors in computing systems, pp 1–14

Pan Y, Steed A (2017) The impact of self-avatars on trust and collaboration in shared virtual environments. PLoS ONE 12(12):e0189078

Pan Y, Steed A (2019) How foot tracking matters: the impact of an animated self-avatar on interaction, embodiment and presence in shared virtual environments. Front Robot AI 6:104

Perez P, Gonzalez-Sosa E, Kachach R et al. (2019) Immersive gastronomic experience with distributed reality. In: Proceedings of IEEE WEVR, pp 1–4

Perez P, Gonzalez-Sosa E, Kachach R et al. (2021) Ecological validity through gamification: an experiment with a mixed reality escape room. In: 2021 IEEE international conference on artificial intelligence and virtual reality (AIVR). IEEE, pp 179–183

Perez P, Gonzalez-Sosa E, Gutierrez J et al. (2022) Emerging immersive communication systems: overview, taxonomy, and good practices for QoE assessment. Front Signal Process 2

Pigny PO, Dominjon L (2019) Using CNNs for users segmentation in video see-through augmented virtuality. In: 2019 IEEE international conference on artificial intelligence and virtual reality (AIVR), pp 229–2295

Rauter M, Abseher C, Safar M (2019) Augmenting virtual reality with near real world objects. In: Proceedings of IEEE VR, pp 1134–1135

Reichheld FF (2003) The one number you need to grow. Harv Bus Rev 81(12):46–55

Slater M, Usoh M (1993) The influence of a virtual body on presence in immersive virtual environments. In: Proceedings of VR, pp 34–42

Sommer P, Schellroth F, Fischer M et al. (2018) Message-oriented middleware for industrial production systems. In: 2018 IEEE 14th international conference on automation science and engineering (CASE). IEEE, Munich, Germany, pp 1217–1223

Steed A, Pan Y, Zisch F et al. (2016) The impact of a self-avatar on cognitive load in immersive virtual reality. In: 2016 IEEE virtual reality (VR), pp 67–76

Tang Y, Wang Z, Lu J, Feng J, Zhou J (2018) Multi-stream deep neural networks for RGB-D egocentric action recognition. IEEE Trans Circuits Syst Video Technol 29(10):3001–3015

Thaler A, Piryanakova I, Stefanucci JK et al. (2018) Visual perception and evaluation of photo-realistic self-avatars from 3D body scans in males and females. Front ICT 5:18

Tome D, Alldieck T, Peluse P et al. (2020) Selfpose: 3D egocentric pose estimation from a headset mounted camera. arXiv:2011.01519

Villegas A, Perez P, Kachach R et al. (2020) Realistic training in VR using physical manipulation. In: 2020 IEEE conference on virtual reality and 3D user interfaces abstracts and workshops (VRW), pp 109–118

Waltemate T, Gall D, Roth D et al. (2018) The impact of avatar personalization and immersion on virtual body ownership, presence, and emotional response. IEEE Trans Vis Comput Graph 24(4):1643–1652

Wittmer BG, Singer MJ (1998) Measuring presence in virtual environments: a presence questionnaire. Presence 7(3):225–240
Witmer BG, Jerome CJ, Singer MJ (2005) The factor structure of the presence questionnaire. Presence Teleoper Virtual Environ 14(3):298–312
Xiang W, Mao H, Athitsos V (2019) Thundernet: a turbo unified network for real-time semantic segmentation. In: 2019 IEEE winter conference on applications of computer vision (WACV). IEEE, pp 1789–1796
Xu C, He J, Zhang X et al. (2019) Towards human motion tracking: multi-sensory IMU/TOA fusion method and fundamental limits. Electronics 8(2):142
Zhang Z (2000) A flexible new technique for camera calibration. IEEE Trans Pattern Anal Mach Intell 22(11):1330–1334

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.