frontiers

Avatars for the Masses: Smartphone-Based Reconstruction of Humans for Virtual Reality

Timo Menzel [1,†], **Erik Wolf** [2,†], **Stephan Wenninger** [1], **Niklas Spinczyk** [1], **Lena Holderrieth** [2], **Carolin Wienrich** [3], **Ulrich Schwanecke** [4], **Marc Erich Latoschik** [2], **Mario Botsch** [1,*]

[1] *Computer Graphics Group, TU Dortmund University, Dortmund, Germany*
[2] *Human-Computer Interaction Group, Julius-Maximilians-Universität Würzburg, Würzburg, Germany*
[3] *Psychology of Intelligent Interactive Systems Group, Julius-Maximilians-Universität Würzburg, Würzburg, Germany*
[4] *Computer Vision and Mixed Reality Group, RheinMain University of Applied Sciences, Wiesbaden, Germany*

Correspondence*:
Mario Botsch
mario.botsch@tu-dortmund.de

† These authors contributed equally to this work

ABSTRACT

Realistic full-body avatars play a key role in representing users in virtual environments, where they have been shown to considerably improve important effects of immersive experiences such as body ownership and presence. Consequently, the demand for realistic virtual humans – and methods for creating them – is rapidly growing. However, despite extensive research into 3D reconstruction of avatars from real humans, an *easy and affordable* method for generating *realistic and VR-capable* avatars is still lacking: Existing methods are either limited to complex capture hardware and/or controlled lab environments, do not provide sufficient visual fidelity, or cannot be rendered at sufficient frame rates for multi-avatar VR applications. To make avatar reconstruction widely available, we developed *Avatars for the Masses* – a client-server-based online service for scanning real humans with an easy-to-use smartphone application that empowers even non-expert users to capture photorealistic and VR-ready avatars. The data captured by the smartphone are transferred to a reconstruction server, where the avatar is generated in a fully automated process. Our advancements in capturing and reconstructing allow for higher-quality avatars even in less controlled in-the-wild environments. Extensive qualitative and quantitative evaluations show our method's avatars to be on par with the ones generated by expensive expert-operated systems. It also generates more accurate replicas in comparison to the current state-of-art in smartphone-based reconstruction, produces much less artifacts and provides a much higher rendering performance in VR in comparison to three representative neural methods. A comprehensive user study confirms similar perception results compared to avatars reconstructed with expensive expert-operated systems, and it underscores a sufficient usability of the overall system. To truly bring avatars to the masses, we will make our smartphone application publicly available for research purposes.

Keywords: Avatar generation, 3D scanning, virtual human, embodiment, virtual body ownership

1 INTRODUCTION

Avatars are digital representations of users that can be dynamically rendered in virtual environments in real time to reflect the behavior of their users (Bailenson and Blascovich, 2004). While avatars can be of almost any conceivable shape and appearance, in this research, we specifically refer to humanoid representations that vary from stylized to realistically reconstructed 3D models. Such avatars may appear generic, lacking distinctive or individual features, or they can be personalized to closely resemble the appearance of their respective user. With the recent surge in virtual reality (VR) research (Skarbez and Jiang, 2024) and the increasing availability of mature head-mounted displays (HMDs) (Sutherland, 1968), avatars have become increasingly important as faithful self-representations of users in almost countless scenarios. These scenarios include metaverse-like social VR environments (Latoschik et al., 2019; Yoon et al., 2019; Aseeri and Interrante, 2021; Mystakidis, 2022) or VR applications to support mental health (Sampaio et al., 2021; Döllinger et al., 2022). Among them are critical applications for which maintaining user identity and conveying realistic emotions are crucial for authentic interactions and a sophisticated user experience (UX). Prior work has shown that realistically personalized full-body avatars, which can look deceptively similar to the user, are superior for the outlined scenarios by increasing the user's sense of presence and embodiment or self-identification with the avatar (Waltemate et al., 2018; Salagean et al., 2023; Fiedler et al., 2024; Kim et al., 2023), or to increase emotional response (Gall et al., 2021; Waltemate et al., 2018).

Unfortunately, many approaches for scanning-based full-body avatar generation rely on complex and expensive multi-camera rigs for photogrammetric reconstruction, such as (Achenbach et al., 2017; Shetty et al., 2024; Ma et al., 2021). Methods for generating avatars from monocular video input make avatar generation more affordable, but early approaches (Alldieck et al., 2018a,b) suffered from insufficient quality, as shown in (Wenninger et al., 2020). Recent avatar reconstructions adapt NeRFs (Mildenhall et al., 2020) or Gaussian Splatting (Kerbl et al., 2023) as underlying representations, for instance (Jiang et al., 2023; Moreau et al., 2024). Although this is an exciting and very promising research direction, our experiments in Section 4 clearly demonstrate that these approaches are not (yet) capable of providing sufficient visual quality and rendering performance for VR applications. So far, the method of Wenninger et al. (2020), which reconstructs mesh-based avatars from smartphone videos, seems to be the most suitable for the affordable reconstruction of photorealistic and VR-capable full-body avatars. However, while the low hardware requirements make avatar reconstruction more affordable, the scanning process requires sufficient experience, the reconstruction process involves commercial products, and the system's operation requires expert knowledge. Consequently, there is still no approach for fast, affordable, and easy-to-operate reconstruction of photorealistic and VR-capable full-body avatars. This prevents the full potential of photorealistic avatars from being realized for many applications.

To bridge this gap and make avatar reconstruction both affordable and widely available to non-expert users, we present *Avatars for the Masses*, an easy-to-use system for smartphone-based person scanning and server-based avatar reconstruction. In particular, our contributions are:

- An easy-to-use smartphone application that visually guides the user through the scanning process, enabling even non-expert users to achieve high-quality results;
- A server-based pipeline that fully automatically reconstructs a photorealistic avatar from smartphone-captured data in about 20 minutes, without relying on commercial components;
- Technical improvements in the capture and reconstruction processes that result in high quality results even in uncontrolled outdoor environments;

- Qualitative and quantitative technical evaluations and comparisons with several state-of-the-art approaches that clearly demonstrate the advantages of our system;
- A user-centric evaluation through a user study that evaluates and confirms both our smartphone app's usability and the resulting avatars' quality (*captured by non-expert first-time users!*).

Our evaluations demonstrate that the proposed system is indeed fast, affordable, and easy to use, and that it achieves avatar quality almost on par with that of complex camera rigs – even in challenging "in-the-wild" capture scenarios. As such, and due to the lack of commercial components, it has the potential to bring avatars to the masses. We will make our system publicly available for research to encourage this.

2 RELATED WORK

In this section, we describe the mechanisms and implications of representing oneself through an avatar in virtual reality (Section 2.1), before discussing different approaches to generate realistic avatars (Section 2.2). In the following, we restrict our discussion to avatars *personalized* (as opposed to generic), *realistic* (as opposed to stylized), and *full-body* (as opposed to head-only or upper-body-only), because these are the most challenging with regard to the outlined desiderata.

2.1 Avatars for Self-Representation in Virtual Reality

The egocentric embodiment of avatars for self-representation in VR (Slater et al., 2010) can positively impact the UX of virtual environments (Mottelson et al., 2023). This includes improving the key psychometric properties of VR, such as the sense of presence (Waltemate et al., 2018; Wolf et al., 2021; Skarbez et al., 2017), or intensifying emotional responses to virtual content (Waltemate et al., 2018; Gall et al., 2021). Other advantages may include improved spatial perception (Mohler et al., 2010; Leyrer et al., 2011), reduced cognitive load (Steed et al., 2016), or higher performance and accuracy (Jung and Hughes, 2016; Pastel et al., 2020) when performing tasks in VR.

A crucial aspect in evaluating the effectiveness of avatar embodiment is the sense of embodiment (SoE), consisting of the feeling of owning (ownership), controlling (agency), and being located within (self-location) a virtual body in a virtual environment (Kilteni et al., 2012; de Vignemont, 2011). Previous work has shown that realistic and personalized avatars increase the SoE towards the avatar (Waltemate et al., 2018; Fiedler et al., 2023; Salagean et al., 2023) and thus contribute to an overall plausible VR experience (Latoschik and Wienrich, 2022).

Photorealistic and personalized avatars are particularly valuable for maintaining the user's identity, which is beneficial in social VR experiences (Yoon et al., 2019; Aseeri and Interrante, 2021; Mystakidis, 2022) or applications supporting mental health (Sampaio et al., 2021; Döllinger et al., 2022; Turbyne et al., 2021). Previous work has also shown that self-related cues through avatar embodiment and personalization significantly increase self-identification with the avatar (Fiedler et al., 2024), potentially maintaining a more accurate self-perception in VR, even in body-swap paradigms (Döllinger et al., 2024). However, a realistic personalization of avatars can also harm UX, as their human-like realism combined with their high affinity to the user can potentially trigger Uncanny Valley effects, leading to negative emotional responses such as eeriness towards the avatars (Mori et al., 2012; Döllinger et al., 2023).

Overall, comprehensive evidence exists for notable effects of photorealistic personalized avatars on important user states. Consequently, we will employ representative psychometric measures for a prominent selection of the aforementioned effects of avatars to evaluate the 3D reconstruction quality achieved with

our developed system. Therefore, Section 5 reports on a user study evaluating our avatars with respect to the sense of embodiment, plausibility, and a potential uncanny valley effect. In addition, this user study also evaluates the general usability and user satisfaction of the smartphone front-end to ensure appropriate ease of use and user satisfaction.

2.2 Generation of Realistic Personalized Avatars

The growing demand for virtual avatars has triggered a lot of research in scanning-based avatar reconstruction in the recent years. We restrict ourselves to realistic full-body avatars and discuss related approaches with respect to our target application requirements: The avatar generation should be affordable and easy to use, the resulting avatars should accurately resemble the scanned person, and the avatars should be suitable for VR applications – meaning they can be rendered from arbitrary camera views and at a sufficiently high frame rate for multi-avatar (social) VR applications.

Many approaches employ complex and expensive rigs of 50–100 cameras to capture high-quality photos or videos of the person to be scanned (Feng et al., 2017; Achenbach et al., 2017; Ma et al., 2021; Kwon et al., 2023; Salagean et al., 2023; Morgenstern et al., 2024; Shetty et al., 2024; Pang et al., 2024). While these methods achieve highly accurate reconstructions, they are restricted to dedicated capture laboratories whose operation requires expert knowledge.

Instead of simultaneously taking images with multiple cameras, other approaches use a single monocular camera (or smartphone) to capture a sequence of images or videos. While this design choice considerably reduces hardware cost and complexity, it increases the capture time, which inevitably causes small movements of the scanned person and reduces geometric accuracy. Early approaches suffer from considerably lower quality compared to camera rigs (Alldieck et al., 2018a,b, 2019), most visible in the face region. Wenninger et al. (2020) address this problem by incorporating close-ups of the head into the avatar reconstruction, producing a quality that is objectively quite close and subjectively very similar to those of camera rigs (Bartl et al., 2021). On the downside, their method is rather complicated to operate, is intended for controlled indoor environments, and relies on commercial components, which prevents widespread use by researchers and non-experts.

More recently, neural geometry representations, such as Neural Radiance Fields (NeRFs) (Mildenhall et al., 2020) or 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) have been extensively adapted to avatar reconstruction. Neural avatar representations (Peng et al., 2021b; Zhao et al., 2022; Jiang et al., 2024; Guo et al., 2023; Xiao et al., 2024; Lin et al., 2024) are capable of reconstructing fine details, since they are not restricted to a fixed mesh topology. Avatars based on NeRFs (Liu et al., 2021; Peng et al., 2021a; Jiang et al., 2023, 2022; Wang et al., 2023; Yu et al., 2023; Wang et al., 2024; Zheng et al., 2022, 2023) or 3DGS (Hu et al., 2024; Shao et al., 2024; Moreau et al., 2024; Li et al., 2024; Habermann et al., 2023) are therefore better suited for reconstructing clothing and hair. However, as our experiments with recent neural avatars show (see Section 4), their generation from image and video data can be very time-consuming (from hours up to days), their rendering is not fast enough for multi-avatar VR applications (where 90 fps at 2k resolution for left/right eye is desired), and their visual fidelity is not sufficient (when viewed from directions not covered by training data). This last point is a particularly challenging limitation, since in multi-user social VR applications there is no control over viewing directions and avatar poses, which can quickly lead to visual artifacts.

We therefore employ a traditional mesh-based representation for virtual avatars and build on the approach of Wenninger et al. (2020), which we advance in several important aspects. First, our smartphone application visually guides the user through the capture process, thereby ensuring high-quality input

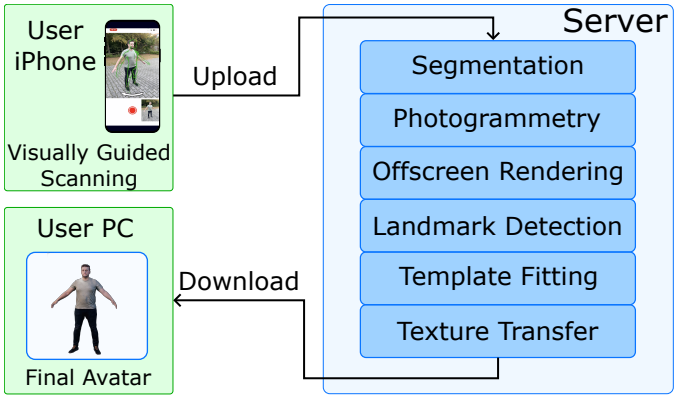

Figure 1. The user scans a subject with our smartphone application (top left). The captured images are uploaded to our processing server (right), where a fully automatic reconstruction pipeline generates a avatar in about 20 minutes. The user can then download his/her avatar into any VR application (bottom left).

data. Second, we technically improve the data acquisition, image pre-processing, and template fitting, leading to more accurate and more robust avatar reconstructions. Third, we replace the commercial components of Wenninger et al. (2020) with carefully selected non-commercial alternatives, allowing us to make our system publicly available. Finally, we evaluate our approach (i) by qualitative and quantitative comparisons to state-of-the-art avatar reconstruction methods, and (ii) in terms of a carefully designed user study, in which first-time users successfully reconstruct and evaluate avatars.

3 AVATAR RECONSTRUCTION

Our approach extends and improves the work by Wenninger et al. (2020). We start with a brief overview of their method in order to point out our specific technical improvements later on. Wenninger et al. (2020) record two videos of the to-be-scanned person: The *body video* circles around the scanning subject twice to capture both the lower and the upper body. The *head video* circles around the face/head in a close-up manner to capture facial details. From these two videos, individual frames are extracted and fed into Agisoft Metashape (Agisoft, 2023), a commercial photogrammetry reconstruction tool, resulting in two point clouds for the body and head. A template mesh is then fitted to the point clouds in a two-step process: The template is first fitted to the body point cloud (for the overall shape) and then to the head point cloud (for fine-tuning facial features). Landmarks detected by OpenPose (Cao et al., 2019) guide the template fitting process. In a final step, the avatar texture is generated from the input images.

Our approach, as outlined in Figure 1, introduces guided smartphone-based data capturing (Section 3.1) and a fully automatic server-based reconstruction pipeline (Section 3.2). In the following, we describe the components of both phases and point out the main contributions and technical improvements compared to Wenninger et al. (2020).

3.1 Smartphone-Based Data Acquisition

Analogous to Wenninger et al. (2020), we capture people by performing (i) a full-body scan in A-pose and (ii) a close-up head scan using a smartphone (see Figure 1, top left). However, our approach differs in the kind of data that is captured (Section 3.1.1) and how the user performs the scanning (Section 3.1.2).

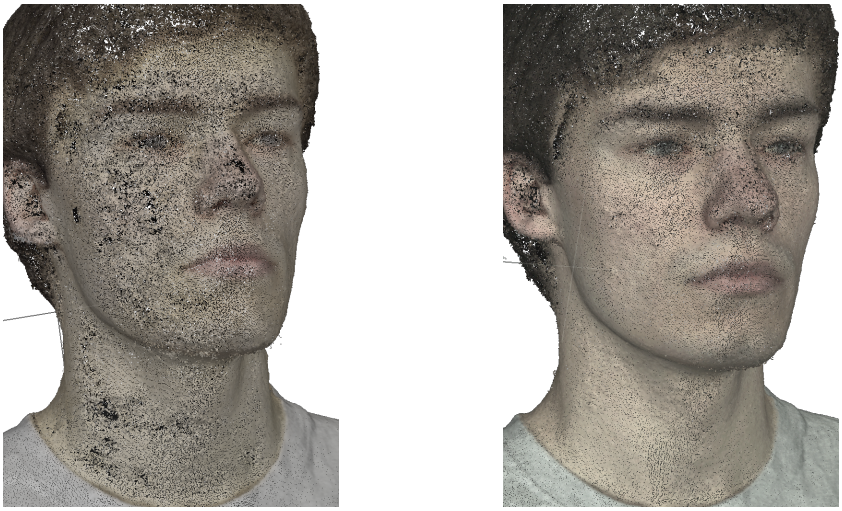

Figure 2. Point clouds reconstructed via photogrammetry from video frames suffer from compression artifacts (left). The higher quality of individually captured images yields more accurate point clouds (right).

3.1.1 Capturing Videos vs. Images

Videos captured with current smartphone cameras are compressed using H.264 or H.265. These algorithms are optimized for viewing each video frame for a fraction of a second only, hence allowing for rather aggressive per-frame compression. In addition, inter-frame compression exploits blockwise similarity of consecutive frames, which further degrades image quality (Wiegand et al., 2003). As photogrammetry algorithms use image gradients to detect feature points, the block edges can negatively influence the quality of the resulting 3D point cloud (Figure 2, left). Furthermore, extracted video frames can be affected by motion blur, which Wenninger et al. (2020) had to handle explicitly.

In contrast, individual photographs can be captured at considerably higher quality, since they suffer much less from motion blur, avoid the inter-frame block compression artifacts, and allow to use less aggressive compression in general. Higher-quality images in turn yield more accurate photogrammetry results (Figure 2, right), which will eventually result in more accurate avatars with fine geometric details and higher-quality textures. Our capture process (described next) therefore records individual photographs instead of videos, at a resolution of 3024×4032 pixels.

In addition to the high-resolution RGB images, we also capture coarse depth images (576×768 pixels) using the smartphone's depth sensor and the phone's orientation (resp. gravity vector). The former helps to determine the reconstructed subject's correct size/scaling, and the latter to determine its correct orientation. Our scanning application is designed for Apple iOS devices, and this additional information is conveniently included in the meta-data of Apple's HEIC image format.

3.1.2 Scanning UI

Our extensive experience with the approach of Wenninger et al. (2020) revealed that the quality of the photogrammetry reconstruction strongly depends on the correct distance and orientation of the camera to the subject. If the camera is too close, some regions (e.g. feet, hands) might not be captured. If the camera is too far away, valuable image resolution is wasted and the point cloud becomes less dense and more noisy. Those errors typically occurred to unexperienced users – despite detailed previous instructions.

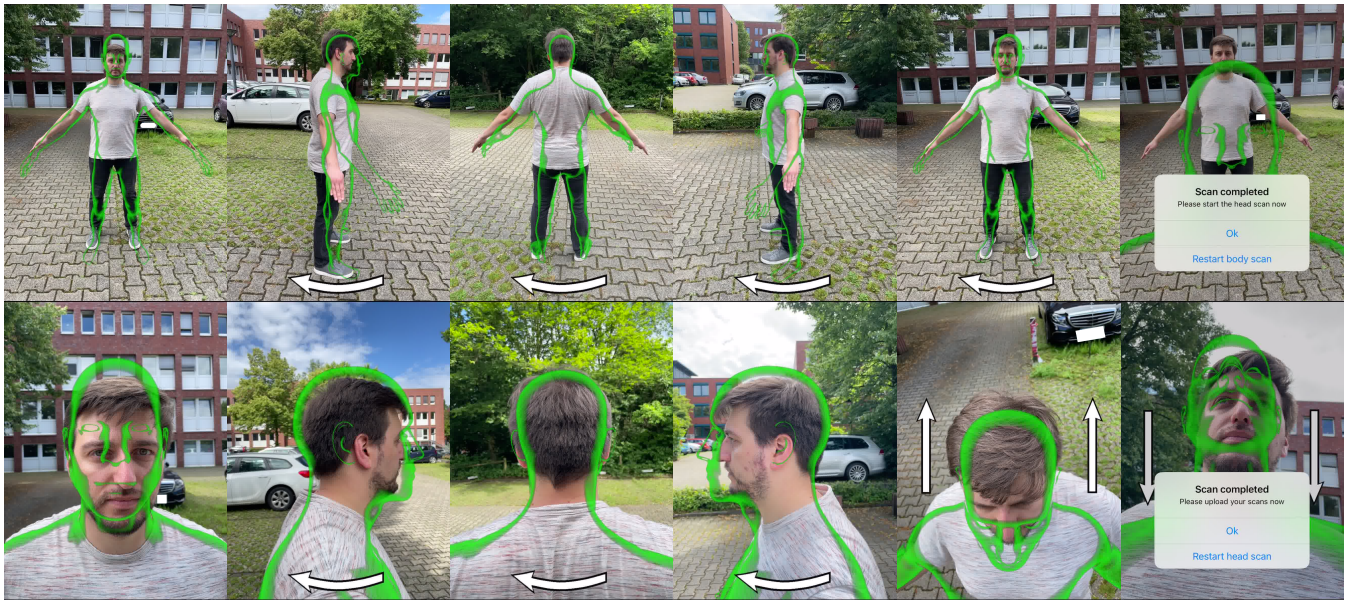

Figure 3. The guided scanning procedure of the smartphone application. Top row, from left to right: Initial overlay before starting the body scan; overlay during the body scan; end of body scan. Bottom row, from left to right: Initial overlay before starting the head scan; overlay during the head scan; end of head scan.

To avoid these errors, our smartphone application visually guides the user through the scanning process: On top of the camera feed, we overlay in green the silhouette of a virtual human model (of average size and shape) as seen from the intended camera position (see Figure 3). The user adjusts the phone/camera to roughly match the silhouette of the subject and the model. It is not necessary to precisely fit the overlay to the subject. The purpose of the overlay only is to guide the user to maintain proper distance and orientation. To also guide the user's movement around the scanned subject, the virtual camera moves around the green virtual model in the same way that the user should move around the scanned subject. In addition, the direction of the movement is indicated by a white arrow.

During the scanning process, the app captures images at a frequency of $1\,\mathrm{Hz}$ and a resolution of 3024×4032 pixels. The speed of the virtual camera's movement is chosen to result in 105 images for the entire scanning process, as this number experimentally turned out to be the best compromise: Fewer images degraded the point cloud quality, more images did not improve the results but increased the computation time. Thanks to this well-controlled capture process, we require just one circle around the subject for the full-body scan and one for the head scan – thus reducing the scanning time to about two-thirds of Wenninger et al. (2020). Since a shorter scanning time reduces artifacts caused by subject movement, it also improves geometric accuracy.

A dialog informs the user when the full-body and head scans are complete (Figure 3, right column), after which the captured images are uploaded to the reconstruction server. All further user instructions or hints during the scanning procedure are displayed in Figure 3. The entire scanning process can also be seen in the accompanying video. To further minimize scanning errors, the app displays a step-by-step tutorial before the scanning process, covering subject preparation (hairstyle, accessories, shoes, and clothes), scan pose requirements (A pose), and scan process explanations (body/head scan procedure).

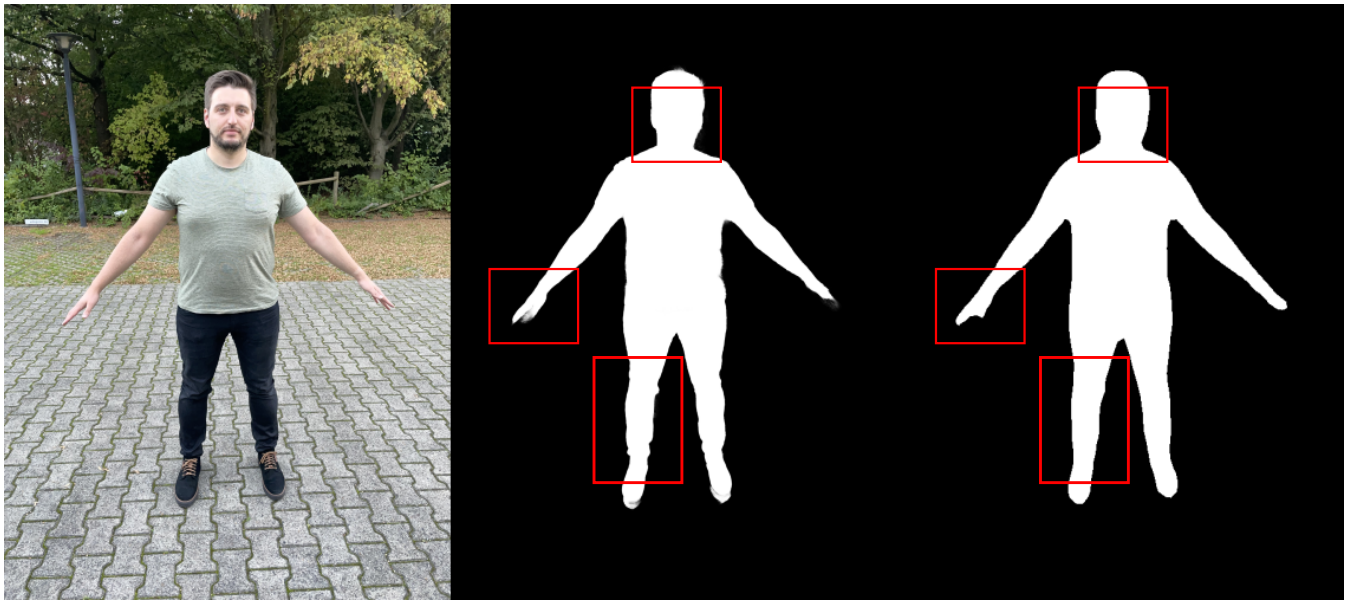

Figure 4. When segmenting foreground and background of the input images, Apple's person segmentation (center) better preserves small details, such as hair and clothing creases, compared to DeepLabV3 (right).

3.2 Server-Based Reconstruction Pipeline

The avatar reconstruction pipeline, whose individual tasks are described in this section, includes several computationally expensive tasks. To speed up the avatar generation process and to reduce the load on the smartphone's resources, the captured images are uploaded to a compute server, where the avatar is automatically reconstructed and can be downloaded by the user.

3.2.1 Image Preprocessing

Our experiments with Wenninger et al. (2020) revealed that their method works well in controlled indoor environments, but in outdoor environments it often gives noticeably worse results. This is due to non-static background, such as leaves moving in the wind or cars driving by. These background movements violate the photogrammetry assumption of a static scene, leading to incorrect extrinsic camera parameters and consequently to errors in the reconstructed 3D point cloud.

To eliminate these problems and thereby make the reconstruction process much more robust with respect to "in-the-wild" capture environments, we segment the input images into foreground and background and mask out the background before passing the images to the photogrammetry process. To this end, we compared DeepLabV3 (Chen et al., 2017) and Apple's person segmentation (Apple Inc., 2023d) (on macOS 15.3.1), and decided for the latter since it produced slightly more accurate and more detailed masks in our experiments (see Figure 4). Moreover, as the image background is excluded from the reconstruction, the number of image features to be matched by the photogrammetry is significantly reduced (accelerating this process by 30%), the resulting point cloud contain considerably fewer points (accelerating later template fitting), and the point clouds contain significantly less noise and outliers (improving accuracy and robustness of the template fitting). Overall, this image preprocessing leads to faster computations and much cleaner point clouds – in particular in uncontrolled outdoor environments (see Figure 5).

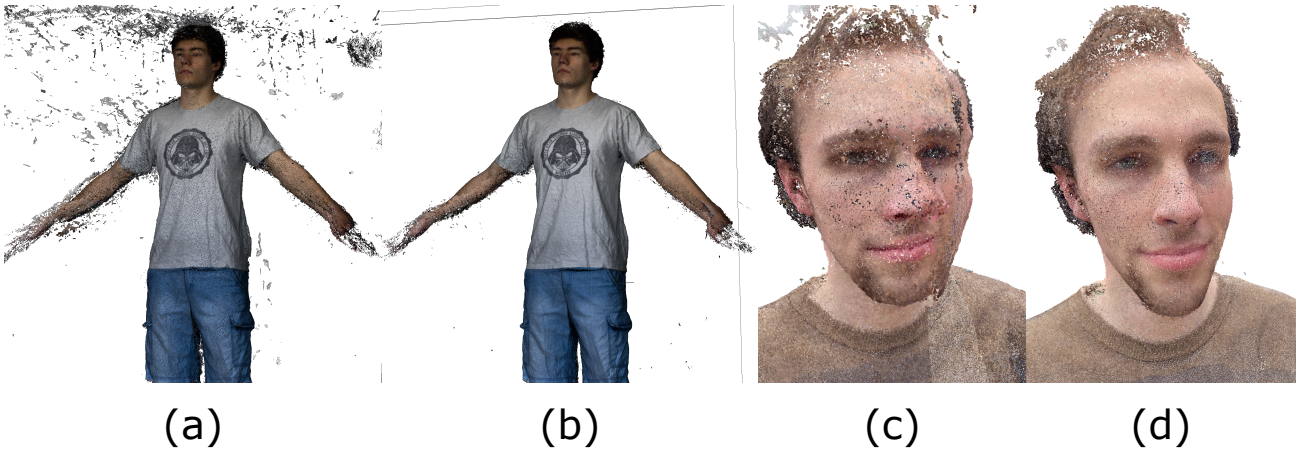

Figure 5. Point clouds reconstructed from unprocessed input images captured in outdoor environments suffer from significant noise, outliers, and misalignments in the face region (a, c). By processing the images through person segmentation and background removal, these artifacts are largely eliminated (b, d).

Figure 6. The (meshed) photogrammetry reconstructions of Agisoft Metashape (left) and Apple's RealityKit (right) are very similar in geometric fidelity, texture quality, and computational performance.

3.2.2 Photogrammetry

The captured and segmented images are passed to the photogrammetry stage, which reconstructs a dense 3D point cloud (see Figure 1). Wenninger et al. (2020) employ Metashape (Agisoft, 2023) for this task, a widely used commercial photogrammetry software. Unfortunately, its license restrictions explicitly prohibit the use in server-based reconstruction scenarios (which we aim for). In order to make our system publicly available for research purposes, we compared several non-commercial alternatives for photogrammetric reconstruction, including MeshRoom (Alice Vision, 2025), COLMAP (Schönberger and Frahm, 2016), and Apple's RealityKit (Apple Inc., 2023b) (on macOS 15.3.1). From these frameworks, Apple's RealityKit consistently produced the highest-quality results, which are very similar in terms of geometric fidelity, texture quality, and processing time to those of Agisoft Metashape (see Figure 6).

Apple's high-level photogrammetry API allows one to specify different configuration options, where we found the settings *raw* detail, *high* feature sensitivity, and *unordered* samples to yield best results. In addition to the high-resolution RGB images, we also provide segmentation masks, coarse depth images, gravity vectors, and EXIF data, which significantly improves the photogrammetry results compared to processing only the RGB images. The depth data and gravity vectors help to determine the correct scale and orientation of the reconstructed object, which make the upcoming steps more reliable.

3.2.3 Landmark Detection

The previous photogrammetry stage produces a (static) high-resolution textured triangle mesh (see Figure 6). This mesh can suffer from artifacts in insufficiently scanned regions and is lacking animation controls (body skeleton, facial blendshapes). The well-established approach is to fit a high-quality template mesh with all required animation controls to the photogrammetry output (point cloud or mesh). This results in a reconstructed avatar mesh that inherits its triangulation, UV texture layout, and animation controls from the template model, while closely resembling the geometric shape and the texture/material of the photogrammetry scan. This template fitting process (described in the next subsection) has to be initialized and guided by a set of landmarks on both the template model (where they are pre-selected once) and the photogrammetry output (where they have to be manually selected or automatically detected).

Wenninger et al. (2020) detect body and face landmarks in the 2D input images using OpenPose (Cao et al., 2019) and (Dlib, 2022), respectively, where they select the best suited (e.g., most frontal) input images based on heuristics. The detected 2D landmarks are re-projected onto the photogrammetry point cloud. While working well in most cases, their approach can fail if wrong images are selected for landmark detection or if a 2D landmark in sparsely sampled regions back-projects to the wrong surface part. We avoid both problems by rendering the photogrammetry mesh (Figure 6) from several camera positions and performing landmark detection on the resulting synthetic images. Our controlled capturing process enables the straightforward selection of suitable camera views, and the back-projection onto the rendered 3D mesh is well-defined for any detected 2D landmark.

We detect 4 hand landmarks (two knuckles on the left and right hands) and 37 face landmarks (eye contours, tip of the nose, and mouth features), which are passed on to the template-fitting stage. To ensure accurate and reliable landmark detection, we compared on a wide range of examples the face/hand landmark detection of OpenPose (Cao et al., 2019), Dlib (2022), Apple's Vision Framework (Apple Inc., 2023c, 2025), and Google's MediaPipe (Lugaresi et al., 2019). Since MediaPipe produced the most reliable results in our experiments, we chose this landmark detector for our reconstruction pipeline.

3.2.4 Template Fitting

The previous two stages result in two high-resolution textured photogrammetry meshes from the body and head scans of the subject standing in A-pose with neutral facial expression, as well as a set of 37 face and 4 body landmarks. We perform curvature-adaptive point sampling on the two photogrammetry meshes to convert them into two point clouds for body and head, respectively. To reconstruct the avatar, we fit a fully rigged statistical template mesh to the photogrammetry data, guided by the landmarks.

Our template mesh was designed by a skilled artist (to be free of license restrictions) and has a slightly higher resolution (23752 vertices) than the template from the Autodesk Character Generator in Wenninger et al. (2020). Its animation rig consists of a full-body skeleton with 59 joints, as well as 52 facial blends that are compatible with ARKit (Apple Inc., 2023a). The template was fit to 1700 scans of the CAESAR database (Robinette et al., 1999) to derive a 30-dimensional PCA subspace of human body shapes.

In a first step, the template is coarsely fitted to the point clouds by iteratively optimizing *alignment* (position, orientation, scale), overall *body shape* (PCA weights), and *body pose* (skeleton joint angles). In the second step, the initial template fit is refined by optimizing all individual *vertex positions*. Both optimization phases minimize the sum of squared distances of photogrammetry points to their closest points on the template mesh in a non-rigid ICP manner, guided by the landmark points. Both steps are regularized to prevent overfitting: the first step by Tikhonov regularization on the PCA weights, the second step by a discrete bending energy (see Achenbach et al. (2017); Wenninger et al. (2020) for details).

Our method differs from Wenninger et al. (2020) in two aspects: Wenninger et al. first fit the template to the body point cloud and then refine the result by fitting it to the head point cloud. Since in their approach the absolute scaling of these point clouds is unknown, the proportions of body to head can be slightly wrong. In contrast, our coarse depth images determine the absolute scale. We also pre-align the body and head point clouds using landmark-guided ICP and then fit the template to both point clouds *simultaneously*. In this process, closest-point correspondences to the head/body regions of the template mesh are computed from the head/body point clouds only, respectively. This approach effectively avoids the wrong body-head proportions (see Figure 9). In addition, since our advanced scanning process yields more accurate point clouds, we require less regularization in the fine-scale fitting step, resulting in more geometric details.

3.2.5 Texture Transfer

After reconstructing the geometric shape of the avatar in the previous step, the final step reconstructs the texture image. The two photogrammetry meshes already feature high-quality textures generated from the input images, but with a rather poor UV texture layout. To have a uniform texture layout for all avatars, we transfer $\texttt{texture}_P$ of the photogrammetry mesh to $\texttt{texture}_A$ of the avatar mesh (having the high-quality texture layout of the template). For each texel $\mathbf{u}_A$ in the avatar's UV layout we determine the corresponding 3D point $\mathbf{x}_A$ on the avatar mesh (based on texture coordinates), find its closest point $\mathbf{x}_P$ on the photogrammetry mesh, and copy its color by its texture coordinate $\mathbf{u}_P$:

$$\texttt{texture}_A\left[\mathbf{u}_A\right] \leftarrow \texttt{texture}_P\left[\mathbf{u}_P\right]$$

Note that we actually fill two textures, from the body and head scan, respectively. These two texture images are then combined into one using Poisson Image Editing (Pérez et al., 2003). This final step of the reconstruction pipeline results in a textured avatar mesh (see Figure 7 for some examples).

4 QUANTITATIVE AND QUALITATIVE EVALUATION

Reconstructing an avatar with our approach starts by capturing a person using our iPhone application (iPhone 12 Pro and iPhone 13 Pro Max in our experiments). The app visually guides the user through the scanning procedure and takes 105 images (45 full-body and 60 head images), which takes about two minutes and is shown in the accompanying video. The captured image data (about 320 MB) is then uploaded to our server, which takes less than one minute over WiFi. The reconstruction is performed in a fully automatic manner on the server (Mac Studio, M1-Max 10-Core CPU, 32-Core integrated GPU, 64 GB RAM) and takes about 19 minutes (1 min segmentation, 7 min photogrammetry, 5 min offscreen rendering, 2 min landmark detection, 4.5 min template fitting and texture generation). The whole process, therefore, takes about 22 minutes only, after which the avatar can then be downloaded in file formats compatible with VR and game engines.

Figure 7. Avatars reconstructed with our appraoch, all being scanned in uncontrolled outdoor settings.

In the following, we compare our results to those of a complex multi-camera rig Achenbach et al. (2017), to the smartphone-based method of Wenninger et al. (2020), and to three recent neural avatar techniques based on NeRFs or 3D Gaussian Splatting (Müller et al., 2022; Shao et al., 2024; Lei et al., 2024).

4.1 Quantitative Comparisons

Following Wenninger et al. (2020), we evaluate the accuracy of our avatar reconstruction by reporting reprojection errors. To this end, we render the resulting textured avatar onto the images captured during the scan process using the camera calibration data from the photogrammetry process (see Figure 8), and then compute the root-mean-square errors over all rendered pixels in CIELab color space, averaged over all images. This metric allows us to measure errors resulting from inaccuracies in both color and geometry.

We perform this evaluation on the 33 subjects that were scanned during the user study described in Section 5. These participants were scanned by (i) another 33 non-expert first-time users of our smartphone application, as well as (ii) an expert using the multi-camera rig at the Embodiment Lab of JMU Würzburg (106 Canon EOS 1300D DSLR cameras, based on Achenbach et al. (2017)). Generating the avatars using the expert-operated multi-camera rig took about 15 minutes. Our pipeline, on the other hand, took around 22 minutes. Despite the tremendous difference in expertise of the scanning person and in cost and complexity of the scanner setup, the results obtained with the camera rig are only slightly better than our smartphone scans (see Figure 8). Averaging over all 33 scans and comparing our RSME (M = 32.83, SD = 4.88) with that of the multi-camera rig (M = 32.29, SD = 6.36) reveals that the error increases by less than 2%, while the financial cost decreases by more than 98%.

4.2 Qualitative Comparisons

Besides the easy-to-use visually guided scanning procedure, our method improves the approach of Wenninger et al. (2020) by several technical contributions, as described in Section 3.2. To evaluate the effect of these contributions, we compare with their method in Figure 9. The two subjects were captured in an outdoor environment by recording videos (for their method) and images (for our method) on the same iPhone 12 Pro. The avatar generation took 15 minutes (their method) and 22 minutes (our method). Our method produces noticeably more accurate results, with more geometric detail and higher-quality textures. This results from more accurate photogrammetry point clouds (due to recording higher-resolution images instead of videos and due to background removal) and from better template

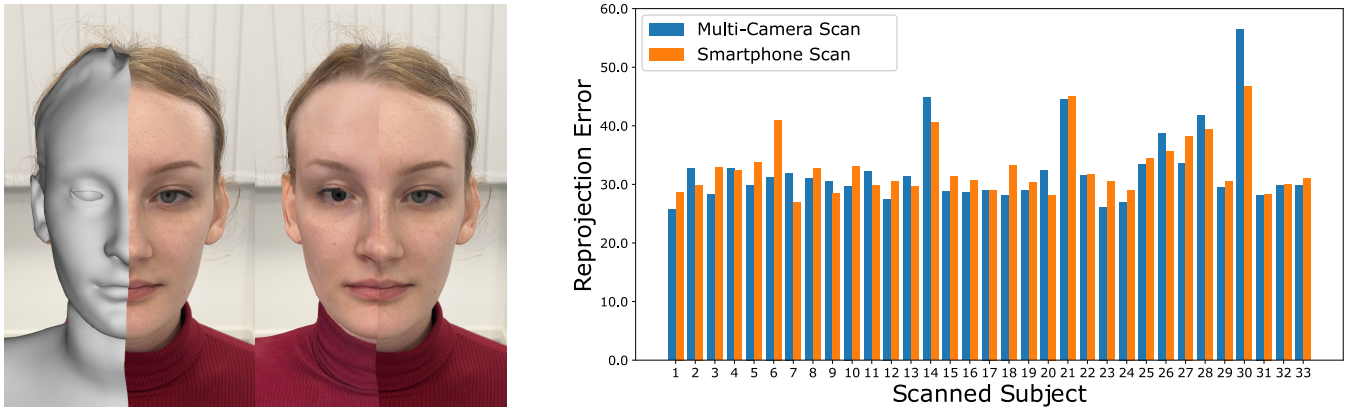

Figure 8. (a): We evaluate the accurate of our approach by reprojecting our avatars (left half of images) into the captured images (right half of images). (b): Despite the significant difference in hardware complexity, the reprojection errors of our smartphone scans are only slightly worse than those of the multi-camera rig.

fitting (due to simultaneously fitting to body and head point clouds, and requiring less regularization). The results of Wenninger et al. (2020) suffer from considerably less geometric details and wrong body-to-head proportions. These differences are even more prevalent for the lower subject, where the head is unnaturally deformed due to camera misalignments in the photogrammetry step (see Figure 5 c).

In order to evaluate whether neural avatars are a viable alternative to mesh-based avatars for VR applications, we experimented with three recent approaches: the NeRF-based InstantAvatar (Jiang et al., 2023) and the 3DGS-based methods SplattingAvatar (Shao et al., 2024) and GART (Lei et al., 2024) – since those methods reported fast training times and high rendering performance. To achieve optimal results, we followed the recommendations of these projects and use their training scripts. Since all three methods can reconstruct avatars from the People Snapshot format (Alldieck et al., 2018b), we recorded equivalent videos (1080 $\times$ 1080 pixels, subject rotating in A-pose) using the same iPhone as for our scans. These videos are then converted to the People Snapshot format using (Alldieck et al., 2018b), and the per-frame SMPL poses are refined using Anim-NeRF (Chen et al., 2021). On this prepared data we ran InstantAvatar using their provided scripts. The data resulting from InstantAvatar then act as input for running SplattingAvatar and GART. We used batch size 4 for InstantAvatar and 2min training of GART, as these produced the best results. The results from different configurations are shown in the supplementary material.

The (required) video pre-processing (landmark detection, segmentation, VideoAvatars pipeline, and Anim-NeRF) took about 17 hours on a compute server with three Nvidia RTX 6000 having 48 GB GPU memory each. Training of InstantAvatar, SplattingAvatar, and GART took another 2–5 minutes, 25 minutes, and 2 minutes, respectively, on a different server with Nvidia RTX A5000 and 24 GB GPU memory. With more than 17 hours, the overall reconstruction time of these methods is 45 times longer than ours.

Figure 10 shows the resulting avatars in a training pose and from a training camera view. While InstantAvatar produces visual clutter, both SplattingAvatar and GART are visually appealing – although more blurry than our reconstruction. However, as shown in Figure 11 and the accompanying video, when it comes to novel poses and/or novel viewpoints, the quality of neural avatars quickly degrades to a level not acceptable for social VR applications. In particular, for multi-avatar VR applications, where users/avatars can take on arbitrary poses and be viewed from arbitrary camera positions, the artifacts shown in Figure 11 are more the rule rather than the exception. Avoiding these generalization problems would require much

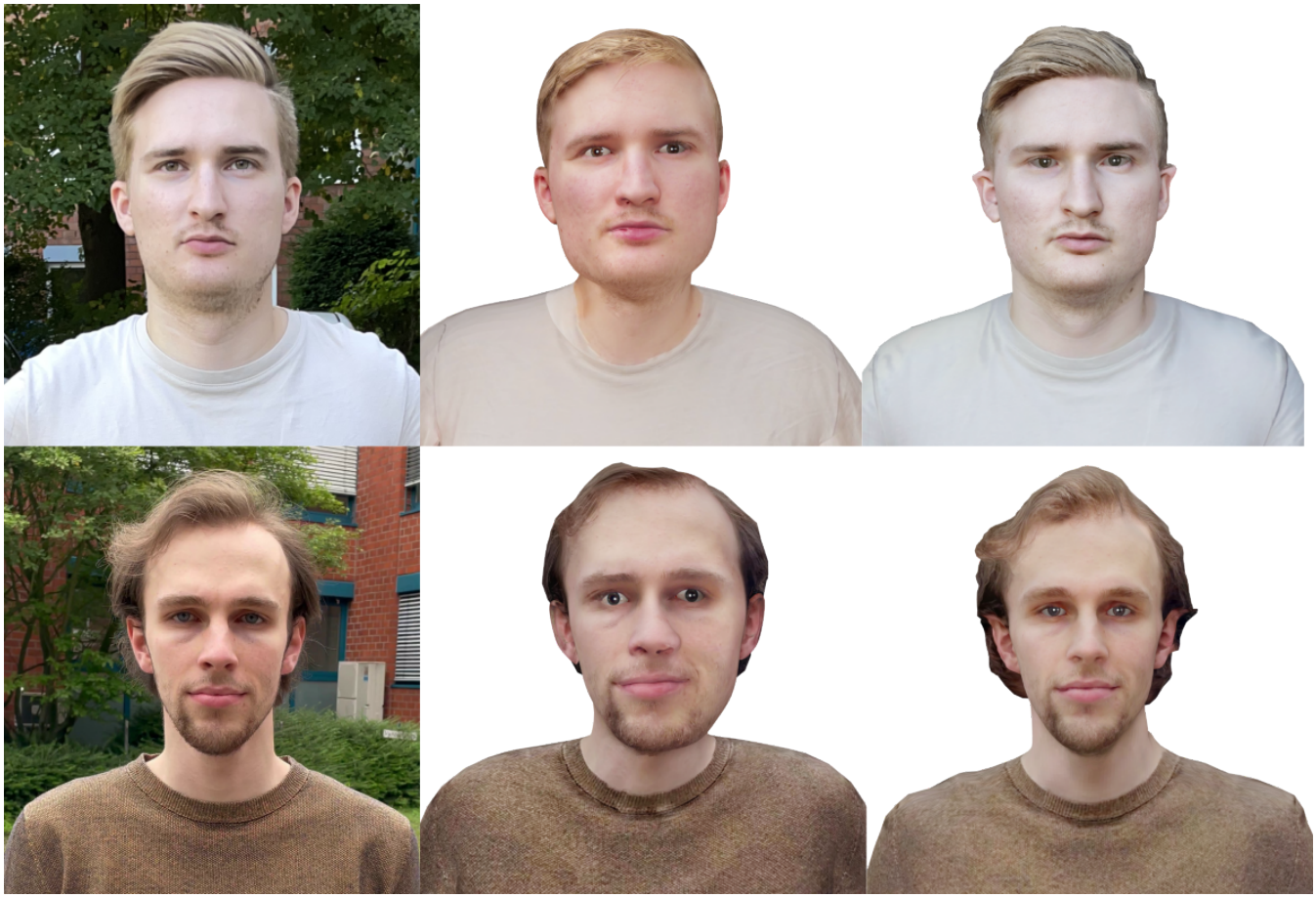

Figure 9. Two subjects captured in an outdoor environment (left), with avatars reconstructed using (Wenninger et al., 2020) (center) and our approach (right). Our avatars are considerably more accurate in terms of geometry and texture, while those of Wenninger et al. (2020) suffer from photogrammetry misalignments, (required) strong regularization, and wrong body-to-head proportions.

more training data, i.e., capturing the subject in significantly more poses and from significantly more camera views, which in turn would make the scanning significantly more complex and the reconstruction significantly more expensive – therefore requiring a complex multi-camera video recording setup. Our mesh-based avatars, in contrast, are sufficiently regularized by the statistical human body template and its animation controls to enable generalization to novel views and novel poses, even when captured from 105 smartphone images only.

In addition to long reconstruction times and suboptimal visual quality, the rendering performance of neural avatars is not (yet) sufficient for VR applications, where currently HMDs require around 90 fps stereoscopic rendering (i.e., 180 fps monoscopic rendering) at about $2k$ resolution per eye. We therefore evaluated the rendering performance at a resolution of 2160×2160 pixels on a VR workstation (AMD Ryzen 9 7950X CPU, RTX 4090 24GB GPU, 64GB RAM). For this monoscopic rendering, InstantAvatar achieved 1.22 fps (819 ms/frame), SplattingAvatar 171 fps (5.9 ms/frame), and GART 245 fps (4.1 ms/frame). While InstantAvatar was far from the required frame rate for VR application, rendering performance of SplattingAvatar and GART were just on the edge of VR applicability. However, contrary to these rather low performance results, our mesh-based representation can be rendered at 4615 fps – therefore comfortably allowing even multiple avatars in the same virtual environment at the same time.

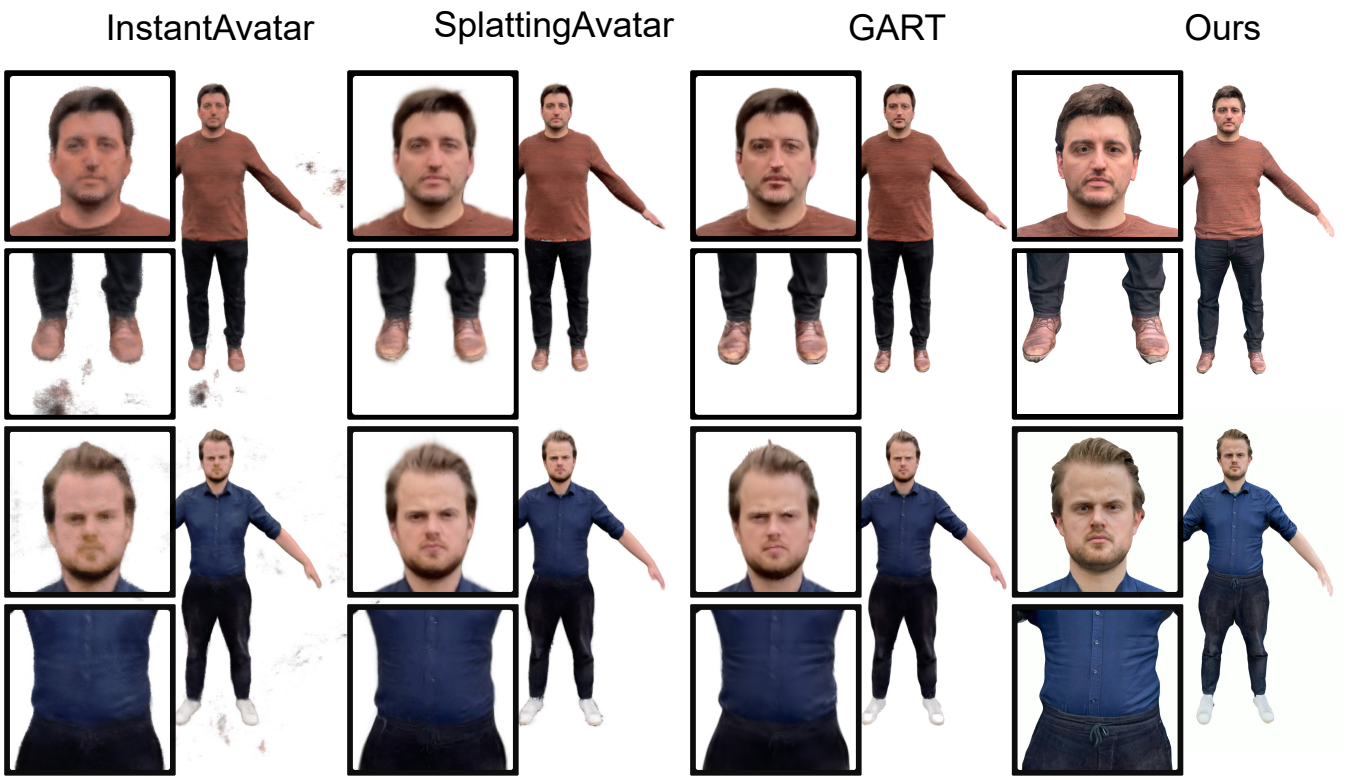

Figure 10. Reconstructed avatars rendered in a pose and view from the training data. From left to right: InstantAvatar (Jiang et al., 2023), SplattingAvatar (Shao et al., 2024), GART (Lei et al., 2024), and ours.

5 USER STUDY

We conducted a comprehensive user study following a multi-method approach. The study evaluated the quality of our generated avatars (called below *smartphone avatars*) by measuring their impact on a prominent selection of well-known and often studied avatar effects. In addition, it also evaluated the general usability and user satisfaction of the smartphone front-end. The purpose was to assess the quality of the avatars subjectively and to improve the user experience of scanning and being scanned with the smartphone app.

To this end, we arranged the participants into dyads, where one participant had to perform a smartphone app scan of another participant. While the scanning participant evaluated the app's usability afterward (in the following called *smartphone app evaluation*), the scanned participant assessed the perception of the scanning processes and the generated avatar (in the following called *avatar evaluation*).

For the smartphone app evaluation, participants performing the smartphone scan were asked to assess the app's usability using standardized questionnaires, allowing for comparison with validated benchmarks. Additionally, we conducted semi-structured interviews to gather more feedback on the user experience of both scanning and being scanned with the smartphone app. The results are used as part of a user-centered design process to improve the app.

For the avatar evaluation, we adopted and extended the approach from Bartl et al. (2021) and utilized a counterbalanced within-subject design comparing our generated smartphone avatars to (a) photorealistically reconstructed personalized avatars from a state-of-the-art expert system (in the following called *camera rig avatar*, see Section 4.1) and (b) gender- and ethnicity-matched generic avatars. We chose condition (a) to compare the quality of our *smartphone avatars* to the quality of personalized avatars frequently used in

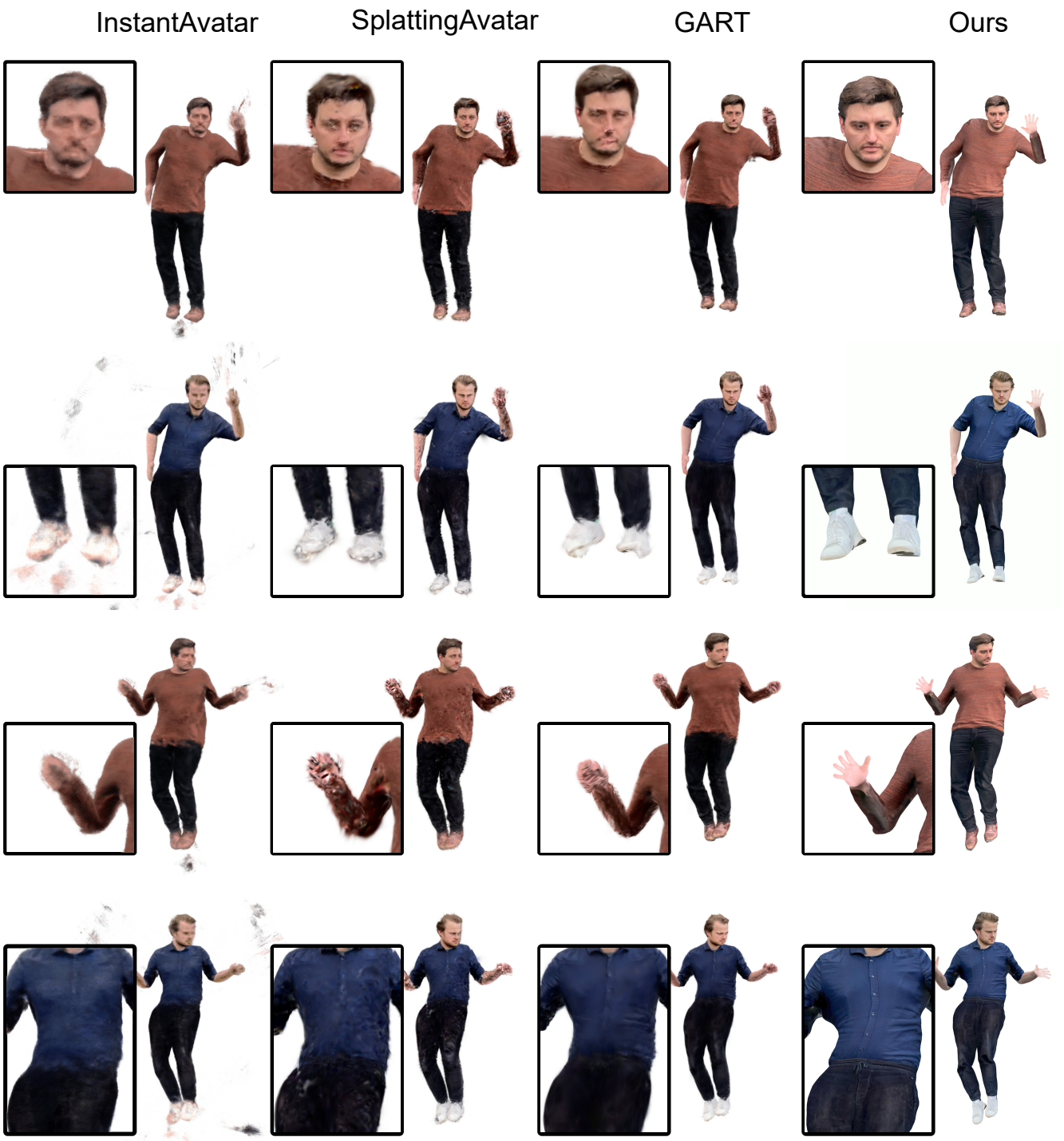

Figure 11. Avatar reconstructions animated in novel poses. From left to right: InstantAvatar (Jiang et al., 2023), SplattingAvatar (Shao et al., 2024), GART (Lei et al., 2024), and our result. Although InstantAvatar, SplattingAvatar, and GART produced visually appealing results in the training poses, the reconstructions get noisier and blurrier in novel poses. Details, e.g., hands and faces, are hardly recognizable anymore and the renderings get even blurrier. Our results in contrast are as sharp and detailed as in the training poses.

recent avatar research but reconstructed by a rather costly and complex technical setup. This allowed us to assess the impact of the proposed method's reconstruction quality on typical well-known and often studied avatar effects in relation to the much lower technical requirements of our method. We chose condition (b)

to assess the quality of both reconstruction methods in terms of self-similarity and self-attribution of the resulting avatars and to measure the overall effect of personalization. An additional in-VR comparison to any of the neural avatars (see Section 4.2) unfortunately was still unreasonable. The given state of the art was inadequate for a VR evaluation due to the low rendering performance but most importantly due to the significant artifacts (see Figure 11 and video), specifically given the arbitrary poses and camera positions typical for the VR exposure.

During individual *one-by-one exposures*, the scanned participants embodied each of the three avatar types successively while engaging in various body-centered movement tasks in front of a virtual mirror within a VR environment. Afterward, they evaluated the avatars regarding (a) sense of embodiment and self-identification, (b) plausibility, and (c) uncanny valley effects. In a final *side-by-side exposure*, participants simultaneously embodied each type of avatar while observing them exclusively from an allocentric perspective in three different virtual mirrors (one for each type) and answering different preference questions. Afterward, we asked the participants why they preferred their chosen avatars.

5.1 Apparatus

5.1.1 Avatars

In the following, we explain the integration of the three different avatar types utilized in our study.

Each participant attending the smartphone app evaluation (in the following called *scanning participant*) used our smartphone app to create a personalized avatar for the corresponding participant attending the avatar evaluation (in the following called *scanned participant*). We maintained uniform lighting conditions to enhance the avatars' comparability with the camera rig avatars. The scanning participant received instructions from the smartphone app tutorial and was directed to guide the scanned participant accordingly. No further post-processing was performed on the smartphone avatars.

We created a personalized camera rig avatar for each participant in the avatar evaluation using the expert body scanner of the Embodiment Lab at the University of Würzburg (see Section 4.1). No further post-processing was performed on the camera rig avatars.

Since avatars that do not match the user's gender and ethnicity have been shown to impact SoE particularly negatively (Do et al., 2024) and consequently would lead to an unequal comparison with personalized avatars that are matched in gender and ethnicity, we decided to match both between user and generic avatars. To this end, we chose the Validated Avatar Library for Inclusion and Diversity (VALID) (Do et al., 2023). Through a LimeSurvey questionnaire, each participant in the avatar evaluation was asked to select the VALID avatar that most closely matched their own gender and ethnicity. As the participants typically attend studies dressed casually, they could choose between 42 casually dressed VALID avatars, consisting of three male and three female avatars, each of seven different ethnicities.

5.1.2 Virtual Reality System

The VR system was realized using Unity 2020.3.25f1 (Unity Technologies, 2020). We utilized a Valve Index head-mounted display (HMD) featuring a resolution of 1440×1600 px per eye and a total field of view of $114.1 \times 109.4°$ (Wolf et al., 2022a). Its refresh rate was set to 90 Hz. Participants' hand and finger movements were tracked through two Index controllers and their built-in proximity sensors. Four SteamVR base stations covered the 3×3 m tracking area. All mentioned components were integrated into the VR system using SteamVR version 2.3 (Valve Corporation, 2024a) and its corresponding Unity plug-in version 2.7.3 (Valve Corporation, 2024b). We routed the HMD's cable to a VR-capable workstation (Intel

Core i7-7700K CPU, NVIDIA GeForce GTX 1080, 16 GB RAM) running the VR system on Windows 10. For body tracking, we utilized the markerless body tracking system from Captury. Body poses were captured using eight FLIR Blackfly S BFS-PGE-16S2C RGB cameras running at 100 Hz, which have been connected via two 4-port 1 GBit/s ethernet frame-grabber to a high-end workstation (NVIDIA GeForce RTX 3080 Ti, 32 GB RAM, AMD Ryzen 9 5900x) running Captury Live in version 259 (Captury, 2023a) on Ubuntu 18 LTS. The body poses were continuously integrated into the VR system using Captury's corresponding Unity plug-in (Captury, 2023b).

5.1.3 Avatar Embodiment

We realized avatar embodiment by retargeting the participant's tracked body pose to the used avatar in real-time following the joint approaches described in previous work (Döllinger et al., 2022; Wolf et al., 2022b). During a short calibration process, in which the participant had to stand rigidly and upright, the embodied avatar was calibrated to continuously follow the position of the HMD and scaled to match the participant's eye height. To avoid sliding feet and inaccuracies in hand and feet positions caused by variations in skeletal structure, segment lengths, or insufficient hand tracking, we utilized an IK-supported end-effector optimization using FinalIK version 2.1. Due to a higher accuracy and sampling rate, hand positions and finger poses were taken from the Index controllers, while elbow, knee, and foot positions were taken from Captury.

5.1.4 Virtual Environment and Tasks

Our virtual environment was based on different Unity assets, which we adapted to create a realistically rendered setting. Figure 12 depicts the virtual environment, accommodating up to three virtual mirrors. Following the guidelines for self-observation mirror placement by Wolf et al. (2022a), each virtual mirror was placed at a distance of 1.5 m from the participant during the study.

During each **one-by-one exposure**, participants embodied one of the three avatars in the virtual environment, where only the middle virtual mirror was shown. They could either observe their embodied avatar directly from an egocentric perspective or look into the virtual mirror to receive an allocentric perspective. Participants were asked to perform various body movement tasks in front of the virtual mirror to promote visuomotor coupling and induce SoE (Slater et al., 2010; González-Franco et al., 2010). The body movement tasks adhered to a structured protocol adapted from Roth and Latoschik (2020) and can be found in the supplements of this work.

During the **side-by-side exposure**, participants embodied all three avatars simultaneously in the virtual environment, where all three virtual mirrors were shown. While they received no egocentric perspective on the avatars, they could observe each avatar through a virtual mirror. The mirrors were labeled with small numbers, and participants responded to four different preference questions by identifying the mirror number displaying their preferred avatar. The assignment of avatars to mirrors changed randomly after each question. The preference questions can be found in the supplements of this work. Figure 12 depicts the side-by-side exposure.

5.2 Measures

5.2.1 Quantitative Measures

We assessed all quantitative measures using previously published questionnaires. When available, we used validated translated German versions of the utilized questionnaires. Otherwise, we used back-and-forth

Figure 12. The three mirrors showing the expert (left), smartphone (middle), and generic (right) avatar of a female participant during the side-by-side exposure.

translations to translate the items into German. Participants answered all questionnaires on a MacBook Pro using LimeSurvey (Limesurvey GmbH, 2024).

We captured the **usability** of the smartphone app using the System Usability Scale (SUS) (Brooke, 1996). It provides a fast and simple way to assess a system's usability using ten questionnaire items each answered on a 5-point Likert scale. The calculated overall score ranges between 0 and 100 (*100 = highest usability*) and can be compared with benchmarks provided by previous work (Bangor et al., 2009; Sauro and Lewis, 2016; Kortum and Sorber, 2015).

For assessing **Sense of Embodiment and Self-Identification** (SoE) towards the avatars, we captured virtual body ownership (VBO) and agency (AG) utilizing the corresponding items of the Virtual Embodiment Questionnaire (VEQ) (Roth and Latoschik, 2020) and self-location (SL) using the additional items introduced by Fiedler et al. (VEQ+) (Fiedler et al., 2023). We used the items capturing self-similarity (SS) and self-attribution (SA) from the VEQ+ to assess self-identification towards the avatars. Each factor measured comprises four items rated on a 7-point Likert scale (*7 = highest VBO, AG, SL, SS, and SA*).

We captured the avatars' **plausibility** utilizing the Virtual Human Plausibility Questionnaire (VHPQ) (Mal et al., 2022, 2024). It consists of seven items that assess the avatars' appearance and behavior plausibility (ABP) and four items for matching the virtual environment (MVE). Each item is rated on a 7-point Likert scale (*7 = highest ABP and MVE*).

We captured tendencies of the avatars' appearance towards the **uncanny valley** using the revised version of the Uncanny Valley Index (UVI) (Ho and MacDorman, 2017). It comprises four items each to assess

the avatars' humanness (HU) and attractiveness (AT) and eight items to capture the avatars' eeriness (EE). While the items are answered on a range between -3 and 3, we report them on a range between 1 and 7 (*7 = highest HU, AT, EE*).

As a control measure, we captured participants' physical symptoms associated with **VR sickness** in a pre-post comparison using the Virtual Reality Sickness Questionnaire (VRSQ) (Kim et al., 2018). It consists of nine items, each of which represents a typical symptom of VR sickness and is answered on a scale between 0 and 3 (*3 = highest symptomatology*). The total score of the VRSQ ranges between 0 and 100 (*100 = highest VR sickness*).

5.2.2 Qualitative Measures

We conducted semi-structured interviews to assess the user experiences related to both scanning and being scanned with the smartphone app. The interview protocols incorporated a retrospective thinking-aloud approach (Bowers and Snyder, 1990; Simon and Ericsson, 1993) to comprehensively analyze the interactions with the smartphone app while not influencing the scan experiences. We further included predefined questions to query positive and negative feelings experienced during the use of the app and while being scanned, the app's functionality and its intended purpose, the impact of the scanning participant on comfort or discomfort when being scanned, and the clarity and comprehensibility of the scanning process. Additionally, participants described aspects of the process they found efficient or challenging and reported any problematic incidents they faced. Finally, participants could suggest improvements to both the scan app functionality and the scanning process and were asked about their scan preferences and if they would participate in a body scan again. Participants in the avatar evaluation were further asked which avatar they preferred regarding self-representation similarity, fidelity, plausibility, and suitability, along with reasons behind their choices. The complete interview protocols and exact phrasing of the preference questions can be found in the supplements of this work.

5.3 Procedures

In the following, we describe the standardized experimental procedures of our smartphone app and avatar evaluations. Figure 13 visualizes both procedures and highlights their intersection during the smartphone app scan. Initially, participants in both procedures received information about the study and privacy, consented to participate, and generated two pseudonymization codes to store personal (i.e., voice recordings and avatars) and evaluation data separately. Subsequently, they proceeded with their respective evaluation procedures.

5.3.1 Smartphone App Evaluation

Each participant in the smartphone app evaluation first completed a tutorial on how to perform a body scan using the smartphone app. As soon as the other participant arrived for the scan in the laboratory, both participants were introduced to each other. The participant performing the scan verified that all requirements for the scan were met and instructed the scanned participant not to speak or move during the scan. To ensure that an assessable avatar was generated, the scanning participant performed two successive scans. After scanning, the scanned participant left the laboratory, and the scanning participant answered the SUS questionnaire using LimeSurvey. Following that, the participant was interviewed and completed demographics. On average, the entire smartphone app evaluation took approximately 41 min.

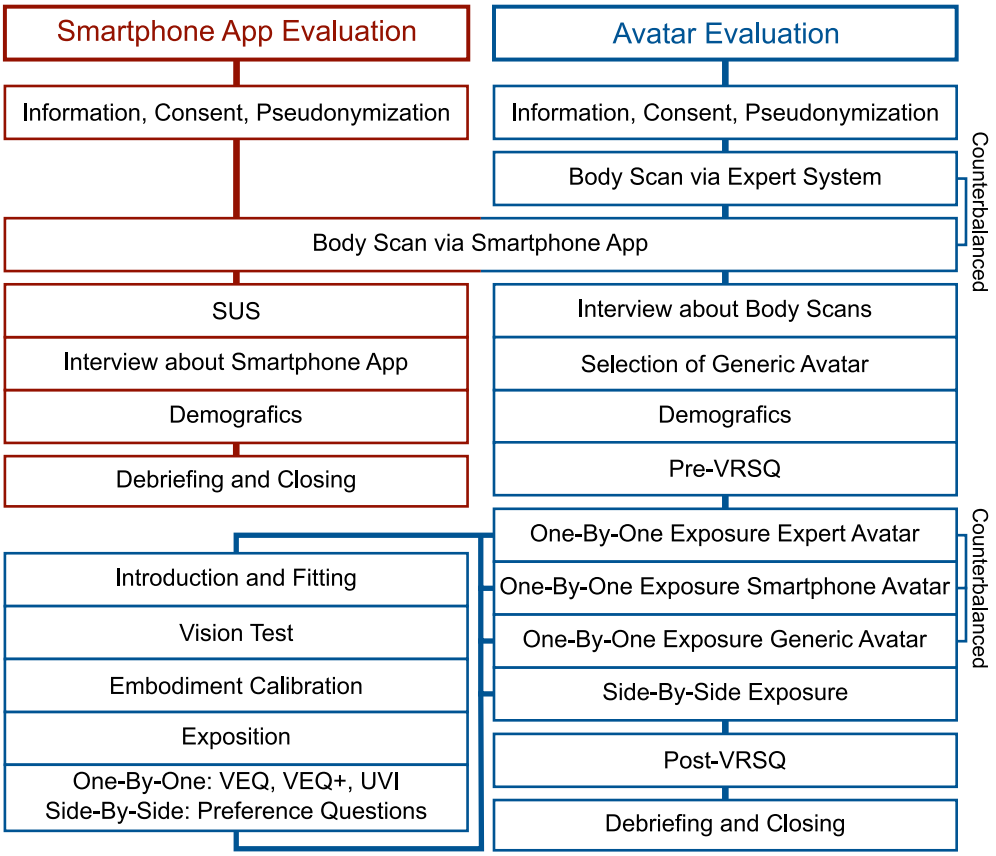

Figure 13. Experimental procedure of a dyad, illustrating the process of evaluating the smartphone app (left) and the avatars (right).

5.3.2 Avatar Evaluation

Each participant in the avatar evaluation first participated in a smartphone and expert scan conducted in a counterbalanced order. After the scans, the participant was interviewed about the scan processes, chose a generic avatar as described above, completed the demographics, and answered the pre-VRSQ. The one-by-one exposures followed in a counterbalanced order, each lasting on average 7.6 min. After each exposure, the participant answered the VEQ, VEQ+, and UVI. The following side-by-side exposure averaged 4.2 min and was accompanied by the preference questions answered verbally in VR. For each exposure, a vision test and the avatar embodiment calibration were performed following the instructions on a virtual whiteboard. In addition, the participant received audio instructions for all tasks. Finally, the participant completed the post-VRSQ. On average, the entire avatar evaluation lasted 103 min.

5.4 Participants

Adhering to the ethical standards of the Declaration of Helsinki, our study received approval from the ethics review board of the Institute Human-Computer-Media (MCM) at the University of Würzburg [1]. We recruited a total of 66 participants organized into 33 dyads using the local participant management system and compensated them either by course credits or cash, both depending on the duration of their participation. In none of the dyads, participants knew each other before the study. All participants had normal or corrected vision and no hearing impairment. Participants evaluating the smartphone app (19

[1] https://www.mcm.uni-wuerzburg.de/forschung/ethikkommission/

female, 14 male) were aged between 19 and 41 ($M = 26.60, SD = 5.48$). None of them had used the smartphone app before. Participants evaluating the avatars (25 female, 8 male) were aged between 20 and 49 ($M = 27.64, SD = 6.90$). While none of them had been scanned with the smartphone app before, nine participants had previously taken part in an expert scan. Most participants in the avatar evaluation (29 White, 2 Asian, 1 MENA) chose a generic avatar that matched their ethnicity. Only one White participant chose a Hispanic avatar. Ten participants used VR for the first time, 20 up to ten times, one more than ten times, and two more than 20 times.

We excluded one dyad from our statistical analysis as one participant used the smartphone app contrary to the instructions, resulting in an unusable avatar. While all participants stated that they had more than five years of experience with the German language, we had to exclude another participant from the avatar evaluation as the experimenter felt that the participant did not understand the questions and instructions correctly, which was confirmed by implausible answers and outliers in the data. Hence, 32 datasets remained for the smartphone app and 31 for the avatar evaluation.

5.5 Data Analysis

We conducted all quantitative analyses using SPSS version 29.0.2.0 (IBM, 2022). Before running the statistical tests, we checked whether our data met the assumption of normality and sphericity for parametric testing. Shapiro-Wilk tests showed clear violations of the normality assumption for both dimensions of the VHPQ and minor violations for VEQ agency and VEQ+ self-location. Mauchly's test for sphericity confirmed homoscedasticity between the groups for all of our measures. Since variance analysis shows robustness to slight violations of normality for groups with $N \geq 30$ (Wilcox, 2022), we decided to perform parametric tests for all measures except those from the VHPQ. All main tests have been performed against an α of .05, while post-hoc tests have been Bonferroni adjusted.

The qualitative feedback has been analyzed following the principles of thematic analysis (Braun and Clarke, 2006). Due to space restrictions, we decided to report the results mainly based on the frequency of certain feedback while mostly refraining from direct quotes.

5.6 Results

5.6.1 Smartphone App Evaluation

The quantitative evaluation of the smartphone app's usability resulted in a reasonably high SUS score ($M = 78.83, SD = 12.23$). We compared the results to absolute benchmarks from existing literature. According to Sauro and Lewis (2016), our smartphone app shows above-average usability. While a score between 77.2 and 78.8 leads to a usability grade of *B+*, a score between 78.9 and 80.7 relates to an *A-*. This grade matches the classifications of the adjective rating scale of Bangor et al. (2009), where a score above 71.4 is considered *good*, while a score above 85.5 would be *excellent*. According to the work of Kortum and Sorber (2015), our smartphone app's usability can almost keep up with the usability of the ten most-used iPhone apps, which have an average SUS score of 79.3.

When analyzing interviews about the usability of the smartphone app, the majority of the 32 participants performing the smartphone app scan found it highly usable. Twenty-nine participants found the app's functionality and purpose easy to understand, while 26 reported that they constantly knew how to use it. As particularly useful features, 20 participants highlighted the overlay for controlling scan distance and movement, 16 participants the initial tutorial, and five participants the arrows indicating the movement direction. Nonetheless, challenges were also noted. Twenty-three participants reported

Table 1. Exact descriptive values for each measure of the avatar evaluation per group and statistical results of the group comparisons.

	Smartphone	Camera Rig	Generic	Group Comparisons
	M (SD)	M (SD)	M (SD)	
Sense of Embodiment				
VEQ Ownership (VBO)	4.10 (1.50)	4.79 (1.28)	3.75 (1.44)	$F(2,60) = 11.011, p < .001, \eta_p^2 = .268$
VEQ Agency (AG)	5.57 (0.96)	5.97 (0.79)	5.79 (0.81)	$F(2,60) = 2.845, p = .066, \eta_p^2 = .087$
VEQ+ Self-Location (SL)	4.11 (1.07)	4.14 (1.03)	3.73 (1.13)	$F(2,60) = 3.502, p = .036, \eta_p^2 = .275$
VEQ+ Self-Similarity (SS)	5.69 (1.11)	5.85 (0.68)	2.69 (1.23)	$F(2,60) = 82.651, p < .001, \eta_p^2 = .734$
VEQ+ Self-Attribution (SA)	4.69 (1.21)	4.99 (0.99)	3.37 (1.16)	$F(2,60) = 31.390, p < .001, \eta_p^2 = .511$
Plausibility				
VHPQ Appearance/Behaviour (ABP)	4.84 (0.82)	5.33 (0.78)	5.27 (0.78)	$\chi^2(2) = 4.581, p = .101, W = .074$
VHPQ Match to VE (MVE)	5.15 (0.98)	5.47 (1.14)	5.80 (0.65)	$\chi^2(2) = 5.782, p = .056, W = .093$
Uncanny Valley				
UVI Humanness (HU)	3.46 (1.16)	3.84 (1.08)	3.36 (0.90)	$F(2,60) = 2.444, p = .095, \eta_p^2 = .075$
UVI Eeriness (EE)	4.05 (0.79)	3.87 (0.97)	3.15 (0.79)	$F(2,60) = 19.313, p < .001, \eta_p^2 = .392$
UVI Attractiveness (AT)	3.90 (1.20)	4.13 (0.84)	4.52 (0.69)	$F(2,60) = 3.264, p = .045, \eta_p^2 = .098$

difficulties maintaining an appropriate moving pace while scanning, with six participants emphasizing this problem, especially for the head scan. Similarly, seven and six participants reported issues with aligning the overlay while moving and keeping the correct distance, respectively. Six participants mentioned the need for high concentration, and 18 felt a bit uncomfortable due to the close proximity to the scanned participant. Six participants considered the relatively long duration of the scan process as unpleasant. To address the mentioned aspects, eight participants suggested a more detailed tutorial, and another four suggested an initial overlay mapping to the height of the scanned participant. To improve the scan process, five participants recommended more interaction with the scanned person, five more additional feedback on pacing their movement during the scan, and another five stressed the need to shorten the scan duration.

In addition to feedback on performing the scan, we obtained reports from the 32 scanned participants on their scanning experience. Overall, the process was clear and manageable, with 30 participants completely understanding the required actions. All participants confirmed their willingness to participate in a smartphone app scan again. However, compared to expert scans, 21 participants noted the smartphone app scan was slower, and 22 found it less comfortable. Prolonged posing discomfort was mentioned by twelve participants, while wardrobe and hairstyle constraints were issues for another four. Fourteen participants anticipated a difference between an expert and a beginner performing the smartphone scan, with four believing the expert would be faster. When asked about suggestions for improvement, four participants indicated that they would accelerate the process to reduce the discomfort of holding the scan pose. Regarding the head scan, four participants suggested a fixation to aid focus, and three to increase the distance between the camera and the head.

5.6.2 Avatar Evaluation

To perform group comparisons on our avatar evaluation data, we calculated either a repeated-measures ANOVA for measures that met the requirements for parametric analysis or Friedman tests as a non-parametric alternative. The descriptive data and the results of the group comparisons can be found in Table 1. For all tests revealing significant differences between groups, we calculated Bonferroni-corrected pairwise post-hoc comparisons that are reported in Figure 14.

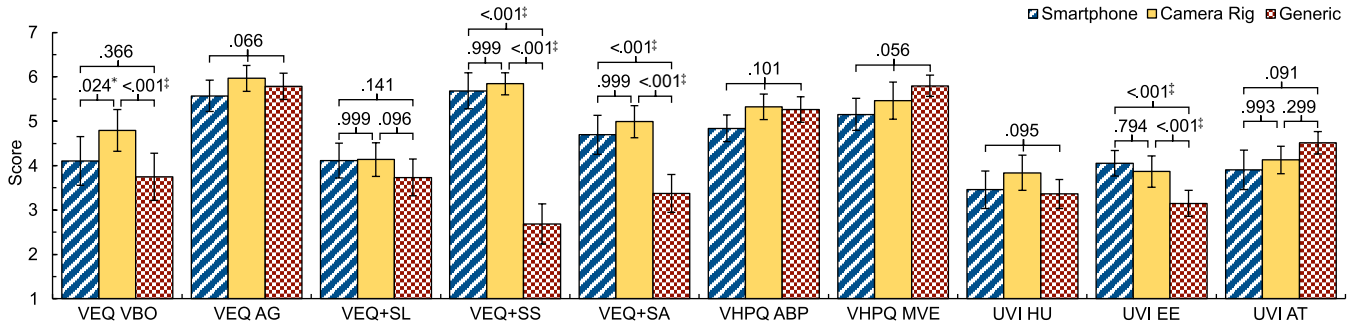

Figure 14. Bar charts for each measure and each group of the avatar evaluation, including statistical test results of the group comparisons and post-hoc tests where applicable. Error bars represent 95 % confidence intervals. Statistical significance indicators: $^{*}\, p < .05$; $^{\dagger}\, p < .01$; $^{\ddagger}\, p < .001$.

During the side-by-side exposure, we asked participants about their preferences regarding self-representation similarity, fidelity, plausibility, and suitability, along with reasons behind their choices. Out of the 31 participants included in the analysis, 16 perceived the smartphone avatars to be more similar to themselves, while 13 preferred the camera rig avatars. Regarding self-representation fidelity, 11 participants preferred the smartphone avatars, 19 chose the camera rig avatars, and one favored the generic one. To feel most plausibly represented in VR, 12 participants chose the smartphone avatars, 18 the camera rig avatars, and one the generic one. When asked which avatar the participants would prefer to be represented in VR, 10 chose the smartphone avatars, 17 the camera rig avatars, and four the generic ones. When asked for their reasoning, participants favoring smartphone avatars mostly mentioned a detailed facial reconstruction and realism as key factors. Those participants who preferred camera rig avatars highlighted the accuracy of body shape reconstruction, noting issues with smartphone avatars' body proportions, particularly the arms. Participants who chose generic avatars consistently did so because of overall dissatisfaction with their personal appearance rather than avatar quality.

6 DISCUSSION

In this section, we discuss the results of the comparisons with the different avatar reconstruction methods and the results of our user study and present the limitations of our work.

6.1 Smartphone App Evaluation

We evaluated the usability of our smartphone app quantitatively using the SUS questionnaire and qualitatively using semi-structured interviews, including a retrospective thinking-aloud approach. The SUS results showed that our smartphone app is already well usable. The qualitative feedback confirmed this impression and highlighted the overlay and tutorial as particularly positive features. However, the qualitative feedback also revealed areas for improvement.

As part of the user-oriented design process, we already incorporated suggested improvements. To address comments regarding the duration of the scan and the pace, we added the option to shorten or extend the scan speed using technical means. The unclear parts of the tutorial have been improved to prepare users for the scan better. Furthermore, we have also added warnings if the scanned person is not sufficiently centered. Other feedback could not be implemented due to technical limitations or requires further research. For example, the distance between the smartphone and the scanned person, especially during the head scan, could only be increased by the loss of detail in the reconstructed avatars. However, since the high

Table 2. Comparison of the different avatar reconstruction methods regarding our requirements. The symbols represent: ✓ completely, ● partially, ✗ not fulfilled.

Method	Easy	Fast	Affordable	Realistic	Full-Body	VR-Ready
Achenbach et al. (2017)	✗	✓	✗	✓	✓	✓
Wenninger et al. (2020)	✗	✓	✓	✓	✓	✓
Jiang et al. (2023)	●	✗	✓	●	✓	✗
Shao et al. (2024)	●	✗	✓	●	✓	✗
Lei et al. (2024)	●	✗	✓	●	✓	✗
Ours	✓	✓	✓	✓	✓	✓

quality of the faces is a significant advantage of our system, we decided to keep the required distance. Furthermore, the interaction between the scanning and scanned person and visual aids (e.g., fixation point) for the scanned person lies outside the influence of our smartphone application.

6.2 Avatar Evaluation

Our quantitative and qualitative comparisons in Section 4 demonstrate that our smartphone application enables even non-experts to reconstruct avatars of a similar quality and accuracy as those produced with an expert-operated multi-camera rig – at a fraction of the price, complexity, and required expertise. Compared to the previous smartphone-based reconstruction (Wenninger et al., 2020), our proposed method is easier to use and gives higher-quality results even in more challenging in-the-wild scenarios.

Our experiments with InstantAvatar (Jiang et al., 2023), SplattingAvatar (Shao et al., 2024), and GART Lei et al. (2024) revealed that neural avatars generalize rather poorly to poses and camera views far from training data – a situation that cannot be avoided in multi-avatar VR applications. Although these generalization problems can be reduced with more training data, this is beyond the capabilities of a simple smartphone-based scanning solution. Also in terms of reconstruction times and rendering performance are neural avatars not yet suitable for VR applications, such that classical mesh-based avatars appear to still be the preferred representation. Table 2 summarizes the fulfillment of our requirements with respect to avatar reconstruction. Although many methods show their strengths in a subset of the criteria, only our system fulfills all of them.

Compared to the work of Waltemate et al. (2018), our user study confirmed that realistic avatars still offer substantial benefits over generic avatars for self-representation, even when the generic avatars are also personalized in gender and ethnicity (Do et al., 2023, 2024). With regard to the comparison, some further notable findings need to be addressed. The statistically significant difference in virtual body ownership between the smartphone and camera rig avatars can potentially be attributed to observed motion artifacts, which can degrade the avatars' appearance. However, smartphone avatars still perform descriptively better than generic avatars. Regarding self-identification, the smartphone and camera rig avatars both show significant advantages to generic avatars, although the smartphone avatars were generated using a significantly cheaper method than the camera rig avatars. For the smartphone avatars, participants emphasized particularly the high similarity of the head. However, results also showed that the eeriness of realistic avatars was significantly higher than generic avatars. This is likely attributable to an Uncanny Valley effect originating from the emotional relatedness to self-personalized avatars, which has also been observed in other research (Mori et al., 2012; Döllinger et al., 2023). When considering the plausibility of

the avatars, it is noticeable that the reconstruction described most realistically had the lowest match with the perceived plausibility. This discrepancy might be attributed to the incongruence between the virtual environment's realistic style and the avatars' photorealistic style (Latoschik and Wienrich, 2022).

6.3 Limitations

Since our method uses photogrammetry software to generate point clouds from images, the input images must contain as little movement as possible. If movement occurs in the background, the segmentation significantly improves the photogrammetry results. However, the motions of the scanned subject violate the photogrammetry assumption, i.e., that the scanned object is rigid and not moving, leading to less accurate point clouds and, therefore, geometric deformations in the final avatar. Figure 7 shows this problem in more detail, as the arms of the second avatar (from left) have visible differences in thickness.

We use a mesh-based representation for our avatars. On the one hand, this enables high-performance rendering and novel pose generation. On the other hand, we represent cloth, hair, and skin by a single textured mesh, which can lead to visual artifacts. An interesting direction for future work would be to combine mesh-based avatars (potentially with multiple layers for skin and cloth) with volumetric details (such as hair) represented by Gaussian Splatting – as this would combine the strengths of both representations.

Our system uses image segmentation to preprocess the input and mask out regions that do not contain people. For that reason, people in the background are a challenging task, as they are not removed. We want to explore the capabilities of the depth sensor to remove people in the background from the masks.

We rely on Apple frameworks for both our scanning client and reconstruction server. For the client, an Android-based app would be possible, but we chose iOS because Apple's photogrammetry yields correctly scaled results thanks to the LiDAR sensor of the pro-level iPhones. On the reconstruction server, only the photogrammetry and the segmentation frameworks are Apple-specific, all other parts of the reconstruction pipeline are cross-platform compatible. We chose Apple's RealityKit and Vision frameworks since in our extensive tests this platform produced the best results while not being limited by restrictive licensing.

The sample in our study consisted of white participants only. As this potentially limits the generalizability of our results, in future work a larger population sample with greater variability in age, sex, and ethnicity should be tested.

7 CONCLUSION

We presented *Avatars for the Masses*, a system that allows non-expert users to scan people and automatically reconstruct realistic VR-ready full-body avatars that achieve similar perception results compared to avatars reconstructed with expensive expert-operated systems. Inspired by the approach of Wenninger et al. (2020), we presented methods to resolve present obstacles that prevent the wide accessibility of realistic full-body avatars. Our custom smartphone application enables laypeople to easily and quickly capture high-quality input images, which, together with background segmentation and an improved template fitting algorithm, result in more convincing reconstructions while reducing restrictions on scanning locations. Our end-to-end solution computes VR-ready avatars that can be easily integrated into existing VR pipelines. To further empower people to create realistic full-body avatars and encourage more avatar-related studies, we will make *Avatars for the Masses* publicly available for research purposes.

REFERENCES

Achenbach, J., Waltemate, T., Latoschik, M. E., and Botsch, M. (2017). Fast Generation of Realistic Virtual Humans. In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology*. doi:10.1145/3139131.3139154

[Dataset] Agisoft (2023). Agisoft Metashape. [visited on 2023-09-29]

[Dataset] Alice Vision (2025). Meshroom. [visited on 2025-02-12]

Alldieck, T., Magnor, M., Bhatnagar, B. L., Theobalt, C., and Pons-Moll, G. (2019). Learning to Reconstruct People in Clothing From a Single RGB Camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*

Alldieck, T., Magnor, M., Xu, W., Theobalt, C., and Pons-Moll, G. (2018a). Detailed Human Avatars from Monocular Video. In *International Conference on 3D Vision (3DV)*. 98–109. doi:10.1109/3DV.2018.00022

Alldieck, T., Magnor, M., Xu, W., Theobalt, C., and Pons-Moll, G. (2018b). Video Based Reconstruction of 3D People Models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*

[Dataset] Apple Inc. (2023a). ARKit Framework - Blendshapes. [visited on 2023-09-27]

[Dataset] Apple Inc. (2023b). RealityKit Framework - Object Capture - PhotogrammetrySession. [visited on 2024-07-05]

[Dataset] Apple Inc. (2023c). Vision Framework - VNFaceLandmarks2D. [visited on 2024-07-05]

[Dataset] Apple Inc. (2023d). Vision Framework - VNGeneratePersonSegmentationRequest. [visited on 2024-07-05]

[Dataset] Apple Inc. (2025). Vision Framework - VNDetectHumanHandPosesRequest. [visited on 2025-02-12]

Aseeri, S. and Interrante, V. (2021). The Influence of Avatar Representation on Interpersonal Communication in Virtual Social Environments. *IEEE Transactions on Visualization and Computer Graphics* 27, 2608–2617. doi:10.1109/TVCG.2021.3067783

Bailenson, J. N. and Blascovich, J. (2004). Avatars. In *Encyclopedia of Human-Computer Interaction* (Great Barrington, MA, USA: Berkshire Publishing Group). 64–68

Bangor, A., Kortum, P., and Miller, J. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of usability studies* 4, 114–123

Bartl, A., Wenninger, S., Wolf, E., Botsch, M., and Latoschik, M. E. (2021). Affordable but not Cheap: A Case Study of the Effects of Two 3D-Reconstruction Methods of Virtual Humans. *Frontiers in Virtual Reality* 2. doi:10.3389/frvir.2021.694617

Bowers, V. A. and Snyder, H. L. (1990). Concurrent versus Retrospective Verbal Protocol for Comparing Window Usability. *Proceedings of the Human Factors Society Annual Meeting* 34, 1270–1274. doi:10.1177/154193129003401720

Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 77–101. doi:10.1191/1478088706qp063oa

Brooke, J. (1996). SUS: A quick and dirty usability scale. In *Usability evaluation in industry* (Taylor & Francis). 4–7

Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., and Sheikh, Y. A. (2019). OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*

[Dataset] Captury (2023a). CapturyLive 259

[Dataset] Captury (2023b). Unity plugin

[Dataset] Chen, J., Zhang, Y., Kang, D., Zhe, X., Bao, L., Jia, X., et al. (2021). Animatable Neural Radiance Fields from Monocular RGB Videos

[Dataset] Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation

de Vignemont, F. (2011). Embodiment, ownership and disownership. *Consciousness and Cognition* 20, 82–93. doi:https://doi.org/10.1016/j.concog.2010.09.004

[Dataset] Dlib (2022). Dlib C++ Library

Do, T. D., Isabella Protko, C., and McMahan, R. P. (2024). Stepping into the Right Shoes: The Effects of User-Matched Avatar Ethnicity and Gender on Sense of Embodiment in Virtual Reality. *IEEE Transactions on Visualization and Computer Graphics* , 1–10doi:10.1109/TVCG.2024.3372067

Do, T. D., Zelenty, S., Gonzalez-Franco, M., and McMahan, R. P. (2023). VALID: A perceptually validated virtual avatar library for inclusion and diversity. *Frontiers in Virtual Reality* 4. doi:10.3389/frvir.2023.1248915

Döllinger, N., Mal, D., Keppler, S., Wolf, E., Botsch, M., Israel, J. H., et al. (2024). Virtual body swapping: A vr-based approach to embodied third-person self-processing in mind-body therapy. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18

Döllinger, N., Beck, M., Wolf, E., Mal, D., Botsch, M., Latoschik, M. E., et al. (2023). "If it's not me it doesn't make a difference" – The impact of avatar customization and personalization on user experience and body awareness in virtual reality. In *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. doi:10.1109/ISMAR59233.2023.00063

Döllinger, N., Wolf, E., Mal, D., Wenninger, S., Botsch, M., Latoschik, M. E., et al. (2022). Resize Me! Exploring the User Experience of Embodied Realistic Modulatable Avatars for Body Image Intervention in Virtual Reality. *Frontiers in Virtual Reality* 3. doi:10.3389/frvir.2022.935449

Feng, A., Suma, E., and Shapiro, A. (2017). Just-in-Time, Viable, 3D Avatars from Scans. In *ACM SIGGRAPH 2017 Talks*. doi:10.1145/3084363.3085045

Fiedler, M. L., Wolf, E., Döllinger, N., Botsch, M., Latoschik, M. E., and Wienrich, C. (2023). Embodiment and personalization for self-identification with virtual humans. In *IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. 799–800. doi:10.1109/VRW58643.2023.00242

Fiedler, M. L., Wolf, E., Döllinger, N., Mal, D., Botsch, M., Latoschik, M. E., et al. (2024). From avatars to agents: Self-related cues through embodiment and personalization affect body perception in virtual reality. *IEEE Transactions on Visualization and Computer Graphics* Accepted for publication

Gall, D., Roth, D., Stauffert, J.-P., Zarges, J., and Latoschik, M. E. (2021). Embodiment in virtual reality intensifies emotional responses to virtual stimuli. *Frontiers in Psychology* 12, 3833. doi:10.3389/fpsyg.2021.674179

González-Franco, M., Pérez-Marcos, D., Spanlang, B., and Slater, M. (2010). The contribution of real-time mirror reflections of motor actions on virtual body ownership in an immersive virtual environment. In *2010 IEEE Virtual Reality Conference (VR)*. 111–114. doi:10.1109/VR.2010.5444805

Guo, C., Jiang, T., Chen, X., Song, J., and Hilliges, O. (2023). Vid2Avatar: 3D Avatar Reconstruction From Videos in the Wild via Self-Supervised Scene Decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12858–12868

Habermann, M., Liu, L., Xu, W., Pons-Moll, G., Zollhoefer, M., and Theobalt, C. (2023). HDHumans: A Hybrid Approach for High-fidelity Digital Humans. *Proc. ACM Comput. Graph. Interact. Tech.* 6. doi:10.1145/3606927

Ho, C.-C. and MacDorman, K. F. (2017). Measuring the uncanny valley effect. *International Journal of Social Robotics* 9, 129–139. doi:10.1007/s12369-016-0380-9

Hu, L., Zhang, H., Zhang, Y., Zhou, B., Liu, B., Zhang, S., et al. (2024). GaussianAvatar: Towards Realistic Human Avatar Modeling from a Single Video via Animatable 3D Gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 634–644

[Dataset] IBM (2022). SPSS Statistics. `https://www.ibm.com/products/spss-statistics`

Jiang, B., Hong, Y., Bao, H., and Zhang, J. (2022). SelfRecon: Self Reconstruction Your Digital Avatar from Monocular Video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*

Jiang, T., Chen, X., Song, J., and Hilliges, O. (2023). InstantAvatar: Learning Avatars From Monocular Video in 60 Seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16922–16932

Jiang, Z., Guo, C., Kaufmann, M., Jiang, T., Valentin, J., Hilliges, O., et al. (2024). Multiply: Reconstruction of multiple people from monocular video in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 109–118

Jung, S. and Hughes, C. E. (2016). The effects of indirect real body cues of irrelevant parts on virtual body ownership and presence. In *Proceedings of the 26th International Conference on Artificial Reality and Telexistence and the 21st Eurographics Symposium on Virtual Environments*. 107–114

Kerbl, B., Kopanas, G., Leimkühler, T., and Drettakis, G. (2023). 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics* 42

Kilteni, K., Groten, R., and Slater, M. (2012). The sense of embodiment in virtual reality. *Presence: Teleoperators & Virtual Environments* 21, 373–387. doi:10.1162/PRES_a_00124

Kim, D. Y., Lee, H. K., and Chung, K. (2023). Avatar-mediated experience in the metaverse: The impact of avatar realism on user-avatar relationship. *Journal of Retailing and Consumer Services* 73. doi:https://doi.org/10.1016/j.jretconser.2023.103382

Kim, H. K., Park, J., Choi, Y., and Choe, M. (2018). Virtual reality sickness questionnaire (VRSQ): Motion sickness measurement index in a virtual reality environment. *Applied Ergonomics* 69, 66–73. doi:https://doi.org/10.1016/j.apergo.2017.12.016

Kortum, P. and Sorber, M. (2015). Measuring the Usability of Mobile Applications for Phones and Tablets. *International Journal of Human–Computer Interaction* 31, 518–529. doi:10.1080/10447318.2015.1064658

Kwon, Y., Liu, L., Fuchs, H., Habermann, M., and Theobalt, C. (2023). DELIFFAS: Deformable Light Fields for Fast Avatar Synthesis. In *Advances in Neural Information Processing Systems*. vol. 36, 40944–40962

Latoschik, M. E., Kern, F., Stauffert, J.-P., Bartl, A., Botsch, M., and Lugrin, J.-L. (2019). Not alone here?! scalability and user experience of embodied ambient crowds in distributed social virtual reality. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 25, 2134–2144

Latoschik, M. E. and Wienrich, C. (2022). Congruence and plausibility, not presence: Pivotal conditions for XR experiences and effects, a novel approach. *Frontiers in Virtual Reality* 3. doi:10.3389/frvir.2022.694433

Lei, J., Wang, Y., Pavlakos, G., Liu, L., and Daniilidis, K. (2024). GART: Gaussian Articulated Template Models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 19876–19887. doi:10.1109/CVPR52733.2024.01879

Leyrer, M., Linkenauger, S. A., Bülthoff, H. H., Kloos, U., and Mohler, B. (2011). The influence of eye height and avatars on egocentric distance estimates in immersive virtual environments. In *Proceedings of the ACM SIGGRAPH Symposium on Applied Perception in Graphics and Visualization*. 67–74. doi:10.1145/2077451.2077464

Li, Z., Zheng, Z., Wang, L., and Liu, Y. (2024). Animatable Gaussians: Learning Pose-dependent Gaussian Maps for High-fidelity Human Avatar Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 19711–19722

[Dataset] Limesurvey GmbH (2024). LimeSurvey: An Open Source survey tool

Lin, W., Zheng, C., Yong, J.-H., and Xu, F. (2024). Relightable and Animatable Neural Avatars from Videos. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 3486–3494. doi:10.1609/aaai.v38i4.28136

Liu, L., Habermann, M., Rudnev, V., Sarkar, K., Gu, J., and Theobalt, C. (2021). Neural actor: neural free-view synthesis of human actors with pose control. *ACM Trans. Graph.* 40. doi:10.1145/3478513.3480528

[Dataset] Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., et al. (2019). MediaPipe: A Framework for Building Perception Pipelines

Ma, S., Simon, T., Saragih, J., Wang, D., Li, Y., Torre, F. L., et al. (2021). Pixel Codec Avatars. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 64–73. doi:10.1109/CVPR46437.2021.00013

Mal, D., Wolf, E., Döllinger, N., Botsch, M., Wienrich, C., and Latoschik, M. E. (2022). Virtual Human Coherence and Plausibility – Towards a Validated Scale. In *IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. 788–789. doi:10.1109/VRW55335.2022.00245

Mal, D., Wolf, E., Döllinger, N., Botsch, M., Wienrich, C., and Latoschik, M. E. (2024). From 2D-screens to VR: Exploring the effect of immersion on the plausibility of virtual humans. In *CHI 24 Conference on Human Factors in Computing Systems Extended Abstracts*. 8

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. (2020). NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*

Mohler, B. J., Creem-Regehr, S. H., Thompson, W. B., and Bülthoff, H. H. (2010). The effect of viewing a self-avatar on distance judgments in an HMD-based virtual environment. *Presence* 19, 230–242. doi:10.1162/pres.19.3.230

Moreau, A., Song, J., Dhamo, H., Shaw, R., Zhou, Y., and Pérez-Pellitero, E. (2024). Human Gaussian Splatting: Real-time Rendering of Animatable Avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 788–798

Morgenstern, W., Bagdasarian, M. T., Hilsmann, A., and Eisert, P. (2024). Animatable virtual humans: Learning pose-dependent human representations in uv space for interactive performance synthesis. *IEEE Transactions on Visualization and Computer Graphics* 30, 2644–2650. doi:10.1109/TVCG.2024.3372117

Mori, M., MacDorman, K. F., and Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics Automation Magazine* 19, 98–100. doi:10.1109/MRA.2012.2192811

Mottelson, A., Muresan, A., Hornbæk, K., and Makransky, G. (2023). A systematic review and meta-analysis of the effectiveness of body ownership illusions in virtual reality. *ACM Trans. Comput.-Hum. Interact.* doi:10.1145/3590767

Müller, T., Evans, A., Schied, C., and Keller, A. (2022). Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Trans. Graph.* 41. doi:10.1145/3528223.3530127

Mystakidis, S. (2022). Metaverse. *Encyclopedia* 2, 486–497. doi:10.3390/encyclopedia2010031

Pang, H., Zhu, H., Kortylewski, A., Theobalt, C., and Habermann, M. (2024). ASH: Animatable Gaussian Splats for Efficient and Photoreal Human Rendering. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1165–1175. doi:10.1109/CVPR52733.2024.00117

Pastel, S., Chen, C.-H., Petri, K., and Witte, K. (2020). Effects of body visualization on performance in head-mounted display virtual reality. *PLOS ONE* 15, 1–18. doi:10.1371/journal.pone.0239226

Peng, S., Dong, J., Wang, Q., Zhang, S., Shuai, Q., Zhou, X., et al. (2021a). Animatable neural radiance fields for modeling dynamic human bodies. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 14294–14303. doi:10.1109/ICCV48922.2021.01405

Peng, S., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Bao, H., et al. (2021b). Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans . In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9050–9059. doi:10.1109/CVPR46437.2021.00894

Pérez, P., Gangnet, M., and Blake, A. (2003). Poisson image editing. *ACM Transactions on Graphics* 22, 313–318

Robinette, K., Daanen, H., and Paquet, E. (1999). The CAESAR project: a 3-D surface anthropometry survey. In *Second International Conference on 3-D Digital Imaging and Modeling (Cat. No.PR00062)*. 380–386. doi:10.1109/IM.1999.805368

Roth, D. and Latoschik, M. E. (2020). Construction of the virtual embodiment questionnaire (VEQ). *IEEE Transactions on Visualization and Computer Graphics* 26, 3546–3556. doi:10.1109/TVCG.2020.3023603

Salagean, A., Crellin, E., Parsons, M., Cosker, D., and Stanton Fraser, D. (2023). Meeting Your Virtual Twin: Effects of Photorealism and Personalization on Embodiment, Self-Identification and Perception of Self-Avatars in Virtual Reality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. doi:10.1145/3544548.3581182

Sampaio, M., Navarro Haro, M. V., De Sousa, B., Vieira Melo, W., and Hoffman, H. G. (2021). Therapists Make the Switch to Telepsychology to Safely Continue Treating Their Patients During the COVID-19 Pandemic. Virtual Reality Telepsychology May Be Next. *Frontiers in Virtual Reality* 1. doi:10.3389/frvir.2020.576421

Sauro, J. and Lewis, J. R. (2016). *Quantifying the user experience: Practical statistics for user research* (Morgan Kaufmann)

Schönberger, J. L. and Frahm, J.-M. (2016). Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*

Shao, Z., Wang, Z., Li, Z., Wang, D., Lin, X., Zhang, Y., et al. (2024). SplattingAvatar: Realistic Real-Time Human Avatars with Mesh-Embedded Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1606–1616

Shetty, A., Habermann, M., Sun, G., Luvizon, D., Golyanik, V., and Theobalt, C. (2024). Holoported Characters: Real-time Free-viewpoint Rendering of Humans from Sparse RGB Cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1206–1215

Simon, H. A. and Ericsson, K. A. (1993). *Protocol analysis: Verbal reports as data* (The MIT Press)

Skarbez, R., Frederick P. Brooks, J., and Whitton, M. C. (2017). A survey of presence and related concepts. *ACM Computing Surveys* 50, 96. doi:10.1145/3134301

Skarbez, R. and Jiang, D. (2024). A Scientometric History of IEEE VR. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. 990–999. doi:10.1109/VR58804.2024.00118

Slater, M., Spanlang, B., Sanchez-Vives, M. V., and Blanke, O. (2010). First person experience of body transfer in virtual reality. *PLOS ONE* 5, e10564. doi:10.1371/journal.pone.0010564

Steed, A., Pan, Y., Zisch, F., and Steptoe, W. (2016). The impact of a self-avatar on cognitive load in immersive virtual reality. In *2016 IEEE Virtual Reality (VR)*. 67–76. doi:10.1109/VR.2016.7504689

Sutherland, I. E. (1968). A head-mounted three-dimensional display. In *Proceedings of the December 9-11, 1968, Fall Joint Computer Conference, Part I*. 757–764. doi:10.1145/1476589.1476686

Turbyne, C., Goedhart, A., de Koning, P., Schirmbeck, F., and Denys, D. (2021). Systematic Review and Meta-Analysis of Virtual Reality in Mental Healthcare: Effects of Full Body Illusions on Body Image Disturbance. *Frontiers in Virtual Reality* 2, 39. doi:10.3389/frvir.2021.657638

[Dataset] Unity Technologies (2020). Unity 2020.3.25f1

[Dataset] Valve Corporation (2024a). Steam VR 2.3

[Dataset] Valve Corporation (2024b). Steam VR Plugin 2.7.3

Waltemate, T., Gall, D., Roth, D., Botsch, M., and Latoschik, M. E. (2018). The Impact of Avatar Personalization and Immersion on Virtual Body Ownership, Presence, and Emotional Response. *IEEE Transactions on Visualization and Computer Graphics* 24, 1643–1652. doi:10.1109/TVCG.2018.2794629

Wang, S., Antic, B., Geiger, A., and Tang, S. (2024). IntrinsicAvatar: Physically Based Inverse Rendering of Dynamic Humans from Monocular Videos via Explicit Ray Tracing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1877–1888

Wang, T., Zhang, B., Zhang, T., Gu, S., Bao, J., Baltrusaitis, T., et al. (2023). RODIN: A Generative Model for Sculpting 3D Digital Avatars Using Diffusion. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4563–4573. doi:10.1109/CVPR52729.2023.00443

Wenninger, S., Achenbach, J., Bartl, A., Latoschik, M. E., and Botsch, M. (2020). Realistic Virtual Humans from Smartphone Videos. In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology*. doi:10.1145/3385956.3418940

Wiegand, T., Sullivan, G., Bjontegaard, G., and Luthra, A. (2003). Overview of the H.264/AVC video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology* 13, 560–576. doi:10.1109/TCSVT.2003.815165

Wilcox, R. R. (2022). *Introduction to robust estimation and hypothesis testing* (Academic Press)

Wolf, E., Döllinger, N., Mal, D., Wenninger, S., Andrea, B., Botsch, M., et al. (2022a). Does Distance Matter? Embodiment and Perception of Personalized Avatars in Relation to the Self-Observation Distance in Virtual Reality. *Frontiers in Virtual Reality* 3. doi:10.3389/frvir.2022.1031093

Wolf, E., Fiedler, M. L., Döllinger, N., Wienrich, C., and Latoschik, M. E. (2022b). Exploring Presence, Avatar Embodiment, and Body Perception with a Holographic Augmented Reality Mirror. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 350–359. doi:10.1109/VR51125.2022.00054

Wolf, E., Merdan, N., Döllinger, N., Mal, D., Wienrich, C., Botsch, M., et al. (2021). The embodiment of photorealistic avatars influences female body weight perception in virtual reality. In *IEEE Virtual Reality and 3D User Interfaces (VR)*. 65–74. doi:10.1109/VR50410.2021.00027

Xiao, J., Zhang, Q., Xu, Z., and Zheng, W.-S. (2024). NECA: Neural Customizable Human Avatar. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20091–20101

Yoon, B., Kim, H.-i., Lee, G. A., Billinghurst, M., and Woo, W. (2019). The Effect of Avatar Appearance on Social Presence in an Augmented Reality Remote Collaboration. In *IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 547–556. doi:10.1109/VR.2019.8797719

Yu, W., Fan, Y., Zhang, Y., Wang, X., Yin, F., Bai, Y., et al. (2023). NOFA: NeRF-Based One-Shot Facial Avatar Reconstruction. In *ACM SIGGRAPH 2023 Conference Proceedings*. doi:10.1145/3588432.3591555

Zhao, H., Zhang, J., Lai, Y.-K., Zheng, Z., Xie, Y., Liu, Y., et al. (2022). High-fidelity human avatars from a single rgb camera. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 15883–15892. doi:10.1109/CVPR52688.2022.01544

Zheng, Z., Huang, H., Yu, T., Zhang, H., Guo, Y., and Liu, Y. (2022). Structured Local Radiance Fields for Human Avatar Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 15893–15903

Zheng, Z., Zhao, X., Zhang, H., Liu, B., and Liu, Y. (2023). AvatarReX: Real-time Expressive Full-body Avatars. *ACM Trans. Graph.* 42. doi:10.1145/3592101