Fusing exocentric and egocentric real-time reconstructions for embodied immersive experiences

1st Stuart Duncan
Information Science
University of Otago
Dunedin, New Zealand
uohci@sjmd.dev

2nd Holger Regenbrecht
Information Science
University of Otago
Dunedin, New Zealand
holger.regenbrecht@otago.ac.nz

3rd Tobias Langlotz
Information Science
University of Otago
Dunedin, New Zealand
tobias.langlotz@otago.ac.nz

***Abstract*—Experiencing one's own body in virtual and mixed reality can enhance applications such as 3D teleconferencing, physical and psychological rehabilitation, and natural 3D user interfaces. Embodied experiences require a dynamic virtual body to represent the user. Typical virtual bodies consist of rigged mesh models which are animated using expensive and cumbersome motion capture systems, or heavy reliance on models of human movement. Models of human appearance and movement are notoriously susceptible to undesirable and "uncanny" appearances, and are often unconvincing as a result. An alternative is to reconstruct the user in real time without relying on a motion capture system or on visual or movement models. With this approach the appearance of the virtual body and its motion are inherently natural by virtue of being directly captured, and embodied experiences are possible even at low levels of detail. However, the necessarily sparse arrangement of reconstruction cameras often produces incomplete virtual bodies. This is partly due to the disparity between the egocentric (first-person) view of the virtual body and the typically exocentric (third-person) perspectives of the reconstruction cameras. In this paper we present a method for reconstructing a more complete view of the user with a minimal number of cameras by combining head-worn egocentric with exocentric depth-sensing cameras, with a focus on the first-person view of the virtual body. We describe our approach to producing the virtual body, including camera registration methods and key technical performance metrics. We also provide insights from a user study with 26 participants indicating that our approach has the potential to increase the sense of embodiment and the perception of the completeness of the virtual body.**

***Index Terms*—virtual reality, mixed reality, embodiment**

I. Introduction

In immersive virtual reality (VR) a user's view of their physical body is obstructed by the HMD and without the inclusion of a *virtual body* the user's lack a visual representation of themselves in the immersive experience. While users can feel a sense of presence and plausibility without visual body representation, the inclusion of an appropriate and dynamic virtual body can enhance the experience in several ways and is even essential for some application scenarios: Embodied users can interact with the VR experience using natural modes of interaction without requiring handheld controllers. In particular, when providing a more natural (larger) field of view, showing a virtual body is crucial [1]. When multiple users share an experience, their virtual bodies can support non-verbal communication in the form of gestures, pose and body language, and even facial expressions supporting the feeling of being together in the same place—a sense of co-presence [2].

Fig. 1. A real-time reconstruction system for embodied experiences, using two stationary exocentric stereo cameras in combination with two egocentric stereo cameras attached to an HMD (left). The first-person view of the user shows a dynamically reconstructed virtual body with data contributed from both the exocentric and egocentric cameras (right).

In psychological and physical therapy, VR applications often rely on the user's perception of their own virtual body to directly affect therapeutic change: Anorexia and body dysmorphia can potentially be addressed by modifying features of a virtual body such as size and shape [3]. Similarly, neurorehabilitation for motor function recovery and chronic pain rely on the artificial movement of body parts, where the visual appearance of the virtual body and the plausibility of the movement are critical [4], [5]. Embodiment is usually characterised by a sense of agency over the virtual body, correct self-location of the body, appropriate visual appearance, and a feeling of ownership, which emerges when an immersive experience supports these factors [6]. Agency requires that the movement of the virtual body closely follows the movement of the user's physical body, without spurious movement or large delays. Self-location typically requires that the virtual body is seen from a first-person (egocentric) perspective, although users can still experience some embodiment in third-person (exocentric) views such as in a virtual mirror. Finally, the sense of ownership can be improved by a *personalised* virtual body which closely resembles the user's physical body. Personalised

979-8-3503-7051-5/23/$31.00 © 2023 IEEE

virtual bodies can also allow users in shared experiences to recognise one another and to believe, to some degree, that they are interacting with real people rather than representative *avatars* [7].

The challenge for many immersive VR systems is to enable embodiment supporting agency, self-location, and in particular stronger ownership by virtual body personalisation. Virtual bodies typically consist of textured mesh models which are rigged for animation, and animated using motion capture system. However, the degree of personalisation is limited by the parameters of the visual model and often personalisation is a tedious manual process. Furthermore, the *plausibility* of the virtual body as a whole also depends on the accuracy of the movement model and the quality of the motion capture data used for animation. Motion capture systems typically track body parts which are instrumented with reflective markers (for conventional motion capture), or those which are visible to the tracking cameras (for recent iterations of VR hardware). In addition to noise and other spurious tracking results, motion of body parts which are not directly tracked must be inferred using a model of human movement which can further reduce plausibility.

Alternatively, one or more depth-sensing (RGBD) cameras can be used to capture the user's visual appearance and their movement simultaneously. This approach requires more processing power and the resulting virtual bodies commonly exhibit less *static* visual and geometric detail. However, *dynamic* details such as deformation of skin and clothing are captured directly and thus are included in the virtual body without modelling. These real-time approaches often use a model or template such as SMPL [8] to produce "cleaner" models and increase the level of visual detail, however this risks smoothing the result and removing some directly captured dynamic detail.

More recently, research has demonstrated embodied experiences using data captured using RGBD cameras directly without models or templates [9]–[11]. This work goes a step further and relies on sampled representations—such as point clouds and voxel grids—rather than using the sampled data to construct mesh models. Despite the lower visual quality, users often prefer these directly sampled representations over mesh models and consistently report that sampled representations support embodied and co-present experiences.

However, a problem for these directly sampled approaches is the lack of data when certain areas of the user's body are consistently hidden, often through self-occlusion by the user themselves (e.g. see figure 2). This becomes even more problematic as often the user has a different view than the (exocentric) camera(s) used for direct capture. While the cameras might not see certain areas (e.g. inside the hands, elbows, feet) they could be relevant for the users and important to enable embodiment. We therefore propose to include first-person *as well as* third-person reconstruction cameras in the system, thereby increasing the chance that the first-person view is always complete while preserving the third-person view required for co-presence.

In this paper, we describe our real-time volumetric capture system for embodied and co-present experiences, along with a user trial to verify its effectiveness for embodied experiences. The capture and rendering system is described elsewhere [citation-anonymised], and here we focus on the extension of the base system to include exocentric *and* egocentric reconstruction cameras. Using our system we conducted a user trial with 26 participants to test the technical effectiveness, the feasibility of the approach and to investigate the contribution of the egocentric cameras to the user's experience. We consider the user's sense of embodiment and the user's perception that their physical body is completely represented by the virtual body. The contribution of this work is a system for using directly captured, sampled representations of users for embodied and co-present experiences, along with a preliminary user trial to verify the efficacy of the system in supporting embodied experiences.

Fig. 2. An example of undesired self-occlusion with only exocentric cameras (top) and complete reconstruction achievable with fused exocentric and egocentric views (bottom)

II. Related Work

Our work touches on several topics within immersive and virtual reality, and aims to support similar experiences to several commercially available systems. In the following, we will first introduce the concepts of embodiment and co-presence,

and describe the technical requirements for supporting these in immersive VR. We then describe some commercially available and research systems intended for embodied and co-present experiences, with an emphasis on the degree to which these systems represent support plausible or authentic representation of the users. In this context we briefly describe research aiming to produce personalised virtual body models using one or more cameras to capture the visual appearance of the users. Within this area, the majority of systems use animated mesh models to represent users. We discuss this approach to provide contrast and comparison with dynamically reconstructed representations which we consider next. Most work on dynamic reconstruction relies on templates or models such as SMPL [8], but user appearance as well as motion is captured in real-time. Here, some dynamic details are captured without modelling, but in fitting data to prior model some captured detail is still removed and a smoother model is substituted. Finally, we present some closely related research work which uses real-time point cloud and sparse voxel reconstructions to represent users to themselves and to one another. User representation by model-free, dynamic, sampled reconstruction is the basis for our existing work, which we discuss before describing the extension of our existing work to include egocentric reconstruction cameras.

A. Defining Embodiment and Co-presence

Embodiment in virtual reality refers to a sense of ownership of and agency over a virtual representation of the user [12]. Agency requires that the user's movement is mapped appropriately to the virtual representation, normally by animating a humanoid model to follow the user's physical movements. Ownership is a composite factor, which typically requires appropriate self location of the virtual body (e.g. it should be seen from a first-person view) and ownership is strengthened with increasing agency and when the visual appearance matches the physical appearance of the user [13].

When dynamically constructing the virtual body, some latency is inevitable, and it is important to minimise this to effectively elicit a sense of agency and ownership [14]. This is an area where motion capture systems have an advantage, because the processing is simpler and the amount of data is significantly reduced when compared to producing a 3D reconstruction. However, research has suggested that embodiment is not significantly impacted by latency up to 210 ms [15], and existing research has shown that dynamic point cloud and sparse voxel grid representations with no specific optimisation to minimise latency do support embodiment [11], [16].

Appropriate self location requires that the virtual body is aligned with the user's physical body in the VR coordinate system so that it is seen from a natural first-person view. When motion capture systems are used, the motion capture tracking system may replace the VR tracking system or the two systems may be registered to one another in order to align the virtual body. Where virtual bodies are reconstructed dynamically, the reconstruction coordinate system must be registered to the VR tracking system to correctly align the virtual body.

B. Systems for Embodiment and Co-Presence

There is now a large body of research, along with several commercial offerings which target embodiment in immersive virtual reality. An easily accessible system is the Meta Horizon[1] platform, where users are represented by a partial virtual body which does not include the legs or lower half of the torso. Horizon virtual bodies can be personalised to some degree, however the process is manual and the range of parameters is limited. Facial animations are synthesized from voice analysis, and eye movements are inferred using an unpublished model. Arms, hands, head, and torso are animated by using the HMDs inside-out tracking system for motion capture. In contrast, some researchers favour more "authentic" representations, and so do we. Yu et al. [11] could show that point cloud reconstructions are superior to animated mesh avatars regarding perceived co-presence, social presence, behavioural impression, and "humanness". Gamelin et al. [9] also could show that 2.5D point cloud representations of users outperform 3D pre-constructed avatars in a remote collaboration setting. Park et al. [16] and Duncan et al. [7] use voxel-based representations of users for indigenous storytelling and could show the technique's feasibility and effectiveness in eliciting presence and co-presence between users, and also with pre-recorded people. Regenbrecht et al. [10] blend real and virtual objects and users as voxel representations in a way that both realms are perceived as indistinguishable.

C. Reconstructing Humans

Instead of personalising an existing avatar, a mesh model may be constructed directly and rigged for animation using a kinematic rigging template. Typically, this approach uses many synchronised cameras to simultaneously capture the user's whole body [17]–[19] or body parts [20]. Complete and highly detailed models can be constructed in this way, however this approach has some limitations; During the capture process the user must hold a static pose that allows a full view of the body or desired body parts, which may be impossible for certain users (e.g. therapeutic applications for treating partial paralysis due to a brain injury). The cost and complexity of the reconstruction system may also be prohibitive [21] and while the reconstruction may be detailed, convincing, and personalised, it is still a static model which lacks many of the dynamic deformation effects which occur in reality. This approach, as well as the personalisation of existing models, also requires motion capture to animate the virtual body, along with the disadvantages (and advantages) of using a VR tracking system or motion capture system to animate the model.

To overcome some of these limitations, virtual bodies can instead be constructed dynamically. However, since the user is free to move around while the virtual body is being captured, the assumption that the user's whole body is completely visible is no longer valid and this can be problematic in many actual applications and is the focus of this work. Some approaches

[1] https://www.meta.com/en-gb/help/accounts/what-is-horizon/

mitigate this issue by increasing the number of reconstruction cameras which in turn often also increases the latency, all things we aim to minimise. Alternative methods use a few cameras in combination with a model that is used to interpolate or infer parts of the body which are not directly observed by the camera. These approaches commonly use the SMPL human body model [8] as a 'template' and as such this body of work is similar to personalising a mesh model with directly captured texture information. More recently, research has demonstrated the use of generative machine learning models to infer the missing parts of the reconstruction [22]–[26]. However, like avatar personalisation, these methods often do not capture unique visual features of the user's physical body and, even if they can, they will do so only once it is visible and substitute generic data if it is not. It also remains challenging to correctly infer the same subtle dynamic deformation effects which are missing from animated static mesh models. We rather approach the problem by considering the expected viewing poses, and aiming to capture the user in such a way that the captured data perceived as complete from these views.

D. Dynamic, Sampled Reconstructions

Rather than following the dominant paradigm of animating a textured mesh surface model, we intend to explore the use of dynamic volumetric reconstructions with *sampled* visual representations. That is, using a discretely sampled representation such as a point cloud or a voxel grid, and rendering it in its native structure without first converting it to a textured mesh model. Research so far supports this direction of inquiry; Point clouds and sparse voxel grids are shown to support embodiment and co-presence in immersive environments, when rendered in their native form [9], [27]. Furthermore, sampled structures in the form of voxel grids are often used as part of a standard reconstruction pipeline for producing mesh models [28], or more computationally or memory efficient internal representation of some data [29]. In addition to the computational advantages of voxel grids over point clouds, the representation inherently encodes a consistent level of geometric detail over the whole grid, while visual detail of dynamically captured data varies with distance to the camera. For these reasons we choose to use a voxel grid as both our internal data representation and the visual representation, instead of converting the grid to a mesh surface model.

Overall, our approach presents a system for embodiment that mitigates the issues of self-occlusions and missing visual information apparent in existing systems which use only exocentric cameras. The key idea is to use dynamic sampled reconstructions captured using exocentric (external to the user) and egocentric (head-worn) cameras to ensure that data is always captured from the user's first-person view, and from selected third-person views. Using this approach we aim to ensure that the virtual body is viewed as complete, without relying on models of human appearance or movement and without increasing the number of (exocentric) cameras since both of these alternatives can negatively affect embodiment.

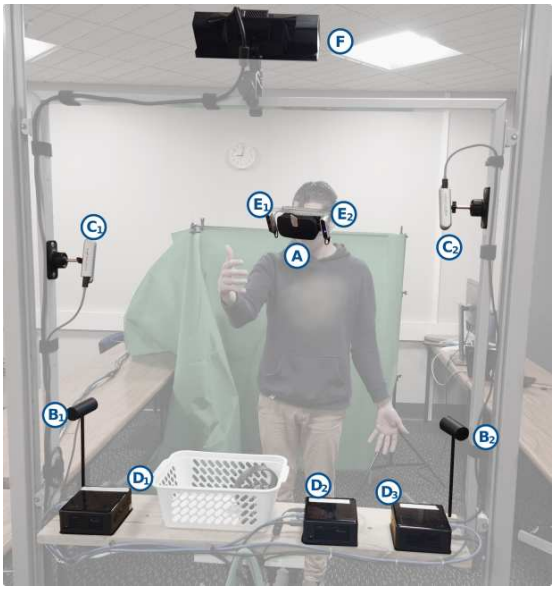

Fig. 3. The volumetric capture system used for the user experiment consists of an Oculus Rift CV1 HMD (A) with two Oculus tracking cameras (B_1, B_2), two egocentric cameras (E_1, E_2), and two exocentric cameras (C_1, C_2). A Kinect (F) is included to track the user's pose for the experiment but does not contribute to the reconstruction. Each egocentric camera and the Kinect are connected to a dedicated Intel NUC computer (D_{1-3}).

III. System Overview

Voxelvideo streams (our real-time volumetric video encoding) of the user consist of a sequence of voxelframes. Each frame is a sparse, coloured voxel grid, which is constructed from a coloured point cloud acquired by a RGBD camera. Occupied voxels each contain one or more coloured points, and the voxel colour is the average colour of the points. The system runs on multiple networked computers to distribute the computational load, and reconstructions are sent between computers by way of a compact serial encoding of the voxelframes.

In this section we present an overview of the network-distributed architecture of the system (subsection III-A), and the octree data structure that we use to construct and encode voxelframes (subsection III-B). These components form the core of the system which, with a suitable camera registration method, supports conventional egocentric-only sampled reconstruction and embodied experiences [citation-anonymised].We then describe our extension of this system to include egocentric (HMD-attached) reconstruction cameras which are registered to the world coordinate system using the VR tracking system to update their registration in real time (section III-C). We integrate our system with Unreal Engine 4[2] (UE4) using the plugin system, which allows integration of voxelvideo streams with arbitrary UE4 projects. Details of our method

[2]https://www.unrealengine.com/

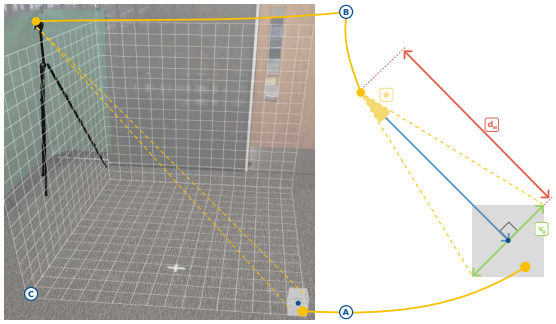

Fig. 4. Physical measurements used to calculate v_g for the furthest voxel (A) from the camera (B) within the extents of the capture volume (C).

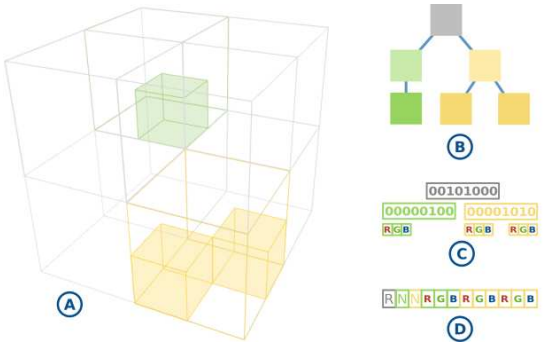

Fig. 5. An exemplary three-level octree showing the volume hierarchy (A) and tree structure of the (B) with three occupied leaf nodes. The serial encoding (C, D) is the breadth-first concatenation of the occupied internal nodes followed by the colours of the leaf nodes correspond to the serial encoding.

of integration, and the limitations and advantages are already described elsewhere [citation-anonymised].

A. System Architecture

Figure 3 shows the arrangement of the physical components of the capture system used for the experiment described in section IV, which we use as an example configuration to describe the system. The arrangement of the exocentric cameras, and the configuration of the hardware of that system is a typical system configuration which we use for embodied experiences [citation-anonymised]; Each camera is connected to a dedicated computer (here Intel NUCs[3] with 8th generation Intel i5 CPUs), and the VR system (an Oculus[4] Rift CV-1) is connected to a rendering computer (featuring an Intel i7, and an NVIDIA RTX 2080Ti GPU). The computers are connected over a dedicated gigabit LAN, with the rendering computer connected via a 10 Gigabit port and network adaptor.

Voxelframes are constructed on each NUC, converted to a compact serial encoding, and sent over UDP to the rendering component running on the rendering computer. The rendering component integrates reconstructions received from each NUC to produce a single coherent voxelframe consisting of data from all the cameras. Combined reconstructions are merged as received, and a combined frame is considered complete once one frame from each camera is integrated. All software components follow a triple buffered producer-consumer pattern which allows sequential stages of the reconstruction and processing pipeline to operate concurrently thereby minimising latency. The concurrent design of the processing pipeline also decouples total latency from the maximum update rate of the reconstruction system which is limited only by the slowest single stage of the pipeline. This arrangement relies on a voxel data structure which can be quickly constructed from a point cloud and converted to/from a compact serial encoding, which we describe in the following section. In section III-D we describe the latency of our system, and the method used to measure the total motion-to-photon latency. This includes measurement of the contribution of several stages of our processing pipeline, which relies mostly on the time taken to construct a voxel grid from the point cloud, and the conversion to and from the serial encoding.

[3]https://www.intel.com/content/www/us/en/products/details/nuc.html
[4]https://oculus.com

B. Voxelframes

We use a fixed-depth sparse voxel octree using a child-pointer structure during tree construction and to recover 3D positions of each occupied leaf node when decoding a serial octree. 3D positions of the leaf nodes must be recovered for rendering, as described in an earlier publication [citation-anonymised]. The grid has fixed spatial bounds which are dependent on the voxel size and the tree depth, which we choose for a particular system configuration. Typically, the tree depth is fixed at 11, which corresponds to a voxel grid with a size of 2^{11} vox^3 =2048 vox^3. This is sufficient for our research, where we allow users to move freely within a cube shaped capture volume of approximately 2.5x2.5x2.5 m^3, and typically use voxel sizes ranging from about 2 mm to 12 mm (where the voxel grid covers ≈4 m^3 and ≈24 m^3 respectively.). Figure 5 illustrates the spatial structure of the octree (limited to a depth of 2 for illustrative purposes) with some occupied voxels along with the corresponding tree structure and serial encoding. A full serial octree is the breadth-first concatenation of the child-occupancy bitfield for all the internal nodes, followed by a contiguous array of voxel colour data. Colour information is stored in three bytes in R-G-B order for both the in-memory and serial encoding.

We construct the octree by first transforming the point cloud from the coordinate system of each camera to the common world coordinate system as described in section III-C. Octrees are constructed by point-wise insertion of all aligned points in the point cloud. A depth threshold and a chroma-key filter are applied to the depth and colour images to reject background points, after which a crop volume is applied (e.g. our 2.56^3 m^3 volume) to remove any remaining irrelevant points. The resulting reconstruction consists of the user as well as any objects that are inside the crop volume. The in-memory octree structure is converted to the serial encoding, and streamed to the render component, where data from multiple cameras is fused into a single coherent representation. Data is fused by merging the octree structures, and averaging the colour of leaf

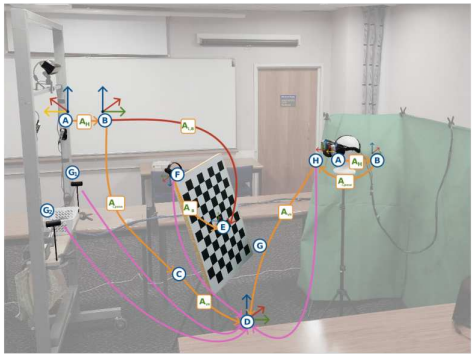

Fig. 6. Ego- and exocentric camera registration uses the same notation as equation 3. Two Oculus tracking cameras (G_1, G_2) define the world-aligned VR coordinate system (D). Captured data is first transformed from right-handed (A) to left-handed (B) camera-centric space, and then either to camera-common-space for exocentric cameras (C) or to HMD-space (H) for egocentric cameras. Data is then transformed to world-space (D).

nodes which are occupied in both trees. This naive approach is sufficient to fuse sparse grids so that the fused reconstruction preserves a single 'hull' of occupied voxels, so long as the residual errors of camera registration and depth estimation are within one half of the voxel width. Residual errors as large as one voxel width are tolerable since a single hull is still produced even though the hull thickness is doubled. For our relatively low levels of detail and the corresponding coarse voxel resolution we found this approach to fusion, and our camera registration method, to be sufficient.

The finest voxel grid resolution for a particular configuration of cameras (or the size of a leaf node) is chosen so that reconstructions produced from the camera point clouds produce a contiguous hull of occupied voxels in the grid. The minimum 'gapless' voxel size, v_g, corresponds to a voxel grid resolution that matches the spatial resolution of the point cloud which in turn depends on the angular resolution, ϕ, of a depth pixel and the maximum expected distance, d_W, between the camera and the subject as illustrated in figure 4

$$v_g = 2\sqrt{2}\, d_w \tan\left(\frac{\phi}{2}\right). \tag{1}$$

When $d < d_w$ multiple points in the point cloud will contribute to a single voxel, and the visual level of detail is reduced. This ensures that level of detail is consistent over the whole capture volume, even though the point cloud density varies with distance from the camera. While v_g is the lower bound on the voxel size, and we can choose larger voxel sizes to trade visual level of detail for latency and to achieve real-time performance on computationally limited hardware.

For a given capture volume and voxel size, $v \leq v_g$, we calculate the minimum octree depth, D, required for an octree to contain the whole volume

$$D = \left\lceil \log_2 \frac{V}{v_g} \right\rceil. \tag{2}$$

C. Camera Registration

To produce a coherent reconstruction from multiple cameras, the cameras must all be registered to the same coordinate space. We use the VR tracking system as the common coordinate system, since the VR system is registered to physical reality. To avoid VR system tracking drift inherent in more recent SLAM-based VR systems we use an Oculus Rift CV-1 since it uses stationary exocentric tracking cameras which serve as fixed reference points. Our camera registration consists of a set of rigid transforms which transform points from the camera coordinate system to the VR coordinate system. The combined registration transform, $\mathbf{A}_i$, for the i^{th} camera is composed of a handedness transform, $\mathbf{A}_H$, a camera pose transform, $\mathbf{A}_{i,pose}$, and a VR-device transform $\mathbf{A}_{VR}$:

$$\mathbf{A}_i = S\mathbf{A}_{VR}\mathbf{A}_{i,pose}\mathbf{A}_H. \tag{3}$$

We calculate the scale factor, S, to convert from camera base units, c_b, to a base unit of one voxel for a given voxel size, v

$$S = \left(\frac{c_b}{v}\right). \tag{4}$$

The handedness transform, $\mathbf{A}_H$, converts points expressed in the right-handed camera space to an equivalent expression in left-handed camera-centric space for compatibility with UE4. For the exocentric cameras, $\mathbf{A}_{VR}$ is the identity transform and $\mathbf{A}_{i,pose}$ transforms directly from camera space to VR space. For egocentric cameras $\mathbf{A}_{i,pose}$ transforms from camera space to the HMD coordinate system, and then $\mathbf{A}_{VR}$ transforms from HMD space to VR space. The HMD coordinate space, and $\mathbf{A}_{VR}$, are defined by the pose of the HMD, and $\mathbf{A}_{VR}$ is updated in real-time by retrieving the pose from the VR tracking system. HMD poses are retrieved at 100 Hz, and at significantly lower latency than the RGBD image acquisition and voxelframe reconstruction process. During the camera registration process we avoid introducing errors due to relative latency by only recording registration data while the target is stationary. During reconstruction, misalignment due to relative latency is mitigated by maintaining a timestamped ring buffer of the most recent 15 HMD poses, and choosing the pose whose age matches the estimated relative latency. We empirically determined the relative latency to be approximately 90 ms by adjusting its value and observing the relative alignment of egocentric and exocentric voxel grids under fast HMD rotation.

Camera registration requires estimating $\mathbf{A}_{i,pose}$ for all the cameras in the system, for which we rely on a chequerboard with a handheld VR controller rigidly attached. We first use hand-eye calibration [30] to estimate the transform between the controller and the coordinate system defined by the chequerboard. Using the calibrated controller-board assembly we then capture a set of point correspondences and apply partial Procrustes analysis to estimate $\mathbf{A}_{i,pose}$. Calibration data for both steps is only captured while the controller-board assembly is stationary, and for egocentric cameras while the

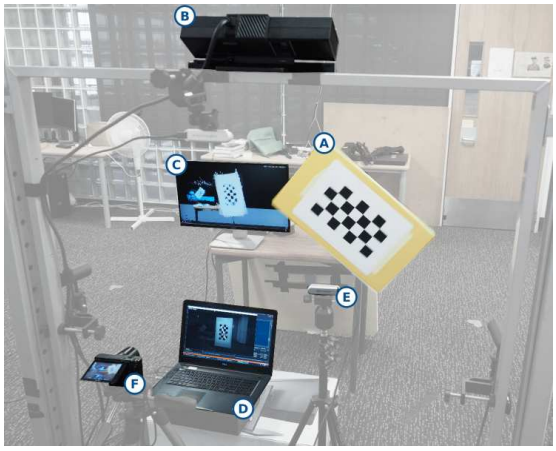

Fig. 7. The chequerboard (A) spins at a constant speed, and is captured by the reconstruction camera (B) and rendered in real-time on the screen (C). The laptop (D) uses the speed-measurement camera (E) to measure the rotational speed of the board. The observation camera (F) captures a snapshot of the physical board and its reconstructed image on the screen, from which the lag angle of the reconstructed board can be estimated.

HMD is stationary. Figure 6 illustrates the camera registration apparatus for our experimental system.

Reconstruction components apply $\mathbf{A}_i$ to all p_j points captured by the camera (after depth threshold and chroma-key filtering)

$$q_j = \lfloor \mathbf{A}_i p_j \rfloor . \tag{5}$$

The resulting set of world-aligned points, q_j, are inserted point-wise into the octree to create the voxelframe. Our method of fusing data from multiple cameras requires the residual alignment error between cameras to be less than half a voxel width otherwise the same surface observed by different cameras will produce multiple occupied hulls in the combined reconstruction. By visual assessment we confirmed that the experimental setup in Figure 3 produces a single hull to 6 mm.

D. System Latency

We estimated the latency to measure the motion-to-photon latency as experienced by the user. To do this we use a chequerboard rotating at a constant speed, and use the lag angle of its reconstruction to estimate the latency. We use two cameras to make measurements: One to estimate the rotational speed of the chequerboard, and one to simultaneously capture the rotation angle of the chequerboard and its reconstruction.

Chequerboard rotational speed, ω, is estimated from two observations of the chequerboard orientation vector, v_1 and v_2, separated by a time interval Δ_t:

$$\omega = \frac{1}{\Delta_t} cos^{-1} \left(\frac{||v_1|| \, ||v_2||}{v_1 \cdot v_2} \right) \tag{6}$$

where the orientation vector is defined as the vector between diagonally opposite internal chequerboard corners. We use an asymmetrical chequerboard so that the direction vector can be tracked without ambiguity over a whole rotation.

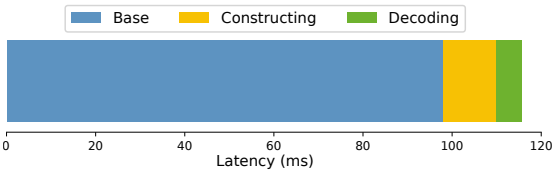

Fig. 8. Motion-to-photon latency consists of base latency (image exposure, USB and network transport, and the rendering pipeline), plus the time taken to construct the serial octree from a point cloud and decode it into the combined reconstruction.

The rotational speed of the chequerboard is estimated in real-time using OpenCV's SolvePnP function to find the chequerboard corners to estimate v_1 and v_2, and to measure Δ_t between the two observations. For a given frame rate and rotational speed we adjust the interval between measurements to ensure that a sufficient number of frames are discarded between v_1 and v_2 so that a rotation the rotation angle is between π and 2π. Rotational speed is then estimated by applying equation 6 and recorded alongside an image which simultaneously captures the physical board and its reconstruction. We then use the GIMP[5] angle measurement tool to manually estimate the relative angle between the board and its reconstruction in the captured image, since the image of the reconstruction is not detailed enough to reliably detect the chequerboard. We recover $\Delta_{t,lag}$ for a given observed rotational speed, ω, and observed relative angle, θ

$$\Delta_t = \frac{\theta}{\omega} \tag{7}$$

To avoid the relative orientation of the screen introducing an angle offset we measure the lag angle while the board is stationary and subtract the stationary lag angle from θ. To avoid rolling shutter distortion we use global-shutter cameras to measure the rotational speed of the board and to capture the images fore estimating the lag angle.

In addition to measuring the total motion-to-photon latency, the capture software is instrumented to measure the time required for each stage to process one frame. A breakdown of the latency into these components is shown in figure 8.

IV. Experimental Evaluation

We conducted a within-subjects experiment to compare the sense of embodiment and perceived body completeness between three conditions: User's reconstructed using exocentric cameras only (X), egocentric cameras only (E), and both egocentric and exocentric cameras (A). 26 participants were recruited from the university staff and students, as well as the public, and all recruited participants completed the experiment. The group consisted of one non-binary, 11 male, and 14 female participants with ages ranging from 21 to 66 and a median age of 33. Twenty-three participants reported prior experience with VR, and the remaining three reported no prior experience.

[5]https://gimp.org

The experimental system consisted of four RealSense D415 cameras, with two attached to the HMD (the egocentric cameras) and two attached to a large metal frame (the exocentric cameras). A Kinect 2 was used to track the user's pose for the purpose of the experiment, but did not contribute to the virtual body since the Kinect does not allow for control over the exposure and colour balance of the image and would introduce colour variation into the virtual body. Each camera was connected to an Intel NUC computer, which was connected via a dedicated gigabit Ethernet switch to the rendering computer (featuring an Intel i7, RTX2080Ti, and a 10 Gigabit Ethernet card). The VR experience was delivered by an Oculus CV-1 with two tracking cameras which were attached to the same metal frame as the RealSense cameras and Kinect to preserve camera registration. Figure 3 illustrates the arrangement of the cameras, the VR system, and the NUCs

Upon arrival, each participant first completed a consent form and a demographic survey. They were then immersed in a simple virtual environment consisting of a floor plane, the default UE4 sky sphere, and a large virtual mirror placed approximately to coincide with the position of the metal frame. A black square on the floor indicates the participant start position. Participants were first asked to identify the virtual mirror, the starting square, and virtual sun. They are then asked to confirm they can see the reflections of those objects in the virtual mirror. They are also invited to move around, and to look down and verbally confirm that they can see their virtual body. Then they were then allowed up to two minutes to familiarise themselves with the system, and asked to indicate when they were ready to begin the experiment.

The experiment begins by asking the participant to stand on the starting square, whereupon they are shown a stick figure illustrating a pose that they are asked to mimic by 'stepping into' into the pose and aligning their arms and legs with the stick figures. An example showing a photograph of the user and their corresponding first-person view is shown in figure 9. They are asked to verbally indicate when they are satisfied with their alignment at which point the skeleton is hidden and the participant is asked to step back onto the starting square. The task is repeated five times with a new stick figure pose each time, after which they remove the HMD and complete a post-immersion questionnaire to assess their sense of embodiment and the perceived completeness of the virtual body. The user then repeats the task with the same five skeletons for the other two conditions, completing the same post-immersion questionnaire each time. Once the participant has experienced all three conditions the experiment is concluded, and the participant is given a chocolate bar as a token of thanks for their participation.

Experimental condition order is randomised, with care taken to ensure that all six possible condition orders appear a similar number of times during the experiment. The same five stick figure poses are used for all conditions and for all participants in randomised order, where the stick figures were produced by capturing the pose of one of the researchers standing in front of the Kinect and holding a pose. Before showing the participant the stick figure it is scaled to match the participants height based on the relative heights of head position of the user and of the skeleton. Stick figure poses were captured prior to the experiment by using the Kinect to capture the experimenter standing in front of the virtual starting square. A range of poses was chosen to ensure that some of them exhibited self-occlusion, with a severe example shown in 2.

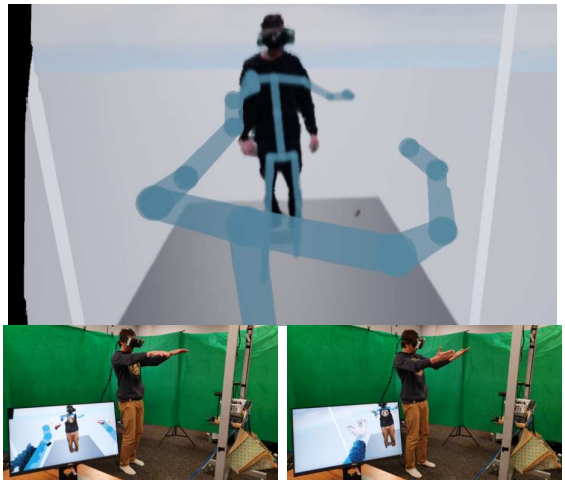

Fig. 9. A participant completing the skeleton matching task showing their first-person view of from the start position (top), and their physical final physical pose including their view with exocentric-only (bottom left) and ego-plus exocentric (bottom right) cameras.

A. Questionnaire

We use the post-immersion embodiment questionnaire described by Gonzalez and Peck [12], but omit the 'response' and 'tactile' sections since we do not use tactile or threat stimuli. Three additional questions are included to assess the degree to which the participants felt that their virtual body was complete:

1) I felt as if my whole body was present in the virtual space.
2) I felt as if the virtual body did not represent my whole body.
3) I felt like parts of the virtual body were missing.

The completeness responses are recorded on the same 7-point Likert scale as the embodiment responses, with a range between -3 (strongly disagree) and +3 (strongly agree). Embodiment scores are calculated as prescribed by Gonzalez and Peck, and completeness scores are calculated from the questions, C_1-C_3, listed above:

$$C = \frac{C_1 + C_2}{2} - C_3 \tag{8}$$

B. Results

Participant responses were assessed for normality using the Shapiro-Wilk test, and for equal variance using the Levene test. Embodiment data was all normal, however only the egocentric-only condition had significantly higher variance than the exocentric-only condition ($p < 0.01$) and the combined condition ($p < 0.05$). We then compared the conditions using a

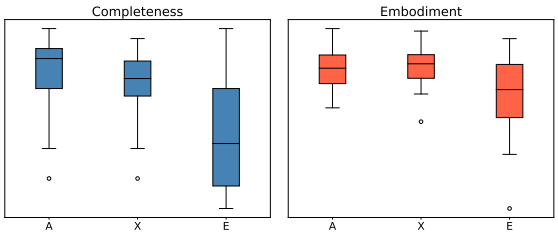

Fig. 10. Virtual body completeness perception, and the overall embodiment score (X=exocentric-only, E=egocentric-only, A=egocentric+exocentric)

Kruskal-Willis test, finding no significant difference between the exocentric and combined conditions, but a significant difference between the egocentric-only condition and both the exocentric ($p < 0.01$) and combined ($p < 0.05$) conditions. Effect size of including egocentric cameras alongside exocentric cameras was small (Cohen's d=0.29 & d=0.12 for completeness and embodiment respectively), while effect sizes for removing the exocentric cameras from a combined configuration (d=1.3 & d=0.81) and for replacing exocentric cameras with egocentric ones (d=1.1 & d=0.89) was large.

After decomposing embodiment scores into their component factors, we found no significant difference in the reported appearance, agency, and location factors of the embodiment score. A significant difference in the sense of ownership was found between the egocentric condition and the other exocentric-only ($p < 0.01$) as well as the egocentric+exocentric ($p < 0.01$).

During the experiment the participants' pose was tracked using the Kinect 2 skeleton tracking, and when the user indicated that they had satisfactorily matched the stick figure pose a snapshot of the skeleton tracking was saved. We used only the head, shoulders, hips, elbows, and knee joints to assess skeleton fit accuracy, since the remaining joints often exhibited tracking instability and position jitter. For each joint of each skeleton of each recorded pose, the magnitude of the error vector between the user joint and the reference joint was calculated. Recorded data was manually filtered to exclude records where the integral Kinect skeleton tracking failed. After filtering, we retained 78% of the data, corresponding to 96 samples for the egocentric and egocentric+exocentric conditions and 90 samples for the exocentric condition. Filtered data was assessed for normality and equal variance using Shapiro-Wilk and Levene analyses. None of the data was normally distributed, but all data had equal variance. As per Kruskal-Willis analysis we found no significant difference between any pair of conditions.

Overall, the results show that embodiment can be effectively supported and that the user's bodies can be reconstructed with low latency in real-time. The combined exocentric and egocentric camera reconstructions could effectively fill the occlusion gaps for the first person view.

V. Conclusion, Discussion, and Future Work

In this work, we presented an approach for supporting embodiment in immersive environments, which could also be used to support co-presence in shared immersive experiences. The key idea is to avoid models and capture the user directly, and to combine data from a few ego- and exocentric cameras to produce a virtual body that appears complete to the user. Because the reconstruction is produced from direct observations of the user's physical body, visual characteristics such as size, proportions, skin tone, and clothing inherently match the user's physical appearance. We have described the extension of our existing volumetric capture system to include egocentric capture cameras. We also conducted a user study with 26 participants evaluating effectiveness and embodiment factors.

Our experimental results show that all three conditions —egocentric cameras, exocentric cameras, and both egocentric and exocentric cameras — are effective in supporting embodied experiences. The user's body is reconstructed in real time with low enough latency to support a sense of agency, and the inclusion of egocentric views in the reconstruction did fill gaps in the exocentric reconstruction. Camera registration and fusion of data from the four RGBD cameras in our experimental system was effective in producing a convincing virtual body.

However, in our user study, we did not detect significant differences in perceived embodiment and completeness between the combined condition and the exocentric-only condition. This could be due to a number of factors, most-likely due to our sample size, the task, and the particular experience with a virtual mirror. We estimated our sample size on the basis of similar studies in the related work and strong expected effect sizes. Similar studies in the field could show significant results with participant numbers as low as a dozen. Our expected strong effect size between conditions was not present, and so in future work we should use larger sample sizes. Our task of taking a certain pose in front of a virtual mirror was an attempt to combine requirements from different potential applications. First, immersive neurorehabilitation relies on a first-person view and often uses a front-facing mirror. In embodied rehabilitation scenarios, for example for post-stroke motor function recovery, but our participants are not stroke survivors and the trial was not conducted in a clinical setting. Secondly, psychological rehabilitation for conditions such as body dysmorphia and anorexia would be an appropriate use of a front-facing virtual mirror. However, in this application the egocentric cameras are unlikely to contribute to the experience since the main focus is the (exocentric) view of the virtual mirror. The primary focus point is the user's view in the mirror, and while looking down from a first-person perspective is desirable it is not critical. And third, a 3D teleconferencing or collaboration application would need both the egocentric and exocentric views but here the user's reflection is only a proxy for the remote user and this makes it impossible to assess the capacity of the system to elicit co-presence between two users.

Our findings — in particular on the effectiveness of sampled dynamic reconstructions and the combination of egocentric and exocentric views for reconstructions — are of value in the design and development of embodied experiences. This agrees with existing work which has demonstrated embodiment with 1-3 exocentric cameras capturing and the user in real time and using point cloud rendering [9], [11] or voxel grids [10], [27]. We add to this that our results suggest that the perceived completeness of the users body may not be significantly effected by temporary incompleteness resulting from self occlusion. However, future studies should avoid combining too many aspects into one study. Instead, task- and context-specific experiments are advisable. A virtual post stroke rehabilitation scenario should be evaluated with people actually suffering from e.g. unilateral motor impairments, and with the guidance of clinicians and other domain experts. 3D teleconferencing scenarios should provide different views to remote and local users, and should be conducted as close to the real application as possible with two embodied users sharing a virtual experience.

We hope that our work inspires researchers and practitioners to experiment with dynamic sampled reconstructions, and with combinations of ego- and exocentric views for reconstruction. Since our primary goal is to assess the effectiveness of the system in eliciting embodiment we have not considered combining egocentric and exocentric views in immersive mixed reality scenarios as suggested by Lindlbauer et al. [29]. We have likewise not yet investigated the effectiveness of this approach for "metaverse-like" applications, where more than two spatially distributed embodied users meet virtually and all experience the same virtual environment, but this appears to be another promising direction of research.

Ethics Statement

The user study was approved by the universities' Ethics Committee. Informed written consent was obtained from all participants prior to their participation.

References

[1] K. Nakano, N. Isoyama, D. Monteiro, N. Sakata, K. Kiyokawa, and T. Narumi, "Head-mounted display with increased downward field of view improves presence and sense of self-location," *IEEE Transactions on Visualization & Computer Graphics*, vol. 27, no. 11, pp. 4204–4214, 2021.

[2] J. S. Casaneuva, "Presence and co-presence in collaborative virtual environments," Master's thesis, University of Cape Town, 2001.

[3] C. Nimcharoen, S. Zollmann, J. Collins, and H. Regenbrecht, "Is That Me?—Embodiment and Body Perception with an Augmented Reality Mirror," in *2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, 2018, pp. 158–163.

[4] C. Heinrich, M. Cook, T. Langlotz, and H. Regenbrecht, "My hands? importance of personalised virtual hands in a neurorehabilitation scenario," *Virtual Reality*, vol. 25, pp. 313–330, 2021.

[5] C. Heinrich, N. Morkisch, T. Langlotz, H. Regenbrecht, and C. Dohle, "Feasibility and psychophysical effects of immersive virtual reality-based mirror therapy," *Journal of NeuroEngineering and Rehabilitation*, vol. 19, no. 1, pp. 1–20, 2022.

[6] K. Kilteni, R. Groten, and M. Slater, "The Sense of Embodiment in Virtual Reality," *Presence: Teleoperators and Virtual Environments*, vol. 21, no. 4, pp. 373–387, 2012.

[7] S. Duncan, N. Park, C. Ott, T. Langlotz, and H. Regenbrecht, "Voxel-Based Immersive Mixed Reality: A Framework for Ad Hoc Immersive Storytelling," *PRESENCE: Virtual and Augmented Reality*, pp. 1–25, Feb. 2023.

[8] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Transactions on Graphics*, vol. 34, no. 6, pp. 248:1–248:16, 2015.

[9] G. Gamelin, A. Chellali, S. Cheikh, A. Ricca, C. Dumas, and S. Otmane, "Point-cloud avatars to improve spatial communication in immersive collaborative virtual environments," *Personal and Ubiquitous Computing*, 2020.

[10] H. Regenbrecht, K. Meng, A. Reepen, S. Beck, and T. Langlotz, "Mixed Voxel Reality: Presence and Embodiment in Low Fidelity, Visually Coherent, Mixed Reality Environments," in *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2017, pp. 90–99.

[11] K. Yu, G. Gorbachev, U. Eck, F. Pankratz, N. Navab, and D. Roth, "Avatars for Teleconsultation: Effects of Avatar Embodiment Techniques on User Perception in 3D Asymmetric Telepresence," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 11, pp. 4129–4139, 2021.

[12] M. Gonzalez-Franco and T. C. Peck, "Avatar Embodiment. Towards a Standardized Questionnaire," *Frontiers in Robotics and AI*, vol. 5, p. 74, 2018.

[13] T. Waltemate, D. Gall, D. Roth, M. Botsch, and M. E. Latoschik, "The impact of avatar personalization and immersion on virtual body ownership, presence, and emotional response," *IEEE transactions on visualization and computer graphics*, vol. 24, no. 4, pp. 1643–1652, 2018.

[14] P. Caserman, M. Martinussen, and S. Göbel, "Effects of end-to-end latency on user experience and performance in immersive virtual reality applications," in *Entertainment Computing and Serious Games: First IFIP TC 14 Joint International Conference, ICEC-JCSG 2019, Arequipa, Peru, November 11–15, 2019, Proceedings 1*. Springer, 2019, pp. 57–69.

[15] T. Waltemate, I. Senna, F. Hülsmann, M. Rohde, S. Kopp, M. Ernst, and M. Botsch, "The impact of latency on perceptual judgments and motor performance in closed-loop interaction in virtual reality," in *Proceedings of the 22nd ACM Conference on Virtual Reality Software and Technology*. Munich Germany: ACM, Nov. 2016, pp. 27–35.

[16] J.-W. N. Park, S. Mills, H. Whaanga, P. Mato, R. W. Lindeman, and H. Regenbrecht, "Towards a Māori Telepresence System," in *2019 International Conference on Image and Vision Computing New Zealand (IVCNZ)*, 2019, pp. 1–6.

[17] T. Waltemate, D. Gall, D. Roth, M. Botsch, and M. E. Latoschik, "The Impact of Avatar Personalization and Immersion on Virtual Body Ownership, Presence, and Emotional Response," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 4, pp. 1643–1652, 2018.

[18] J. Achenbach, T. Waltemate, M. E. Latoschik, and M. Botsch, "Fast generation of realistic virtual humans," in *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*, ser. VRST '17. Gothenburg, Sweden: Association for Computing Machinery, 2017, pp. 1–10.

[19] O. Schreer, I. Feldmann, T. Ebner, S. Renault, C. Weissig, D. Tatzelt, and P. Kauff, "Advanced Volumetric Capture and Processing," *SMPTE Motion Imaging Journal*, vol. 128, no. 5, pp. 18–24, 2019.

[20] C. Heinrich, M. Cook, T. Langlotz, and H. Regenbrecht, "My hands? importance of personalised virtual hands in a neurorehabilitation scenario," *Virtual Reality*, Jul 2020. [Online]. Available: https://doi.org/10.1007/s10055-020-00456-4

[21] A. Bartl, S. Wenninger, E. Wolf, M. Botsch, and M. E. Latoschik, "Affordable but not cheap: A case study of the effects of two 3d-reconstruction methods of virtual humans," *Frontiers in Virtual Reality*, p. 123, 2021.

[22] M. Habermann, W. Xu, M. Zollhöfer, G. Pons-Moll, and C. Theobalt, "LiveCap: Real-Time Human Performance Capture From Monocular Video," *ACM Transactions on Graphics*, vol. 38, no. 2, pp. 14:1–14:17, 2019.

[23] T. Yu, J. Zhao, Z. Zheng, K. Guo, Q. Dai, H. Li, G. Pons-Moll, and Y. Liu, "DoubleFusion: Real-time Capture of Human Performances with Inner Body Shapes from a Single Depth Sensor," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019.

[24] S. Wang, X. Zuo, C. Du, R. Wang, J. Zheng, and R. Yang, "Dynamic Non-Rigid Objects Reconstruction with a Single RGB-D Sensor," *Sensors*, vol. 18, no. 3, p. 886, 2018.

[25] T. Yu, K. Guo, F. Xu, Y. Dong, Z. Su, J. Zhao, J. Li, Q. Dai, and Y. Liu, "BodyFusion: Real-Time Capture of Human Motion and Surface Geometry Using a Single Depth Camera," in *2017 IEEE International Conference on Computer Vision (ICCV)*. Venice: IEEE, 2017, pp. 910–919.

[26] M. Dou, P. Davidson, S. R. Fanello, S. Khamis, A. Kowdle, C. Rhemann, V. Tankovich, and S. Izadi, "Motion2fusion: Real-time volumetric performance capture," *ACM Transactions on Graphics*, vol. 36, no. 6, pp. 246:1–246:16, 2017.

[27] H. Regenbrecht, J.-W. N. Park, C. Ott, S. Mills, M. Cook, and T. Langlotz, "Preaching Voxels: An Alternative Approach to Mixed Reality," *Frontiers in ICT*, vol. 6, p. 7, 2019.

[28] O. Schreer, I. Feldmann, S. Renault, M. Zepp, M. Worchel, P. Eisert, and P. Kauff, "Capture and 3D Video Processing of Volumetric Video," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 4310–4314.

[29] D. Lindlbauer and A. D. Wilson, "Remixed Reality: Manipulating Space and Time in Augmented Reality," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ser. CHI '18. Montreal QC, Canada: ACM Press, 2018, pp. 1–13.

[30] R. Tsai and R. Lenz, "A new technique for fully autonomous and efficient 3D robotics hand/eye calibration," *IEEE Transactions on Robotics and Automation*, vol. 5, no. 3, pp. 345–358, 1989.