Timo Menzel[1], Erik Wolf[1], Stephan Wenninger[1], Niklas Spinczyk[1], Lena Holderrieth[1], Ulrich Schwanecke[1], Marc Erich Latoschik[1], and Mario Botsch[1]

[1]Affiliation not available

August 30, 2024

Abstract

Realistic full-body avatars play a key role in representing users in virtual environments, where they have been shown to considerably improve body ownership and presence. Driven by the growing demand for realistic virtual humans, extensive research on scanning-based avatar reconstruction has been conducted in recent years. Most methods, however, require complex hardware, such as expensive camera rigs and/or controlled capture setups, thereby restricting avatar generation to specialized labs. We propose WILDAVATARS, an approach that empowers even non-experts without access to complex equipment to capture realistic avatars in the wild. Our avatar generation is based on an easy-to-use smartphone application that guides the user through the scanning process and uploads the captured data to a server, which in a fully automatic manner reconstructs a photorealistic avatar that is ready to be downloaded into a VR application. To increase the availability and foster the use of realistic virtual humans in VR applications we will make WILDAVATARS publicly available for research purposes.

WILDAVATARS: Smartphone-Based Reconstruction of Full-Body Avatars in the Wild

Timo Menzel[†], Erik Wolf[†], Stephan Wenninger, Niklas Spinczyk, Lena Holderrieth
Ulrich Schwanecke, Marc Erich Latoschik, Mario Botsch

***Abstract*—Realistic full-body avatars play a key role in representing users in virtual environments, where they have been shown to considerably improve body ownership and presence. Driven by the growing demand for realistic virtual humans, extensive research on scanning-based avatar reconstruction has been conducted in recent years. Most methods, however, require complex hardware, such as expensive camera rigs and/or controlled capture setups, thereby restricting avatar generation to specialized labs. We propose WILDAVATARS, an approach that empowers even non-experts without access to complex equipment to capture realistic avatars in the wild. Our avatar generation is based on an easy-to-use smartphone application that guides the user through the scanning process and uploads the captured data to a server, which in a fully automatic manner reconstructs a photorealistic avatar that is ready to be downloaded into a VR application. To increase the availability and foster the use of realistic virtual humans in VR applications we will make WILDAVATARS publicly available for research purposes.**

***Index Terms*—Virtual reality, virtual human, avatar, agent, embodiment, virtual body ownership, agency, self-location, body image, body weight perception**

I. INTRODUCTION

Avatars are digital representations of users that can be dynamically rendered in real-time within virtual environments to reflect their users' behavior [1]. In our research, we specify the term *avatar* further to 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 resemble the appearance of their respective user closely. With the recent surge of virtual reality (VR) research [2] and the increasing availability of mature head-mounted displays (HMDs) [3], avatars have become more and more important as faithful self-representations of users in almost countless scenarios. These scenarios include metaverse-like social VR environments [4], [5], [6], [7] or VR applications for supporting mental health [8], [9], where maintaining the users' identity and conveying realistic emotions are crucial for authentic interactions and a sophisticated user experience (UX). Prior work has shown that realistically personalized avatars, which can look deceptively similar to the user, are superior for the outlined scenarios by strengthening the users' sense of presence and embodiment or increasing their self-identification with the avatar [10], [11], [12], [13].

Therefore, extensive research on the generation of realistic avatars has been conducted in recent years, aiming to realize convincing and plausible virtual representations. While some approaches focus on the full-body reconstruction of humans [14], [15], [16], [17], [18], others only focus on certain body parts [19], [20], [21], [22], [23], [24]. However, most of these approaches are subject to considerable limitations, including the need for expensive hardware or complicated scan routines, which make them inaccessible to non-experts, place restrictions on the person to reconstruct or the scanning environment that reduces potential use-cases, or long computation times that hinder the immediate use of avatars in virtual environments.

To simplify the generation of realistically personalized avatars and make them accessible to a broader audience, we present WILDAVATARS. This user-centered system consists of a customized smartphone application for image capturing and an accompanying server for avatar reconstruction. Our smartphone application visually guides the user through the scanning process, making it easy to follow the instructions. The captured images are uploaded to a processing server to create a high-quality avatar of the scanned person. Using photogrammetry, landmark detection, and an automatic template-fitting approach, we generate realistic, personalized full-body avatars in less than 20 minutes. In addition, we present solutions to minimize the restrictions on the scanning location and further enable non-professionals to create virtual people of themselves using only a smartphone.

To evaluate the quality of our proposed system, we conducted a user study following a multi-method approach to assess both the smartphone app's usability and the quality of the created avatars. To this end, we arranged participants into dyads, where one participant used the smartphone app to create an avatar of another participant. Afterwards, the

- *Timo Menzel, Stephan Wenninger, Niklas Spinczyk, and Mario Botsch are with the Computer Graphics Group, TU Dortmund University, 44227 Dortmund, Germany. E-Mail: {timo.menzel, stephan.wenninger, niklas.spinczyk, mario.botsch}@tu-dortmund.de.*
- *Erik Wolf, Lena Holderrieth and Marc Erich Latoschik are with the Human-Computer Interaction (HCI) Group, Julius-Maximilians-Universität Würzburg (JMU), 97074 Würzburg, Germany. E-Mail: {erik.wolf, lena.holderrieth, marc.latoschik}@uni-wuerzburg.de.*
- *Erik Wolf is also with Human-Computer Interaction (HCI) Group, Universität Hamburg, 22605 Hamburg, Germany. E-Mail: erik.wolf@uni-hamburg.de.*
- *Ulrich Schwanecke is with the Computer Vision and Mixed Reality Group, RheinMain University of Applied Sciences, 65195 Wiesbaden, Germany. E-Mail: ulrich.schwanecke@hs-rm.de.*

This research has been funded by the German Federal Ministry of Education and Research (BMBF) in the projects VIA-VR (16SV8446) and ViTraS (16SV8219/16SV8225), and by the "Stiftung Innovation in der Hochschullehre" through the project "Hybrid Learning Center" (FBM2020-EA-690-01130). Erik Wolf gratefully acknowledges a Meta Research PhD Fellowship.

(Corresponding author: Timo Menzel)

[†] These authors contributed equally

participant who carried out the scan assessed the smartphone app's usability as part of qualitative interviews and qualitative benchmarks, while the scanned participant reported on the experience of being scanned before evaluating the generated avatar in VR in comparison to avatars originating from an expert system and to gender- and ethnicity-matched generic avatars.

II. Related Work

Various approaches for creating realistic avatars have been proposed. There are differences in the type of capture device and input data or in the representation of the resulting avatars.

A. Input Data

Some of the methods employ expensive camera rigs to capture accurate and high-quality images as input for reconstructing humans [11], [14], [19], [25], [26], [27], [28]. These methods can provide impressive results but are usually unsuitable for non-experts due to the high cost of the scanning and computation hardware or the complex recording process and operation.

There are approaches to minimize hardware costs by using monocular videos as input [15], [16], [17], [18], [29], [30], [31], [32]. These methods reduce the number of required cameras to just one, which increases the accessibility and affordability for non-experts. On the one hand, relying on videos is a good possibility to record every side of a subject for a convincing reconstruction. On the other hand, video data from consumer devices usually is more compressed than image data and contains more blurring. In consequence, compromises in geometric accuracy and texture resolution are unavoidable, as we show in Section III-A1.

Cao et al. [33] and Ichim et al. [20] reconstruct avatars using smartphones. In both methods, the user captures several images from different perspectives. While Ichim et al. compute an avatar in under an hour, Cao et al. need up to six hours for the computation. However, both approaches only generate virtual head avatars, which restricts the applicability in different VR scenarios.

Recent developments try to use just one image as input for the reconstruction [22], [23], [34]. These approaches use prior knowledge to fill in the missing information. However, although learning-based methods can reduce the impact of this missing data and provide plausible results, the fundamentally ill-posed nature of the problem leads to less accurate results than methods that use more input.

To address the mentioned shortcomings, our approach also uses a single camera instead of a multi-camera rig but captures consecutive images instead of videos with the aim of reducing the amount of blurring and artifacts in the captured images. Since our proposed system is designed to be available for everybody, we use a commodity smartphone instead of a special camera rig. Furthermore, our visual guidance and timer-based capturing process help the user to easily capture the necessary images for our pipeline while preserving the conveniences of video capture.

B. Representations of Avatars

Avatars can be represented in various ways. Many methods use explicit mesh-based representations to reconstruct realistic avatars [14], [16], [17], [18], [29]. These methods usually require a pre-defined rigged template model whose mesh is deformed to closely match the shape of the captured person. The advantage of such template-based approaches is that the resulting avatars can be used in existing graphics pipelines and game engines (e.g., Unity or Unreal) and are, therefore, suitable for current AR/VR applications. However, these methods restrict the animation possibilities since clothing and hair are often represented by the same surface and cannot be simulated in a physically correct way.

In contrast to explicit representations, implicit representations [35], [36], [37], [38] allow to model fine details since they are not restricted by a fixed topology. Neural Radiance Fields (NeRF) [15], [30], [39], [40], [41], [42], [43] are therefore better suited for animating clothing and hair due to the fine details in color and shape. Recently, Gaussian Splatting-based methods [44], [45], [46], [47], [48] have become popular. Starting from a point cloud or a mesh, these methods use a mixture of Gaussians to represent realistic avatars. Nonetheless, computing these implicit representations often is a time-consuming task (computation times between two hours and ten days for [31], [43], [44], [45], [46], [48]) and the more expensive rendering leads to lower frame rates (43 fps [31], 0.025 fps [38], 0.05 fps [41], 0.2 fps [42], 25 fps [43] and 10 fps [47]) depending on the complexity of the rendered scene, e.g., the number of avatars in social VR scenarios. This makes those approaches less useful for VR applications where high frame rates are crucial to prevent motion sickness [49].

C. Avatars for Self-Representation in Virtual Reality

The egocentric embodiment of avatars for self-representation in VR [50] can have a considerably positive impact on the UX of virtual environments [51]. These effects include the enhancement of VR's psychometric key features, such as the sense of presence [10], [52], [53], or intensifying emotional responses to virtual content [10], [54]. Other advantages may include improved spatial perception [55], [56], reduced cognitive load [57] or higher performance and accuracy [58], [59] 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 truly owning (ownership), controlling (agency), and being located within (self-location) a virtual body in a virtual environment [60], [61]. Prior work has shown that the realistic personalization of avatars can increase the SoE towards the avatar [10], [62], [11] and thereby contribute to an overall plausible VR experience [63].

Having a realistically personalized body becomes particularly valuable when it helps maintain the user's identity, as beneficial in social VR experiences [5], [6], [7] or applications supporting mental health [8], [9], [64]. Prior work has also shown that self-related cues through both avatar embodiment and personalization significantly increase self-identification with the avatar [12], potentially maintaining a more accurate self-perception in VR. However, a realistic personalization of avatars might also impact UX negatively, as their human-like realism in connection with their high affinity to the user can potentially invoke Uncanny Valley effects, leading to negative emotional responses like eeriness towards the avatars [65], [66]. Therefore, we evaluate our avatars in relation to the here presented psychometric measures to preclude significant negative effects and ensure a positive UX when being used for self-representation in VR.

III. Technical System

This section describes our proposed method and then shows results, including quantitative and qualitative comparisons with two state-of-the-art methods and our ablation studies.

A. Method

Our proposed approach is inspired by the smartphone-based avatar reconstruction of Wenninger et al. [18]. They propose to record a body and a head scan video, from which they extract individual video frames. These video frames are then used to generate two dense point clouds using Agisoft Metashape [67], a widely used commercial photogrammetry software. After that, OpenPose [68] is used to automatically detect landmarks that serve as input together with the corresponding head and body point clouds for a template fitting approach that uses non-rigid registration to deform a template model separately to the head or body of the scanned person. The resulting head and body mesh are then combined. Subsequently, the resulting geometry is textured using a graph-cut algorithm to map the colors of the video frames to the corresponding part in the mesh. While we follow a similar approach to generate point clouds from visual input data and subsequently apply template fitting to personalize an avatar template, we support the data acquisition using a custom smartphone application (Section III-A1) and add a preprocessing step (Section III-A2) to reduce prior restrictions on scanning locations. Instead of fitting in two separate steps, we present a method to register head and body point clouds, which allows us to perform a simultaneous fitting of both parts (Section III-A3). Figure 1 shows an overview of our pipeline.

1) Data Acquisition: Analogously to Wenninger et al. [18], we record people by performing a full-body scan in A-pose and a close-up head scan using a smartphone (see Figure 1, top left). However, similar to Ichim et al. [20], we capture individual images (105 images at 3024 × 4032 px resolution) during the scanning process instead of extracting individual frames from a video stream. Video captured with most current smartphone cameras is compressed using H.264 or H.265 compression algorithms. These compression algorithms are designed to view each frame only for a fraction of a second. This allows for 50% storage savings at the cost of various compression artifacts [69]. Furthermore, video compression uses inter-frame compression, i.e., utilizing the similarity of consecutive frames. H.264 and H.265 compression uses motion detection to compress movement between frames. For this, the image is grouped into regions of 4 × 4 px to 16 × 16 px, and their movement between consecutive frames is stored [69]. While this allows for very efficient video compression, these blocks are typically not contained in the next frame exactly, further degrading image quality. This increase in image resolution and quality leads to more accurate point clouds (see Figure 2) and, therefore, allows more detailed textures and finer details in the final avatar mesh.

To reduce user-induced errors in the scan results and make the creation of realistic avatars available to non-expert users as well, we developed a custom application for iOS that features (i) timer-based image capture, (ii) visual guidance through the scanning process, and (iii) automatic upload to our processing server. Our application captures images at a rate of one image per second. In addition, we display a green overlay during the scanning process, which rotates to show the user the next camera position (see Figure 3). Arrows below or next to the overlay further indicate the movement direction. The user is informed via a dialog when the scan is complete (see Figure 3, right column). The entire scan procedure can be seen in the accompanying video.

2) Data Preprocessing: We use computationally expensive tasks like image segmentation and photogrammetry to create an avatar from the captured images. To avoid relying on

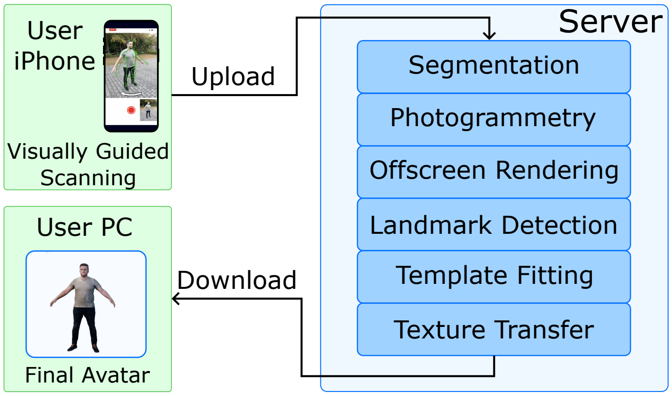


Fig. 1. Our avatar fitting pipeline. A user scans a subject with our custom smartphone application. After scanning, the images are uploaded to our processing server, where we automatically reconstruct a realistic avatar, which is ready to be downloaded into any VR application. The entire reconstruction pipeline takes about 22 minutes.

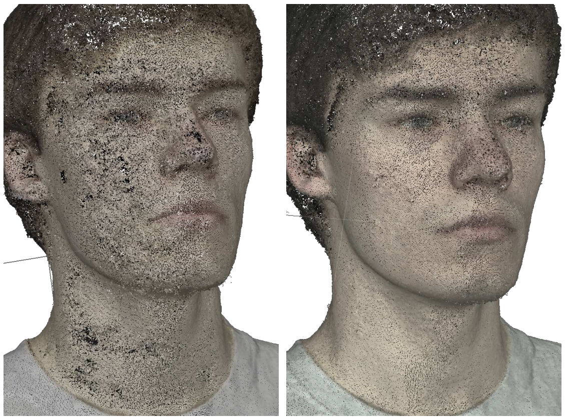
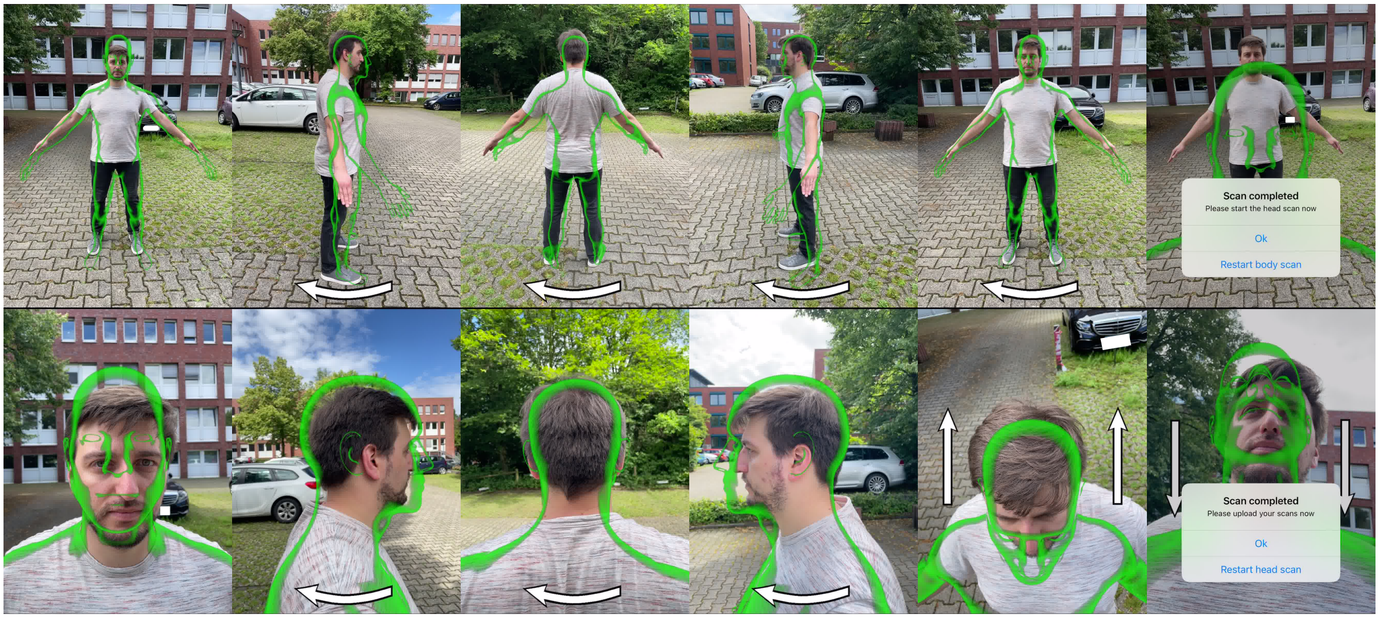


Fig. 2. Reconstructed point clouds from video frames (left) and images (right).


Scan completed
Please start the head scan now

Scan completed
Please upload your scans now

Fig. 3. Scanning procedure of our iOS avatar scanning app. Top left: Initial overlay before starting the body scan; Top middle: Overlay during the body scan with a direction arrow; Top right: End of body scan alert; Bottom left: Initial overlay before starting the head scan; Bottom middle: Overlay during the head scan with direction arrow(s); Bottom right: End of head scan alert.

the client's computational resources, we upload the captured images to a server for reconstruction. After finishing the reconstruction, the server notifies the user via email, including a download link for the VR-ready avatar. In the following, we describe the different steps of the reconstruction pipeline on our server.

Background Segmentation: Scanning people using a smartphone in uncontrolled outdoor settings presents a pressing issue. The presence of moving objects in the environment, such as driving cars or the leaves of trees in the wind, significantly impacts the accuracy of the results. These movements in the background impact the photogrammetry step in our pipeline by violating the photogrammetry assumption, i.e., scanning rigid non-moving objects, leading to false extrinsic camera calibrations. Segmenting the input images into foreground and background simplifies the problem for the photogrammetry algorithm since features and movements in the background can no longer affect the feature extraction. Therefore, the photogrammetry algorithm aligns the camera positions only relative to the scanned subject. Furthermore, the segmentation significantly reduces the number of features that must be matched against each other. Besides, the background of the images is excluded from the reconstructed point cloud and, thus, removes noise (see Figure 4, top). While slight movements in the background would often result in duplicate point clouds, by only aligning the cameras relative to the subject, the amount of noise and duplicate parts is reduced (see Figure 4, bottom).

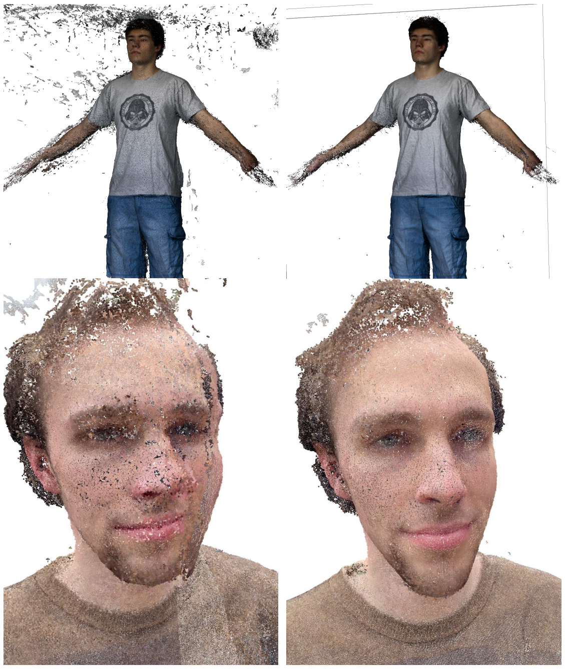


Fig. 4. Segmenting the images into foreground and background reduces noise in reconstructed point clouds. Point clouds without segmenting the input images (left), point clouds reconstructed from segmented input images (right).

We compared the *DeepLabV3* [70] implementation from PyTorch with Apple's person segmentation framework (on macOS 14.5) [71]. While both methods reliably segmented the captured person, Apple's segmentation framework produced slightly better results in our test cases. Thus, we chose this system to create the segmentation masks (see Figure 1 and Figure 5).

Photogrammetry: The captured images and the generated binary masks are passed to a photogrammetry software (see Figure 1, top middle). Wenninger et al. reconstructed point clouds using Agisoft Metashape [67], a widely used commercial tool for that task. Since we want to make our system

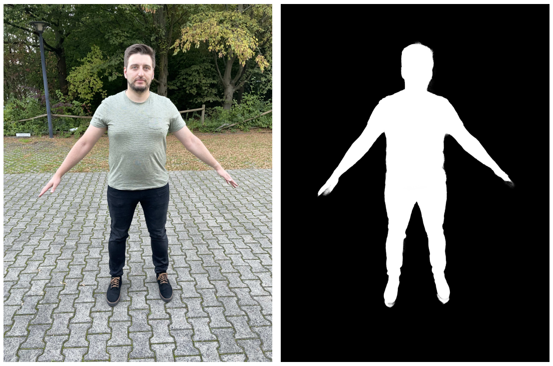


Fig. 5. Exemplary result of Apple's person segmentation framework. Input image (left) and segmentation mask (right).

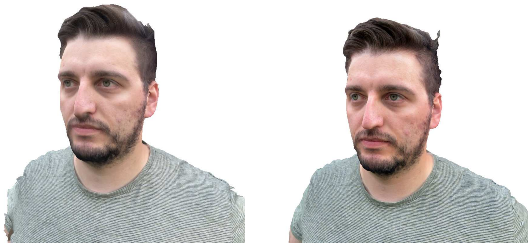


Fig. 6. Photogrammetry results from Agisoft Metashape (left) and Apple's photogrammetry toolkit (right).

publicly available for research purposes as a client-server solution, we cannot use this software due to license restrictions. Therefore, we decided to use Apple's photogrammetry toolkit [72] (on macOS 14.5).

It provides a high-level photogrammetry API that allows users to specify four different configuration settings. In our approach, we use the detail setting *raw*, feature sensitivity setting *high*, and sample ordering *unordered*. Additionally, we provide segmentation masks, depth data, gravity vectors, and EXIF data, which improved the photogrammetry results significantly. The depth data, captured using the smartphone's depth sensor (576 px × 768 px), helps to scale the reconstructed object to the correct world size, and the gravity vectors help to compute the object's correct orientation.

Although Metashape offers many more different settings throughout the photogrammetry pipeline, we can produce results of similar quality with Apple's photogrammetry toolkit using the described settings, as shown in Figure 6.

Landmark Detection: Wenninger et al. [18] automatically detect landmarks that guide their template fitting approach and are particularly important to accurately fit the face region. Using the estimated camera calibration, their approach automatically detects landmarks in the 2D images and projects them back onto the computed 3D point cloud. Unfortunately, we can only retrieve a textured mesh from Apple's photogrammetry pipeline, while the intrinsic camera calibration is unavailable as output data. To still be able to detect landmarks on the photogrammetry result, we instead render the resulting textured mesh from various camera positions. This allows us to run standard 2D landmark detection to detect 37 landmarks (eye contours, the tip of the nose, and mouth features) on the rendered images and back-project the landmarks onto the photogrammetry result using the information from the rendered depth buffers. We compared Dlib [73], Apple's face landmarks framework [74], and MediaPipe [75]. We use MediaPipe in our pipeline as it worked best in our test cases.

3) Template Fitting: We perform curvature-adaptive point sampling on the reconstructed photogrammetry mesh and then add the back-projected landmarks as individual points to the resulting point clouds. These point clouds are the input for a template fitting step, where a statistical, animatable template model is deformed to match the scanned data. The implemented template fitting approach follows the method presented in [18], with some notable differences explained in the following.

Template: The main mesh of our template model is defined by a set of $N = 23{,}752$ vertices $\mathcal{V}$ and their positions $\mathcal{X} = \left(\mathbf{x}_1, \ldots, \mathbf{x}_N\right)$, a skinned rig consisting of 59 joints controlled by joint angles $\boldsymbol{\theta} \in \mathbb{R}^{179}$, a 30-dimensional parametric shape model $\left(\mathbf{P}, \boldsymbol{\mu}\right)$ consisting of principal component matrix $\mathbf{P}$ and mean $\boldsymbol{\mu}$ controlled by shape parameters $\boldsymbol{\alpha}$, and 52 blendshapes modeled after the ARKit blendshape set [76]. Additionally, the template mesh includes auxiliary meshes and blendshapes for mouth, teeth, and eyes. The pose of the template model can be adjusted by using the standard linear blend skinning [77] equation $\texttt{skin}(\mathcal{X}, \boldsymbol{\theta})$, which takes unposed vertex positions $\mathcal{X}$ and the joint angles θ to compute posed vertex positions $\mathcal{X}'$.

Registration: The reconstructed head and body point clouds are not in the same coordinate system and may have different scaling due to inaccuracies in the captured depth data. Therefore, as a first step, we align the head to the body scan by employing ICP with scaling guided by the previously computed set of matching landmarks in both point clouds. This step ensures that the proportions of the head and body are respected, which is not the case in the approach of Wenninger et al. [18], as they first fit their template to the body scan and then register the head scan afterward (see Figure 9).

Our template fitting approach is then implemented as a two-step energy minimization problem. First, we optimize

- the alignment from point clouds to the template given by affine transformation composed of rotation $\mathbf{R} \in SO(3)$, scaling $s \in \mathbb{R}$, and translation $\mathbf{t} \in \mathbb{R}^3$,
- the pose of the template given by joint angles $\boldsymbol{\theta} \in \mathbb{R}^{179}$,
- the coarse shape of the deformable model defined by PCA parameters $\boldsymbol{\alpha} \in \mathbb{R}^{30}$,

by minimizing the following cost function:

$$E_{\text{init}}(\boldsymbol{\alpha}, \boldsymbol{\theta}, \mathbf{R}, s, \mathbf{t}) = E_{\text{fit}}(\boldsymbol{\alpha}, \boldsymbol{\theta}, \mathbf{R}, s, \mathbf{t}) + E_{\text{reg}}(\boldsymbol{\alpha})\,, \qquad (1)$$

where E_{fit} is responsible for fitting the model to the data by penalizing the weighted squared per-correspondence distances between the template mesh and the point clouds:

$$E_{\text{fit}}(\boldsymbol{\alpha}, \boldsymbol{\theta}, \mathbf{R}, s, \mathbf{t}) = \frac{1}{\sum w_c} \sum_{c \in \mathcal{C}} w_c \left\| \, s\mathbf{R}\mathbf{p}_c + \mathbf{t} - \texttt{skin}_c(\mathbf{P}\boldsymbol{\alpha} + \boldsymbol{\mu}, \boldsymbol{\theta}) \, \right\|^2 , \qquad (2)$$

where $\texttt{skin}_c$ is the standard linear blend skinning equation applied to the corresponding point on the template surface (expressed via barycentric coordinates), $\mathcal{C}$ is a set of correspondences, weighted by per-correspondence weights w_c. It contains both the automatically detected landmarks and closest point correspondences. We compute the closest point correspondences from scan to template since this computation direction provides more accurate fitting [78]. The closest point on the template mesh is found for every point in the point clouds, defined by barycentric coordinates for a specific triangle. Note that this set contains correspondences from *head* and *body* point clouds, i.e., shape and pose are simultaneously optimized for both point cloud targets. To this end, we exclude all correspondences between the body point cloud and the head region of the template model and vice versa. $E_{\text{reg}}(\boldsymbol{\alpha})$ is a weighted Tikhonov regularization term that prevents the shape model from overfitting:

$$E_{\text{reg}}(\boldsymbol{\alpha}) = \frac{\lambda_{\text{reg}}}{d} \parallel \boldsymbol{\Gamma}\boldsymbol{\alpha} \parallel^2 , \tag{3}$$

where d is the dimension of the shape model, $\boldsymbol{\Gamma} = \texttt{diag}(1/\sigma_1, ..., 1/\sigma_d)$ and σ_i are the eigenvalues of the PCA matrix. We minimize Equation (1) by block coordinate descent, i.e., we alternatingly optimize the alignment, pose, and shape parameters. The alignment is computed in a closed-form manner [79], the pose is optimized using inverse kinematics [80], and we optimize the shape parameters subject to the per-correspondence distances between template mesh and point clouds are minimized in the least squares sense.

This optimization results in a coarse registration of the template model, namely the resulting vertex positions $\bar{\mathcal{X}}$, the estimated pose parameters $\bar{\boldsymbol{\theta}}$, rotation $\mathbf{R}$, scaling s, translation $\mathbf{t}$ and the shape parameters $\boldsymbol{\alpha}$. Since the parametric shape model is not expressive enough to accurately model clothing or hair, we further refine this coarse registration with a fine-scale physics-based deformation. Formally, we minimize

$$E_{\text{fine}}(\mathcal{X}) = E_{\text{fit}}(\mathcal{X}) + \mu_{\text{reg}} E_{\text{reg}}\left(\mathcal{X}, \bar{\mathcal{X}}\right) , \tag{4}$$

where

$$E_{\text{fit}}(\mathcal{X}) = \frac{1}{\sum w_c} \sum_{c \in \mathcal{C}} w_c \parallel \mathbf{p}_c - \texttt{skin}_c\big(\mathcal{X}, \bar{\boldsymbol{\theta}}\big) \parallel^2$$

penalizes the squared per-correspondence distances similar to Equation (2) and

$$E_{\text{reg}}\big(\mathcal{X}, \bar{\mathcal{X}}\big) = \frac{1}{\sum_e A_e} \sum_{e \in \mathcal{E}} A_e \parallel \Delta \mathbf{x}(e) - \mathbf{R}_e \Delta \bar{\mathbf{x}}(e) \parallel^2$$

penalizes geometric distortion from the result of the initial registration by measuring the squared deviation of the per-edge Laplacian weighted by per-edge areas A_e (see [14], [18] for details).

Again, we iteratively minimize Equation (4) by block coordinate descent by alternatingly solving for new vertex positions and per-edge rotations $\mathbf{R}_e$. We use the per-correspondence weights w_c to give less influence to correspondences in the hand region, as it might be unreliably reconstructed in the photogrammetry step, and to gradually decrease the landmarks' influence. The regularization weight μ_{reg} is also gradually decreased from $\mu_{\text{reg}} = 1$ to $\mu_{\text{reg}} = 10^{-9}$. Note that due to the measures taken to ensure a more reliable photogrammetry result (Section III-A1 and Section III-A2), we can decrease the regularization weight further than [18], resulting in more geometric details in the final reconstruction.

Texture Generation: As mentioned in Section III-A2, Apple's photogrammetry toolkit does not provide intrinsic camera calibrations. Therefore, we cannot project the resulting mesh onto the captured images to generate the texture. Instead, we transfer the texture from the photogrammetry's resulting mesh to the reconstructed avatar. For each texel, we determine the point on the avatar using the texture coordinates. We then calculate the closest point to the mesh from the photogrammetry and then use its texture coordinates to determine the texel in the mesh's texture. This corresponding texel is then transferred to the avatar texture.

B. Results

The avatar reconstruction with our system takes about 22 minutes. Due to our guided scanning process, the scanning time is set at two minutes. In the process, 105 images (45 full-body and 60 head images, approx. 320 MB) are captured using an iPhone. The upload to our server (Mac Studio, M1-Max 10-Core CPU, 32-Core integrated GPU, 64 GB RAM) was carried out via a WiFi connection and took on average less than one minute in our experiments; the reconstruction on our server took about 19 minutes. Figure 7 shows renderings of our reconstructed avatars. All subjects were scanned using our custom smartphone application running on an iPhone 12 Pro, and the avatars were automatically computed on our processing server.

To quantitatively evaluate the reconstructed avatars, we compare to the multi-camera rig approach of Achenbach et al. [14]. Furthermore, we present qualitative comparisons to the smartphone-based scanning of Wenninger et al. [18].

Quantitative Comparisons: For quantitative comparisons, we scanned 33 subjects and created personalized avatars using our smartphone application running on an iPhone 13 Pro Max and using the expert body scanner of the Embodiment Lab at the University of Würzburg (see Section IV). The scanner consists of a circular rig equipped with 106 Canon EOS 1300D DSLR cameras and uses the approach of Achenbach et al. [14]. Scans were performed by an expert system operator who guided the scanned participant. The work of Bartl et al. [81] provides a detailed description of the whole system. We compute the reprojection error, similar to Wenninger et al. [18], by first rendering the resulting textured avatar onto the images captured during the scan process (see Figure 8b). Second, the root-mean-square errors are computed on all rendered pixels per image in CIELab color space. Third, the errors are averaged over all images, resulting in a metric that allows us to measure the error resulting from inaccuracies in color and geometry. Note that we do not have access to intrinsic camera calibration. Therefore, we use Agisoft Metashape to compute extrinsic and intrinsic camera calibration and use those to reproject our avatars. Figure 8a shows the resulting reprojection errors. In most cases, the results obtained with the complex camera rig are slightly better than the errors using

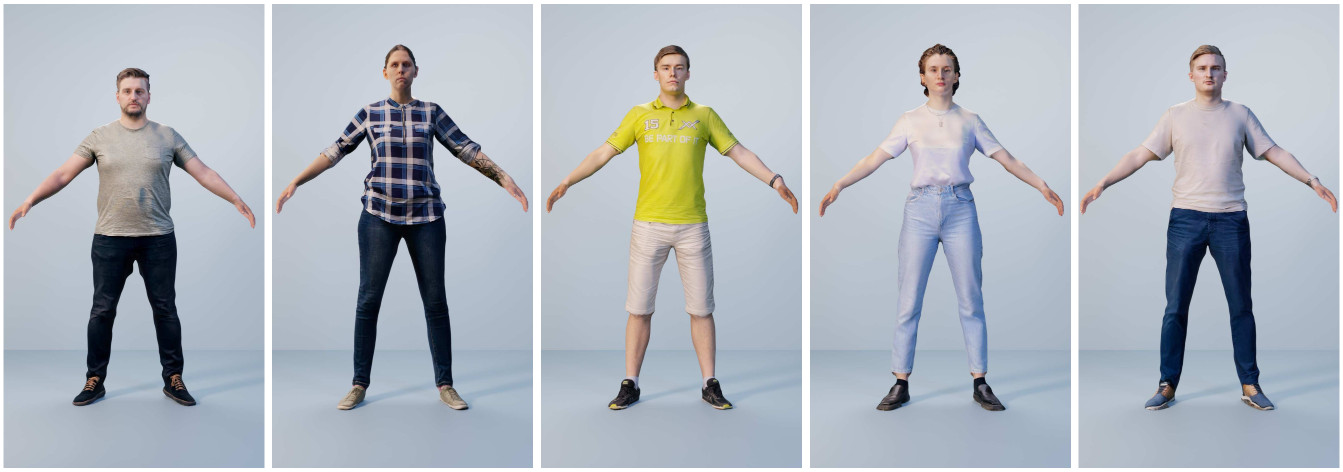
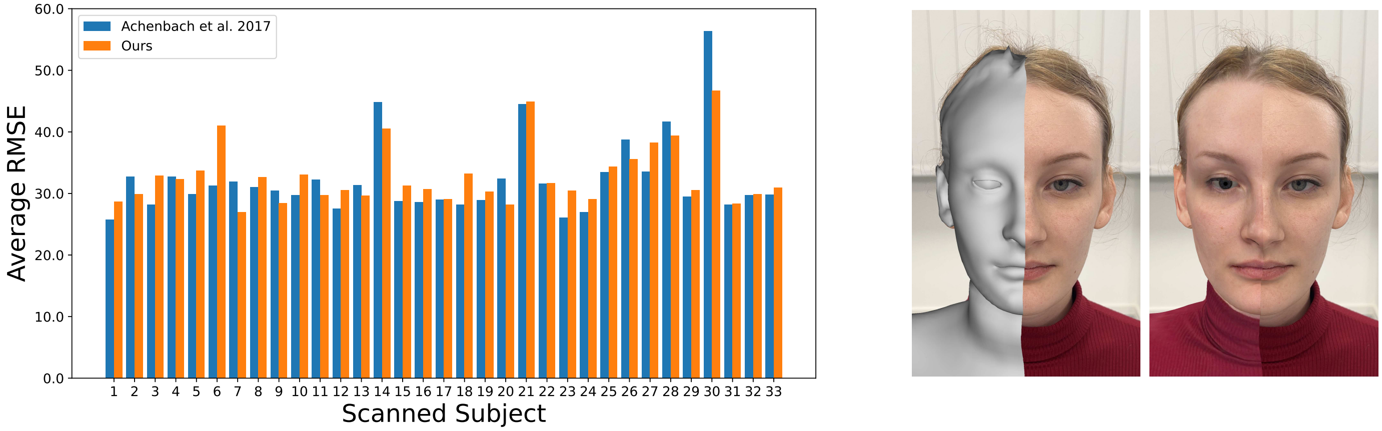


Fig. 7. Exemplary results of our avatar reconstruction pipeline. All subjects were scanned in an uncontrolled outdoor setting.



(a)

(b)

Fig. 8. (a): Root-mean-square reprojection errors of [14] and our method. (b): Reprojection of a reconstructed avatar (left half of images) onto an input image (right half of images).

our method. Although our method requires only a smartphone camera instead of an expensive camera rig, our overall average RMSE (μ = 32.83, σ = 4.88) is quite similar to the multi-camera rig reconstruction (μ = 32.29, σ = 6.36).

Qualitative Comparison with Wenninger et al.: Comparisons with Wenninger et al. [18] are shown in Figure 9. The input images for our method were captured using our custom smartphone application running on an iPhone 12 Pro, and the input videos for the method of Wenninger et al. were recorded with the same iPhone using the system camera app. The left column shows reconstructions in an indoor setting, and the right column shows reconstructions in an outdoor setting. Our avatar pipeline produces similar results regardless of the scanning location. The main difference between our two reconstructions is the baked-in lighting in the texture that occurs due to the different lighting conditions.

The reconstructions of Wenninger et al. look similar to the real person, but there are greater differences in the shape of the head and head-to-body proportions. The latter being even more prevalent in the outdoor setting. Comparing the results of Wenninger et al. to our method indicates that our pipeline produces more detailed geometry and a more detailed texture. This is due to the higher resolution input data and the improvements in photogrammetry and fitting steps. The differences in geometry are especially visible in hair regions, shoulders, and the overall shape of the face. The more detailed geometry is a result of the higher-resolution input data (see Section III-A1) and preprocessing (see Section III-A2), which leads to better photogrammetry results. Therefore, it is possible to fit the avatar more accurately. The better head-to-body proportions in our results are due to the alignment of the head scan to the body scan, as this allows the proportions to be respected via a simultaneous fitting of the point clouds.

IV. User Study

We conducted a user study following a multi-method approach to comprehensively evaluate the developed smartphone app along with the generated avatars (in the following called

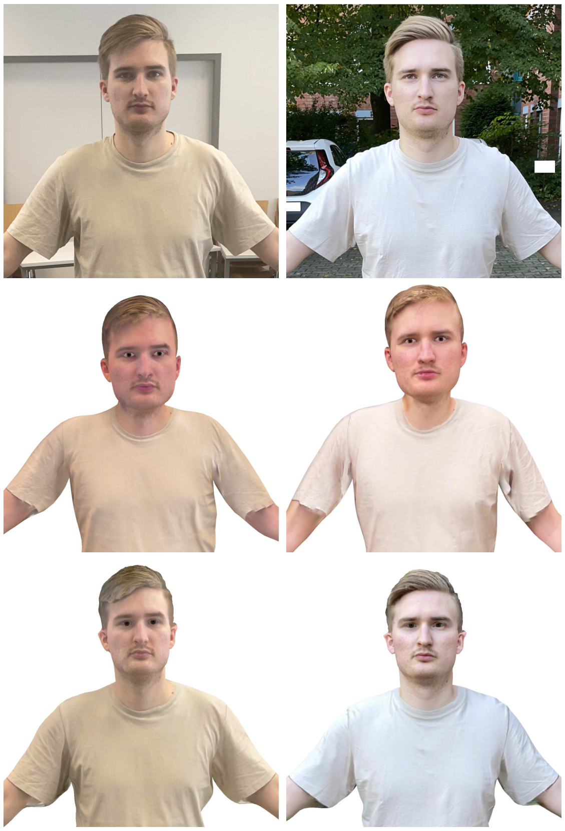


Fig. 9. Comparisons with the avatar reconstruction of Wenninger et al. [18] (middle) and our reconstructions (bottom) in an indoor setting (left) and an outdoor setting (right). Input images captured with our smartphone application (top).

smartphone avatars). The purpose was to improve the user experience of scanning and being scanned with the smartphone app and to assess the quality of the avatars subjectively. To this end, we arranged 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 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 *expert avatar*, see Section III-B) and (b) gender- and ethnicity-matched generic avatars. 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.

A. Apparatus

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

a) Smartphone Avatars: 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 expert 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.

b) Expert Avatars: We created a personalized expert avatar for each participant in the avatar evaluation using the expert body scanner of the Embodiment Lab at the University of Würzburg (see Section III-B). No further post-processing was performed on the expert avatars.

c) Generic Avatars: Since avatars not matching the user's gender and ethnicity have been shown to impact SoE particularly negatively [82] 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) [83]. Through a LimeSurvey questionnaire, each participant in the avatar evaluation was asked to select the VALID avatar that matched their own gender and ethnicity most. As the participants typically attend studies dressed casually, they could choose between 42 casually dressed VALID avatars, composed of three male and three female avatars, each from seven different ethnicities.

2) Virtual Reality System: The VR system was realized using Unity 2020.3.25f1 [84]. We utilized a Valve Index head-mounted display (HMD) featuring a resolution of $1440 \times 1600\,\text{px}$ per eye and a total field of view of $114.1 \times 109.4°$ [85]. 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 $3 \times 3\,\text{m}$ tracking area. All mentioned components were integrated into the VR system using SteamVR version 2.3 [86] and its corresponding Unity plug-in version 2.7.3 [87]. 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 [88] on Ubuntu 18 LTS. The body poses were continuously integrated into the VR system using Captury's corresponding Unity plug-in [89].

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 [9], [90]. During a short calibration process, where 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.

4) Virtual Environment and Tasks: Our virtual environment was based on different Unity assets, which we adapted to create a realistically rendered setting. Figure 10 depicts the virtual environment, accommodating up to three virtual mirrors. Following the guidelines for self-observation mirror placement by Wolf et al. [85], each virtual mirror was positioned at a distance of 1.5 m to the participant during the study.

a) One-By-One Exposure: 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 [50], [91]. The body movement tasks adhered to a structured protocol adapted from Roth and Latoschik [92] and can be found in the supplements of this work.

b) Side-By-Side Exposure: 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 an individual 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 10 depicts the side-by-side exposure.

B. Measures

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 translations to translate items into German. Participants answered all questionnaires on a Mac Book Pro using LimeSurvey [93].

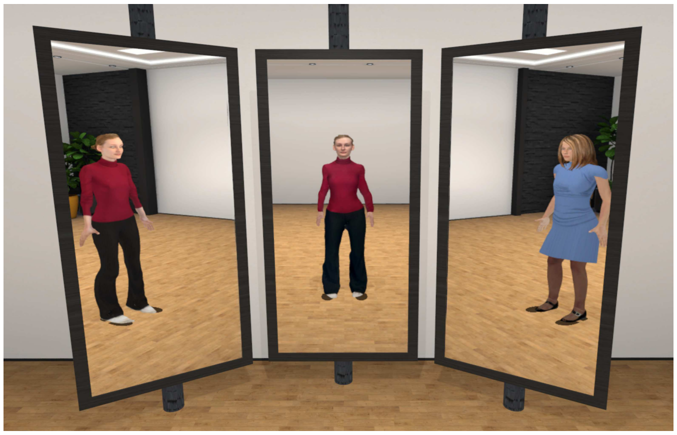


Fig. 10. The three mirrors showing the expert (left), smartphone (middle), and generic (right) avatar of a female participant during the side-by-side exposure.

a) Usability: We captured the usability of the smartphone app using the System Usability Scale (SUS) [94]. 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 [95], [96], [97].

b) Sense of Embodiment and Self-Identification: For assessing SoE towards the avatars, we captured virtual body ownership (VBO) and agency (AG) utilizing the corresponding items of the Virtual Embodiment Questionnaire (VEQ) [92] and self-location (SL) using the additional items introduced by Fiedler et al. (VEQ+) [62]. For assessing self-identification towards the avatars, we used the items capturing self-similarity (SS) and self-attribution (SA) from the VEQ+. Each measured factor comprises four items rated on a 7-point Likert scale (*7 = highest VBO, AG, SL, SS, and SA*).

c) Plausibility: We captured the avatars' plausibility utilizing the Virtual Human Plausibility Questionnaire (VHPQ) [98], [99]. 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*).

d) Uncanny Valley: We captured tendencies of the avatars' appearance towards the uncanny valley using the revised version of the Uncanny Valley Index (UVI) [100]. 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 it on a range between 1 and 7 (*7 = highest HU, AT, EE*).

e) VR Sickness: 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) [101]. 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*).

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 [102], [103] 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 the 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 enhancements to both the functionality of the scan app 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 in terms of 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.

C. Procedures

In the following, we describe the standardized experimental procedures of our smartphone app and avatar evaluations. Figure 11 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 separately store personal (i.e., voice recordings and avatars) and evaluation data. Subsequently, they proceeded with their respective evaluation procedures.

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 evaluateable avatar was generated, the scanning participant carried out two scans successively. 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.

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 displayed 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.

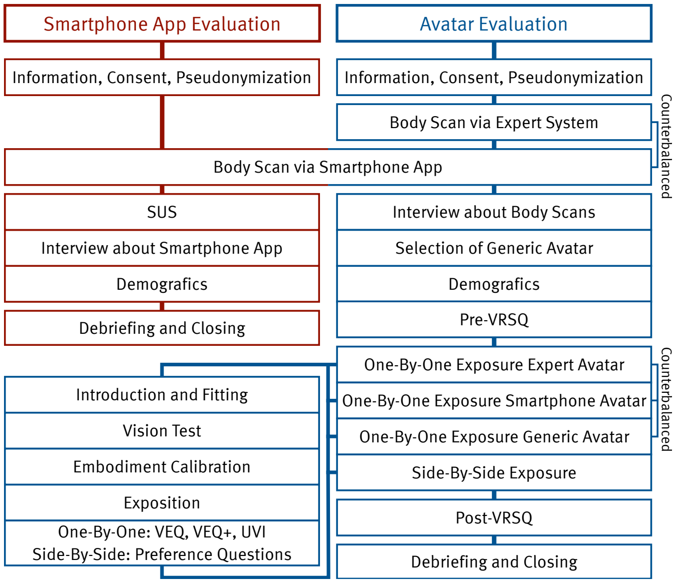

Smartphone App Evaluation
Avatar Evaluation
Information, Consent, Pseudonymization
Information, Consent, Pseudonymization
Body Scan via Expert System
Body Scan via Smartphone App
Counterbalanced
SUS
Interview about Smartphone App
Demografics
Debriefing and Closing
Interview about Body Scans
Selection of Generic Avatar
Demografics
Pre-VRSQ
Introduction and Fitting
Vision Test
Embodiment Calibration
Exposition
One-By-One: VEQ, VEQ+, UVI
Side-By-Side: Preference Questions
One-By-One Exposure Expert Avatar
One-By-One Exposure Smartphone Avatar
One-By-One Exposure Generic Avatar
Side-By-Side Exposure
Counterbalanced
Post-VRSQ
Debriefing and Closing

Fig. 11. Experimental procedure of a dyad, illustrating the process of evaluating the smartphone app (left) and the avatars (right).

D. 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 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

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

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.

E. Data Analysis

We conducted all quantitative analyses using SPSS version 29.0.2.0 [104]. 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 $N \geq 30$ [105], 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 [106]. Due to space restrictions, we decided to report the results mainly based on the frequency of certain feedback while mostly refraining from direct quotes.

F. Results

1) Smartphone App Evaluation: The quantitative evaluation of the smartphone app's usability resulted in a reasonably high SUS score ($M = 78.83, SD = 12.23$). As we merely evaluated the first version of the app without a comparative condition, we compared the results to absolute benchmarks from existing literature. According to Sauro and Lewis [96], 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. [95], 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 [97], 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 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 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.

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 I. For all tests revealing significant differences between groups, we calculated Bonferroni-corrected pairwise post-hoc comparisons that are reported in Figure 12.

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 expert avatars. Regarding self-representation fidelity, 11 participants preferred the smartphone avatars, 19 chose the expert avatars, and one favored the generic one. To feel most plausibly represented in VR, 12 participants chose the smartphone avatars, 18 the expert avatars, and one the generic one. When asked which

TABLE I
EXACT DESCRIPTIVE VALUES FOR EACH MEASURE OF THE AVATAR EVALUATION PER GROUP AND STATISTICAL RESULTS OF THE GROUP COMPARISONS.

	Smartphone Avatars	Expert Avatars	Generic Avatars	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)	$F(2,60) = 3.264, p = .045, \eta_p^2 = .098$

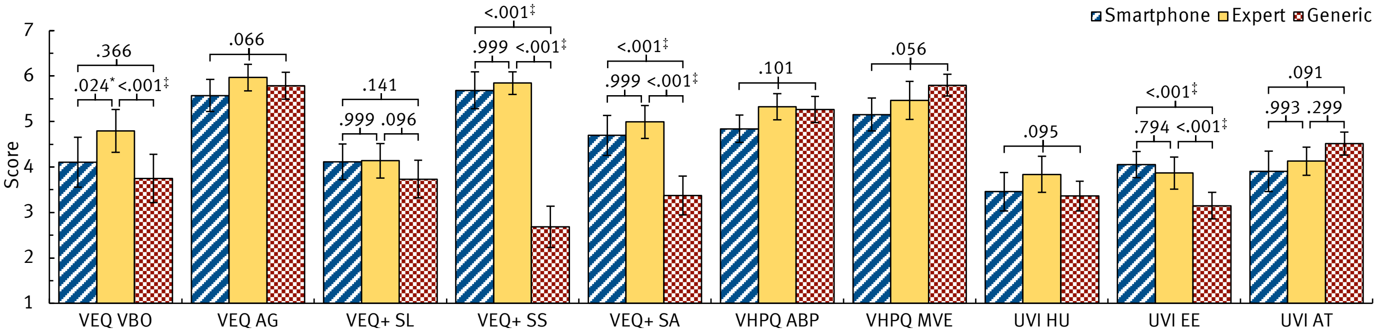


Fig. 12. Bar charts for each measure and each group of the avatar evaluation, including statistical test results of the group comparisons and post-hoc tests where applicable. Error bars represent 95 % confidence intervals. Statistical significance indicators: $^* p < .05$; $^\dagger p < .01$; $^\ddagger p < .001$.

avatar the participants would prefer to be represented in VR, 10 chose the smartphone avatars, 17 the expert 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 expert 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.

V. Discussion

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

A. 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 within technical means. Unclear parts in the tutorial have been improved to prepare users better for the scan. 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 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.

B. Avatar Evaluation

As described in Section III-B, the expert-operated system of Achenbach et al. [14] and our proposed novice-operated system produced photorealistic avatars of similar quality. The calculated reprojection errors were comparably low. These results highlight the potential and advantages of our method, enabling non-expert users to generate photorealistic, personalized avatars without requiring expensive hardware. The comparison with the method of Wenninger et al. [18] further showed that our introduced WILDAVATARS system outperforms existing smartphone reconstruction solutions, even when dealing with more complex input data due to less strict restrictions on scanning locations and subject appearances. The result is a more convincing avatar reconstruction, with more detailed geometry and better texture quality.

In comparison to the work of Waltemate et al.[10], 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 [83], [82]. Regarding the comparison, some further notable findings need to be addressed. The statistically significant difference in virtual body ownership between the smartphone and expert avatars can potentially be attributed to observed motion artifacts, which can degrade the avatars' appearance. However, the smartphone avatars perform descriptively still better than the generic avatars. Regarding self-identification, the smartphone and expert avatars both show significant advantages to generic avatars, although the smartphone avatars were generated using a significantly cheaper method than the expert 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 [65], [66]. 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 [63].

C. Limitations

Our study and our system have some limitations that we would like to describe in the following:

Motion Artifacts: 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.

Mesh-Based Representation: We use a mesh-based model for our avatars. Therefore, clothing, hair, and skin are all represented in the same surface, which can create wrong impressions. To compensate for this, one could incorporate Gaussian Splatting techniques to create a more realistic impression [31], [44], [45].

Crowded Background: 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 discard people in the background from the masks.

VI. CONCLUSION

We presented WILDAVATARS, a system that allows non-expert users to scan people and automatically reconstruct realistic VR-ready full-body avatars that achieve similarly good perception results compared to avatars reconstructed with expensive state-of-the-art expert systems. We proposed methods to reduce restrictions and limitations on scanning locations and provide helpful feedback by visually guiding users through the scanning process. Our system will be publicly available for research purposes, enabling the realization of avatar-related studies. Since the system was designed using a user-centered process, it is easily understandable and allows non-experts to use photorealistic avatars without difficult-to-use equipment.

REFERENCES

[1] J. N. Bailenson and J. Blascovich, "Avatars," in *Encyclopedia of Human-Computer Interaction*. Great Barrington, MA, USA: Berkshire Publishing Group, 2004, pp. 64–68.

[2] R. Skarbez and D. Jiang, "A Scientometric History of IEEE VR," in *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, 2024, pp. 990–999.

[3] I. E. Sutherland, "A head-mounted three-dimensional display," in *Proceedings of the December 9-11, 1968, Fall Joint Computer Conference, Part I*, 1968, pp. 757–764.

[4] M. E. Latoschik, F. Kern, J.-P. Stauffert, A. Bartl, M. Botsch, and J.-L. Lugrin, "Not alone here?! scalability and user experience of embodied ambient crowds in distributed social virtual reality," *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, vol. 25, no. 5, pp. 2134–2144, 2019.

[5] B. Yoon, H.-i. Kim, G. A. Lee, M. Billinghurst, and W. Woo, "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)*, 2019, pp. 547–556.

[6] S. Aseeri and V. Interrante, "The Influence of Avatar Representation on Interpersonal Communication in Virtual Social Environments," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 5, pp. 2608–2617, 2021.

[7] S. Mystakidis, "Metaverse," *Encyclopedia*, vol. 2, no. 1, pp. 486–497, 2022.

[8] M. Sampaio, M. V. Navarro Haro, B. De Sousa, W. Vieira Melo, and H. G. Hoffman, "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*, vol. 1, 2021.

[9] N. Döllinger, E. Wolf, D. Mal, S. Wenninger, M. Botsch, M. E. Latoschik, and C. Wienrich, "Resize Me! Exploring the User Experience of Embodied Realistic Modulatable Avatars for Body Image Intervention in Virtual Reality," *Frontiers in Virtual Reality*, vol. 3, 2022.

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

[11] A. Salagean, E. Crellin, M. Parsons, D. Cosker, and D. Stanton Fraser, "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*, 2023.

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

[13] D. Y. Kim, H. K. Lee, and K. Chung, "Avatar-mediated experience in the metaverse: The impact of avatar realism on user-avatar relationship," *Journal of Retailing and Consumer Services*, vol. 73, 2023.

[14] J. Achenbach, T. Waltemate, M. E. Latoschik, and M. Botsch, "Fast Generation of Realistic Virtual Humans," in *Proceedings of the ACM Symposium on Virtual Reality Software and Technology*, 2017.

[15] T. Jiang, X. Chen, J. Song, and O. Hilliges, "InstantAvatar: Learning Avatars From Monocular Video in 60 Seconds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 16 922–16 932.

[16] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll, "Detailed Human Avatars from Monocular Video," in *International Conference on 3D Vision (3DV)*, 2018, pp. 98–109.

[17] ——, "Video Based Reconstruction of 3D People Models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[18] S. Wenninger, J. Achenbach, A. Bartl, M. E. Latoschik, and M. Botsch, "Realistic Virtual Humans from Smartphone Videos," in *Proceedings of the ACM Symposium on Virtual Reality Software and Technology*, 2020.

[19] S. Ma, T. Simon, J. Saragih, D. Wang, Y. Li, F. L. Torre, and Y. Sheikh, "Pixel Codec Avatars," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 64–73.

[20] A. E. Ichim, S. Bouaziz, and M. Pauly, "Dynamic 3D Avatar Creation from Hand-Held Video Input," *ACM Trans. Graph.*, vol. 34, no. 4, 2015.

[21] P. Grassal, M. Prinzler, T. Leistner, C. Rother, M. Niebner, and J. Thies, "Neural Head Avatars from Monocular RGB Videos," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[22] T. Khakhulin, V. Sklyarova, V. Lempitsky, and E. Zakharov, "Realistic one-shot mesh-based head avatars," in *Computer Vision – ECCV 2022*, Cham, 2022, pp. 345–362.

[23] P. Caselles, E. Ramon, J. Garcia, X. G. i Nieto, F. Moreno-Noguer, and G. Triginer, "SIRA: Relightable Avatars from a Single Image," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023, pp. 775–784.

[24] W. Zielonka, T. Bolkart, and J. Thies, "Instant Volumetric Head Avatars," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 4574–4584.

[25] A. Feng, E. Suma, and A. Shapiro, "Just-in-Time, Viable, 3D Avatars from Scans," in *ACM SIGGRAPH 2017 Talks*, 2017.

[26] W. Morgenstern, M. T. Bagdasarian, A. Hilsmann, and P. Eisert, "Animatable virtual humans: Learning pose-dependent human representations in uv space for interactive performance synthesis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, no. 5, pp. 2644–2650, 2024.

[27] A. Shetty, M. Habermann, G. Sun, D. Luvizon, V. Golyanik, and C. Theobalt, "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)*, 2024, pp. 1206–1215.

[28] Y. Kwon, L. Liu, H. Fuchs, M. Habermann, and C. Theobalt, "DELIFFAS: Deformable Light Fields for Fast Avatar Synthesis," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 40 944–40 962.

[29] T. Alldieck, M. Magnor, B. L. Bhatnagar, C. Theobalt, and G. Pons-Moll, "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)*, 2019.

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

[31] J. Wen, X. Zhao, Z. Ren, A. G. Schwing, and S. Wang, "GoMAvatar: Efficient Animatable Human Modeling from Monocular Video Using Gaussians-on-Mesh," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 2059–2069.

[32] S. Hu, T. Hu, and Z. Liu, "GauHuman: Articulated Gaussian Splatting from Monocular Human Videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 20 418–20 431.

[33] C. Cao, T. Simon, J. K. Kim, G. Schwartz, M. Zollhoefer, S.-S. Saito, S. Lombardi, S.-E. Wei, D. Belko, S.-I. Yu, Y. Sheikh, and J. Saragih, "Authentic Volumetric Avatars from a Phone Scan," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 4, 2022.

[34] T. Liao, X. Zhang, Y. Xiu, H. Yi, X. Liu, G. Qi, Y. Zhang, X. Wang, X. Zhu, and Z. Lei, "High-Fidelity Clothed Avatar Reconstruction from a Single Image," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 8662–8672.

[35] Z. Jiang, C. Guo, M. Kaufmann, T. Jiang, J. Valentin, O. Hilliges, and J. Song, "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)*, 2024, pp. 109–118.

[36] C. Guo, T. Jiang, X. Chen, J. Song, and O. Hilliges, "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)*, 2023, pp. 12 858–12 868.

[37] J. Xiao, Q. Zhang, Z. Xu, and W.-S. Zheng, "NECA: Neural Customizable Human Avatar," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 20 091–20 101.

[38] W. Lin, C. Zheng, J.-H. Yong, and F. Xu, "Relightable and Animatable Neural Avatars from Videos," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 4, pp. 3486–3494, 2024.

[39] T. Wang, B. Zhang, T. Zhang, S. Gu, J. Bao, T. Baltrusaitis, J. Shen, D. Chen, F. Wen, Q. Chen, and B. Guo, "RODIN: A Generative Model for Sculpting 3D Digital Avatars Using Diffusion," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 4563–4573.

[40] W. Yu, Y. Fan, Y. Zhang, X. Wang, F. Yin, Y. Bai, Y.-P. Cao, Y. Shan, Y. Wu, Z. Sun, and B. Wu, "NOFA: NeRF-Based One-Shot Facial Avatar Reconstruction," in *ACM SIGGRAPH 2023 Conference Proceedings*, 2023.

[41] S. Wang, B. Antic, A. Geiger, and S. Tang, "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)*, 2024, pp. 1877–1888.

[42] Z. Zheng, H. Huang, T. Yu, H. Zhang, Y. Guo, and Y. Liu, "Structured Local Radiance Fields for Human Avatar Modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 15 893–15 903.

[43] Z. Zheng, X. Zhao, H. Zhang, B. Liu, and Y. Liu, "AvatarReX: Real-time Expressive Full-body Avatars," *ACM Trans. Graph.*, vol. 42, no. 4, 2023.

[44] L. Hu, H. Zhang, Y. Zhang, B. Zhou, B. Liu, S. Zhang, and L. Nie, "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)*, 2024, pp. 634–644.

[45] Z. Shao, Z. Wang, Z. Li, D. Wang, X. Lin, Y. Zhang, M. Fan, and Z. Wang, "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)*, 2024, pp. 1606–1616.

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

[47] Z. Li, Z. Zheng, L. Wang, and Y. Liu, "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)*, 2024, pp. 19 711–19 722.

[48] M. Habermann, L. Liu, W. Xu, G. Pons-Moll, M. Zollhoefer, and C. Theobalt, "HDHumans: A Hybrid Approach for High-fidelity Digital Humans," *Proc. ACM Comput. Graph. Interact. Tech.*, vol. 6, no. 3, 2023.

[49] J. Wang, R. Shi, W. Zheng, W. Xie, D. Kao, and H.-N. Liang, "Effect of Frame Rate on User Experience, Performance, and Simulator Sickness in Virtual Reality," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 5, pp. 2478–2488, 2023.

[50] M. Slater, B. Spanlang, M. V. Sanchez-Vives, and O. Blanke, "First person experience of body transfer in virtual reality," *PLOS ONE*, vol. 5, no. 5, p. e10564, 2010.

[51] A. Mottelson, A. Muresan, K. Hornbæk, and G. Makransky, “A systematic review and meta-analysis of the effectiveness of body ownership illusions in virtual reality,” *ACM Trans. Comput.-Hum. Interact.*, 2023.

[52] E. Wolf, N. Merdan, N. Döllinger, D. Mal, C. Wienrich, M. Botsch, and M. E. Latoschik, “The embodiment of photorealistic avatars influences female body weight perception in virtual reality,” in *IEEE Virtual Reality and 3D User Interfaces (VR)*, 2021, pp. 65–74.

[53] R. Skarbez, J. Frederick P. Brooks, and M. C. Whitton, “A survey of presence and related concepts,” *ACM Computing Surveys*, vol. 50, no. 6, p. 96, 2017.

[54] D. Gall, D. Roth, J.-P. Stauffert, J. Zarges, and M. E. Latoschik, “Embodiment in virtual reality intensifies emotional responses to virtual stimuli,” *Frontiers in Psychology*, vol. 12, p. 3833, 2021.

[55] B. J. Mohler, S. H. Creem-Regehr, W. B. Thompson, and H. H. Bülthoff, “The effect of viewing a self-avatar on distance judgments in an HMD-based virtual environment,” *Presence*, vol. 19, no. 3, pp. 230–242, 2010.

[56] M. Leyrer, S. A. Linkenauger, H. H. Bülthoff, U. Kloos, and B. Mohler, “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*, 2011, pp. 67–74.

[57] A. Steed, Y. Pan, F. Zisch, and W. Steptoe, “The impact of a self-avatar on cognitive load in immersive virtual reality,” in *2016 IEEE Virtual Reality (VR)*, 2016, pp. 67–76.

[58] S. Jung and C. E. Hughes, “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*, 2016, pp. 107–114.

[59] S. Pastel, C.-H. Chen, K. Petri, and K. Witte, “Effects of body visualization on performance in head-mounted display virtual reality,” *PLOS ONE*, vol. 15, no. 9, pp. 1–18, 2020.

[60] K. Kilteni, R. Groten, and M. Slater, “The sense of embodiment in virtual reality,” *Presence: Teleoperators & Virtual Environments*, vol. 21, no. 4, pp. 373–387, 2012.

[61] F. de Vignemont, “Embodiment, ownership and disownership,” *Consciousness and Cognition*, vol. 20, no. 1, pp. 82–93, 2011.

[62] M. L. Fiedler, E. Wolf, N. Döllinger, M. Botsch, M. E. Latoschik, and C. Wienrich, “Embodiment and personalization for self-identification with virtual humans,” in *IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 2023, pp. 799–800.

[63] M. E. Latoschik and C. Wienrich, “Congruence and plausibility, not presence: Pivotal conditions for XR experiences and effects, a novel approach,” *Frontiers in Virtual Reality*, vol. 3, 2022.

[64] C. Turbyne, A. Goedhart, P. de Koning, F. Schirmbeck, and D. Denys, “Systematic Review and Meta-Analysis of Virtual Reality in Mental Healthcare: Effects of Full Body Illusions on Body Image Disturbance,” *Frontiers in Virtual Reality*, vol. 2, p. 39, 2021.

[65] M. Mori, K. F. MacDorman, and N. Kageki, “The uncanny valley [from the field],” *IEEE Robotics Automation Magazine*, vol. 19, no. 2, pp. 98–100, 2012.

[66] N. Döllinger, M. Beck, E. Wolf, D. Mal, M. Botsch, M. E. Latoschik, and C. Wienrich, “”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)*, 2023.

[67] Agisoft, “Agisoft Metashape,” 2023, [visited on 2023-09-29]. [Online]. Available: https://www.agisoft.com

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

[69] T. Wiegand, G. Sullivan, G. Bjontegaard, and A. Luthra, “Overview of the H.264/AVC video coding standard,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, 2003.

[70] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” 2017.

[71] Apple Inc., “Vision Framework - VNGeneratePersonSegmentationRequest,” 2023, [visited on 2024-07-05]. [Online]. Available: https://developer.apple.com/documentation/vision/vngeneratepersonsegmentationrequest

[72] ——, “RealityKit Framework - Object Capture - PhotogrammetrySession,” 2023, [visited on 2024-07-05]. [Online]. Available: https://developer.apple.com/documentation/realitykit/photogrammetrysession

[73] Dlib, “Dlib C++ Library,” 2022. [Online]. Available: https://dlib.net

[74] Apple Inc., “Vision Framework - VNFaceLandmarks2D,” 2023, [visited on 2024-07-05]. [Online]. Available: https://developer.apple.com/documentation/vision/vnfacelandmarks2d

[75] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, “MediaPipe: A Framework for Building Perception Pipelines,” 2019.

[76] Apple Inc., “ARKit Framework - Blendshapes,” 2023, [visited on 2023-09-27]. [Online]. Available: https://developer.apple.com/documentation/arkit/arfaceanchor/2928251-blendshapes

[77] T. Magnenat, R. Laperriere, and D. Thalmann, “Joint-dependent local deformations for hand animation and object grasping,” 1988, p. 26–33.

[78] J. Achenbach, E. Zell, and M. Botsch, “Accurate Face Reconstruction through Anisotropic Fitting and Eye Correction,” in *Vision, Modeling & Visualization*, D. Bommes, T. Ritschel, and T. Schultz, Eds., 2015.

[79] B. K. P. Horn, “Closed-Form Solution of Absolute Orientation Using Unit Quaternions,” *Journal of the Optical Society of America A*, vol. 4, no. 4, pp. 629–642, 1987.

[80] A. Aristidou, J. Lasenby, Y. Chrysanthou, and A. Shamir, “Inverse Kinematics Techniques in Computer Graphics: A Survey,” *Computer Graphics Forum*, vol. 37, no. 6, pp. 35–58, 2018.

[81] A. Bartl, S. Wenninger, E. Wolf, M. Botsch, and M. E. Latoschik, “Affordable but not Cheap: A Case Study of the Effects of Two 3D-Reconstruction Methods of Virtual Humans,” *Frontiers in Virtual Reality*, vol. 2, 2021.

[82] T. D. Do, C. Isabella Protko, and R. P. McMahan, “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*, pp. 1–10, 2024.

[83] T. D. Do, S. Zelenty, M. Gonzalez-Franco, and R. P. McMahan, “VALID: A perceptually validated virtual avatar library for inclusion and diversity,” *Frontiers in Virtual Reality*, vol. 4, 2023.

[84] Unity Technologies, “Unity 2020.3.25f1,” 2020. [Online]. Available: https://unity.com/

[85] E. Wolf, N. Döllinger, D. Mal, S. Wenninger, B. Andrea, M. Botsch, M. E. Latoschik, and C. Wienrich, “Does Distance Matter? Embodiment and Perception of Personalized Avatars in Relation to the Self-Observation Distance in Virtual Reality,” *Frontiers in Virtual Reality*, vol. 3, 2022.

[86] Valve Corporation, “Steam VR 2.3,” 2024. [Online]. Available: https://store.steampowered.com/app/250820/SteamVR/

[87] ——, “Steam VR Plugin 2.7.3,” 2024. [Online]. Available: https://assetstore.unity.com/packages/tools/integration/steamvr-plugin-32647

[88] Captury, “CapturyLive 259,” 2023. [Online]. Available: https://captury.com/real-time-processing/

[89] ——, “Unity plugin,” 2023. [Online]. Available: https://captury.com/resources/

[90] E. Wolf, M. L. Fiedler, N. Döllinger, C. Wienrich, and M. E. Latoschik, “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)*, 2022, pp. 350–359.

[91] M. González-Franco, D. Pérez-Marcos, B. Spanlang, and M. Slater, “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)*, 2010, pp. 111–114.

[92] D. Roth and M. E. Latoschik, “Construction of the virtual embodiment questionnaire (VEQ),” *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 12, pp. 3546–3556, 2020.

[93] Limesurvey GmbH, “LimeSurvey: An Open Source survey tool,” LimeSurvey GmbH, Hamburg, Germany, 2024. [Online]. Available: https://www.limesurvey.org

[94] J. Brooke, “SUS: A quick and dirty usability scale,” in *Usability evaluation in industry*. Taylor & Francis, 1996, pp. 4–7.

[95] A. Bangor, P. Kortum, and J. Miller, “Determining what individual SUS scores mean: Adding an adjective rating scale,” *Journal of usability studies*, vol. 4, no. 3, pp. 114–123, 2009.

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

[97] P. Kortum and M. Sorber, “Measuring the Usability of Mobile Applications for Phones and Tablets,” *International Journal of Human–Computer Interaction*, vol. 31, no. 8, pp. 518–529, 2015.

[98] D. Mal, E. Wolf, N. Döllinger, M. Botsch, C. Wienrich, and M. E. Latoschik, “Virtual Human Coherence and Plausibility – Towards a Validated Scale,” in *IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 2022, pp. 788–789.

[99] ——, "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*, 2024, p. 8.
[100] C.-C. Ho and K. F. MacDorman, "Measuring the uncanny valley effect," *International Journal of Social Robotics*, vol. 9, no. 1, pp. 129–139, 2017.
[101] H. K. Kim, J. Park, Y. Choi, and M. Choe, "Virtual reality sickness questionnaire (VRSQ): Motion sickness measurement index in a virtual reality environment," *Applied Ergonomics*, vol. 69, pp. 66–73, 2018.
[102] V. A. Bowers and H. L. Snyder, "Concurrent versus Retrospective Verbal Protocol for Comparing Window Usability," *Proceedings of the Human Factors Society Annual Meeting*, vol. 34, no. 17, pp. 1270–1274, 1990.
[103] H. A. Simon and K. A. Ericsson, *Protocol analysis: Verbal reports as data*. The MIT Press, 1993.
[104] IBM, "SPSS Statistics," https://www.ibm.com/products/spss-statistics, 2022.
[105] R. R. Wilcox, *Introduction to robust estimation and hypothesis testing*. Academic Press, 2022.
[106] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, 2006.

Timo Menzel is a PhD candidate and research associate at the Computer Graphics & Geometry Processing Group at TU Dortmund University, Germany. He finished degrees in Informatics for the Natural Sciences (B.Sc.) and Intelligent Systems (M.Sc.) at Bielefeld University in 2018 and 2020, respectively.

Erik Wolf is a PhD candidate at the Human-Computer Interaction (HCI) Group of the Julius-Maximilians-Universität Würzburg, Germany, and a research associate at the Human-Computer Interaction (HCI) Group of the Universität Hamburg, Germany. In 2022, he received the Meta PhD Research Fellowship in "AR/VR Future Technologies" for his PhD research centering on individual-, application-, and system-related factors influencing the perception of virtual humans in AR/VR. He finished degrees in Human-Computer-Systems (B.Sc.) and Human-Computer Interaction (M.Sc.) at the Julius-Maximilians-Universität Würzburg in 2017 and 2020, respectively.

Stephan Wenninger is a PhD candidate and research associate at the Computer Graphics & Geometry Processing Group at TU Dortmund University, Germany. He finished degrees in Cognitive Sciences (B.Sc.) at the University of Tübingen and Intelligent Systems (M.Sc.) at Bielefeld University in 2015 and 2018, respectively.

Niklas Spinczyk is a student and research assistant at TU Dortmund University, Germany. He received a degree in Computer Science (B.Sc.) in 2022.

Lena Holderrieth is a student and research assistant at Julius-Maximilians-Universität Würzburg, Germany. She received a degree in Media Informatics (B.Sc.) at Technische Hochschule Nürnberg Georg Simon Ohm in 2022.

Ulrich Schwanecke is full professor and head of the Computer Vision and Mixed Reality Group at RheinMain University of Applied Sciences. He received his Master's degree (Dipl.-Math.) in Mathematics and Computer Science from the Johannes Gutenberg University Mainz in 1997. From 1997 to 2000, he worked as a research assistant at Technische Universität Darmstadt, from where he received his PhD (Dr. rer. nat.) in 2000. Subsequently, he held a postdoctoral position for one year at the Max-Planck Institute for Computer Science in Saarbrücken. From 2001 to 2003, he worked as a researcher at Daimler AG in Ulm. Since October 2003, he is a full professor for Computer Graphics and Vision at RheinMain University of Applied Sciences in Wiesbaden.

Marc Erich Latoschik is full professor and head of the Human-Computer Interaction (HCI) Group at the University of Würzburg. He studied mathematics and computer science at the University of Paderborn, the New York Institute of Technology, and the Bielefeld University, where he received his PhD in multimodal gesture and speech interaction for virtual reality in 2001. Marc has an extensive background in the computer industry, but finally decided to devote all his time to research and joined academia in 1996. He headed the AI & VR Lab at Bielefeld University until 2007, became a professor for media informatics at the University of Applied Sciences (HTW) in Berlin, and founded the Intelligent Graphics Group at Bayreuth University in 2009 before he finally took over the HCI chair at Würzburg in 2011. His research is focussing on immersive interactive XR interfaces with a specific emphasise on combining Artificial Intelligence and Computer Graphics.

Mario Botsch is full professor in the Computer Science Department at TU Dortmund University. He is heading both the Chair of Computer Graphics and the Computer Graphics & Geometry Processing Group. He studied mathematics and received his M.Sc. from the University of Erlangen-Nürnberg (1999). After that, he started his PhD at the Max Planck Institute for Informatics, before moving to RWTH Aachen and finishing his PhD in computer science (2005). Afterwards, he was PostDoc at ETH Zurich (2005-2008). Before joining TU Dortmund in fall 2020, he was professor for Computer Graphics at Bielefeld University (2008-2020).