EgoAvatar: Egocentric View-Driven and Photorealistic Full-body Avatars

JIANCHUN CHEN, Max Planck Institute for Informatics, Germany and Saarbrücken Research Center for Visual Computing, Interaction and AI, Germany

JIAN WANG, Max Planck Institute for Informatics, Germany and Saarbrücken Research Center for Visual Computing, Interaction and AI, Germany

YINDA ZHANG, Google Inc., United States of America

ROHIT PANDEY, Google Inc., United States of America

THABO BEELER, Google Inc., Switzerland

MARC HABERMANN, Max Planck Institute for Informatics, Germany and Saarbrücken Research Center for Visual Computing, Interaction and AI, Germany

CHRISTIAN THEOBALT, Max Planck Institute for Informatics, Germany and Saarbrücken Research Center for Visual Computing, Interaction and AI, Germany
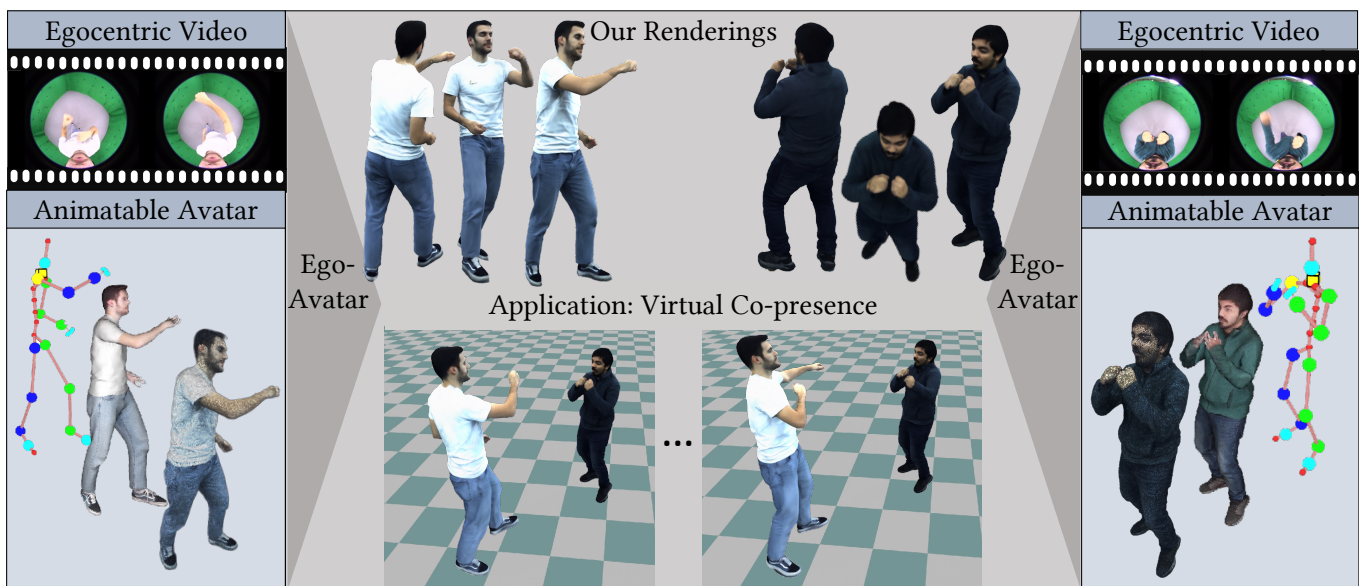
Fig. 1. We propose *EgoAvatar*, which takes an egocentric video stream capturing a real human in motion as input and sub-sequentially recovers the skeleton motion, explicit surface mesh, and Gaussian splats representing the geometry and appearance of the avatar. Once the parameters of our character model are recovered, *EgoAvatar* allows us to re-render the full-body avatar in a free viewpoint, which potentially enables a virtual and full-body co-presence.

Authors' Contact Information: Jianchun Chen, Max Planck Institute for Informatics, Germany and Saarbrücken Research Center for Visual Computing, Interaction and AI, Germany, jchen@mpi-inf.mpg.de; Jian Wang, Max Planck Institute for Informatics, Germany and Saarbrücken Research Center for Visual Computing, Interaction and AI, Germany, jianwang@mpi-inf.mpg.de; Yinda Zhang, Google Inc., United States of America, yindaz@google.com; Rohit Pandey, Google Inc., United States of America, rohitpandey@google.com; Thabo Beeler, Google Inc., Switzerland, tbeeler@google.com; Marc Habermann, Max Planck Institute for Informatics, Germany and Saarbrücken Research Center for Visual Computing, Interaction and AI, Germany, mhaberma@mpi-inf.mpg.de; Christian Theobalt, Max Planck Institute for Informatics, Germany and Saarbrücken Research Center for Visual Computing, Interaction and AI, Germany, theobalt@mpi-inf.mpg.de.

Immersive VR telepresence ideally means being able to interact and communicate with digital avatars that are indistinguishable from and precisely reflect the behaviour of their real counterparts. The core technical challenge is two fold: Creating a digital double that faithfully reflects the real human and tracking the real human solely from egocentric sensing devices that are lightweight and have a low energy consumption, e.g. a single RGB camera. Up to date, no unified solution to this problem exists as recent works solely focus on egocentric motion capture, only model the head, or build avatars from multi-view captures. In this work, we, for the first time in literature, propose a person-specific egocentric telepresence approach, which jointly models the photoreal digital avatar while also driving it from a single egocentric video. We first present a character model that is animatable, i.e. can be

solely driven by skeletal motion, while being capable of modeling geometry and appearance. Then, we introduce a personalized egocentric motion capture component, which recovers full-body motion from an egocentric video. Finally, we apply the recovered pose to our character model and perform a test-time mesh refinement such that the geometry faithfully projects onto the egocentric view. To validate our design choices, we propose a new and challenging benchmark, which provides paired egocentric and dense multi-view videos of real humans performing various motions. Our experiments demonstrate a clear step towards egocentric and photoreal telepresence as our method outperforms baselines as well as competing methods. For more details, code, and data, we refer to our project page.

1 Introduction

With recent developments in VR headset technology, wearable computing is becoming more and more a reality with great potential for applications like immersive and social telepresence that no longer relies on complicated, stationary, and spatially constrained external capture setups, e.g. multi-camera rigs. However, this comes with the challenge of faithfully reproducing, animating and re-rendering the real human in the virtual world from only limited body-mounted sensor data. More precisely, this requires 1) turning a real human into an animatable, full-body, and photoreal digital avatar and 2) faithfully tracking the persons motion from limited sensor data of the headset to drive the avatar.

Both of these points pose a significant research challenge and many recent works proposed promising solutions. Creating animatable and photoreal digital twins from real world measurements of a human, e.g. multi-view video [Bagautdinov et al. 2021; Habermann et al. 2023, 2021; Kwon et al. 2023; Liu et al. 2021; Luvizon et al. 2024; Pang et al. 2023; Xiang et al. 2022, 2021] or single view video [Jiang et al. 2022; Weng et al. 2022] has been extensively researched recently, achieving high-quality results while also enabling real-time performance [Zhu et al. 2023]. Follow-up works [Remelli et al. 2022; Shetty et al. 2023; Sun et al. 2024; Xiang et al. 2023] have shown how to drive such avatars from sparse signals like a few stationary and external cameras. Nevertheless, such a hardware setup constrains the person to remain within a fixed and very limited capture volume observed by the external cameras. In contrast, other works focused on egocentric motion capture [Akada et al. 2022; Cha et al. 2018; Kang et al. 2023; Rhodin et al. 2016; Wang et al. 2023a, 2022, 2021, 2023c] using head-mounted cameras while ignoring rendering functionality. However, none of the above works tackles the problem of driving a photoreal full body avatar from single egocentric video. Few works [Elgharib et al. 2020; Wei et al. 2019] focused on driving a head avatar from egocentric video. However, they cannot drive the entire human body, which is significantly more challenging due

to body articulation leading to large pose variation, diversity in clothing, and drastically different appearance properties, i.e. clothing material vs. skin reflectance properties. The most closely related work is EgoRenderer [Hu et al. 2021], which is the only one attempting to solve this problem, but their quality is far from photoreal and suffers from severe temporal jitter. In summary, none of the prior works can drive a *photorealistic full-body avatar from a single egocentric RGB camera feed*.

To address this, we present *EgoAvatar*, the first approach for driving and rendering a photoreal full-body avatar solely from a single egocentric RGB camera. Our proposed system enables operation in unconstrained environments while maintaining high-visual fidelity, thus, paving the way for immersive telepresence without requiring complicated (multi-view) camera setups during inference (see Fig. 1) freeing the user from spatial constraints. Here, the term *full-body* refers to modeling the entire clothed human body while we do not account for hand gestures and facial expression.

More precisely, given a multi-view video of the person, we first learn a drivable photorealistic avatar, i.e. a function that takes the skeletal motion in form of joint angles as input and predicts the avatar's motion- and view-dependent geometry and appearance. First, a geometry module, *MotionDeformer*, predicts the surface deformation with respect to a static template mesh given the skeletal motion. Subsequently, our appearance module, *GaussianPredictor*, takes the respective posed and deformed template geometry and learns 3D Gaussian parameters in the texture space parameterized by the template's UV atlas, where each texel is encoding a 3D Gaussian. During training, we splat the Gaussians into image space and supervise them with the multi-view video.

We then capture a second multi-view video of the subject, this time wearing the headset, from which we train a personalized egocentric view-driven skeleton keypoint detector, dubbed *EgoPoseDetector*, to extract the subject's pose in the form of 3D joint predictions. Our inverse kinematics (IK) module, called *IKSolver*, completes and refines those noisy and incomplete joint predictions to produce temporally stable joint angles of the virtual character. From those, the *MotionDeformer* provides initial estimate of the deformed surface, which is further refined with the proposed *EgoDeformer* to ensure a faithful reprojection into the egocentric view. Lastly, the optimized surface is used to predict the final appearance using our *GaussianPredictor*. In summary, our contributions are threefold:

- We propose *EgoAvatar*, the first full-body photoreal and egocentrically driven avatar approach, which, at inference, simply takes as input a monocular video stream from a head-mounted down-facing camera and faithfully re-creates realistic full-body appearance that can be re-rendered in a free-viewpoint video.
- We further introduce a carefully designed avatar representation, egocentric tracking pipeline, and an egocentric geometry refinement, which in conjunction outperform all design alternatives.
- We propose the first dataset, which provides paired 120 camera multi-view and monocular egocentric videos capturing the full body human at 4K resolution.
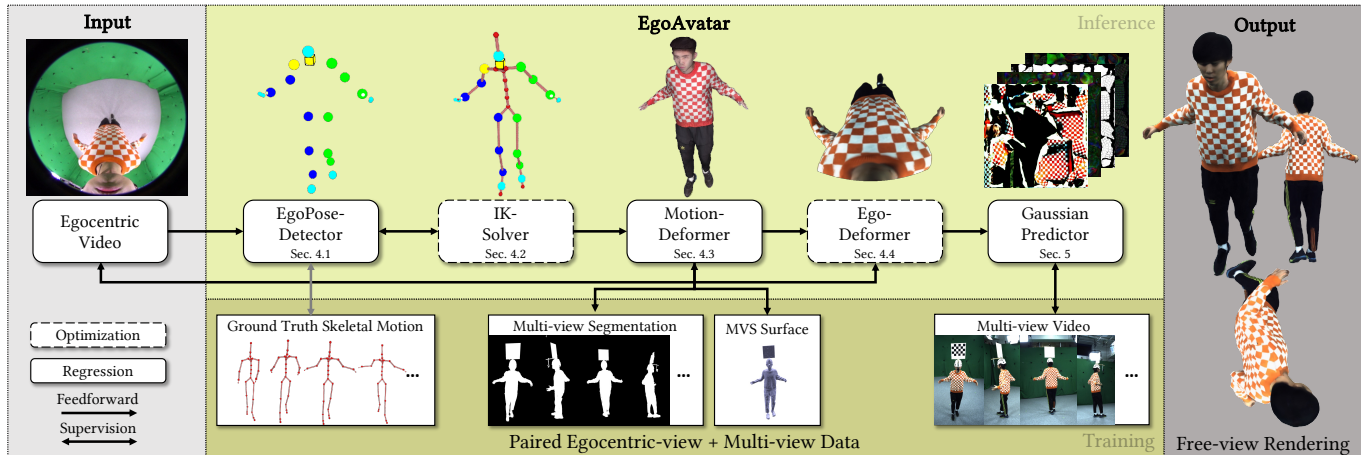
Fig. 2. **Overview of *EgoAvatar*.** Taking as input a single egocentric RGB video, we first detect the skeletal pose in form of 3D keypoints (Sec. 4.1) and then solve for the skeleton parameters, i.e. joint angles, using our *IKSolver* (Sec. 4.2). The motion signal drives the mesh-based avatar via our *MotionDeformer* that is pre-trained on multi-view videos of the actor performing various motions (in Sec. 4.3). At inference time, our *EgoDeformer* further enhances the egocentric view alignment of the predicted avatar (Sec. 4.4). Finally, our *GaussianPredictor* generates dynamic Gaussian parameters in the UV space of the character's mesh, which model the motion- and view-dependent appearance of the avatar (Sec. 5). Given the recovered Gaussian parameters representing our character, we can render free viewpoint videos of the avatar that is solely driven from an egocentric RGB video of the real human using Gaussian splatting.

Our experiments demonstrate, both, quantitatively and qualitatively a clear improvement over the state of the art while for the first time demonstrating a unified approach for egocentrically-driven, photoreal, full-body, and free-view human rendering.

2 Related Work

Motion-driven Full-body Avatars. Motion-driven full-body clothed avatars [Bagautdinov et al. 2021; Habermann et al. 2021] represent the human with explicit meshes and learn the dynamic human geometry and appearance from multi-view video. Some works [Liu et al. 2021; Peng et al. 2021; Weng et al. 2022; Zheng et al. 2023] propose SMPL-based [Loper et al. 2015] human priors combined with neural radiance fields [Mildenhall et al. 2020] for human avatar synthesis. However, it is difficult to perform faithful avatar animation and rendering without an explicit modeling of clothing geometry. In consequence, these works typically fail to recover high-frequency cloth details or loose types of apparel. To better capture clothing dynamics, later works [Xiang et al. 2022, 2021] model clothes as a separate mesh layers, but they require a sophisticated clothing registration and segmentation pipeline. In this work, we closely follow DDC [Habermann et al. 2021] for our skeleton motion-dependent geometry representation of the avatar, which consists of a single mesh where deformations are modeled in a coarse-to-fine manner. However, similar to some recent works [Pang et al. 2023], we represent the appearance as Gaussian splats [Kerbl et al. 2023] in contrast to DDC's dynamic texture maps, which drastically improves the rendering quality. Importantly, all these works solely concentrate on creating an animatable avatar, but they do not target driving it from sparse egocentric observations, which is the focus of our work.

Sparse View-driven Full-body Avatars. Motion-driven avatars can provide visually plausible geometry and rendering quality, but they typically fall short in faithfully re-producing the *actual and true* geometry and appearance that are present in the images, which is a critical requirement for virtual telepresence applications. This is primarily caused by the one-to-many mapping [Liu et al. 2021], i.e. one skeletal pose can result in different surface and appearance configurations as they are typically influenced by more than the skeletal pose, e.g. external forces and initial clothing states. Therefore, researchers seek for affordable sensor as additional input signal to generate authentic avatars, e.g. sparse view RGB [Kwon et al. 2021, 2022; Remelli et al. 2022; Shetty et al. 2023] and RGBD [Xiang et al. 2023] driving signals. Without explicit modeling of pose and body geometry, LookinGood [Martin-Brualla et al. 2018] can also re-render human captured from monocular/sparse RGBD sensors. However, such methods rely on stationary multi-camera rigs, which are complicated to setup and heavily constrain the physical capture space. Few works [Elgharib et al. 2020; Jourabloo et al. 2022] have focused on driving avatars from egocentric camera setups, but they solely reconstruct and render the face or head rather than the full human body. Egocentric *full-body* avatar synthesis is a significantly more challenging task due to the highly non-rigid surface deformation of clothing, the severe self-occlusion, and the complex material patterns, i.e. cloth vs. skin reflectance. EgoRenderer [Hu et al. 2021] is the only work that attempted to solve this task. However, their results are far from photoreal and contain a observable amount of temporal jitter due to their SMPL-based character representation and inaccurate skeletal pose prediction. In contrast, our hybrid character representation, i.e. deformable mesh and Gaussian splatting, achieves significantly higher rendering quality while our personalized pose detector and inverse kinematics solver also demonstrates drastically improved motion tracking quality.

Egocentric Human Pose and Shape Estimation. Motivated by the current advances in AR/VR , we can divide the egocentric setup into two categories: Front-facing and down-facing camera setups. Front-facing cameras [Li et al. 2023; Luo et al. 2021; Yuan and Kitani 2019] have limited visibility on the human body movement, which contradicts our goal of high-fidelity body and clothing capture. Instead, we follow the down-facing camera setup [Tome et al. 2019; Xu et al. 2019] to recover the full-body egocentric 3D human pose. Despite the direct observation of the human body, conventional pose trackers [Martinez et al. 2017] suffer from severe self-occlusion, unstable camera motion, and fisheye camera distortions. To cope with the aforementioned challenges, researchers [Wang et al. 2021, 2023c] try to localize the world-space pose with the help of SLAM and scene depth estimation as constraints. A recently proposed fisheye vision transformer model [Wang et al. 2023a] with a diffusion-guided pose detector greatly improves the generalization ability and the performance of egocentric pose estimation leveraging large-scale synthetic data. For improved accuracy, we finetune the model of Wang et al. [2023a] on person-specific egocentric data. We highlight that this model predicts keypoints while not considering free-view rendering and photorealism at all. In contrast, we are interested in driving our photoreal character and propose a dedicated inverse kinematics solver, which recovers joint angles from keypoint estimates.

3 Character Model

We employ a parametric model following DDC [Habermann et al. 2021] to represent the explicit body and clothing shape. For each character, we obtain a template mesh $T_0 \in \mathbb{R}^{4890 \times 3}$, a skeleton S and skinning weights $\mathcal{W}$ from a body scanner. The skeleton S is controlled by 54 degrees of freedom (DoF) $\theta \in \mathbb{R}^{54}$ including joint rotations, global rotations, and translations. The canonical shape T is parameterized as a non-rigid deformation $T(\cdot)$ of the 4890 vertices template shape T_0 using a 489 nodes embedded graph [Sumner et al. 2007]. The embedded deformation is parameterized by per-node rotations $\alpha \in \mathbb{R}^{489 \times 3 \times 3}$ and translation $t \in \mathbb{R}^{489 \times 3}$, and per-vertex offsets $d \in \mathbb{R}^{4890 \times 3}$. Driven by motion parameters θ, we first transform the skeleton via Forward Kinematics (FK) $\mathcal{J}(\theta)$ and then animate the canonical character model T into posed space M via the Linear Blend Skinning (LBS) [Magnenat et al. 1988] function $W(\cdot)$. Formally, our character model is defined as

$$M = W(T(T_0, \alpha, t, d), \mathcal{J}(\theta), \mathcal{W}), \tag{1}$$

which allows us to effectively represent surface deformations from coarse-to-fine as well as character skinning.

4 Motion-driven Avatar Geometry Recovery

This section introduces our pipeline (see also Fig. 2) that recovers the explicit full-body avatar model M from an egocentric RGB video of the real human. To this end, we first predict 3D joint positions (Sec. 4.1) using our personalized *EgoPoseDetector*. Then, we introduce an inverse kinematics solver (Sec. 4.2), dubbed *IKSolver*, which recovers the DoFs of the skeleton from the 3D joint predictions. Next, we introduce a data-driven *MotionDeformer* (Sec. 4.3), which maps skeletal pose to motion-dependent surface deformations effectively capturing details beyond pure skinning-based deformation and, thus, serving as a strong prior. Due to the ambiguity of fine-grained cloth

deformation, which is not solely dependent on the skeletal motion, an optimization is derived at test time to further align the avatar's geometry with egocentric image silhouettes (Sec. 4.4).

4.1 *EgoPoseDetector* for Pose Detection

Inspired by the current success of learning-based 3D human pose estimation, we adapt the vision transformer-based egocentric pose detector by Wang et al. [2023a] that is pretrained on a large synthetic dataset. Given an image I of frame τ, the neural network localizes 25 body joints $J \in \mathbb{R}^{25 \times 3}$. To further boost the accuracy for the pose detections, we customize a person-specific detector

$$J_{\text{local}}^\tau = \mathcal{F}_{\text{ViT}}(I^\tau) \tag{2}$$

by fine-tuning the generic detection network on subject-specific data, i.e. egocentric images and respective keypoints, using our paired egocentric- and multi-view data (see also Sec. 6.1). Notably, the 3D keypoint detections are predicted with respect to the egocentric camera coordinate system, which we then transform to the global coordinate system, denoted as J^τ, assuming the head-mounted camera can be tracked in 3D space. We highlight that this is a reasonable assumption as recent head-mounted displays offer highly accurate head tracking.

4.2 *IKSolver* for Skeletal Motion Estimation

With the estimated body joint positions J^τ, our inverse kinematics solver optimizes the skeletal pose parameters θ^τ of frame τ. Here, we perform a coarse-to-fine optimization, where we first solve for the global rotation and translation, and then jointly refine these parameters with the joint angles. Each stage iteratively minimizes the energy

$$\arg\min_{\theta^\tau} E_{\text{Data}}(\theta^\tau) + E_{\text{Temporal}}(\theta^\tau) + E_{\text{DoFLimit}}(\theta^\tau) + E_{\text{Reg}}(\theta^\tau). \tag{3}$$

Concretely, our data term

$$E_{\text{Data}}(\theta^\tau) = \sum_{j=0}^{24} ||\mathcal{J}_j(\theta^\tau) - J_j^\tau||_2 \tag{4}$$

aligns the skeleton joint j with their prediction. Further, we introduce some regularizers

$$E_{\text{Temporal}}(\theta^\tau) = \sum_{j=0}^{24} ||\mathcal{J}_j(\theta^\tau) - \mathcal{J}_j(\theta^{\tau-1})||_2 \tag{5}$$

$$E_{\text{DoFLimit}}(\theta^\tau) = \sum_{d=0}^{53} ||\max(\theta_d^\tau - \theta_{\max,d}, -\theta_d^\tau + \theta_{\min,d}, 0)||_2 \tag{6}$$

to produce temporally smooth motions and plausible joint angles, to account for 1) noisy joint predictions especially for self-occluded parts, and 2) indeterminate DoFs that are not directly supervised through joint positions (*i.e.* upper arm rotation). Here, $\theta_{\max}$ and $\theta_{\min}$ are anatomically inspired joint angle limits that were empirically determined. Lastly, we introduce a simple, yet effective regularization

$$E_{\text{Reg}}(\theta^\tau) = \sum_{d=0}^{53} ||\theta_d - \bar{\theta}_d||_2 \tag{7}$$

ensuring that the optimized pose is close to the mean pose $\bar{\theta}$ of the training motions. This avoids implausible angle twists, i.e. poses that may coincide with the keypoint detections, but that have implausible angle configurations.

4.3 *MotionDeformer* for Clothed Avatar Animation

The aim of this stage is to produce the character surface M of the clothed human avatar from the optimized motion θ^τ. Simply posing the template T_0 using LBS is not sufficient to recover dynamic clothing details as skinning mostly models piece-wise rigid deformations. Hence, we aim at modeling the fine-grained clothing deformations conditioned on the normalized motion input $\hat{\theta}^\tau = \{\theta^i : i \in \{\tau - 2, \ldots, \tau\}\}$. Particularly, we exploit two structure-aware graph neural networks [Habermann et al. 2021]

$$\alpha^\tau, t^\tau = \mathcal{F}_{\text{EG}}(\hat{\theta}^\tau, T_0) \tag{8}$$

$$d^\tau = \mathcal{F}_\delta(\hat{\theta}^\tau, T(T_0, \alpha^\tau, t^\tau, 0)) \tag{9}$$

to sequentially recover the low frequency embedded graph parameters α, t and high frequency vertex offset d. The network is supervised with rendering, mask, and Chamfer losses against multi-view images, foreground segmentations and multi-view stereo reconstructions. Importantly, we leverage the *unpaired* training sequence depicting the subject without the head-mounted camera. With the predicted α, t, d, we recover the surface geometry of our avatar M^τ using Eq. 1 for frame τ. Thus, we have a strong pose-dependent surface deformation prior, which models surface deformations beyond typical skinning that we leverage in the next stage.

4.4 *EgoDeformer* for Mesh Refinement

While the previous stage produces a plausible surface shape M that preserves the major geometry details, it misses the stochastic motion-independent cloth movements and contains minor motion prediction errors, which results in a misalignment between the mesh projection $\Pi(M)$ and the egocentric image I. We, thus, employ a test-time optimization to enhance the geometric accuracy. Specifically, we optimize a coordinate-based Multi-Layer Perceptron (MLP) $\mathcal{F}_{\text{MLP}, \Psi}$ to represent the smooth non-rigid clothing deformation, i.e. the final vertex position of deformed mesh M' are $v + \mathcal{F}_{\text{MLP}, \Psi}(v)$, where $v \in \mathbb{R}^3$ is the vertex coordinate of M.

We optimize the MLP weights Ψ using

$$\arg\min_{\Psi} E_{\text{Sil}}(\Psi) + E_{\text{Lap}}(\Psi) + E_{\text{Arap}}(\Psi), \tag{10}$$

where

$$E_{\text{Sil}}(\Psi) = ||I_{\text{Sil}} - \Pi(M')||_2 \tag{11}$$

describes a silhouette loss that minimizes the disparity of the projection $\Pi(\cdot)$ of the deformed mesh M' against egocentric silhouette images I_{Sil} segmented using [Kirillov et al. 2023]. However, recovering pixel-perfect geometry from single view is highly ill-posed, especially due to rich wrinkles. Therefore, we do not explicitly reconstruct wrinkles with photo-metric energy in the test time, but rather learn to generate plausible wrinkles as dynamic Gaussian splats. We further introduce a Laplacian smoothness term E_{Lap} [Desbrun et al. 1999] and part-based as-rigid-as-possible term E_{Arap} [Sorkine and Alexa 2007] to regularize the deformation. To further increase

temporal consistency, we apply a low pass filter as post-processing step. For further details, we refer to the supplemental material.

5 Gaussian-based dynamic appearance

So far, we were only concerned with recovering the geometry of the human. In this section, we introduce our approach to render high-resolution dynamic appearance of the 3D avatar by fusing pretrained body textures with egocentric observations. Given a tracked mesh, deep textures [Habermann et al. 2021; Lombardi et al. 2018] successfully model the dynamic textures and shading by predicting a motion and view-dependent UV texture. However, mesh-based rendering considers only the first intersection of each ray over the triangulated surface, therefore leading to artifacts in thin structures, e.g. hair, and for regions where mesh tracking is often inaccurate, e.g. hands. Instead, mesh-driven volume rendering [Lombardi et al. 2021] provides extra flexibility in compensating for mesh tracking errors and modeling complex surface deformations. Motivated by recent advances of 3D Gaussian Splatting (3DGS) [Kerbl et al. 2023], we leverage 3D Gaussian spheres on top of the mesh surface as primitives, potentially enabling high-resolution rendering.

3DGS represents the 3D object as a collection of Gaussian spheres and renders them in a fully differentiable manner given a virtual camera view. Specifically, each Gaussian sphere is parameterized by a centroid $x \in \mathbb{R}^3$, a color $c \in \mathbb{R}^{48}$ in the form of 3-order Spherical Harmonics coefficients, a Quaternion rotation $\phi \in \mathbb{R}^4$, a scaling $s \in \mathbb{R}^3$, and an opacity $o \in \mathbb{R}$. We formulate the final mesh-driven 3D Gaussians as

$$G(x'; x, \Sigma) = \exp^{\frac{1}{2}(x'-x)^T \Sigma^{-1}(x'-x)}, \tag{12}$$

where the covariance matrix Σ is parameterized by the predicted scaling s and rotation ϕ. 3D Gaussian splats are rendered into a camera view following

$$C = \sum_{i \in \mathcal{N}} c'_i o'_i \prod_{j=i}^{i-1} (1 - \alpha_j). \tag{13}$$

Each pixel C blends N rasterized Gaussian projections. Here, the color and opacity of each Gaussian per pixel are $c'_i = G(x'; x_i, \Sigma_i)c_i$ and $o'_i = G(x'; x_i, \Sigma_i)o_i$.

Dynamic Appearance Modeling. With the recovery of the motion-driven avatar shape, we uniformly sample 3D Gaussians in the UV space of the underlying character mesh, i.e. each texel that is covered by a triangle in the UV map represents a Gaussian. In particular, we first obtain the 3D position, rotation and normal of the deformed mesh M in the training time and M' in the test time. Then, for each texel i, we use barycentric interpolation to fetch the initial position $\hat{x}_i$, rotation $\hat{\phi}_i$ and normal $\hat{n}_i$, respectively. Inspired by the dynamic texture module in DDC, we leverage a UNet [Ronneberger et al. 2015]

$$\{\Delta x_i, \Delta\phi_i, c_i, s_i, o_i\} = \mathcal{F}_{\text{C-UNet}}(n_i, x_0). \tag{14}$$

to predict Gaussian parameters and offsets in UV space, taking as input the global encoding of the mesh geometry (namely, normal map $\hat{n}_i$) and the skeleton root position x_0. Given the differentiable rendering equation R as described in Eq.12 and 13, we produce the

final rendering I_r as

$$I_r = R(\Delta x_i + x_i, < \Delta \phi_i, \phi_i >, c_i, s_i, o_i), \quad (15)$$

where $< \cdot, \cdot >$ denotes the Quaternion multiplication operation.

Training Losses. We supervise our model with multi-view 4K images using L1 and SSIM [Wang et al. 2004] losses. Following Xiang et al. [2023], we also leverage the ID-MRF [Wang et al. 2018] loss to ensure perceptually realistic renderings. Our loss reads as

$$L = L_{\text{L1}} + L_{\text{SSIM}} + L_{\text{ID}-\text{MRF}}. \quad (16)$$

Training Scheme. We separate Gaussians for the head and body region. All Gaussians are jointly trained on the sequence without the head-mounted device. Then, the body Gaussians are fine-tuned on the additional sequence with the head-mounted device. At inference, we assemble the two parts of Gaussian splats for full-body rendering.

6 Experiments

Next, we describe our evaluation protocol on our new benchmark dataset (Sec. 6.1). Then, we evaluate our approach qualitatively (Sec. 6.2) and provide quantitative comparisons (Sec. 6.3). Last, we ablate our individual design choices (Sec. 6.4) and demonstrate robustness to in-the-wild capture conditions (Sec. 6.5). We also refer to the supplemental document and video for more details.

6.1 Experimental Setup

Dataset. To train and evaluate our approach, we record three individual sequences in a 120-camera multi-view studio for each subject. Concretely, we acquire two training sequences where the subjects performs various motions, one with head-mounted camera (used in Sec. 4.1 and 5) and the other one without head-mounted device (used in Sec. 4.3 and 5). Further, we recorded a test sequence with the head-mounted device where the subjects perform unseen movements. In total, we collect three different subjects for evaluation, containing one garment with rich texture, one garment with plain texture, and one garment with rich wrinkles. Additionally, we collect egocentric videos of subjects in outdoor environments to qualitatively evaluate the robustness to in-the-wild conditions. Ground truth skeletal poses and surface geometry are obtained with markerless motion capture [TheCaptury 2020] and implicit surface reconstruction [Wang et al. 2023b] , respectively. If not stated otherwise, we report results using the ground truth head pose for all methods, including competing ones.

Baselines. We compare our work to a motion-driven avatar representation [Habermann et al. 2021], referred to as DDC, and sparse view-driven methods; DVA [Remelli et al. 2022] and HPC [Shetty et al. 2023]. All methods are trained on 4K 120-view studio camera streams plus one synchronized 640p egocentric view camera stream following our training protocol. Since existing methods require ground truth motion as input during test time, we evaluate baseline methods with our skeleton motion estimate (see Sec. 4.2) for fair comparisons.

Metrics. To quantitatively evaluate the novel view synthesis quality, we report Peak Signal-to-Noise Ratio (PSNR) averaged over five hold-out views (uniformly distributed across the walls and ceiling)
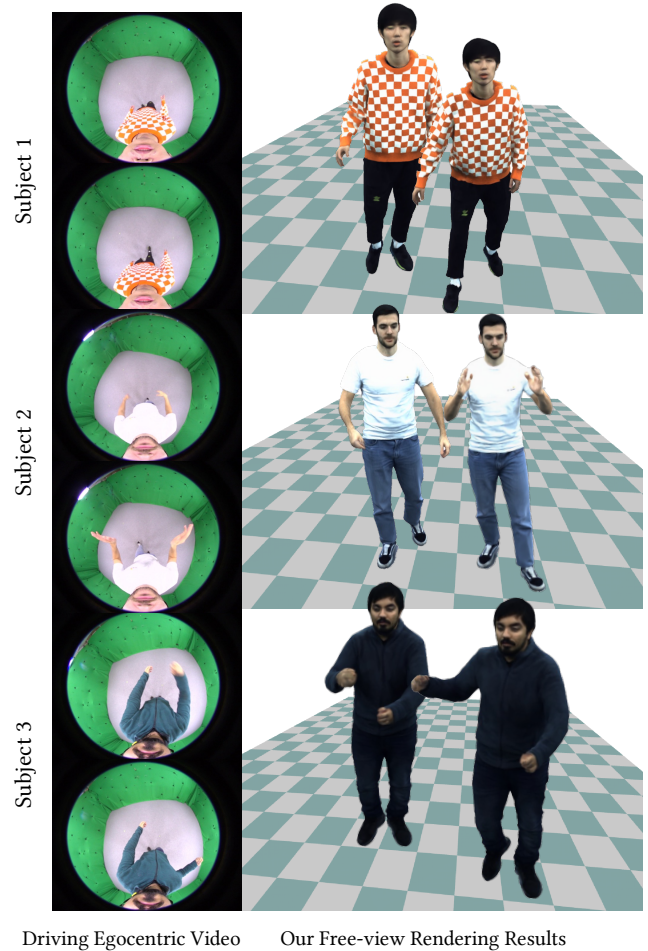
Fig. 3. **Qualitative Results.** On the left, we show frames of the egocentric driving video depicting the real human. On the right, we render the virtual avatar closely following the egocentric driving signal. We highlight the high level of detail and photorealism, e.g. the high-frequency texture on the orange pullover. Moreover, our method faithfully models the dynamic geometry and appearance effects, e.g. wrinkles and shadows on the shirt.

and over all test frames. We further measure Learned Perceptual Image Patch Similarity (LPIPS) [Zhang et al. 2018a] and FID [Heusel et al. 2017] for evaluating perceptual and distributional accuracy. To evaluate the accuracy of our predicted motion, we report the Mean Per Joint Position Error (MPJPE). In terms of surface accuracy, we report the Point-to-Surface Distance (P2SD) between the recovered geometry and the ground truth surface. As we are not interested in recovering the head-mounted device itself, we ignore the head region when computing quantitative metrics.

6.2 Qualitative Results

In Fig. 3, we provide qualitative results of our method for three different subjects. Note that our virtual avatar closely follows the motion of the real human, solely requiring the egocentric video. Moreover,

Table 1. **Quantiative Evaluation.** We quantitatively compare our method to recent animatable [Habermann et al. 2021] and sparse image-driven [Remelli et al. 2022; Shetty et al. 2023] methods. We report PSNR (higher is better), LPIPS (lower is better), and FID (lower is better) across novel views and frames from the test set, excluding head and helmet from the calculation. As HPC [Shetty et al. 2023] produces black background color without alpha masks, we segment the background by thresholding. In terms of LPIPS and FID, our method outperforms prior works by a significant margin, confirming the clear improvement in terms of visual quality that can be obvserved in Fig. 4. The comparable performance in terms of PSNR can be explained by the fact that the PSNR metric is not a perception-based metric and that it favors blurred results over sharp yet slightly misaligned renderings [Zhang et al. 2018b].

Method	Subject 1			Subject 2			Subject 3		
	PSNR ↑	LPIPS (×1000) ↓	FID ↓	PSNR ↑	LPIPS (×1000) ↓	FID↓	PSNR ↑	LPIPS (×1000) ↓	FID↓
DDC [2021]	20.81	<u>45.89</u>	37.30	**23.35**	<u>46.56</u>	<u>39.28</u>	<u>21.40</u>	<u>42.92</u>	<u>28.63</u>
DVA [2022]	**21.09**	46.83	84.56	22.86	48.11	71.92	20.99	43.50	61.63
HPC [2023]	19.80	49.76	<u>32.58</u>	21.14	54.98	87.19	20.82	46.56	34.20
Ours	<u>20.92</u>	**42.53**	**19.28**	<u>23.34</u>	**44.22**	**36.26**	**21.52**	**40.08**	**26.17**

Table 2. **Ablation Study on our *EgoPoseDetector*.** Here, we ablate the personalization step of the pose detector (see Sec. 4.1). After finetuning on person-specific data, we can observe a clear improvement in terms of tracking quality. MPJPE is reported in centimeters.

Method	Subject 1	Subject 2	Subject 3
pre finetune	4.98	4.59	4.04
Ours (post finetune)	**3.72**	**2.91**	**2.64**

Table 3. **Ablation Study.** We quantitatively study the influence of the regularization term E_{Reg} in our *IKSolver* (Sec. 4.2) as well as our *MotionDeformer* (Sec. 4.3) and *EgoDeformer* (Sec. 4.4) for the test-time mesh refinement. To assess rendering quality we report the previously introduced metrics, and to evaluate the tracking quality we report the Mean Per Joint Position Error (MPJPE) and the Point-to-Surface Distance (P2SD) with respect to the ground truth skeletal motion and 3D surface, respectively. Geometric distances are reported in centimeters. As a reference, we also report metrics when using the ground truth motion. Note that all our design choices consistently improve the results, thus, proofing their contribution to the overall accuracy of our method.

Method	Tracking		Rendering		
	MPJPE ↓	P2SD ↓	PSNR ↑	LPIPS ↓	FID ↓
w/o E_{Reg} in *IKSolver*	3.47	1.85	20.59	44.20	21.05
w/o *Motion-Deformer*	3.18	2.19	20.27	47.37	36.45
w/o *Ego-Deformer*	3.18	1.74	20.77	43.22	**19.09**
Ours	**3.18**	**1.67**	**20.92**	**42.53**	19.28
w/ GT motion	0.00	1.05	22.88	37.98	17.71

our method recovers high-fidelity and photorealistic details that can be clearly seen in the free-view results. For more qualitative results, we refer to our supplemental video.

6.3 Comparison

In Tab. 1, we provide quantitative comparisons evaluating the novel view synthesis quality on the test sequences of three subjects. Our method shows a clear improvement in terms of LPIPS and FID scores against all baselines, demonstrating that our method outperforms them in terms of perceptually realistic renderings. Concerning the PSNR, we highlight that this metric is less sensitive or even favours blurred results over sharper but spatially misaligned ones [Zhang et al. 2018b]. This explains why our method sometimes is second best despite its clearly superior visual quality.

Fig. 4 visually demonstrates the superiority of our method in terms of reproducing the sharp boundary of complicated textures (e.g. column 1), consistent high-frequency wrinkles (e.g. columns 2 and 4), and realistic cloth shading (e.g. column 3). In contrast, DDC suffers from a lack of aforementioned details. DVA and HPC assume an external multi-view setting where occlusions are less frequent than the egocentric setting. Thus, their image-projection-based feature formulation fails to provide good appearance conditioning in our egocentric setup, which results in blurry and flickering renderings. In contrast, we intentionally do not condition the appearance module on such image features leading to higher rendering quality.

6.4 Ablation Studies on EgoView Avatar Tracking

We carry out ablation studies to validate the key components proposed in Sec. 4 concerning egocentric motion estimation and surface recovery. Results in Tab. 3 are reported on Subject 1. For reference, we also report metrics when using the ground truth motion.

Personalizing the Pose Predictor (Sec. 4.1). In Tab. 2, we evaluate the influence of personalizing the egocentric pose predictor, i.e. finetuning on subject specific data. It can be clearly seen that test time accuracy across all subjects is clearly improved despite the minimal overhead for fine-tuning.

Regularization in IKSolver (Sec. 4.2). The first and the third rows of Tab. 3 show that using the averaged motion $\bar{\theta}$ as a simple motion prior effectively improves the motion tracking accuracy by 0.29cm and 0.11cm respectively in terms of MPJPE. This is also visually confirmed in Fig. 5 where we can see that with similar joint marker position, our predicted motion, especially for underdetermined skeletal degrees of freedom, better follows the training distribution, which prevents catastrophic skinning failure under challenging poses and provides more natural foot poses, without introducing physics prior.

MotionDeformer (Sec. 4.3). From the comparison between the second and the third row of Tab. 3, we can see that the model w/o *MotionDeformer* significantly underperforms in, both, tracking and

rendering quality. This demonstrates that LBS-based character animation cannot well model the complex and highly non-linear clothing deformation. In contrast, the learning-based *MotionDeformer* predicts reasonable clothing animation result even under challenging body movement, e.g. in column 2, Fig. 5.

EgoDeformer (Sec. 4.4). The improvement between the third and forth rows of Tab. 3 illustrate the benefit of introducing our *EgoDeformer* module in, both, geometric reconstruction and rendering. Due to the severe self-occlusion, the enhancements from *EgoDeformer* mainly focus on the upper body. Fig. 6 compares the egocentric alignment and rendering quality pre- and post-deformation. The most observable improvements come from the forearm region, where *EgoDeformer* better captures fine-grained body and clothing dynamics compared to the *MotionDeformer*-only baseline.

6.5 Robustness Testing under Novel Illumination

The robustness of our egocentric video-driven avatar approach against novel scenarios is essential. Therefore, we test our approach on three novel outdoor scenarios that significantly differ from our studio lighting and environment. Since the ground truth head pose cannot be easily acquired in this setting, we assume a static head pose and only provide qualitative root-aligned results. Note that, both, the estimated pose as well as the rendering quality in Fig. 7 are plausible despite being under very different illumination conditions.

7 Limitations and Future Work

While our method presents a significant step towards full-body egocentric video-driven avatars, there are still open questions and challenges to be addressed in the future. Currently, our character model solely models outgoing radiance as a function of pose and surface preventing relighting the avatar. Thus, in the future, we plan to explore decomposing the character's outgoing radiance into radiance transfer functions and illumination. Moreover, our skeletal pose tracking is purely based on kinematics and ignores physics and the avatar's surroundings. Thus, in case of human-object interaction scenarios, our method might recover geometry that penetrates the object's surface or result in complete tracking failure. In the future, we plan to explore physics-based motion capture incorporating scene constrains. Finally, we currently do not model and track hand gestures or facial expressions. Thus, future work may also look into more expressive capture and rendering for which we believe our approach builds a solid foundation.

8 Conclusion

In this work, we present *EgoAvatar*, the first unified approach to animate and render a photoreal full-body avatar driven solely from a monocular egocentric video feed. To this end, we learn an animatible avatar representation from multi-view video and introduce a personalized egocentric pose and surface tracking pipeline. During inference, given a single RGB egocentric video of the real human, *EgoAvatar* can recover the skeletal pose and 3D geometry as well as Gaussian appearance parameters of our avatar allowing us to render photorealistic free-view videos at unprecedented quality. We believe our work presents a significant step towards immersive telepresence on-the-go as well as other applications in VR and AR such as online tutoring, film making, and gaming.

References

Hiroyasu Akada, Jian Wang, Soshi Shimada, Masaki Takahashi, Christian Theobalt, and Vladislav Golyanik. 2022. UnrealEgo: A New Dataset for Robust Egocentric 3D Human Motion Capture. In *European Conference on Computer Vision (ECCV)*.

Timur Bagautdinov, Chenglei Wu, Tomas Simon, Fabian Prada, Takaaki Shiratori, Shih-En Wei, Weipeng Xu, Yaser Sheikh, and Jason Saragih. 2021. Driving-signal aware full-body avatars. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–17.

Young-Woon Cha, True Price, Zhen Wei, Xinran Lu, Nicholas Rewkowski, Rohan Chabra, Zihe Qin, Hyoungnun Kim, Zhaoqi Su, Yebin Liu, Adrian Ilie, Andrei State, Zhenlin Xu, Jan-Michael Frahm, and Henry Fuchs. 2018. Towards Fully Mobile 3D Face, Body, and Environment Capture Using Only Head-worn Cameras. *IEEE Transactions on Visualization and Computer Graphics* 24, 11 (2018), 2993–3004. https://doi.org/10.1109/TVCG.2018.2868527

Mathieu Desbrun, Mark Meyer, Peter Schröder, and Alan H Barr. 1999. Implicit fairing of irregular meshes using diffusion and curvature flow. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. 317–324.

Mohamed Elgharib, Mohit Mendiratta, Justus Thies, Matthias Nießner, Hans-Peter Seidel, Ayush Tewari, Vladislav Golyanik, and Christian Theobalt. 2020. Egocentric Videoconferencing. *ACM Transactions on Graphics* 39, 6, Article 268 (Dec 2020).

Marc Habermann, Lingjie Liu, Weipeng Xu, Gerard Pons-Moll, Michael Zollhoefer, and Christian Theobalt. 2023. Hdhumans: A hybrid approach for high-fidelity digital humans. *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 6, 3 (2023), 1–23.

Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. 2021. Real-time deep dynamic characters. *ACM Transactions on Graphics (ToG)* 40, 4 (2021), 1–16.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

Tao Hu, Kripasindhu Sarkar, Lingjie Liu, Matthias Zwicker, and Christian Theobalt. 2021. Egorenderer: Rendering human avatars from egocentric camera images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14528–14538.

Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. 2022. NeuMan: Neural Human Radiance Field from a Single Video. https://arxiv.org/abs/2203.12575

Amin Jourabloo, Fernando De la Torre, Jason Saragih, Shih-En Wei, Stephen Lombardi, Te-Li Wang, Danielle Belko, Autumn Trimble, and Hernan Badino. 2022. Robust Egocentric Photo-Realistic Facial Expression Transfer for Virtual Reality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20323–20332.

Taeho Kang, Kyungjin Lee, Jinrui Zhang, and Youngki Lee. 2023. Ego3dpose: Capturing 3d cues from binocular egocentric views. In *SIGGRAPH Asia 2023 Conference Papers*. 1–10.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics* 42, 4 (2023).

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. 2023. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 4015–4026.

Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. 2021. Neural human performer: Learning generalizable radiance fields for human performance rendering. *Advances in Neural Information Processing Systems* 34 (2021), 24741–24752.

Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. 2022. Neural Image-based Avatars: Generalizable Radiance Fields for Human Avatar Modeling. In *The Eleventh International Conference on Learning Representations*.

Youngjoong Kwon, Lingjie Liu, Henry Fuchs, Marc Habermann, and Christian Theobalt. 2023. DELIFFAS: Deformable Light Fields for Fast Avatar Synthesis. *Advances in neural information processing systems* (2023).

Jiaman Li, Karen Liu, and Jiajun Wu. 2023. Ego-Body Pose Estimation via Ego-Head Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and*

Pattern Recognition. 17142–17151.

Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. 2021. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM transactions on graphics (TOG)* 40, 6 (2021), 1–16.

Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. 2018. Deep appearance models for face rendering. *ACM Transactions on Graphics (ToG)* 37, 4 (2018), 1–13.

Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. 2021. Mixture of Volumetric Primitives for Efficient Neural Rendering. *ACM Trans. Graph.* 40, 4, Article 59 (jul 2021), 13 pages. https://doi.org/10.1145/3450626.3459863

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2015. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics* 34, 6 (2015), 1–16.

Zhengyi Luo, Ryo Hachiuma, Ye Yuan, and Kris Kitani. 2021. Dynamics-regulated kinematic policy for egocentric pose estimation. *Advances in Neural Information Processing Systems* 34 (2021), 25019–25032.

Diogo Luvizon, Vladislav Golyanik, Adam Kortylewski, Marc Habermann, and Christian Theobalt. 2024. Relightable Neural Actor with Intrinsic Decomposition and Pose Control. In *European Conference on Computer Vision (ECCV)*.

Thalmann Magnenat, Richard Laperrière, and Daniel Thalmann. 1988. Joint-dependent local deformations for hand animation and object grasping. In *Proceedings of Graphics Interface'88*. Canadian Inf. Process. Soc, 26–33.

Ricardo Martin-Brualla, Rohit Pandey, Shuoran Yang, Pavel Pidlypenskyi, Jonathan Taylor, Julien Valentin, Sameh Khamis, Philip Davidson, Anastasia Tkach, Peter Lincoln, et al. 2018. LookinGood: enhancing performance capture with real-time neural re-rendering. *ACM Transactions on Graphics (TOG)* 37, 6 (2018), 1–14.

Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. 2017. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision.* 2640–2649.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *European Conference on Computer Vision.* Springer, 405–421.

Haokai Pang, Heming Zhu, Adam Kortylewski, Christian Theobalt, and Marc Habermann. 2023. ASH: Animatable Gaussian Splats for Efficient and Photoreal Human Rendering. (2023). arXiv:2312.05941 [cs.CV]

Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. 2021. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 9054–9063.

Edoardo Remelli, Timur Bagautdinov, Shunsuke Saito, Chenglei Wu, Tomas Simon, Shih-En Wei, Kaiwen Guo, Zhe Cao, Fabian Prada, Jason Saragih, et al. 2022. Drivable volumetric avatars using texel-aligned features. In *ACM SIGGRAPH 2022 Conference Proceedings.* 1–9.

Helge Rhodin, Christian Richardt, Dan Casas, Eldar Insafutdinov, Mohammad Shafiei, Hans-Peter Seidel, Bernt Schiele, and Christian Theobalt. 2016. EgoCap: egocentric marker-less motion capture with two fisheye cameras. *ACM Trans. Graph.* 35, 6, Article 162 (dec 2016), 11 pages. https://doi.org/10.1145/2980179.2980235

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18.* Springer, 234–241.

Ashwath Shetty, Marc Habermann, Guoxing Sun, Diogo Luvizon, Vladislav Golyanik, and Christian Theobalt. 2023. Holoported Characters: Real-time Free-viewpoint Rendering of Humans from Sparse RGB Cameras. arXiv:2312.07423 [cs.CV]

Olga Sorkine and Marc Alexa. 2007. As-rigid-as-possible surface modeling. In *Symposium on Geometry processing*, Vol. 4. Citeseer, 109–116.

Robert W. Sumner, Johannes Schmid, and Mark Pauly. 2007. Embedded deformation for shape manipulation. *ACM Trans. Graph.* 26, 3 (jul 2007), 80–es. https://doi.org/10.1145/1276377.1276478

Guoxing Sun, Rishabh Dabral, Pascal Fua, Christian Theobalt, and Marc Habermann. 2024. MetaCap: Meta-learning Priors from Multi-View Imagery for Sparse-view Human Performance Capture and Rendering. In *ECCV*.

TheCaptury. 2020. Captury motion capture redefined: Go markerless. https://captury.com/

Denis Tome, Patrick Peluse, Lourdes Agapito, and Hernan Badino. 2019. xr-egopose: Egocentric 3d human pose from an hmd camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 7728–7738.

Jian Wang, Zhe Cao, Diogo Luvizon, Lingjie Liu, Kripasindhu Sarkar, Danhang Tang, Thabo Beeler, and Christian Theobalt. 2023a. Egocentric Whole-Body Motion Capture with FisheyeViT and Diffusion-Based Motion Refinement. *arXiv preprint arXiv:2311.16495* (2023).

Jian Wang, Lingjie Liu, Weipeng Xu, Kripasindhu Sarkar, Diogo Luvizon, and Christian Theobalt. 2022. Estimating egocentric 3d human pose in the wild with external weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 13157–13166.

Jian Wang, Lingjie Liu, Weipeng Xu, Kripasindhu Sarkar, and Christian Theobalt. 2021. Estimating egocentric 3d human pose in global space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 11500–11509.

Jian Wang, Diogo Luvizon, Weipeng Xu, Lingjie Liu, Kripasindhu Sarkar, and Christian Theobalt. 2023c. Scene-aware Egocentric 3D Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 13031–13040.

Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. 2023b. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 3295–3306.

Yi Wang, Xin Tao, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. 2018. Image inpainting via generative multi-column convolutional neural networks. *Advances in neural information processing systems* 31 (2018).

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

Shih-En Wei, Jason Saragih, Tomas Simon, Adam W. Harley, Stephen Lombardi, Michal Perdoch, Alexander Hypes, Dawei Wang, Hernan Badino, and Yaser Sheikh. 2019. VR facial animation via multiview image translation. *ACM Trans. Graph.* 38, 4, Article 67 (jul 2019), 16 pages. https://doi.org/10.1145/3306346.3323030

Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. 2022. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition.* 16210–16220.

Donglai Xiang, Timur Bagautdinov, Tuur Stuyck, Fabian Prada, Javier Romero, Weipeng Xu, Shunsuke Saito, Jingfan Guo, Breannan Smith, Takaaki Shiratori, et al. 2022. Dressing avatars: Deep photorealistic appearance for physically simulated clothing. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–15.

Donglai Xiang, Fabian Prada, Timur Bagautdinov, Weipeng Xu, Yuan Dong, He Wen, Jessica Hodgins, and Chenglei Wu. 2021. Modeling clothing as a separate layer for an animatable human avatar. *ACM Transactions on Graphics (TOG)* 40, 6 (2021), 1–15.

Donglai Xiang, Fabian Prada, Zhe Cao, Kaiwen Guo, Chenglei Wu, Jessica Hodgins, and Timur Bagautdinov. 2023. Drivable Avatar Clothing: Faithful Full-Body Telepresence with Dynamic Clothing Driven by Sparse RGB-D Input. *SIGGRAPH Asia 2022 Conference Papers* (2023), 1–9.

Weipeng Xu, Avishek Chatterjee, Michael Zollhoefer, Helge Rhodin, Pascal Fua, Hans-Peter Seidel, and Christian Theobalt. 2019. Mo^{2}Cap^2: Real-time mobile 3d motion capture with a cap-mounted fisheye camera. *IEEE transactions on visualization and computer graphics* 25, 5 (2019), 2093–2101.

Ye Yuan and Kris Kitani. 2019. Ego-pose estimation and forecasting as real-time pd control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 10082–10092.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018a. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 586–595.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018b. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.

Zerong Zheng, Xiaochen Zhao, Hongwen Zhang, Boning Liu, and Yebin Liu. 2023. AvatarRex: Real-time Expressive Full-body Avatars. *ACM Transactions on Graphics (TOG)* 42, 4 (2023).

Heming Zhu, Fangneng Zhan, Christian Theobalt, and Marc Habermann. 2023. TriHuman : A Real-time and Controllable Tri-plane Representation for Detailed Human Geometry and Appearance Synthesis. arXiv:2312.05161 [cs.CV]
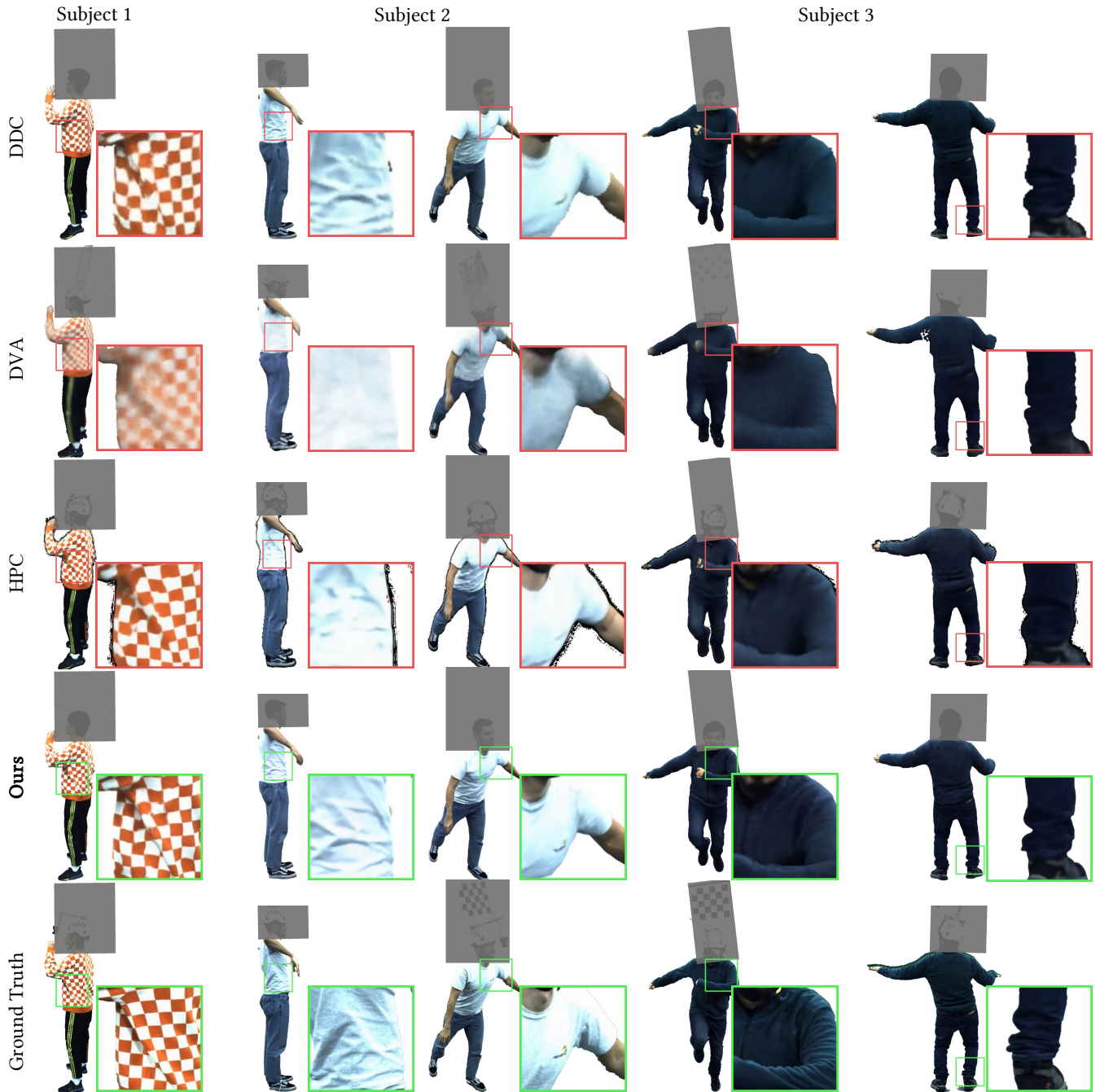
Fig. 4. **Qualitative Comparisons.** We compare our method to recent animatible [Habermann et al. 2021] and sparse image-driven [Remelli et al. 2022; Shetty et al. 2023] methods in terms of novel view synthesis on three testing sequences showing different subjects. As none of these methods is able to predict the skeletal pose from egocentric video, we provide *our* pose estimate for a fair comparison. For image-driven methods, we supply the egocentric video as driving signal. Due to the different underlying 3D representation, we do not perform post-processing, i.e. head avatar exchange, for baseline methods. However, we apply a semi-transparent mask on the region we exclude from quantitative comparison. We highlight the clear improvement in terms of visual quality that our method can achieve compared to prior works, which primarily stems from our carefully designed character representation (see Sec. 3, 4.3, 4.4, and 5). We increase the brightness of subject 3 for better visualization.
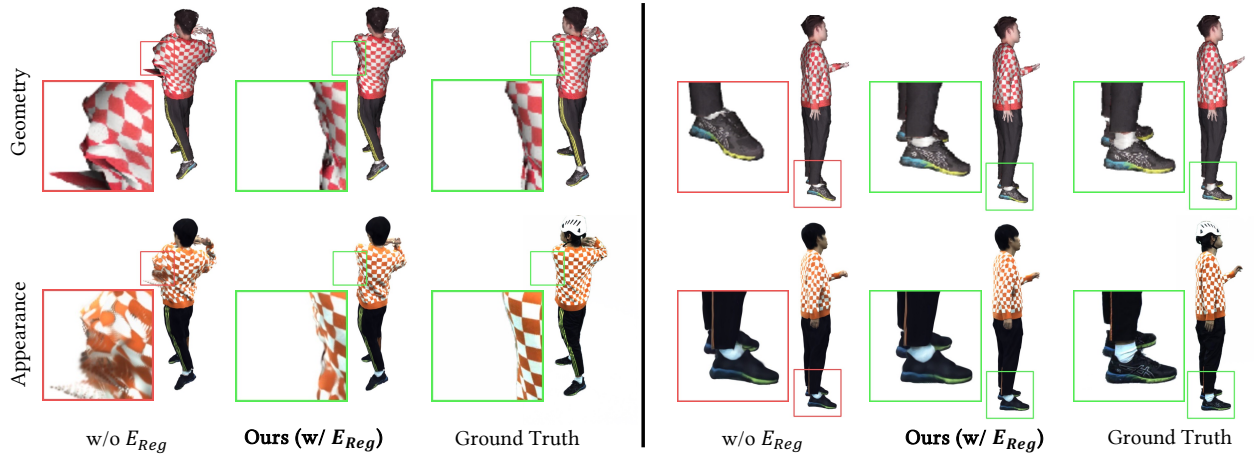
Fig. 5. **Ablation Study of our *IKSolver*.** Without our regularization term E_{Reg}, our *IKSolver* (see Sec. 4.2) might converge to twisted angles along the longitudinal bone axis. While such poses may perfectly describe the 3D joint detections, they typically lead to high mesh distortions (see insets). Our simple, yet effective, regularization prevents such cases and steers the optimization towards a better solution leading to significantly reduced mesh distortions.
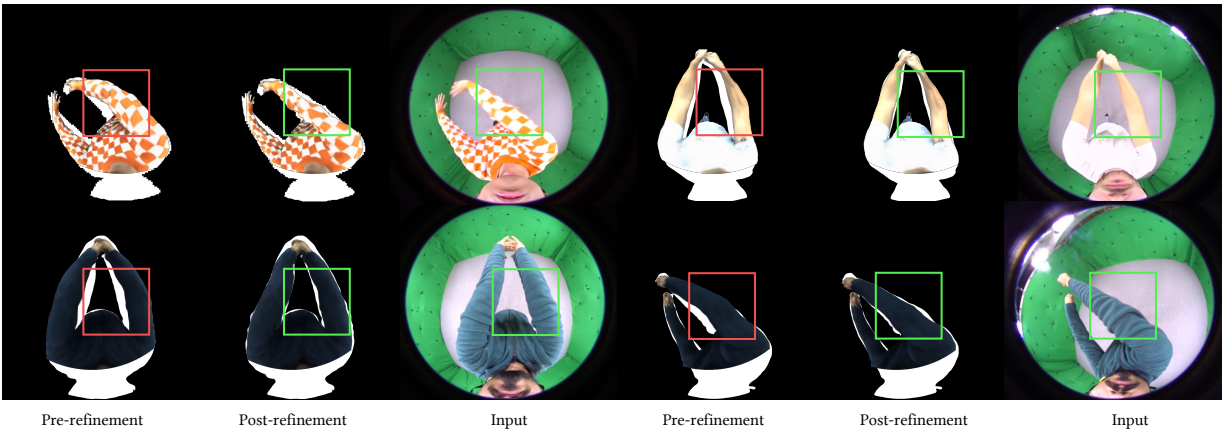

Fig. 6. **Ablation Study of our *EgoDeformer*.** We render our result in the egocentric view and overlay it with the ground truth segmentation mask. Note that after our proposed refinement step, the avatar overlays significantly better with the ground truth. Thus, our final avatar more faithfully reflects the true driving signal.
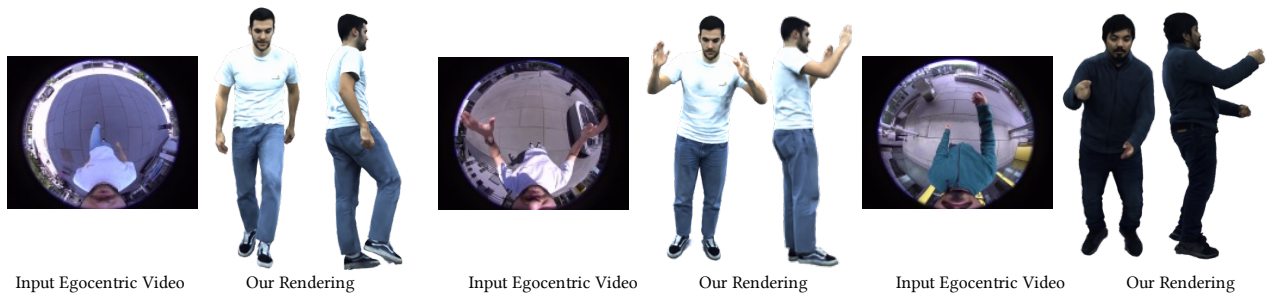

Fig. 7. **Qualitative In-the-Wild Results.** We captured egocentric videos of the subjects in uncontrolled outdoor scenarios of varying illumination conditions and environments. Note that our tracking and rendering pipeline is robust to such changes and the recovered pose, geometry, and appearance faithfully reflect the egocentric driving signal.