electronics

MDPI

Review

Taxonomy and Survey of Current 3D Photorealistic Human Body Modelling and Reconstruction Techniques for Holographic-Type Communication

Radostina Petkova *, Ivaylo Bozhilov, Desislava Nikolova, Ivaylo Vladimirov and Agata Manolova

Faculty of Telecommunications, Technical University of Sofia, 8 Kliment Ohridski blvd., 1000 Sofia, Bulgaria; ibojilov@tu-sofia.bg (I.B.); dnikolova@tu-sofia.bg (D.N.); ivladimirov@tusofia.bg (I.V.); amanolova@tu-sofia.bg (A.M.)

* Correspondence: rapetkova@tu-sofia.bg

Abstract: The continuous evolution of video technologies is now primarily focused on enhancing 3D video paradigms and consistently improving their quality, realism, and level of immersion. Both the research community and the industry work towards improving 3D content representation, compression, and transmission. Their collective efforts culminate in the striving for real-time transfer of volumetric data between distant locations, laying the foundation for holographic-type communication (HTC). However, to truly enable a realistic holographic experience, the 3D representation of the HTC participants must accurately convey the real individuals' appearance, emotions, and interactions by creating authentic and animatable 3D human models. In this regard, our paper aims to examine the most recent and widely acknowledged works in the realm of 3D human body modelling and reconstruction. In addition, we provide insights into the datasets and the 3D parametric body models utilized by the examined approaches, along with the employed evaluation metrics. Our contribution involves organizing the examined techniques, making comparisons based on various criteria, and creating a taxonomy rooted in the nature of the input data. Furthermore, we discuss the assessed approaches concerning different indicators and HTC.

Keywords: human body modelling; human body reconstruction; holographic-type communication; 3D avatars; deep-based human body reconstruction

Citation: Petkova, R.; Bozhilov, I.; Nikolova, D.; Vladimirov, I.; Manolova, A. Taxonomy and Survey of Current 3D Photorealistic Human Body Modelling and Reconstruction Techniques for Holographic-Type Communication. *Electronics* **2023**, *12*, 4705. https://doi.org/10.3390/electronics12224705

Academic Editor: Beiwen Li

Received: 21 October 2023
Revised: 15 November 2023
Accepted: 17 November 2023
Published: 19 November 2023

1. Introduction

Technological advancements have initiated the era of HTC. As explained in [1], HTC involves the transition from one person's actual location to another without the need for a physical traversal of the intervening space. However, the actual enablement of a truly immersive holographic experience necessitates the creation of convincing Mixed Reality (MR) environments, incorporating virtual elements and lifelike human avatars. The support of natural interactions between the virtual participants (the avatars), manipulated by real individuals, is one of the greatest aspects distinguishing future HTC from conventional voice and video-based modes of communication. Moreover, besides HTC, many other applications in the field of healthcare, such as remote consultations, surgical training, remote collaboration, remote rehabilitation, etc. [2–8]; education, including remote collaborative learning, remote guest speakers, anatomy education, etc. [9–13]; entertainment, such as interactive storytelling, Augmented Reality (AR)/Virtual Reality (VR) gaming, live concerts, etc. [14–18]; and e-commerce, in particular virtual try-on [19–21] will benefit from a visually authentic and interactive human appearance. Both academia and industry attempt to automate the detailed acquisition of 3D human pose and shape. The availability of sophisticated 3D acquisition equipment and powerful reconstruction algorithms has made realistic avatar generation possible. In fact, significant advancements are made in this field as digital avatars progressively acquire greater lifelike qualities, leading to increased trust

among individuals. However, replicating real social interactions, including eye contact, body language, and conveying emotions through nonverbal cues (such as touch) and social signals (such as coexistence, closure, and intimacy), remains a formidable challenge, even with the current state of technological advancements. Therefore, the creation of lifelike human avatars that accurately represent both human appearance and behavior is a prominent subject within the realm of holographic experiences.

So far, a multitude of distinct methods for 3D human modelling and reconstruction have been developed, and tremendous efforts are still ongoing in this direction. However, the existing methods exhibit significant diversity in terms of whether they employ a parametric model, the chosen reconstruction approach, the dataset utilized, and, most notably, the input data type. Previous surveys have placed emphasis on variations in the parametric modelling of the 3D human body shape [22], as well as on the types of reconstruction approaches, such as traditional, regression-based, or optimization-based methods, among others [23–26]. In contrast, beyond reviewing parametric models, datasets, and evaluation metrics, our work endeavors to provide a clear distinction among diverse 3D modelling approaches based on the type of input data. To this end, we survey existing methods of 3D human body modelling and reconstruction techniques and establish a taxonomy categorizing the methods into image-based, video-based, and depth-based approaches.

The survey was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses statement (PRISMA) [27] and is shown in Figure 1.

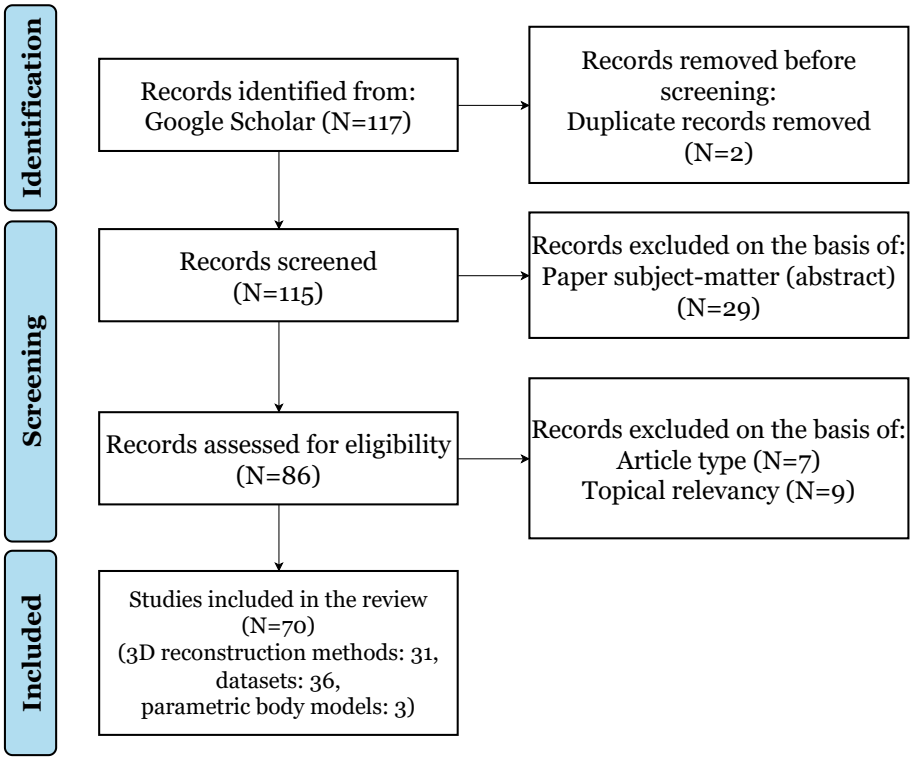

Figure 1. Summary of PRISMA flowchart of the article selection process.

During the screening process, we searched the Google Scholar database using the word combinations: human body modelling, human body reconstruction, 3D avatars, image-based reconstruction, video-based reconstruction, depth-based reconstruction, parametric human body models, statistical body models, etc. The search resulted in 117 records, of which 2 were excluded due to record duplication. Then, 29 more were excluded on the basis of the papers' subject matter according to their abstracts. We assessed the remaining 81 papers as eligible for this study and examined their content in detail. Finally, 70 works were considered the most relevant on the topic and were included in the review.

The structure of our paper is illustrated in Figure 2. Section 1 contains a brief introduction to the topic and gives the motivation behind this work. Section 2 introduces some popular parametric human body models that are often used for the task of 3D human body reconstruction. Section 3 presents the available datasets used for 3D human body reconstruction. In Section 4, a few evaluation metrics that are usually employed for performance assessment are described. Section 5 proposes a taxonomy for 3D human body modelling and reconstruction techniques based on the input data and examines existing methods in the field. The parametric models, datasets, and evaluation metrics used are highlighted. Section 6 discusses the reviewed papers. Finally, this work is concluded in Section 7.

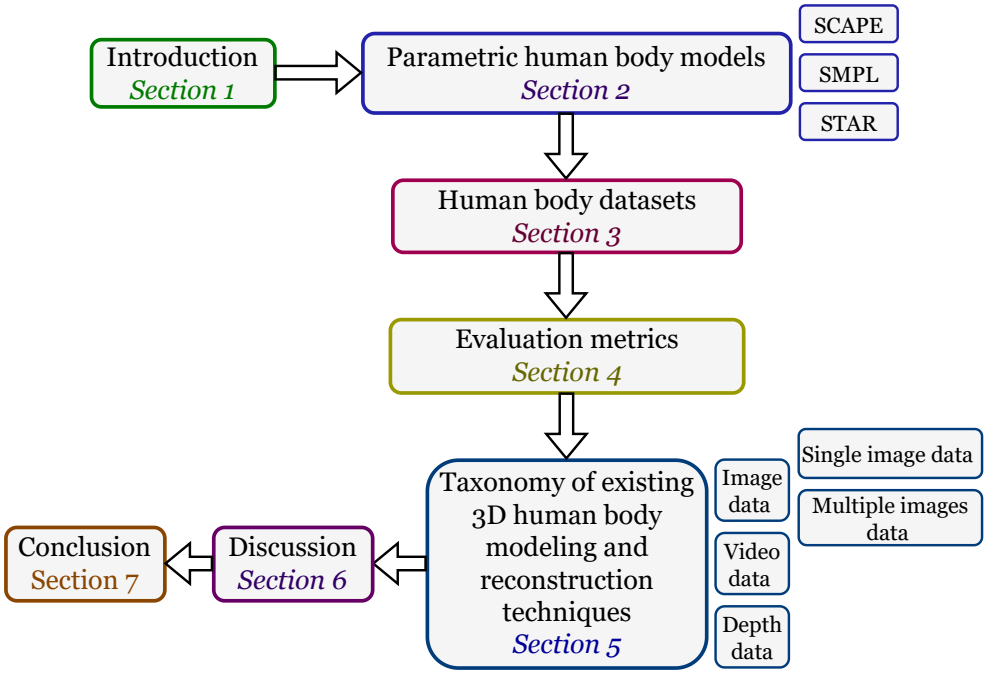

Figure 2. Paper structure.

2. Parametric Human Body Models

Building lifelike human avatars and the subsequent animation is one of the main challenges facing HTC. There is a need for creating accurate representations of HTC participants, which necessitates the detailed reconstruction of 3D digital human models. The generation of such models requires individual- or population-based anthropometric data. Anthropometric measurements are used to describe a person's physical appearance. They are estimates of the distances (both linear and curved) between anatomical landmarks or circumferences at specific human body regions of interest. Height (stature), weight (body mass), upright sitting height, triceps skinfold (upper arm girth), arm circumference (upper arm girth), abdominal circumference (waist circumference), calf circumference, knee height, and elbow breadth are all common anthropometric measurements [28]. The anthropometric database must be extremely thorough in order to be credible for a specific group and to account for multivariate coherences.

Constructing an accurate human body model from various types of input data, such as single images, multi-view images, videos, or depth maps, is a great challenge. Existing methods for fitting a pose to the input data typically rely on parametric, yet statistical, human body models. Such an approach usually requires the indication of body joints, which is mostly carried out manually, but automatic and semi-automatic methods [29] also exist. Further, deep neural networks have been recently used to compute statistical models' parameters [30]. These types of modelling techniques have become an integral component in the recent methods for 3D human body reconstruction and animation. Here, we present

two of the most popular statistical body models and one that is a promising improvement of the second presented model.

2.1. SCAPE

The authors of [31] introduced SCAPE (Shape Completion and Animation of People). It presents a data-driven approach for creating a human body model that considers variations in both shape and posture. Their methodology involves the development of two distinct models for body deformation: one that accounts for deformations resulting from changes in an individual's pose and another that captures deformations across various body shapes among different individuals. To accomplish this, a specific dataset of human body scans is collected. It comprises a pose dataset containing multiple pose scans of a specific individual and a body shape dataset containing scans of multiple individuals in similar poses.

SCAPE considers the pose and shape deformations over each of the mesh triangles, p_k, with triangle points, respectively, $x_{k,1}$, $x_{k,2}$, and $x_{k,3}$. Particularly, deformations are applied over the triangle edges $\hat{v}_{k,j} = x_{k,j} - x_{k,1}$, where $j = 2, 3$. A specific triangle's deformation is given in Equation (1).

$$v_{k,j} = R_{l[k]} S_k Q_k \hat{v}_{k,j}, \tag{1}$$

where $v_{k,j}$ corresponds to the edges of the transformed triangle, $R_{l[k]}$ is related to the rigid rotation matrix that is the same for all triangles in the mesh that belong to the specific body part $l[k]$, and Q_k is a linear transformation matrix that is associated with the non-rigid pose-induced deformations and is specific to each triangle. S_k is another linear transformation matrix that corresponds to body-shape-induced deformations. Both deformation matrices, Q_k and S_k, are not known but can be obtained by using the preliminary model-learned parameters $\{\boldsymbol{a}_k\}$ and $\{U, \mu\}$, such that $Q_k = \mathcal{Q}_{\boldsymbol{a}_k}(\Delta r_{l[k]})$ and $S_k = \mathcal{S}_{U,\mu}(\beta)$, where $\Delta r_{l[k]}$ corresponds to a joint angle that is representative of the relative joint rotations of two rigid parts adjacent to the same joint and β corresponds to the body shape parameters. Both the joint rotations and the body shape parameters are provided by the user. On the other hand, $\{\boldsymbol{a}_k\}$ corresponds to the learned SCAPE model parameters that are related to the pose-induced body deformations, and $\{U, \mu\}$ are the learned PCA parameters that capture the space of model shape deformations. Finally, given all the transformation matrices, R, Q, and S, associated with a specific pose and body shape, a completely new body mesh, Y, can be synthesized according to Equation (2), where $y_{j,k}$ corresponds to the specific triangle points of the generated model.

$$E_H[Y] = \sum_k \sum_{j=2,3} \left\| R_{l[k]} S_k Q_k \hat{v}_{j,k} - (y_{j,k} - y_{1,k}) \right\|^2 \tag{2}$$

Figure 3 illustrates a block diagram detailing the SCAPE body model generation process. It delineates three separate blocks representing the input parameters, namely the SCAPE mean template shape; the user-defined parameters, R and β; and the learned SCAPE parameters, $\{\boldsymbol{a}_k\}$ and $\{U, \mu\}$. The last two sets of parameters are essential for constructing the pose-induced (Q_k) and the shape-induced (S_k) deformation matrices. Within the block SCAPE model generation, these matrices, along with $R_{l[k]}$, modify the SCAPE template mesh and thus generate the new body model.

Although SCAPE has high fidelity, it lacks the ability to capture a strong correlation between body shape and muscle deformation, for which a more expressive model is needed. This may be due to the fact that the model is learned separately for pose and shape variations. However, developing a method that simultaneously uses scans from different people with different poses would require a different approach.

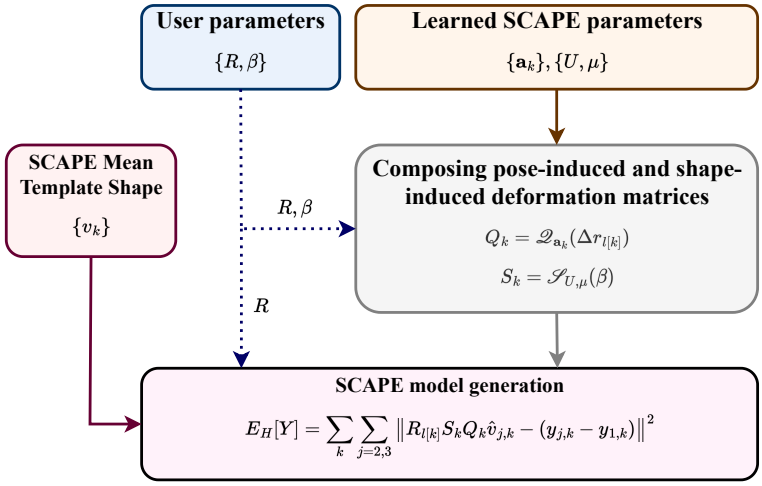

Figure 3. Block diagram of the SCAPE body model generation process.

2.2. SMPL

More recent methods for 3D human body reconstruction use an SMPL (Skinned Multi-Person Linear) Model [32]. Similar to SCAPE, the body model is also examined in two different aspects: identity-dependent shape and non-rigid pose-dependent shape. In contrast to SCAPE, where mesh triangles are primarily utilized, an SMPL considers a vertex-based skinning approach. A mean template mesh, $\overline{\mathbf{T}} \in \mathbb{R}^{3N}$, in the zero pose $\boldsymbol{\theta}^*$ facilitates the model, where N is the total number of vertices. The model is also defined by the following functions. A blend shape function, $B_S(\boldsymbol{\beta}, \mathcal{S}) : \mathbb{R}^{|\boldsymbol{\beta}|} \rightarrow \mathbb{R}^{3N}$, takes as an input the shape parameters, $\boldsymbol{\beta}$, and a set of learned body shape parameters, $\mathcal{S}$, and as an output a blend shape sculpting the subject identity according to Equation (3):

$$B_S(\boldsymbol{\beta}; \mathcal{S}) = \sum_{n=1}^{|\boldsymbol{\beta}|} \beta_n \mathbf{S_n} \tag{3}$$

The function $J(\boldsymbol{\beta}) : \mathbb{R}^{|\boldsymbol{\beta}|} \rightarrow \mathbb{R}^{3K}$ predicts the location of the K skeletal joints with respect to the subject-specific body shape according to Equation (4):

$$J(\boldsymbol{\beta}, \mathcal{J}, \overline{\mathbf{T}}, \mathcal{S}) = \mathcal{J}(V_{shaped}) \tag{4}$$

where $V_{shaped} = \overline{\mathbf{T}} + B_S(\boldsymbol{\beta}; \mathcal{S})$. A pose-dependent blend shape function, $B_P(\boldsymbol{\theta}, \mathcal{P}) : \mathbb{R}^{|\boldsymbol{\theta}|} \rightarrow \mathbb{R}^{3N}$, takes as input the pose parameters, $\boldsymbol{\theta}$, and a set of learned body pose parameters, $\mathcal{P}$, and as an output blend shapes effected by pose-dependent deformations, considering Equation (5).

$$B_P(\boldsymbol{\theta}, \mathcal{P}) = \sum_{n=1}^{9K} (R_n(\boldsymbol{\theta}) - R_n(\boldsymbol{\theta}^*)) \mathbf{P}_n \tag{5}$$

Then, a blend skinning function, $W(\cdot)$, rotates the mesh vertices around the estimated joint centers with respect to the set of learned blend weights $\mathcal{W} \in \mathbb{R}^{3N \times K}$. The resulting model is described by $M(\boldsymbol{\beta}, \boldsymbol{\theta}, \overline{\mathbf{T}}, \mathcal{S}, \mathcal{J}, \mathcal{P}, \mathcal{W}) : \mathbb{R}^{|\boldsymbol{\beta}| \times |\boldsymbol{\theta}|} \rightarrow \mathbb{R}^{3N}$ and is finally defined in Equation (6)

$$M(\boldsymbol{\beta}, \boldsymbol{\theta}, \overline{\mathbf{T}}, \mathcal{S}, \mathcal{J}, \mathcal{P}, \mathcal{W}) = W(T_P(\boldsymbol{\beta}, \boldsymbol{\theta}), J(\boldsymbol{\beta}; \mathcal{J}, \overline{\mathbf{T}}, \mathcal{S}), \boldsymbol{\theta}, \mathcal{W}) \tag{6}$$

where $\{\overline{\mathbf{T}}, \mathcal{W}, \mathcal{S}, \mathcal{J}, \mathcal{P}\}$ is the full set of SMPL model parameters. Except for the mean template model, $\overline{\mathbf{T}}$, the rest are the learnable model pose $(\mathcal{W}, \mathcal{J}, \mathcal{P})$ and model shape $(\mathcal{S})$ parameters obtained during the training. In contrast, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are passed from the user and control the learned parameters, generating a completely new body model. $T_P(\boldsymbol{\beta}, \boldsymbol{\theta}, \overline{\mathbf{T}}, \mathcal{S}, \mathcal{P})$

accounts for the offset from the template model caused due to identity-dependent and pose-dependent shape deformations Equation (7).

$$T_P(\boldsymbol{\beta}, \boldsymbol{\theta}, \overline{\mathbf{T}}, \mathcal{S}, \mathcal{P}) = \overline{\mathbf{T}} + B_S(\boldsymbol{\beta}, \mathcal{S}) + B_P(\boldsymbol{\theta}, \mathcal{P}) \tag{7}$$

Figure 4 visualizes a block diagram of the SMPL human body model generation process. The figure defines three separate blocks for input parameters, namely the SMPL mean template shape, $\overline{\mathbf{T}}$, the user parameters, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, and the learned SMPL parameters, $\mathcal{W}, \mathcal{S}, \mathcal{J}, \mathcal{P}$. The user parameters and the SMPL learned parameters are used to generate the shape blend shapes, the joint location of the new body shape, and the pose blend shapes. In the block shape and pose correction, the obtained shape and pose blend shapes are added to the SMPL template mesh in order to create a template offset, T_P. The generation of a new body model is indicated in the SMPL model generation block, where the offset of the template mesh, the predicted joint locations, the pose parameters, and the blend weights are passed to a standard blend skinning function $W(\cdot)$.

Since its creation in 2015, the SMPL model has been extensively utilized in various reconstruction algorithms due to its open-source nature, compatibility with diverse datasets, and widespread popularity, making it a cornerstone in 3D human body research.

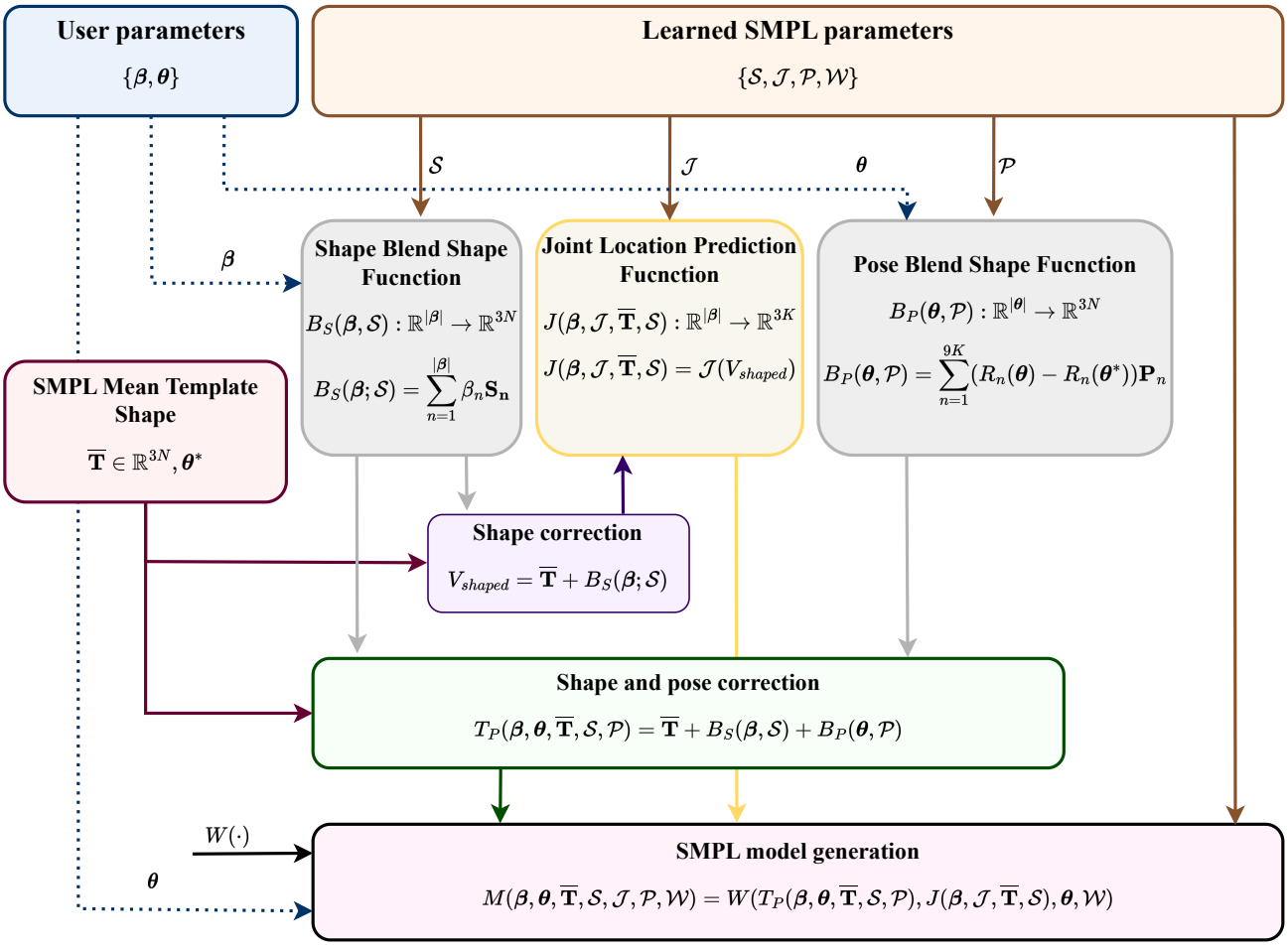

Figure 4. Block diagram of the SMPL body model generation process.

2.3. STAR

Although the SMPL model is widely adopted due to its intuitive parametrization, it suffers from several drawbacks, indicated by [33]. The first limitation that is considered is the huge number of parameters due to the use of global blend shapes. Since each vertex in the mesh is related to every joint in the kinematic tree, the pose-corrective offsets may

capture a spurious long-range correlation, resulting in less realistically generated models. The authors of [33] define the STAR (Sparse Trained Articulated Human Body Regressor), where subsets of mesh vertices that are influenced by specific joint movements are learned. This is reflected by applying per-joint pose correctives and obtaining better results according to deformation realism and a reduced number of model parameters. A second limitation of the SMPL model is the separate examination of the pose-dependent deformation and the body shape. The authors of the STAR argue for the simultaneous consideration of both body pose and BMI (Body Mass Index) by learning shape-dependent pose-corrective blend shapes. Third, the SMPL training dataset is presented by not so many body scans, limiting the shape space. In contrast, the STAR model is trained with additional body scans, resulting in better model generalization.

Similar to the SMPL model, the STAR is a vertex-based model that also factors the body shape into the subject's identity shape and pose-dependent deformations. However, contrary to the SMPL model, the authors assume that pose-corrective deformation is a function of both body pose, $\boldsymbol{\theta} \in \mathbb{R}^{|\boldsymbol{\theta}|}$, and shape, $\boldsymbol{\beta} \in \mathbb{R}^{|\boldsymbol{\beta}|}$. Additionally, during training, a subset of vertices that is relevant to a specific joint, j, is learned, so the pose-corrective blend shape function is applied to it. A template model, $\overline{\mathbf{T}} \in \mathbb{R}^{3N}$, where N is the total number of vertices, is subject to deformation by a shape-corrective blend shape function, B_S, as a meaning of the subject's identity and a pose-corrective blend shape function, B_P, as a meaning of the subject's pose with assumed realism in shape.

The shape-corrective blend shape function $B_S(\boldsymbol{\beta}; \mathcal{S}) : \mathbb{R}^{|\boldsymbol{\beta}|} \rightarrow \mathbb{R}^{3N}$ is defined in Equation (8):

$$B_S(\boldsymbol{\beta}; \mathcal{S}) = \sum_{n=1}^{|\boldsymbol{\beta}|} \beta_n \boldsymbol{S_n} \tag{8}$$

where $\boldsymbol{\beta}$ are the shape coefficients and $\mathcal{S}$ is a set of learned parameters that express the principal components capturing the shape variability space.

Further, the pose-corrective blend shape function with respect of the subject's identity pose, $B_P(\mathbf{q}, \beta_2; \mathcal{K}, \mathbf{A}) : \mathbb{R}^{|\mathbf{q}|\times 1} \rightarrow \mathbb{R}^{3N}$, and BMI is defined in Equation (9):

$$B_P(\mathbf{q}, \beta_2; \mathcal{K}, \mathbf{A}) = \sum_{j=1}^{K-1} B_P^j(\mathbf{q}_{ne}(j), \beta_2; \mathcal{K}_j, \mathbf{A}_j) \tag{9}$$

In this case, a pose-corrective function is applied for each joint, j, in the kinematic tree independently by $B_P^j(\mathbf{q}_{ne}(j), \beta_2; \mathcal{K}_j, \mathbf{A}_j)$, where K is the total number of joints (the root joint is not considered), $\mathbf{q}_{ne}(j) \subset \mathbf{q}$ is a subset that contains a single joint, j, and its direct neighbors in the kinematic three, β_2 corresponds to the PCA coefficient of the second principal component, which is highly related to the BMI, $\mathcal{K}_j \in \mathbb{R}^{3N\times|\mathbf{q}_{ne}(j)|+1}$ is a linear regressor weight matrix, and $\mathbf{A}_j$ corresponds to the activation weights for each vertex. The last two terms are learned during training.

The template mesh with an added pose- and shape-corrective offset, T_P, is defined in Equation (10):

$$T_P(\boldsymbol{\beta}, \boldsymbol{\theta}, \overline{\mathbf{T}}, \mathcal{S}, \mathcal{K}, \mathbf{A}) = \overline{\mathbf{T}} + B_S(\boldsymbol{\beta}, \mathcal{S}) + B_P(\mathbf{q}, \beta_2; \mathcal{K}, \mathbf{A}) \tag{10}$$

Finally, a standard skinning function, W, is applied considering the transformed mesh, T_P, the full set of predicted body joints, $J(\boldsymbol{\beta}, \mathcal{J}, \overline{\mathbf{T}}, \mathcal{S}) = \mathcal{J}(\overline{\mathbf{T}} + B_S(\boldsymbol{\beta}; \mathcal{S}))$, $J \in \mathbb{R}^{3K}$, and a learned set of blend weight parameters, $\mathcal{W}$. The STAR model is defined in Equation (11):

$$M(\boldsymbol{\beta}, \boldsymbol{\theta}, \overline{\mathbf{T}}, \mathcal{S}, \mathcal{J}, \mathcal{K}, \mathbf{A}, \mathcal{W}) = W(T_p(\boldsymbol{\beta}, \boldsymbol{\theta}, \overline{\mathbf{T}}, \mathcal{S}, \mathcal{K}, \mathbf{A}), \boldsymbol{\theta}, \mathcal{W}). \tag{11}$$

A block diagram of the STAR human body model generation process is given in Figure 5. Since the STAR model builds on the SMPL model, both block diagrams look quite similar. However, STAR modifications are highlighted in the light red blocks. First, the input learned parameters include $\mathcal{K}$ and $\mathbf{A}$ instead of $\mathcal{P}$. Then, the pose blend shape

function takes different input parameters and is applied for each joint, j, in the kinematic tree. The generation of the template offset and the new STAR model in the respective blocks are similar to those of the SMPL model, with a difference in the function parameters. The remaining few blocks are the same as those of the SMPL model.

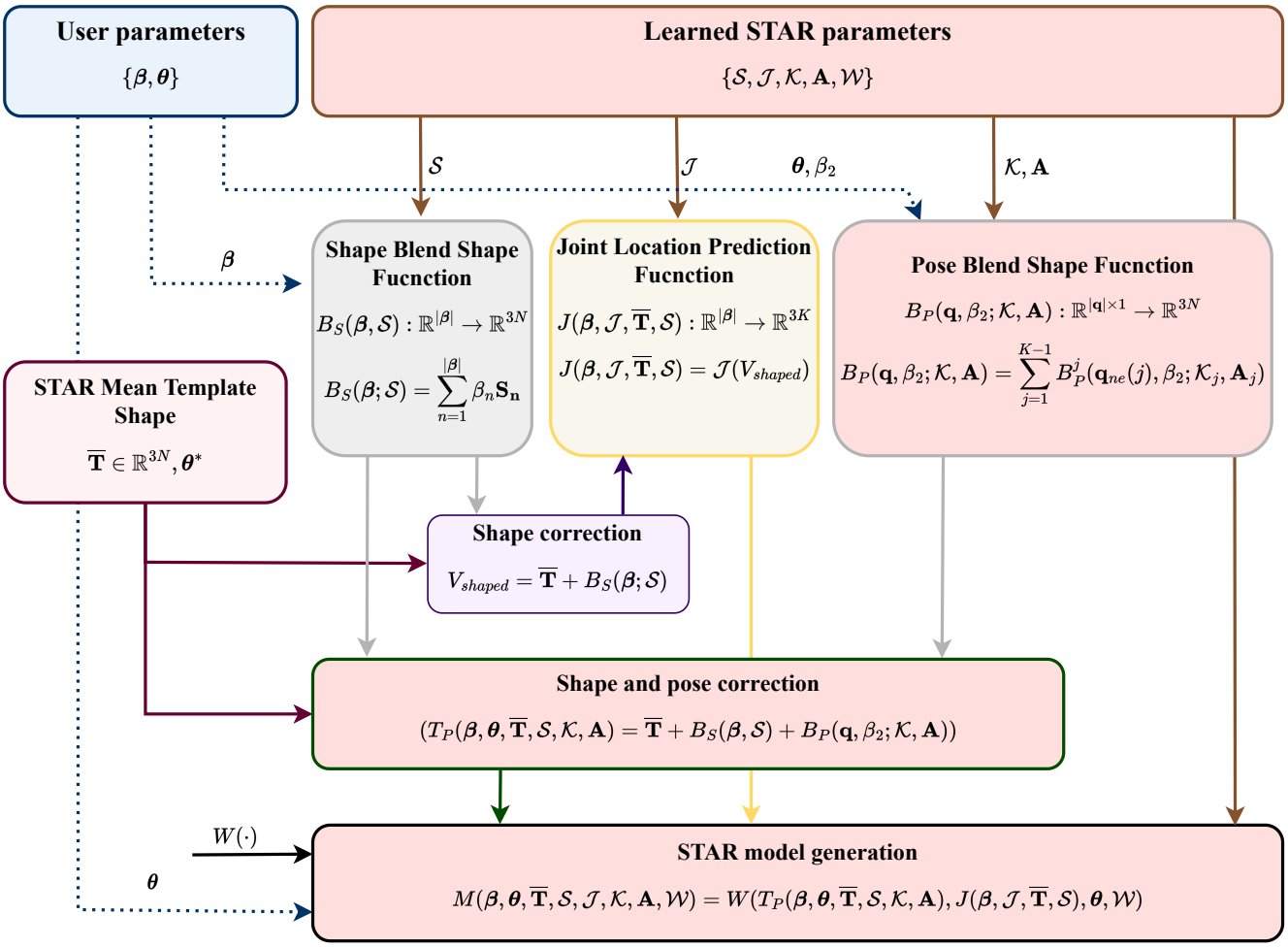

Figure 5. Block diagram of the STAR body model generation process.

3. Human Body Datasets

Human bodies are flexible, moving in various ways and deforming their clothing and muscles. Another complicating issue, like the occlusion of different body parts during movement, may necessitate comprehensive scene modelling in addition to the peoples in the scenario. Such image understanding scenarios push, for example, the avatar body animation system's ability to use prior knowledge and structural correlations by constraining estimates of unseen body components using limited visible information. Insufficient data coverage is one of the most significant issues for trainable systems. So, many researchers have concentrated their efforts on creating publicly available datasets that can be used to build operational systems for realistic scenarios.

One of the largest motion capture datasets is the Human 3.6 M dataset [34]. It consists of 3.6 million fully visible human poses and corresponding images. All of them are captured by a high-speed motion capture system. The recording setup consists of 15 sensors (4 calibrated high-resolution progressive scan cameras that acquire video data at 50 Hz, 1 time-of-flight sensor, and 10 motion cameras), using hardware and software synchronization. This allows for accurate capture and synchronization. The dataset contains activities performed by 11 professional actors (6 male, 5 female) in 17 scenarios—taking photos, discussing, smoking, talking on the phone, etc. Also, accurate 3D joint positions and joint

angles from high-speed motion capture systems are provided. Other useful additions are 3D laser scans of the actors, high-resolution videos, and accurate background subtraction.

The MPI-INF-3DHP [35] is a 3D human body pose estimation dataset. It consists of both constrained indoor and complex outdoor scenes. The dataset comprises eight actors (4 male, 4 female) enacting eight distinct activity sets, each lasting about a minute. With a diverse camera setup, 14 cameras in total, over 1.3 M frames have been obtained, with 500 k originating from cameras at chest height. The dataset provides genuine 3D annotations and a skeleton compatible with the "universal" skeleton of Human 3.6 M. To bridge the gap between studio and real-world conditions, chroma-key masks are available, facilitating extensive scene augmentation. The test set, enriched with various motions, camera viewpoints, clothing varieties, and outdoor settings, aims to challenge and benchmark pose estimation algorithms.

The Synthetic Humans for Real Tasks (SURREAL) dataset [36] contains 6.5 M frames of synthetic humans, organized into 67,582 continuous sequences. The SMPL [32] body model is employed to generate these synthetic bodies, with body deformations distinguished by pose and intrinsic shape. Created in 2017, it is the first large-scale person dataset to generate depth, body parts, optical flow, 2D/3D pose, surface, normal, and ground truth for Reed Green Blue (RGB) video input. The provided images are photorealistic renderings of people in different shapes, textures, viewpoints, and poses.

Dynamic Fine Alignment Using Scan Texture (DFAUST) [37] is considered a 4D dataset. It consists of high-resolution 3D scans of moving non-rigid objects, captured at 60 fps. A new mesh registration method is proposed. It uses both 3D geometry and texture information to register all scans in a sequence according to a common reference topology. The method makes use of texture constancy across short and long time intervals, as well as dealing with temporal offsets in shape and texture.

Microsoft Common Objects in Context (MS COCO) [38] is a large-scale object detection, segmentation, and captioning dataset. It consists of many other objects, but also humans and human photos. The dataset offers recognition in context, superpixel stuff segmentation, and 250,000 people with key points.

Leeds Sports Pose (LSP) [39] and its extended version—LSPe [40]— are human body joint detection datasets. The LSPe dataset contains 10,000 images gathered from Flickr searches for the tags 'parkour', 'gymnastics', and 'athletics' and uses poses that are challenging to estimate. Each image has a corresponding annotation that might not be highly accurate because it is gathered from Amazon Mechanical Turk. Each image is annotated with up to 14 visible joint locations.

The Bodies Under Flowing Fashion (BUFF) dataset, as delineated by the authors of [41], offers over 11,000 3D human body models engaged in complex movements. It is distinctive in its inclusion of videos featuring individuals in clothing paired with 3D models devoid of clothing textures. This dataset emerges from a multi-camera active stereo system, utilizing 22 pairs of stereo cameras, color cameras, speckle projectors, and white-light LED panels operating at varied frame rates. This system outputs 3D meshes averaging around 150 K vertices, capturing subjects in two distinct clothing styles. Of the initial six subjects, the data from one were withheld, resulting in a public release of 11,054 scans. To derive a semblance of "ground truth", the subjects were captured in minimal attire, with the dataset's accuracy showcased by the proximity of more than half of the scan points to the mean of the estimates. The BUFF dataset efficiently captures detailed aspects of human movement while also considering the impact of different clothing on a body's shape and motion.

The HumanEva datasets [42], comprising HumanEva-I and HumanEva-II, offer a blend of video recordings and motion capture data from subjects performing predefined actions. HumanEva-I encompasses data from four subjects executing six distinct actions, each with synchronized video and motion capture, and one with only motion capture. This dataset leverages seven synchronized cameras, utilizing multi-view video data coupled with pose annotations. On the other hand, HumanEva-II focuses on two subjects, both

of whom are also present in HumanEva-I, performing an extended "Combo" sequence, resulting in roughly 2500 synchronized frames. The data, collected under controlled indoor conditions, capture the intricacies of natural movement, albeit with challenges posed by illumination and grayscale imagery.

The UCLA Human–Human–Object Interaction (HHOI) dataset [43] is a novel RGB-D (Reed Green Blue—Depth) video collection detailing both human–human and human–object–human interactions, captured using a Microsoft Kinect v2 sensor. Comprising three human–human interactions—-hand shakes, high-fives, and pull ups-—and two human–object–human interactions—-throw and catch and hand over a cup—the dataset features an average of 23.6 instances for each interaction. These instances are performed by eight actors, recorded from multiple angles, and spanning 2–7 s at a frame rate of 10–15 fps. While objects within the dataset are discerned using background subtraction on both RGB and depth images, the Microsoft Kinect v2's skeleton estimation is also utilized. The dataset, divided into four distinct folds for training and testing, ensures no overlap of actor combinations between the sets. The training algorithm demonstrates robust convergence within 100 iterations, operating on a standard 8-core 3.6 GHz and yielding an average synthesis speed of 5 fps using an unoptimized Matlab code.

To address prevalent challenges in viewpoint invariant pose estimation, a novel technical solution has been presented in [44]. It integrates local pose details into a learned, viewpoint-invariant feature space. This approach enhances the iterative error feedback model to incorporate higher-order temporal dependencies and adeptly manage occlusions via a multi-task learning methodology. Complementing this endeavor is the introduction of the Invariant Top View (ITOP) dataset, a comprehensive collection of 100 K depth images capturing 20 individuals across 15 diverse actions, encompassing a wide range of views, from front, top, to side, inclusive of occluded body segments. Each image in the ITOP dataset is meticulously labeled with precise 3D joint coordinates relative to the camera's perspective. With its unique blend of front/side and top views—-the latter captured from ceiling-mounted cameras—-the ITOP dataset stands as a significant resource for benchmarking and furthering advancements in viewpoint-independent pose estimation.

The 3D Human Body Model dataset established by the authors of [45] is a synthetic dataset that consists of 20,000 three-dimensional models of human bodies in static poses and an equal gender distribution. It is generated with the STAR parametric model [33]. While generating the models, two primary considerations were maintained: the natural Range of Motion (ROM) for each joint and the prevention of self-intersections in the 3D mesh. Existing research on the human ROM was referenced to define the limitations of joint rotations. Despite adhering to ROM constraints, certain non-idealities sometimes result in self-intersections in areas like the pelvic region, knees, and elbows. To address this, each vertex of the mesh is associated with a specific bone group. Self-intersections between non-adjacent bone groups are considered forbidden, and an algorithm flags such meshes as invalid.

The 3D Poses in the Wild Dataset (3DPW) [46] offers a unique perspective by capturing scenarios in challenging outdoor environments. This extensive dataset encompasses more than 51,000 frames featuring seven different actors donning 18 distinct clothing styles. The data collection process involves the use of a handheld smartphone camera to record the actions of one or two actors. Notably, 3DPW enhances its utility by providing highly accurate mesh ground truth annotations. These annotations are generated by fitting the SMPL model to the raw ground truth markers.

The Max Planck Institute for Informatics (MPII) dataset serves to evaluate the accuracy of articulated human pose estimation. This dataset comprises approximately 25,000 images, featuring annotations for over 40,000 individuals, including their body joints. These images were systematically compiled, capturing a wide array of everyday human activities. In total, the dataset encompasses 410 different human activities, with each image labeled according to the specific activity depicted. The images are extracted from YouTube videos.

In addition to the well-known datasets mentioned earlier, there exist several others employed in the 3D reconstruction methods reviewed. A comprehensive list of these datasets, including the ones previously mentioned, is presented in Table 1. These datasets are compared based on the availability of various types of data, such as RGB images, frame sequences, depth maps, multi-view perspectives, 2D poses, 3D poses, and 3D body meshes. The respective papers that utilize each dataset are also referenced. The dash indicates missing or unspecified types of data. In instances where RGB frame sequences are provided, it is automatically assumed that RGB images are also accessible.

Table 1. Datasets and reviewed papers that have referred to them

Dataset	RGB Image	Frame Sequence	Depth	Multi-View	2D Pose	3D Pose	3D Mesh	Used in References
Human 3.6 M [34]	✓	✓	✓	✓	✓	✓	✓	[47–58]
MPI-INF-3DHP [35]	✓	✓	-	✓	-	✓	-	[47–50,55,59]
Synthetic Humans for Real Tasks (SURREAL) [36]	✓	✓	✓	-	✓	✓	✓	[49,58,60,61]
Dynamic Fine Alignment Using Scan Texture (DFAUST) [37]	✓	✓	✓	✓	-	✓	✓	[58,62,63]
Microsoft Common Objects in Context (MS COCO) [38]	✓	-	-	-	-	-	-	[47–50,52,57]
Leeds Sports Pose (LSP) [39]	✓	-	-	-	✓	-	-	[47–50,55]
Leeds Sports Pose Extended (LSPe) [40]	✓	-	-	-	✓	-	-	[47–50,52,55]
Bodies Under Flowing Fashion (BUFF) [41]	✓	✓	-	✓	-	-	✓	[62,64]
HumanEva [42]	✓	✓	-	✓	-	✓	-	[51,54,56]
Human–Human–Object Interaction (HHOI) [43]	✓	✓	✓	✓	-	✓	-	[51]
Invariant Top View (ITOP) [44]	-	✓	✓	✓	-	✓	-	[65]
3D Poses in the Wild Dataset (3DPW) [46]	✓	✓	-	-	-	✓	✓	[50,52,55,59]
People Snapshot [62]	✓	✓	-	-	-	✓	✓	[66,67]
Unite the People [68]	✓	-	-	-	✓	✓	✓	[48,60]
Max Planck Institute for Informatics (MPII) [69]	✓	✓	-	-	✓	✓	✓	[47,48,50,52,55]
Polarization Human Shape and Pose Dataset (PHSPD) [70]	✓	✓	✓	✓	✓	✓	✓	[61]
Thuman2.0 [71]	-	-	-	-	-	-	✓	[72]
3D Occlusion Human (3DOH) [73]	✓	-	-	✓	✓	✓	✓	[52]
Expressive Hands and Faces (EHF) [74]	✓	-	-	-	✓	-	✓	[75]
Sports Shape and Pose 3D (SSP3D) [76]	✓	✓	-	-	✓	✓	✓	[75]
Articulated dataset [77]	✓	✓	-	✓	-	✓	✓	[63,78]
Clothed Auto Person Encoding (CAPE) [79]	✓	✓	-	✓	-	✓	✓	[78]
WCPA [80]	✓	-	✓	-	-	✓	✓	[81]
Human Multiview Behavioral Imaging (HUMBI) [82]	✓	✓	-	✓	✓	✓	✓	[53]
Carnegie Mellon University (CMU) [83]	✓	✓	-	✓	✓	-	-	[56]
Dynacap [84]	✓	✓	-	✓	✓	✓	✓	[85]
DeepCap [86]	✓	✓	-	✓	-	✓	✓	[85]
Multiperson Pose Test Set in 3D (MuPoTS-3D) [87]	✓	✓	-	✓	✓	✓	-	[59]
Advanced Industrial Science and Technology (AIST) [88]	✓	✓	-	✓	-	-	-	[59]
University of British Columbia (UBC 3V) [89]	-	✓	✓	✓	-	✓	-	[65]
MonoPerfCap [90]	✓	✓	-	✓	-	-	✓	[64]
SAIL-VOS 3D (S3D) [91]	✓	✓	✓	✓	-	-	✓	[92]
ZJU-MoCap [93]	✓	✓	-	✓	-	-	-	[67,85]

4. Evaluation Metrics

Applying evaluation metrics is crucial to quantitatively assessing the reconstruction quality of generated models in the field of 3D human body reconstruction. Here, we briefly introduce some of the most commonly utilized metrics for this purpose. These metrics provide a standardized and objective means of evaluating the accuracy and fidelity of reconstructed 3D human body models, ensuring a reliable assessment of the reconstruction process.

- Mean Per-Joint Position Error (MPJPE)

The MPJPE [58] is a common metric that evaluates the performance of human pose estimation algorithms. It measures the mean distance in mm between the skeleton joints of the ground truth 3D pose and the joints from the estimated pose. The formulation is provided in Equation (12):

$$E_{MPJPE} = \frac{1}{N_S} \sum_{i=1}^{N_S} \left\| m_{f,S}^{(f)}(i) - m_{g_t,S}^{(f)}(i) \right\|_2 \tag{12}$$

where N_S corresponds to the total number of skeleton joints, $m_{f,S}^{(f)}(i)$ is a function that returns the coordinates of the i-th joint of skeleton S in frame f, and $m_{g_t,S}^{(f)}(i)$ is a function that refers to the i-th joint of the skeleton in the ground truth frame. A commonly used modification of the MPJPE is the Procrustes-aligned MPJPE (PA-MPJPE), which is calculated in a similar way with the difference that the reconstructed model and the ground truth one are previously aligned using the Procrustes algorithm.

- Mean Average Vertex Error (MAVE)

The MAVE [56] is used to find the averaged distance between the vertices of the reconstructed 3D human model and the vertices of the ground truth data. It is defined by

$$E_{MAVE} = \frac{1}{N} \sum_{i=1}^{N} \sqrt{\left\| \vartheta_i - \bar{\vartheta}_l \right\|_2^2} \tag{13}$$

where N is the total number of vertices of the 3D model, ϑ_i is a vertex from the predicted 3D human body model, and $\bar{\vartheta}_i$ is a vertex from the corresponding ground truth data.

- Chamfer Distance

The symmetric point-to-point Chamfer distance measures the similarity between two point clouds P and Q. A common formulation is given in Equation (14):

$$d_{CD}(P,Q) = \frac{1}{|P|} \sum_{x \in P} \min_{y \in Q} ||x-y||_2^2 + \frac{1}{|Q|} \sum_{y \in Q} \min_{x \in P} ||x-y||_2^2 \tag{14}$$

where x and y are points from P and Q, respectively, and $|P|$ and $|Q|$ are the total number of points in P and Q. The utilization of the min function refers to measuring the distance of a point from one point cloud to its nearest neighbors in the other point cloud. A modified version of the Chamfer distance, where the sum is replaced by a max function, is provided in Equation (15):

$$d_{MCD}(P,Q) = max \left\{ \frac{1}{|P|} \sum_{x \in P} \min_{y \in Q} ||x-y||_2^2, \frac{1}{|Q|} \sum_{y \in Q} \min_{x \in P} ||x-y||_2^2 \right\} \tag{15}$$

- Vertex-to-Surface Distance (VSD)

The VSD metric quantifies the average distance between the vertices of a point cloud and their corresponding points on the surface of a triangular mesh. These surface points

can either be the vertices of the mesh or points that reside on its faces or edges. The authors of [94] incorporate this metric into their 3D model-fitting algorithm, employing a lifted optimization technique. Points on the surface are defined via surface coordinates, represented as $u = \{p, v, w\}$. Here, $p \in \mathbb{N}$ denotes the triangle's index where the point is situated, and $v \in [0, 1]$, $w \in [0, 1 - v]$ are the coordinates within the unit triangle. Therefore, the 3D coordinates of a point on the surface can be defined as shown in Equation (16), where $\mathbf{v}_1, \mathbf{v}_2$ and $\mathbf{v}_2$ are the vertices of the p-th triangle.

$$S(u) = (1 - v - w)\mathbf{v}_1 + v\mathbf{v}_2 + w\mathbf{v}_3 \tag{16}$$

The distance between a point from a point cloud and its correspondence on the surface of the mesh can be further calculated as described in Equation (17), where D is the number of points x_i in the point cloud and $U = \{u_i\}_{i=1}^{D}$ are the surface coordinates of those points.

$$E_{VDS}(U) = \frac{1}{D} \sum_{i=1}^{D} ||S(u_i) - x_i||^2 \tag{17}$$

Within the confines of the algorithm described in [94], the mesh under consideration is parametric. This implies that its vertex coordinates are contingent on the parameter vector, θ. Given that the algorithm adopts lifted optimizations, both θ and the surface coordinates, U, are optimized concurrently.

5. Taxonomy of Existing 3D Human Body Modelling and Reconstruction Techniques

In this section, we identify three distinct types of visual data utilized as input for 3D human body modeling and reconstruction: single or multiple images, video data, and depth map data. Correspondingly, Figure 6 presents a taxonomy of existing 3D human body modeling and reconstruction techniques based on the input data. In essence, 3D human body modeling is a task in computer vision and computer graphics aimed at generating a 3D photorealistic representation of the human body. This task often involves, but is not limited to, processes such as data acquisition, 2D pose estimation through the detection of 2D body joints, camera calibration and triangulation in the case of multiple views, 3D pose estimation to derive the pose of the future 3D model, model shape optimization when using parametric body models, surface reconstruction and texturing for obtaining a detailed 3D representation of the model's surface, and post-processing and refinement to increase the quality of the reconstructed 3D model. However, it is challenging to compose a concrete framework of operations that applies universally across different input data types and reconstruction approaches. Nevertheless, several key steps, which are visualized in Figure 7, are commonly accomplished in most of the examined methods, including 3D pose estimation, coarse 3D shape estimation, 3D shape refinement, and texture recovery. The blocks surrounded with dash lines indicate that the utilization of the parametric body model and the appliance of 3D pose estimation is sometimes omitted by some of the examined algorithms, most of which are deep learning based. The figure also illustrates the different types of input data.

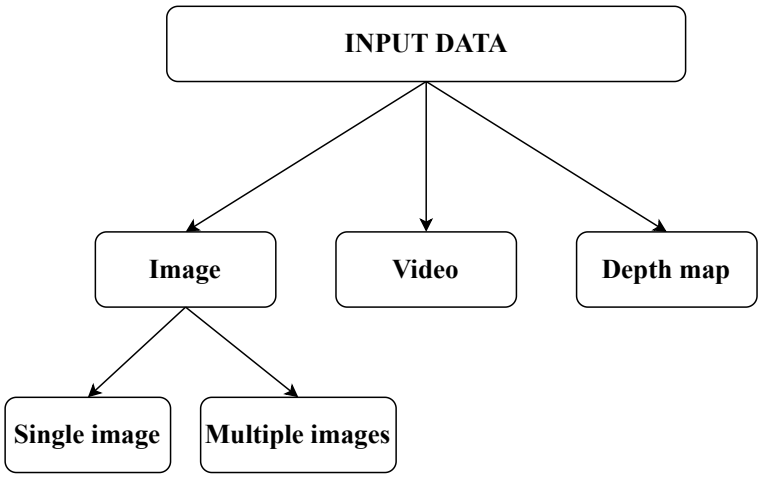

Figure 6. Taxonomy of 3D human body modelling and reconstruction techniques based on input data.

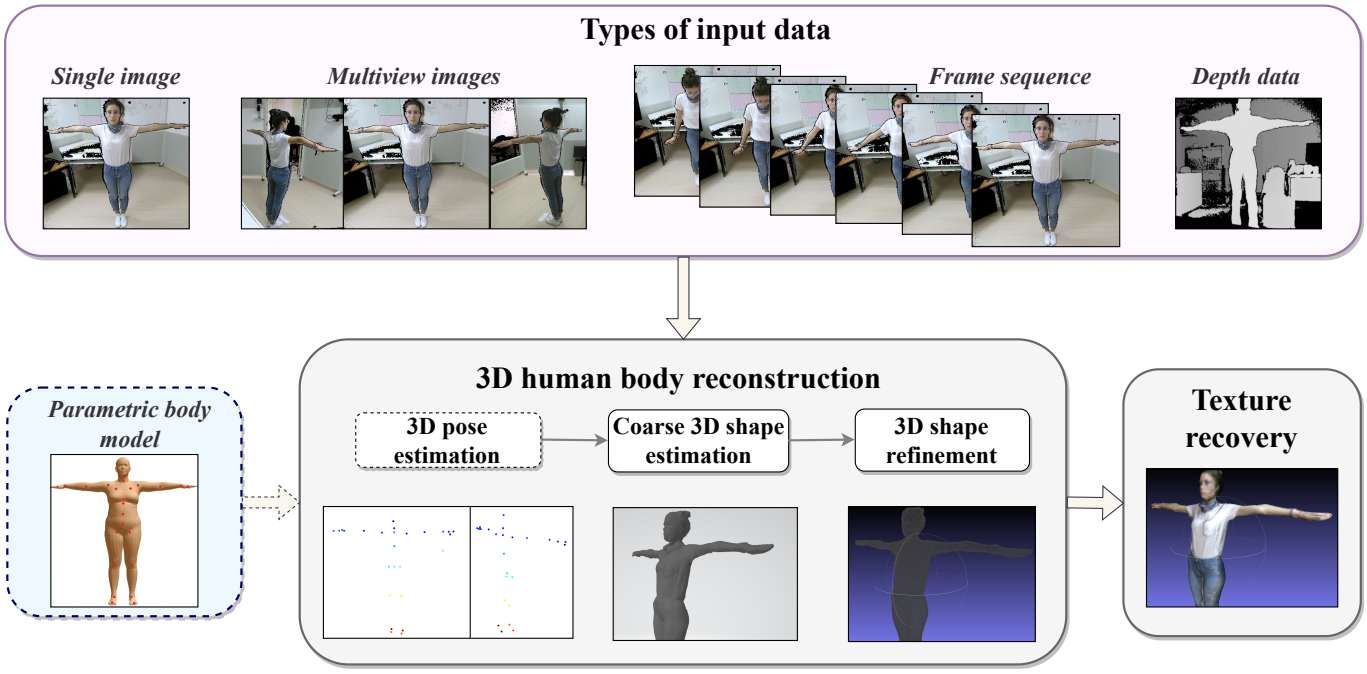

Figure 7. A 3D human body model reconstruction. (Parametric body model [32])

5.1. Image Data

In this subsection, we will explore significant works focused on 3D model reconstruction from 2D images. Initially, we will delve into 3D reconstruction techniques that leverage a single image as input, followed by methods based on multiple images.

5.1.1. Single Image

The authors of [95] focus on estimating both the shape and the pose of a person from a single image, with only a rough estimate of the height. They use a database of over 2400 subjects and utilize SCAPE to build their own parametric human body model. Further, the authors apply "Silhouettes, Edges, and Shading" to create an output 3D model that reflects both the pose and the shape of the human from the input 2D image. The authors derive that shadowing may significantly improve the estimation of human exact measurements.

In [47], an end-to-end framework for reconstructing a full 3D mesh of a human body from a single RGB image is developed. The authors utilize the SMPL model to encode the mesh of a 3D human body. An accent is placed on the 3D body representation, the exploita-

tion of iterative 3D regression with feedback, and on a factorized adversarial prior. Many 2D (LSP, LSPe, MPII, and MS COCO) and 3D (MPI-INF-3DHP and Human 3.6 M) datasets are utilized. The authors report great results for some of the lowest MPJPE values.

BodyNet [60] is an end-to-end trainable neural network that generates a 3D body shape from a single RGB image. The input 2D images are taken from the SURREAL and Unite the People datasets. All of the selected images depict people with different clothing and are snapped from different camera viewpoints. There are four main aspects in the proposed workflow: volumetric inference for 3D human shape, multi-view re-projection loss on the silhouette, multi-task learning with intermediate supervision, and fitting a parametric body model. The SMPL model is fit to the output of BodyNet for evaluation purposes. The approach yields cutting-edge results, and there is a strong belief that it can serve as a trainable building block for upcoming techniques utilizing 3D body data. Surface error, Voxel, and Silhouette Intersection over Union (IoU) are exploited for evaluation metrics.

The authors of [48] retrieve beyond skinned detailed human body shape models from a single image in a coarse-to-fine manner. They combine the robustness of parametric body models with the flexibility of free-form deformations by proposing a novel learning-based framework. Specifically, the SMPL model is utilized for obtaining an initial parametric mesh model whose surface is further defined by performing non-rigid 3D deformation on the mesh. A deep learning approach is applied to each stage of the proposed network. Initially, an SMPL mesh is estimated from the input image. Then, all other stages serve as refinement phases that predict the mesh deformation to finally produce a detailed human shape. The authors use three datasets together to conduct their experiments: the WILD dataset—used for training and testing and constructed by 5 free pre-existing human datasets—Human 3.6 M, MS COCO, LSP, LSPe, MPII, MPI-INF-3DHP, and the Unite the People dataset; the RECON dataset—used for evaluation and constructed by the authors through traditional multi-view 3D reconstruction techniques; and the SYN dataset—also used for evaluation and constructed by the authors by rendering synthetic human mesh models from the PVHM dataset [96]. The authors report results that outperform other SMPL-based approaches when running their three custom-created datasets. However, further improvements are needed to reduce errors in the depth direction.

In [61], a method for 3D human shape reconstruction from a polarization image is proposed. The method is based on a dedicated deep learning approach called Structure from Polarization. It consists of two main stages. The first stage estimates the surface normal from a single polarization image. The second stage estimates the human body shape and pose using the already available surface normal and the raw polarization image. Body shape refinement is also considered. For the body shape and pose estimation, the SMPL model is utilized. The authors use the synthetical SURREAL dataset, as well as one real-world dataset—the Polarization Human Shape and Pose Dataset (PHSPD). Empirical results are derived by using the Mean Angle Error (MAE) for the normal estimation evaluation and the MPJPE metric.

The authors of [49] introduce a method for recovering a complete 3D mesh of a human body from a single image. They develop a deep learning approach based on a generative adversarial network, which consists of a specially designed shape–pose-based generator and a multi-source discriminator. The SMPL model is an important part of the shape–pose-based generator that outputs the generated human body mesh. The training is performed on multiple different datasets, namely LSP, LSPe, MS COCO, MPI-INF-3dHP, Motion and Shape capture (MoSh), SURREAL, and Human 3.6 M. The proposed method is evaluated through pose and segmentation evaluation metrics. Specifically, for the pose evaluation, the MPJPE is utilized.

In [50], the issues related to the absence of high-resolution images for the task of 3D human model reconstruction are addressed by developing an RSC-Net (RSC stands for resolution-aware network, a self-supervision loss, and a contrastive learning scheme), a deep-based resolution-aware network that is able to handle images with arbitrary resolution. An accent is placed on the 3D human pose and shape representation, as SMPL

is utilized. The authors also present a temporal recurrent module that is able to extend the single-image model to low-resolution videos. The model is trained using multiple datasets, such as the Human 3.6 M, MPI-INF-3DHP, LSP, LSPe, MPII, and MS COCO datasets. The evaluation results obtained using the MPJPE and PA-MPJPE demonstrate better performance among other algorithms.

The work of [51] proposes a pose grammar for a 3D human body model recovering in a natural way. The method takes an estimated 2D pose as an input and learns a generalized 2D–3D mapping function for 3D pose estimation. The proposed deep grammar network consists of two important components: a base 3D pose network that encodes appearance and geometry features from the input image and the detected 2D pose and a 3D pose grammar network, based on bi-directional recurrent neural network that encodes human body dependencies and relations. The authors use the Human 3.6 M, HumanEva [42], and HHOI databases. In order to generate additional training samples, they also utilize a novel Pose Sample Simulator. The results are given as a comparison between the estimated pose and the ground truth in mm through the Average Euclidean Distance.

In [72], a novel 3D object representation, specifically aimed at enhancing the efficiency and accuracy of monocular real-time 3D human reconstruction, named the Fourier Occupancy Field (FOF) is introduced. Specifically, the FOF presents a 3D object as a 2D field, converting the occupancy field along the z-axis into a Fourier series and retaining only the initial few terms. It demonstrates the capability to represent high-quality 3D human geometries using a 2D map aligned with images, bridging the gap between 2D image data and 3D geometries. Experimental validation was conducted using both publicly available datasets (Thuman2.0 and Twindom) and real-world captured data. The VSD and Chamfer distance are used for evaluating the results.

The authors of [52] introduce POCO (pose and shape estimation with confidence). It is a novel framework aimed at addressing the challenge of 3D human pose and shape estimation from 2D images, while also providing a measure of uncertainty in its estimations. The model infers both body parameters, specifically leveraging SMPL parameters, and the accompanying regression uncertainty in a single feed-forward network pass. POCO is based on a dual conditioning strategy that includes an image-conditioned base density function and a pose-conditioned scale. The model is trained on the MS COCO, Human 3.6 M, MPI-INF-3D, MPII, and LSPe datasets. An evaluation is conducted on the 3DPW, 3D Occlusion Human (3DOH), and 3DPW-OCC datasets. The MPJPE, PA-MPJPE, and Per-vertex error (PVE) are used as evaluation metrics.

In [75], a methodology for estimating whole-body human parameters from a single image, addressing the challenges of monocular human body estimation in wild conditions, is presented. The method, called "KBody", employs a predict-and-optimize approach that seeks to balance three traits, pose, shape, and pixel alignment, while also effectively managing partial images. KBody's methodology aims to improve fitting quality via the introduction of virtual joints, which are tailored to fit estimated data and facilitate a harmonious interaction with silhouette constraints. Further, to manage images with missing information, the method utilizes an appearance-prior approach, completing them in a structurally plausible way. A variant of SMPL, SMPL-X [74], is utilized. Performance is assessed via the Procrustes-aligned vertex-to-vertex error (PA-V2V), scale-corrected per-vertex Euclidean error in a neutral pose (PVE-T-SC) [76], and IoU. The expressive Hands and Faces (EHF) and Sports Shape and Pose 3D (SSP3D) datasets are used for conducting the experiments.

Table 2 summarizes the examined papers for 3D human body modeling and reconstruction based on single-image input data and compares them by assets, constraints, the utilization of parametric models, datasets, and evaluation metrics.

Table 2. Comparison of 3D modelling approaches based on single-image input.

Research	Year	Main Focus	Assets	Constraints	Parametric Model	Dataset	Evaluation Metric
[95]	2009	Computing parametric body shape from shading	Minimal human intervention; extracting body shape from paintings	Limited lighting conditions; impossibility of recovering hair and clothes	SCAPE	Not specified	Not specified
[47]	2018	Human mesh recovery of 3D joint angles and body shape	Precise joints locations; no requirement for 2D to 3D paired data	Additional processing requirement for obtaining better results; impossibility of recovering hair and clothes	SMPL	LSP, LSPe, MPII, MS COCO, Human 3.6 M, MPI-INF-3DHP	MPJPE, PA-MPJPE
[60]	2018	Inference of volumetric human body shape directly from a single image	Fully automated end-to-end prediction system; functioning as a trainable building block	Impossibility of recovering hair; low results accuracy after the segmentation step	SMPL (only for evaluation purposes)	SURREAL, Unite the People	Voxel IoU, Silhouette IoU, Surface error
[48]	2019	Coarse-to-fine refinement of parametric 3D model composed from a single image	Exploitation of custom build datasets	Pose ambiguities; Large errors in body mesh prediction	SMPL	Human 3.6 M, MS COCO, LSP, LSPe, MPII, MPI-INF-3DHP, Unite the People, RECON, SYN	Silhouette IoU, 2D joint error, 3D error (MAVE but with nearest neighbors)
[61]	2020	3D human body model reconstruction from a polarized image using synchronized cameras	Providing geometric details of the surface; Obtaining more reliable depth maps	Need of a polarization cameras; Limited datasets with polarized images	SMPL	SURREAL, PHSPD	MAE, MPJPE
[49]	2021	Introduction of Generative Adversarial Networks for human mesh reconstruction	Detailed body shape; Real-time solution; Possibility for use with video data; Discriminator used for reality check	Impossibility of recovering hair and clothes; Getting faster and better results with a pretrained generator	SMPL	LSP, LSPe, MS COCO, MPI-INF-3dHP, MoSh, SURREAL, Human 3.6 M	MPJPE
[50]	2021	3D model reconstruction from low-resolution images	Possibility for training with all kinds of image resolutions; textured 3D model in color	Impossibility of recovering long or voluminous hair; easily affected by noise	SMPL	Human 3.6 M, MPI-INF-3DHP, LSP, LSPe, MPII, MS COCO	MPJPE, PA-MPJPE
[51]	2021	Introduction of a pose grammar for achieving better 3D human body model representation	Enforcing high-level constraints over human poses	Not specifying body shape recovering; requirement for different types of data for achieving better results	Not used	Human 3.6 M, HumanEva, HHOI	Average Euclidean Distance
[72]	2022	3D geometry representation for monocular real-time and accurate human reconstruction	FOF for representing high-quality 3D human geometries using a 2D map aligned with images	FOF inability for representing too-thin objects	SMPL (partially)	Thuman2.0, Twindom	VSD, Chamfer distance

Table 2. *Cont.*

Research	Year	Main Focus	Assets	Constraints	Parametric Model	Dataset	Evaluation Metric
[52]	2023	3D pose and shape estimation with confidence	3D human pose and shape estimation from 2D images, while providing a measure of pose uncertainty	Not providing shape uncertainty	SMPL	MS COCO, Human 3.6 M, MPI-INF-3D, MPII, LSPe, 3DPW, 3DOH, 3DPW-OCC	MPJPE, PA-MPJPE, PVE
[75]	2023	Framework for estimating whole-body human parameters from a single image	3D estimation on human body shape and pose; handling images with missing information	Image-based appearance-prior technique for completion coming with limitations for non-frontal facing images	SMPL-X	EHF, SSP3D	PA-V2V, PVE-T-SC, IoU

5.1.2. Multiple Images

In [97], a CNN-based approach for accurate 3D human body reconstruction from silhouettes is proposed. The authors contribute with the creation of extensive, realistic synthetic data at a larger scale; the adoption of a multi-task learning strategy for the prediction of multiple outputs, including shape, 3D joint positions, pose angles, and body volume; and the introduction of a novel network architecture that incorporates known body measurements (e.g., height) and per-pixel segmentation confidence as additional inputs. The SMPL parametric model is utilized. The authors conduct their experiments on the CAESAR dataset [98] and assess the achieved results by leveraging the mean distance as an evaluation metric.

An approach using multiple images and angles for 3D human body modelling is developed by [99], and a method for automatic 3D character reconstruction from frontal and lateral monocular 2D RGB views is proposed. The template mesh of the SMPL model is used in the first stage for obtaining a body model from the frontal view. Then, this modified SMPL model is inputted into a second stage, where it is further modified by the lateral view. The method focuses on two main aspects: the shape and the texture of the model. Their custom dataset consists of front-view and side-view photos of people.

In the work presented in [78], the authors suggest a method for reconstructing 3D human body models from multiple images. This approach involves learning an implicit function for representing 3D shapes, relying on multi-scale features derived from multi-stage end-to-end neural networks. Since the approach excludes the utilization of a geometrical prior derived from parametric human body models, the current approach is considered model-free. The experiments are conducted over two datasets, which are the Articulated dataset and Clothed Auto Person Encoding (CAPE) dataset [79]. A quantitative evaluation is performed on both datasets using the VSD, Chamfer distance, and IoU.

The authors of [53] introduce a multi-view human body mesh translator (MMT) model for 3D human body mesh estimation. Specifically, it is a non-parametric deep-learning-based model that leverages a vision transformer. It performs feature-level fusion, which combines multi-view features to generate contextualized embeddings for the purpose of decoding the output mesh representation. Consequently, the MMT takes multiple images as input and fuses their features in both the encoding and the decoding stages. As a result, a representation embedded with global information for the human body model is composed. Experiments are conducted on the Human 3.6 M and Human Multiview Behavioral Imaging (HUMBI) datasets, and the model performance is assessed by utilizing the MPJPE, PA-MPJPE, and MPVE evaluation metrics.

In [66], a novel meta-optimization technique is introduced that is specifically designed to navigate scenarios wherein accurate initial guesses (e.g., certain poses and shapes at

specific camera angles) are not available for rendering and reconstructing 3D human figures. The covariance matrix adaptation annealing method is utilized, allowing for the easy incorporation of domain knowledge of hierarchical human anatomy. The SMPL model is used. The authors conducted their experiments on the People Snapshot Dataset [62] and Human 3.6. Further, reprojection errors and the MPJPE are employed for the result evaluation.

In [81], the authors focus on 3D clothed human body reconstruction based on multiple views and poses. They benefit from the geometry prior provided by the SMPLX model in order to learn the latent codes of a posed mesh by taking multiple images as an input. WCPA dataset is used for training and testing and quantitative evaluation is performed via calculating the Chamfer distance of different strategies on the test dataset.

Table 3 summarizes the examined papers for 3D human body modeling and reconstruction based on multiple-image input data and compares them by assets, constraints, the utilization of parametric models, datasets, and evaluation metrics.

Table 3. Comparison of 3D modelling approaches based on multiple images input.

Research	Year	Main Focus	Assets	Constraints	Parametric Model	Dataset	Evaluation Metric
[97]	2019	3D human body reconstruction from silhouettes	A supervised-learning-based approach utilizing CNNs for 3D human body recovering	Need to increase the range of acceptable poses and camera viewpoints while maintaining the same performance	SMPL	CAESAR	Mean distance
[99]	2020	Reconstruction from two points of view: frontal and lateral	Processing one body side at a time; tackling the problem of self-occlusions	Negative effect of lighting over the accuracy of the reconstructed model; custom dataset	SMPL	Custom dataset	Not specified
[78]	2021	3D human body reconstruction from multiple images	Learning model-free implicit function for 3D human body representation relying on multi-scale features	Not optimised generalization results due to training set limitations	Not used	Articulated dataset, CAPE	VSD, Chamfer distance, IoU
[53]	2022	3D human body mesh recovery via MMT model	Utilization of a non parametric deep-learning-based model (MMT) leveraging a Vision Transformer and applying feature-level fusion	Exploited evaluation metrics—still rough to appropriately assess the reconstruction ability of the model; slower performance than the parametric models	Not used	Human 3.6 M, HUMBI	MPJPE, PA-MPJPE, MPVE
[66]	2022	Meta-optimization technique for 3D human rendering and reconstruction	The approach—designed for scenarios where accurate initial guesses are not available	Long execution time and slow convergence	SMPL	People-Snapshot Dataset, Human3.6.	Reprojection errors, MPJPE
[81]	2022	3D clothed human body reconstruction based on multiple views and pose	Deep-learning approach incorporating the SMPLX model and non-parametric implicit function learning	Not specified	SMPLX	WCPA	Chamfer distance

5.2. Video Data

The process of 3D human body reconstruction from video data involves the process of generating a 3D model of a person's body by analyzing a video sequence. This technique leverages multiple frames of a person's movements to create an accurate and detailed representation of their body shape and posture.

The authors of [54] implement a fully convolutional model for 3D pose estimation from sequences of 2D key points. Exploiting temporal convolutions is very important for modelling time dependencies among series. The authors employ dilated convolutions that facilitate the modelling of long-term time dependencies while maintaining efficiency. Further, a semi-supervised training approach is introduced for settings where labeled data of 3D ground truth pose is not available. Two video datasets for training the model are utilized: Human 3.6 M and HumanEva-I. The evaluation of the proposed method is performed mainly by computing the MPJPE and its variants.

In [62], a method for reconstructing a textured 3D human body model from a single monocular video of a moving person is introduced. The goal of the authors is to generate a personalized human avatar of the captured subject that correctly reflects its body shape, hair, and clothes. Building a textured map and an underlying skeleton rigged to the surface is also considered. The method combines three important steps: pose reconstruction using the SMPL model, consensus shape estimation via transforming the collection of dynamic body poses into a common reference frame, and frame refinement and texture map generation. The experiments are performed on the DFAUST and BUFF datasets. Since these datasets consist of 3D scans of moving people, a virtual camera rendering 2D video sequences is implemented. The VSD is used as a quantitative evaluation metric.

The approach for 3D human modelling that the authors of [55] propose is to combine the accuracy of the optimization-based methods with the promptness of the deep-based regression methods. They introduce SPIN: SMPL optimization in the loop. This approach utilizes a deep neural network to initialize an iterative optimization routine for fitting the SMPL parametric model to 2D joints within the training loop. Already-fitted estimates of the model are subsequently used to supervise the network. The experiments are conducted on multiple datasets, such as Human 3.6 M, MPI-INF-3DHP, LSP, LSPe, 3DPW, MPII, COCO. The mean reconstruction error, MPJPE, Area Under the Curve (AUC), and Percentage of Correct Key points (PCK) evaluation metrics are exploited.

MotioNet [56] is a deep neural network that performs a 3D human motion reconstruction from a monocular video. The authors claim that this method is the first data-driven approach that directly outputs a kinematic skeleton, which can be used for motion representation. The motion datasets that are used in this work are the Carnegie Mellon University (CMU) (containing 2605 captured elementary actions and dance moves performed by 144 subjects), Human 3.6 M, and HumanEva datasets. The results are organized by the specific motion, and the current approach declares one of the lowest MPJPEs.

In [57], a 3D uplifting model for the purposes of trying on clothes virtually in real-time is introduced. This functionality is applicable to e-commerce and other fashion-related purposes. To keep the system's universality, the authors developed it to be compatible with conventional devices like smartphones and tablets. This implies that their approach is limited to monocular cameras or RGB video streams only. The framework consists of the following steps: skeleton reconstruction and pose estimation, human body recovery and adjustment to the estimated pose, garment mapping and reshaping, the projection of the result to a real-time image, and pose refinement and model alignment for proper body overlay. The datasets used during the implementation are MS COCO and Human 3.6 M. The achieved quality and quickness of the results depend on the visual characteristics of the input data, such as the image contrast and the color palette.

Implementing methods for 3D human body model reconstruction from low-resolution video is valuable. The authors of [50] upgrade their method for reconstructing models from a single image with arbitrary resolution quality to reconstruction from video, again with arbitrary resolution quality. The RSC-Net, used for single-image inputs, is extended

by incorporating a temporal post-processing step in order to handle video inputs. The 3DPW, Human 36 M, InstaVariety [100], and MPI-INF-3DHP datasets of video sequences are utilized for conducting the experiments. The quantitative evaluation is performed using the MPJPE, PA-MPJPE, and acceleration error.

The authors of [85] propose a methodology for modelling animatable human avatars with dynamic garments. Their method is implemented by applying a Neural Radiance Field (NeRF)-based representation and managing cloth deformations on multiple hierarchical levels, all while utilizing a conditional variational auto-encoder to discern node-related variables for facilitating realistic and dynamic animation. The model is trained end-to-end using only RGB videos. The SMPL model is utilized. The datasets that are used are Dynacap, DeepCap, and ZJU-MoCap. The Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) are used for the quantitative evaluation.

In [67], a method for human reconstruction and synthesis from monocular RGB videos is explored, which is found challenging due to issues like clothing texture, occlusions, and pose changes. The authors counter the common use of NeRFs and implicit methods, which are often chosen for their ability to represent clothed humans. Their approach is based on the optimization of an SMPL+D mesh and the utilization of a multi-resolution texture representation using only RGB images, binary silhouettes, and sparse 2D key points as inputs. The method demonstrates enhanced capability in capturing geometric details compared to traditional visual hull mesh-based methods. It also shows notable improvements and speedups in novel pose synthesis compared to NeRF-based methods, without the latter's typical, unwanted artifacts. Experiments are conducted on the ZJU-MoCap, People-Snapshot, and Self-Recon datasets. For the geometry reconstruction evaluation, the Chamfer distance and VSD metric are utilized.

The authors of [59] employ PoseBERT, a transformer-based module for temporal 3D human modelling using monocular RGB videos. The SMPL parametric model is utilized. The AMASS dataset [101] is used for training the model. For evaluation purposes, the 3DPW, MPI-INF-3DHP, Multiperson Pose Test Set in 3D (MuPoTS-3D), and Advanced Industrial Science and Technology (AIST) datasets are employed. The MPJPE, PA-MPJPE, and MPVE are the main metrics used for assessing the achieved results.

Table 4 summarizes the examined papers for 3D human body modeling and reconstruction based on video input data and compares them by assets, constraints, the utilization of parametric models, datasets, and evaluation metrics.

Table 4. Comparison of 3D modelling approaches based on video input

Research	Year	Main Focus	Assets	Constraints	Parametric Model	Dataset	Evaluation Metric
[54]	2018	Temporal convolutions and semi-supervised training on video for 3D pose estimation	Training with a small amount of labeled data; using the model when motion capture is challenging	Complicated because of the number of the model's layers; not estimating body shape	Not used	Human 3.6 M, HumanEva-I	MPJPE
[62]	2018	Obtaining a textured 3D model from a monocular video of a moving person	Reconstructing 3D model with detailed hair, body, clothes, and kinematic skeleton	Limited reconstruction of self-occluded zones; less accurate results due to fast movements	SMPL	DFAUST, BUFF	VSD
[55]	2019	SMPL model optimization in the loop	Self-improving training process; possibility for training in the absence of 3D annotations	High complexity and impossibility of real-time implementations	SMPL	Human 3.6 M, MPI-INF-3DHP, LSP, LSPe, 3DPW, MPII, COCO	Mean reconstruction error, MPJPE, AUC, PCK

Table 4. *Cont.*

Research	Year	Main Focus	Assets	Constraints	Parametric Model	Dataset	Evaluation Metric
[56]	2020	Generating an accurate skeleton from monocular videos	Creation of a more natural human motion movement; tackling the self-occlusion problem	No body shape estimation; large effects of camera movements over the global positioning of the joints	Not used	CMU, Human 3.6 M, HumanEva	MPJPE
[57]	2020	System functionality for real-time garment overlay	Possibility of real-time reconstruction; framework for garment overlay	Impossibility of recovering hair and clothes; limited body mesh projection optimization	SMPL	MS COCO, Human 3.6 M	Not specified
[50]	2021	3D human model reconstruction from low-resolution videos	Applicable to low-resolution videos; better accuracy compared to the case with an image as input	Impossibility of recovering long or voluminous hair; limited textured 3D model in color; slower implementation	SMPL	3DPW, Human36M, InstaVariety, MPI-INF-3DHP	MPJPE, PA-MPJPE, ACC.
[59]	2022	3D pose estimation from monocular RGB videos	Exploiting a generic transformer module	Performance degradation in case of fast human motion or long-term occlusions	SMPL	AMASS, 3DPW, MPI-INF-3DHP, MuPoTS-3D, AIST	MPJPE, PA-MPJPE, MPVE
[85]	2022	A methodology for modelling animatable human avatars with dynamic garments	Recreating appearance and motion by leveraging neural scene representation while explicitly accounting for the motion hierarchy of clothes	Method performance depending on pose variance of the training data; assumption of accurate body pose estimation for the training images	SMPL	Dynacap, DeepCap, ZJU-MoCap	PSNR, SSIM
[67]	2023	A methodology for human geometry and realistic textures recovering from a monocular RGB video	SMPL+D mesh optimization and utilization of a multi-resolution texture representation using RGB images, binary silhouettes, and sparse 2D keypoints	Not specified	SMPL	ZJU-MoCap, People-Snapshot, Self-Recon	Chamfer distance, VSD

5.3. Depth Map Data

For the purposes of 3D human body reconstruction, some approaches [58,63–65,92] exploit depth maps that are generated by specific systems. These kinds of systems use structured light or the Time of Flight principle to measure the depth of an object, which shows how far from the system the object is. Released in 2010 for gaming purposes, Microsoft Kinect v1 has led a large number of academics to explore its possibilities outside of just the video gaming experience. It uses a structured light method in which the radiation is a known sparkle pattern on the scene. Dissimilar to it, Microsoft Kinect V2 utilizes the Time of Flight method, in which the entire scene is flooded with light, and the depth is determined by the time it takes each photon particle to return to the sensor [102].

In [63], a deep-learning-based approach for human body reconstruction from a single RGB image in a calibration-free context is proposed. The novelty of the method is the way the system is trained, in which a multi-view analysis of depth images is benefited from.

By switching between two modes, RGB and D, or, in other words, by alternating between RGB and depth data, the functionality of the deep model is improved by letting it learn the space of pose and shape deformations. The authors use MPI-INF-3DHP, DFAUST, and the Articulated datasets. They also collected their own samples by using a calibrated multi-Kinect setup (five mesh sequences containing around 250 frames each). The evaluation metric that they employ is the IoU. Finally, the authors derive that the use of multiple views and depth information in the training process is critical for obtaining an accurate reconstruction of the human body.

The authors of [65] developed a method for 3D human pose estimation from depth maps. They name their model Deep Depth Pose, and it can adopt a 3D pose as a linear combination of adapted prototype poses. Particularly, the model is defined as a CNN that takes as an input the depth map containing a person and a set of predefined prototype human body poses and returns the computed 3D positions of the person's body joints. The architecture is also enabled to work with multiple views, providing more accurate estimations. The datasets used for training the model are ITOP and University of British Columbia (UBC 3V) Hard-pose. The evaluation metrics applied are the average error, computed by comparing the estimated and the ground truth 3D joint locations (MPJPE), and the mean average precision (mAP), and AUC). The experimental results indicate great method performance and achievement of state-of-the-art outcomes on both datasets according to precision and decrease in the average error.

In [58], a technique for 3D human body pose and shape estimation based on a single depth image is introduced. Considering the joint features and original depth features, the method incorporates a spatial attention feature extractor that is able to capture local spatial features of the depth images and the 3D joints. The authors implement a weakly supervised mechanism based on the SMPL model for achieving better efficiency on real-depth data. They also add a differentiable rendering layer utilized for the transformation of the 3D models into silhouettes and depths. The experiments are conducted on the SURREAL, Human 3.6 M, and DFAUST databases, as well as on some real depth images. The results demonstrate the high efficiency of the proposed technique by comparing the reconstruction errors (mm) and MAVEs with those of many other state-of-the-art methods on the above datasets.

The authors of [64] propose PeelHuman—a novel shape representation of the human body that is robust to severe self-occlusions. To achieve this, they compose an end-to-end pipeline method called PeelGAN that reconstructs a textured human model from a single RGB image using an adversarial approach. Two types of depth data are also exploited: Peeled Depth and RGB maps. The authors also introduce their own 3D dataset consisting of 2000 multiple human body model sequences that vary in shape and pose and are recorded by a calibrated setup of four Microsoft Kinect sensors. Except for their custom dataset, two additional datasets are also exploited: BUFF and MonoPerfCap. A comparison to other methods is carried out by calculating the Chamfer distance.

In [92], a novel method, termed the "occupancy planes (OPlanes) representation", is introduced. It is an approach for the 3D reconstruction from a single-view RGB-D image that involves the creation of multiple image-like planes (OPlanes). These planes slice through the camera's view frustum, indicating occupancy at every pixel location for the respective 3D point. Notably, OPlanes allows for the adaptive adjustment of the number and location of the occupancy planes during both inference and training, thus providing a resolution flexibility that is superior to that of conventional voxel grid representations. Evaluations of the introduced approach were conducted on the SAIL-VOS 3D (S3D) dataset using the IoU, Chamfer distance, and Normal Consistency, revealing improvements over the preceding reconstruction efforts, particularly in scenarios featuring occluded or partially visible humans.

Table 5 summarizes the examined papers for 3D human body modeling and reconstruction based on single image depth map data and compares them by assets, constraints, utilization of parametric models, datasets, and evaluation metrics.

Table 5. Comparison of 3D modelling approaches based on depth map data

Research	Year	Main Focus	Assets	Constraints	Parametric Model	Dataset	Evaluation Metric
[63]	2018	Textured 3D human body reconstruction from a single RGB image and co-learning with Microsoft Kinect depth images	Considering depth information during the training process	Partially handling non-rigid deformations	Not used (not used for reconstruction, but SMPL dataset is used)	MPI-INF-3DHP, DFAUST, the Articulated dataset	IoU
[65]	2018	3D human pose estimation from depth maps using a Deep Learning approach	Possibility for using data from both single and multiple viewpoints; no demands on pixel-wise segmentation and temporal information	No body shape estimation; additional noise when using images in the wild	Not used	ITOP and UBC 3V Hard-pose	MPJPE, mAP, AUC
[58]	2020	3D human body pose and shape estimation from a single depth image	Possibility of using the model with real depth data achieved by the incorporated weakly supervised mechanism	Complicated with several functionality stages	SMPL (in the training process)	SURREAL, Human 3.6 M, DFAUST	MAVE
[64]	2020	Creating 3D human body representation from a set of Peeled Depth and RGB maps	Tackling severe self-occlusions; handling images wide assortment of shapes, poses, and textures	Not providing full body shape	Not used	BUFF, MonoPerfCap, Custom dataset	Chamfer distance
[92]	2023	3D reconstruction from a single-view RGB-D image through creating multiple image-like planes (OPlanes)	Exploitation of spatial correlations between adjacent locations within a plane, appropriate particularly for occluded or partially visible humans	Not specified	Not used	S3D	IoU, Chamfer distance, Normal Consistency

6. Discussion

Several significant conclusions can be drawn from a comprehensive analysis of the reviewed papers. Based on the nature of the input data, i.e., single or multiple images, video data, or depth maps, we can categorize three primary approaches to 3D human body reconstruction, which form the foundation of our taxonomy. While there may be variations within these approaches, and even among different methods within a specific approach, they share some essential elements in the processing context. These common features include data acquisition and the necessity of employing datasets, 3D pose estimation, 3D shape estimation, and possibly texture recovery, regardless of whether a parametric body model serves as a geometric prior. The evaluation of the proposed methods and the resulting outcomes are another common aspect among the examined works.

Following another criterion, the reviewed papers can be categorized into another taxonomy that distinguishes between two types of reconstruction methods employed: optimization-based approaches, where a parametric body model is iteratively fitted to certain observations (single image [95], multiple images [66,99], and video [57,62]) and regression-based approaches, which involve training deep neural models to directly produce 3D representations of human bodies, although they can also incorporate geometric priors (single image [47–52,60,61,72,75],

multiple images [53,78,81,97], video data [50,54,56,59,85], and depth map [58,63–65,92]). Notably, some approaches, like the one designed by the authors of [55] (video input), employ a combination of both types of methods by leveraging regression-based prediction to achieve a strong initialization prior to the subsequent optimization step.

We can also categorize the presented methods based on the resulting 3D-type representation of the human body model. While the majority of works focus on reconstructing a 3D body mesh, there are alternative representations that rely on implicit neural representations [81,85]. Additionally, the authors of [72] introduce a unique 3D geometry representation termed the Fourier Occupancy Field.

Further, our investigation revealed the existence of around 35 different datasets currently utilized for the task of 3D human body reconstruction. Notably, some well-known datasets, such as Human 3.6 M, MPI-INF-3DHP, MS COCO, LSP, LSPe, 3DPW, MPII, and SURREAL, are frequently employed in the reviewed papers. Additionally, many of the methods make use of multiple datasets to ensure more generalized results. However, some works advocate for the expansion of data usage to enhance method performance [51,61,78,97,99].

In terms of the utilization of parametric body models, nearly all the papers rely on the SMPL model (single image [47–50,52,61,72], multiple images [66,97,99], video [50,57,59,62,67,85], and depth map data [58]), with two of them using its modified version, SMPL-X (single image [75], and multiple images [81]). Except for [95] (a single image), which uses SCAPE, the remaining methods do not depend on statistical body models for a geometric prior, which unsurprisingly are all regression-based (single image [51,60], multiple images [53,78], video [54,56], and depth map data [63–65,92]).

Regarding the use of quantitative evaluation metrics, it was observed that the most commonly used metrics for pose estimation evaluations are the MPJPE and PA-MPJPE. Additionally, the VSD and Chamfer distance are frequently exploited for comparing a reconstructed 3D human model with the ground truth.

According to the quality of the avatar's appearance, recovering color, texture, hair, and garments is essential. However, only a few works attempt to recover these aspects. Specifically, there are limited works addressing the recovery of hair [62], color and texture [50,63,67], and clothes [62,81]. Conversely, some works acknowledge these challenges and declare the impossibility of recovering hair [47,49,50,57,60,95] and clothing [57]. Nevertheless, the challenge of accurately capturing real individuals' interactions remains largely unresolved, as the majority of works primarily focus on achieving realistic avatar appearances.

Reconstructing a complete body model also involves addressing issues related to missing information and self-occlusions. A subset of papers [56,64,75,92,99] confront these challenges, achieving promising results. Another challenge arises from the limitations imposed by lighting conditions [95,99], which can hinder the reconstruction process. Consequently, certain works either do not succeed in reconstructing the full body shape [64] or do not prioritize body shape recovery at all [51,52,54,56,65]. However, their contributions to body pose estimation are included in this review, as they represent a crucial step in the 3D reconstruction process.

Finally, the real-time implementation of algorithms for human body reconstruction presents a notable challenge among the reviewed papers. Only two of the examined methods declare real-time implementation [49,57]. Conversely, other works report extended execution times [53,55,66] and cite a substantial processing overhead [47,50,58]. This high computational demand poses a significant obstacle, especially in applications like HTC. An approach such as the one in [103] may be applicable, but it is not included in this review because it is based on a conceptual idea that is still not implemented in practice.

7. Conclusions

In this paper, we conducted a comprehensive review of existing methodologies for 3D human body recovery. Our approach to establishing a taxonomy for 3D reconstruction techniques was primarily based on the nature of the input data. Specifically, we categorized

3D human body reconstruction into three distinct categories: those reliant on single or multiple image data, video data, and depth map data. Each of the methods that we examined was thoroughly assessed in the context of the datasets employed, the utilization of geometric priors through parametric body models, and the evaluation metrics applied. Additionally, we provided insights into the strengths and limitations associated with each approach. Subsequently, we performed an in-depth analysis of the reviewed methods.

In conclusion, while considerable progress has been made in the field of 3D human body recovery and reconstruction, it is yet to be fully optimized for applications like HTC. Achieving realism in avatars must extend beyond merely replicating the appearance of real individuals. There is a need for further research and development to enable avatars to effectively convey authentic emotions and interactions, all of which must occur in real time.

Author Contributions: Conceptualization, methodology, R.P., I.B., D.N. and I.V.; writing—original draft preparation, R.P. and I.B.; writing—review and editing, A.M.; supervision, A.M.; project administration, A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the contract KP-06-H37/8 from 6 December 2019 for the research project: "Inference algorithms for semantic knowledge extraction based on deep architectures for context-aware holographic communication" of the Bulgarian Research Fund of the Ministry of Education and Science, Bulgaria.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Written informed consent has been obtained from the patient(s) to publish this paper.

Acknowledgments: The authors acknowledge the support of the R&D Teleinfrastructure Laboratory at the Technical University of Sofia, Bulgaria.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

3DOH	3D Occlusion Human
3DPW	3D Poses in the Wild
AIST	Advanced Industrial Science and Technology
AR	Augmented Reality
AUC	Area Under the Curve
BMI	Body Mass Index
BUFF	Bodies Under Flowing Fashion
CAPE	Clothed Auto Person Encoding
CMU	Carnegie Mellon University
DFAUST	Dynamic Fine Alignment Using Scan Texture
EHF	Expressive Hands and Faces
FOF	Fourier Occupancy Field
HHOI	Human-Human-Object Interaction
HTC	Holographic-Type Communication
HUMBI	HUman Multiview Behavioral Imaging
IoU	Intersection over Union
ITOP	Invariant-Top View
LSP	Leeds Sports Pose
LSPe	Leeds Sports Pose extended
MAE	Mean Angle Error
mAP	mean Average Precision
MAVE	Mean Average Vertex Error
MMT	Multi-view human body Mesh Translator
MoSh	Motion and Shape capture

MPII	Max Planck Institute for Informatics
MPJPE	Mean Per Joint Position Error
MR	Mixed Reality
MS COCO	MicroSoft Common Objects in COntext
MuPoTS-3D	Multiperson Pose Test Set in 3D
NeRF	Neural Radiance Field
OPlanes	Occupancy Planes
PA-MPJPE	Procrustes Aligned MPJPE
PA-V2V	Procrustes-Aligned Vertex-to-Vertex
PCA	Principal Component Analysis
PCK	Percentage of Correct Key points
PHSPD	Polarization Human Shape and Pose Dataset
POCO	Pose and shape estimation with confidence
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PSNR	Peak Signal-to-Noise Ratio
PVE	Per-Vertex Error
PVE-T-SC	Per-Vertex Euclidean error in a neutral (T) pose
RGB	Reed Green Blue
RGB-D	Reed Green Blue—Depth
ROM	Range of Motion
RSC	Resolution-aware network, a Self-supervision loss, and a Contrastive learning scheme
S3D	SAIL-VOS 3D
SCAPE	Shape Completion and Animation of People
SSIM	Structural Similarity Index
SMPPL	Skinned Multi-Person Linear Model
SSP3D	Sports Shape and Pose 3D
STAR	Sparse Trained Articulated Human Body Regressor
SURREAL	Synthetic hUmans foR REAL tasks
UBC	University of British Columbia
VR	Virtual Reality
VSD	Vertex to Surface Distance

References

1. Manolova, A.; Tonchev, K.; Poulkov, V.; Dixir, S.; Lindgren, P. Context-aware holographic communication based on semantic knowledge extraction. *Wirel. Pers. Commun.* **2021**, *120*, 2307–2319. [CrossRef]
2. Haleem, A.; Javaid, M.; Khan, I.H. Holography applications toward medical field: An overview. *Indian J. Radiol. Imaging* **2020**, *30*, 354–361. [CrossRef] [PubMed]
3. Jumreornvong, O.; Yang, E.; Race, J.; Appel, J. Telemedicine and medical education in the age of COVID-19. *Acad. Med.* **2020**, *95*, 1838–1843. [CrossRef] [PubMed]
4. Nayak, S.; Patgiri, R. 6G communication technology: A vision on intelligent healthcare. *Health Inform. Comput. Perspect. Healthc.* **2021**, 1–18. [CrossRef]
5. Ahmad, H.F.; Rafique, W.; Rasool, R.U.; Alhumam, A.; Anwar, Z.; Qadir, J. Leveraging 6G, extended reality, and IoT big data analytics for healthcare: A review. *Comput. Sci. Rev.* **2023**, *48*, 100558.
6. Ahad, A.; Tahir, M. Perspective—6G and IoT for Intelligent Healthcare: Challenges and Future Research Directions. *ECS Sens. Plus* **2023**, *2*, 011601. [CrossRef]
7. Bucioli, A.A.; Cyrino, G.F.; Lima, G.F.; Peres, I.C.; Cardoso, A.; Lamounier, E.A.; Neto, M.M.; Botelho, R.V. Holographic real time 3D heart visualization from coronary tomography for multi-place medical diagnostics. In Proceedings of the 2017 IEEE 15th International Conference on Dependable, Autonomic and Secure Computing, 15th International Conference on Pervasive Intelligence and Computing, 3rd International Conference on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), Orlando, FL, USA, 6–10 November 2017; pp. 239–244.
8. Sirilak, S.; Muneesawang, P. A new procedure for advancing telemedicine using the HoloLens. *IEEE Access* **2018**, *6*, 60224–60233. [CrossRef]
9. Choi, P.J.; Oskouian, R.J.; Tubbs, R.S.; Choi, P.J.K. Telesurgery: Past, present, and future. *Cureus* **2018**, *10*, e2716. [CrossRef]
10. Barkhaya, N.M.M.; Abd Halim, N.D. A review of application of 3D hologram in education: A meta-analysis. In Proceedings of the 2016 IEEE 8th International Conference on Engineering Education (ICEED), Kuala Lumpur, Malaysia, 7–8 December 2016; pp. 257–260.
11. Ramachandiran, C.R.; Chong, M.M.; Subramanian, P. 3D hologram in futuristic classroom: A review. *Period. Eng. Nat. Sci.* **2019**, *7*, 580–586. [CrossRef]

12. Ahmad, A.S.; Alomaier, A.T.; Elmahal, D.M.; Abdlfatah, R.F.; Ibrahim, D.M. EduGram: Education Development Based on Hologram Technology. *Int. J. Online Biomed. Eng.* **2021**, *17*, 32–49. [CrossRef]
13. Yoo, H.; Jang, J.; Oh, H.; Park, I. The potentials and trends of holography in education: A scoping review. *Comput. Educ.* **2022**, *186*, 104533. [CrossRef]
14. Hughes, A. Death is no longer a deal breaker: The hologram performer in live music. *Future Live Music* **2020**, 114–128. Available Online: https://books.google.bg/books?id=QB3LzQEACAAJ (accessed on 20 October 2023).
15. Matthews, J.; Nairn, A. Holographic ABBA: Examining Fan Responses to ABBA's Virtual "Live" Concert. *Pop. Music Soc.* **2023**, 1–22. [CrossRef]
16. Rega, F.; Saxena, D. Free-roam virtual reality: A new avenue for gaming. In *Advances in Augmented Reality and Virtual Reality*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 29–34.
17. Fanini, B.; Pagano, A.; Pietroni, E.; Ferdani, D.; Demetrescu, E.; Palombini, A. Augmented Reality for Cultural Heritage. In *Springer Handbook of Augmented Reality*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 391–411.
18. Banfi, F.; Pontisso, M.; Paolillo, F.R.; Roascio, S.; Spallino, C.; Stanga, C. Interactive and Immersive Digital Representation for Virtual Museum: VR and AR for Semantic Enrichment of Museo Nazionale Romano, Antiquarium di Lucrezia Romana and Antiquarium di Villa Dei Quintili. *ISPRS Int. J. Geo Inf.* **2023**, *12*, 28. [CrossRef]
19. Meng, Y.; Mok, P.Y.; Jin, X. Interactive virtual try-on clothing design systems. *Comput. Aided Des.* **2010**, *42*, 310–321. [CrossRef]
20. Santesteban, I.; Otaduy, M.A.; Casas, D. Learning-based animation of clothing for virtual try-on. *Proc. Comput. Graph. Forum* **2019**, *38*, 355–366. [CrossRef]
21. Zhao, F.; Xie, Z.; Kampffmeyer, M.; Dong, H.; Han, S.; Zheng, T.; Zhang, T.; Liang, X. M3d-vton: A monocular-to-3d virtual try-on network. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 13239–13249.
22. Cheng, Z.Q.; Chen, Y.; Martin, R.R.; Wu, T.; Song, Z. Parametric modeling of 3D human body shape—A survey. *Comput. Graph.* **2018**, *71*, 88–100. [CrossRef]
23. Chen, L.; Peng, S.; Zhou, X. Towards efficient and photorealistic 3d human reconstruction: A brief survey. *Vis. Inform.* **2021**, *5*, 11–19. [CrossRef]
24. Correia, H.A.; Brito, J.H. 3D reconstruction of human bodies from single-view and multi-view images: A systematic review. *Comput. Methods Programs Biomed.* **2023**, *239*, 107620. [CrossRef]
25. Tian, Y.; Zhang, H.; Liu, Y.; Wang, L. Recovering 3D human mesh from monocular images: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, 1–25. [CrossRef]
26. Sun, M.; Yang, D.; Kou, D.; Jiang, Y.; Shan, W.; Yan, Z.; Zhang, L. Human 3D avatar modeling with implicit neural representation: A brief survey. In Proceedings of the 2022 14th International Conference on Signal Processing Systems (ICSPS), Zhenjiang, China, 18–20 November 2022; pp. 818–827.
27. Page, M.; McKenzie, J.; Bossuyt, P.; Boutron, I.; Hoffmann, T.; Mulrow, C.; Shamseer, L.; Tetzlaff, J.; Akl, E.; Brennan, S.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Syst. Rev.* **2021**, *10*, 89. [PubMed]
28. Christoff, N. Modeling of 3D Human Body for Photorealistic Avatar Generation: A Review. In Proceedings of the iCEST, Ohrid, North Macedonia, 27–29 June 2019.
29. Zhou, S.; Fu, H.; Liu, L.; Cohen-Or, D.; Han, X. Parametric reshaping of human bodies in images. *ACM Trans. Graph.* **2010**, *29*, 1–10. [CrossRef]
30. Omran, M.; Lassner, C.; Pons-Moll, G.; Gehler, P.; Schiele, B. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In Proceedings of the 2018 International Conference on 3D Vision (3DV), Verona, Italy, 5–8 September 2018; pp. 484–494.
31. Anguelov, D.; Srinivasan, P.; Koller, D.; Thrun, S.; Rodgers, J.; Davis, J. Scape: Shape completion and animation of people. *ACM SIGGRAPH Pap.* **2005**, 408–416. [CrossRef]
32. Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; Black, M.J. SMPL: A skinned multi-person linear model. *ACM Trans. Graph.* **2015**, *34*, 1–16. [CrossRef]
33. Osman, A.A.; Bolkart, T.; Black, M.J. Star: Sparse trained articulated human body regressor. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Part VI 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 598–613.
34. Ionescu, C.; Papava, D.; Olaru, V.; Sminchisescu, C. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 1325–1339.
35. Mehta, D.; Rhodin, H.; Casas, D.; Fua, P.; Sotnychenko, O.; Xu, W.; Theobalt, C. Monocular 3d human pose estimation in the wild using improved cnn supervision. In Proceedings of the 2017 International Conference on 3D Vision (3DV), Qingdao, China, 10–12 October 2017; pp. 506–516.
36. Varol, G.; Romero, J.; Martin, X.; Mahmood, N.; Black, M.J.; Laptev, I.; Schmid, C. Learning from synthetic humans. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 109–117.
37. Bogo, F.; Romero, J.; Pons-Moll, G.; Black, M.J. Dynamic FAUST: Registering human bodies in motion. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6233–6242.

38. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Part V 13; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
39. Johnson, S.; Everingham, M. Clustered pose and nonlinear appearance models for human pose estimation. *Proc. BMVC* **2010**, *2*, 5.
40. Johnson, S.; Everingham, M. Learning effective human pose estimation from inaccurate annotation. In Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011; pp. 1465–1472.
41. Zhang, C.; Pujades, S.; Black, M.J.; Pons-Moll, G. Detailed, accurate, human shape estimation from clothed 3D scan sequences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4191–4200.
42. Sigal, L.; Balan, A.O.; Black, M.J. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *Int. J. Comput. Vis.* **2010**, *87*, 4–27.
43. Shu, T.; Ryoo, M.S.; Zhu, S.C. Learning social affordance for human-robot interaction. *arXiv* **2016**, arXiv:1604.03692.
44. Haque, A.; Peng, B.; Luo, Z.; Alahi, A.; Yeung, S.; Fei-Fei, L. Towards viewpoint invariant 3d human pose estimation. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Part I 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 160–177.
45. Bozhilov, I.; Tonchev, K.; Manolova, A.; Petkova, R. 3d human body models compression and decompression algorithm based on graph convolutional networks for holographic communication. In Proceedings of the 2022 25th International Symposium on Wireless Personal Multimedia Communications (WPMC), Herning, Denmark, 30 October–2 November 2022; pp. 532–537.
46. Von Marcard, T.; Henschel, R.; Black, M.J.; Rosenhahn, B.; Pons-Moll, G. Recovering accurate 3d human pose in the wild using imus and a moving camera. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 601–617.
47. Kanazawa, A.; Black, M.J.; Jacobs, D.W.; Malik, J. End-to-end recovery of human shape and pose. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 7122–7131.
48. Zhu, H.; Zuo, X.; Wang, S.; Cao, X.; Yang, R. Detailed human shape estimation from a single image by hierarchical mesh deformation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4491–4500.
49. Gao, R.; Wen, M.; Park, J.; Cho, K. Human Mesh Reconstruction with Generative Adversarial Networks from Single RGB Images. *Sensors* **2021**, *21*, 1350.
50. Xu, X.; Chen, H.; Moreno-Noguer, F.; Jeni, L.A.; De la Torre, F. 3D human pose, shape and texture from low-resolution images and videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 4490–4504. [CrossRef]
51. Xu, Y.; Wang, W.; Liu, T.; Liu, X.; Xie, J.; Zhu, S.C. Monocular 3d pose estimation via pose grammar and data augmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 6327–6344. [CrossRef] [PubMed]
52. Dwivedi, S.K.; Schmid, C.; Yi, H.; Black, M.J.; Tzionas, D. POCO: 3D Pose and Shape Estimation with Confidence. *arXiv* **2023**, arXiv:2308.12965.
53. Jiang, X.; Nie, X.; Wang, Z.; Liu, L.; Liu, S. Multi-view Human Body Mesh Translator. *arXiv* **2022**, arXiv:2210.01886.
54. Pavllo, D.; Feichtenhofer, C.; Grangier, D.; Auli, M. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7753–7762.
55. Kolotouros, N.; Pavlakos, G.; Black, M.J.; Daniilidis, K. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2252–2261.
56. Shi, M.; Aberman, K.; Aristidou, A.; Komura, T.; Lischinski, D.; Cohen-Or, D.; Chen, B. Motionet: 3d human motion reconstruction from monocular video with skeleton consistency. *ACM Trans. Graph.* **2020**, *40*, 1–15.
57. Makarov, I.; Chernyshev, D. Real-time 3D model reconstruction and mapping for fashion. In Proceedings of the 2020 43rd International Conference on Telecommunications and Signal Processing (TSP), Milan, Italy, 7–9 July 2020; pp. 133–138.
58. Liu, L.; Wang, K.; Yang, J. 3D Human Body Shape and Pose Estimation from Depth Image. In Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision (PRCV), Shenzhen, China, 14–17 October 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 410–421.
59. Baradel, F.; Brégier, R.; Groueix, T.; Weinzaepfel, P.; Kalantidis, Y.; Rogez, G. PoseBERT: A Generic Transformer Module for Temporal 3D Human Modeling. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, 1–16. [CrossRef]
60. Varol, G.; Ceylan, D.; Russell, B.; Yang, J.; Yumer, E.; Laptev, I.; Schmid, C. Bodynet: Volumetric inference of 3d human body shapes. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 20–36.
61. Zou, S.; Zuo, X.; Qian, Y.; Wang, S.; Xu, C.; Gong, M.; Cheng, L. 3D human shape reconstruction from a polarization image. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Part XIV 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 351–368.
62. Alldieck, T.; Magnor, M.; Xu, W.; Theobalt, C.; Pons-Moll, G. Video based reconstruction of 3d people models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8387–8397.
63. Venkat, A.; Jinka, S.S.; Sharma, A. Deep textured 3d reconstruction of human bodies. *arXiv* **2018**, arXiv:1809.06547.

64. Jinka, S.S.; Chacko, R.; Sharma, A.; Narayanan, P. Peeledhuman: Robust shape representation for textured 3d human body reconstruction. In Proceedings of the 2020 International Conference on 3D Vision (3DV), Fukuoka, Japan, 25–28 November 2020; pp. 879–888.
65. Marin-Jimenez, M.J.; Romero-Ramirez, F.J.; Munoz-Salinas, R.; Medina-Carnicer, R. 3D human pose estimation from depth maps using a deep combination of poses. *J. Vis. Commun. Image Represent.* **2018**, *55*, 627–639.
66. Lu, Y.; Chen, G.; Pang, C.; Zhang, H.; Zhang, B. Subject-Specific Human Modeling for Human Pose Estimation. *IEEE Trans. Hum.-Mach. Syst.* **2022**, *53*, 54–64. [CrossRef]
67. Jena, R.; Chaudhari, P.; Gee, J.; Iyer, G.; Choudhary, S.; Smith, B.M. Mesh Strikes Back: Fast and Efficient Human Reconstruction from RGB videos. *arXiv* **2023**, arXiv:2303.08808.
68. Lassner, C.; Romero, J.; Kiefel, M.; Bogo, F.; Black, M.J.; Gehler, P.V. Unite the people: Closing the loop between 3d and 2d human representations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6050–6059.
69. Andriluka, M.; Pishchulin, L.; Gehler, P.; Schiele, B. 2d human pose estimation: New benchmark and state of the art analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 3686–3693.
70. Zou, S.; Zuo, X.; Qian, Y.; Wang, S.; Guo, C.; Xu, C.; Gong, M.; Cheng, L. Polarization human shape and pose dataset. *arXiv* **2020**, arXiv:2004.14899.
71. Yu, T.; Zheng, Z.; Guo, K.; Liu, P.; Dai, Q.; Liu, Y. Function4D: Real-time human volumetric capture from very sparse consumer rgbd sensors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 5746–5756.
72. Feng, Q.; Liu, Y.; Lai, Y.K.; Yang, J.; Li, K. FOF: Learning fourier occupancy field for monocular real-time human reconstruction. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 7397–7409.
73. Zhang, T.; Huang, B.; Wang, Y. Object-occluded human shape and pose estimation from a single color image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 7376–7385.
74. Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A.A.; Tzionas, D.; Black, M.J. Expressive body capture: 3d hands, face, and body from a single image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10975–10985.
75. Zioulis, N.; O'Brien, J.F. KBody: Towards General, Robust, and Aligned Monocular Whole-Body Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 6214–6224.
76. Sengupta, A.; Budvytis, I.; Cipolla, R. Synthetic training for accurate 3d human pose and shape estimation in the wild. *arXiv* **2020**, arXiv:2009.10013.
77. Vlasic, D.; Baran, I.; Matusik, W.; Popović, J. Articulated mesh animation from multi-view silhouettes. *ACM Siggraph Pap.* **2008**, 1–9. [CrossRef]
78. Li, Z.; Oskarsson, M.; Heyden, A. Learning to Implicitly Represent 3D Human Body From Multi-scale Features and Multi-view Images. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 8968–8975.
79. Ma, Q.; Yang, J.; Ranjan, A.; Pujades, S.; Pons-Moll, G.; Tang, S.; Black, M.J. Learning to dress 3d people in generative clothing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6469–6478.
80. ECCV 2022 WCPA Challenge: From Face, Body and Fashion to 3D Virtual Avatars. Available online: https://tianchi.aliyun.com/competition/entrance/531958/introduction (accessed on 20 October 2023).
81. Chen, J.; Yi, W.; Wang, T.; Li, X.; Ma, L.; Fan, Y.; Lu, H. Pixel2ISDF: Implicit Signed Distance Fields Based Human Body Model from Multi-view and Multi-pose Images. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–24 October 2022; Springer: Berlin/Heildeberg, Germany, 2022; pp. 366–375.
82. Yu, Z.; Yoon, J.S.; Lee, I.K.; Venkatesh, P.; Park, J.; Yu, J.; Park, H.S. Humbi: A large multiview dataset of human body expressions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2990–3000.
83. De la Torre, F.; Hodgins, J.; Bargteil, A.; Martin, X.; Macey, J.; Collado, A.; Beltran, P. Guide to the Carnegie Mellon University Multimodal Activity (Cmu-Mmac) Database. 2009. Available online: https://www.ri.cmu.edu/pub_files/pub4/de_la_torre_frade_fernando_2008_1/de_la_torre_frade_fernando_2008_1.pdf (accessed on 20 October 2023).
84. Habermann, M.; Liu, L.; Xu, W.; Zollhoefer, M.; Pons-Moll, G.; Theobalt, C. Real-time deep dynamic characters. *ACM Trans. Graph.* **2021**, *40*, 1–16.
85. Zheng, Z.; Huang, H.; Yu, T.; Zhang, H.; Guo, Y.; Liu, Y. Structured local radiance fields for human avatar modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 8–24 June 2022; pp. 15893–15903.
86. Habermann, M.; Xu, W.; Zollhofer, M.; Pons-Moll, G.; Theobalt, C. Deepcap: Monocular human performance capture using weak supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5052–5063.

87. Mehta, D.; Sotnychenko, O.; Mueller, F.; Xu, W.; Sridhar, S.; Pons-Moll, G.; Theobalt, C. Single-shot multi-person 3d pose estimation from monocular rgb. In Proceedings of the 2018 International Conference on 3D Vision (3DV), Verona, Italy, 5–8 September 2018; pp. 120–130.
88. Tsuchida, S.; Fukayama, S.; Hamasaki, M.; Goto, M. AIST Dance Video Database: Multi-Genre, Multi-Dancer, and Multi-Camera Database for Dance Information Processing. *Proc. ISMIR* **2019**, *1*, 6.
89. Shafaei, A.; Little, J.J. Real-time human motion capture with multiple depth cameras. In Proceedings of the 2016 13th Conference on Computer and Robot Vision (CRV), Victoria, BC, Canada, 1–3 June 2016; pp. 24–31.
90. Xu, W.; Chatterjee, A.; Zollhöfer, M.; Rhodin, H.; Mehta, D.; Seidel, H.P.; Theobalt, C. Monoperfcap: Human performance capture from monocular video. *ACM Trans. Graph.* **2018**, *37*, 1–15.
91. Hu, Y.T.; Wang, J.; Yeh, R.A.; Schwing, A.G. Sail-vos 3D: A synthetic dataset and baselines for object detection and 3D mesh reconstruction from video data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1418–1428.
92. Zhao, X.; Hu, Y.T.; Ren, Z.; Schwing, A.G. Occupancy planes for single-view rgb-d human reconstruction. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 3633–3641.
93. Peng, S.; Zhang, Y.; Xu, Y.; Wang, Q.; Shuai, Q.; Bao, H.; Zhou, X. 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, Nashville, TN, USA, 20–25 June 2021; pp. 9054–9063.
94. Shen, J.; Cashman, T.J.; Ye, Q.; Hutton, T.; Sharp, T.; Bogo, F.; Fitzgibbon, A.; Shotton, J. The phong surface: Efficient 3D model fitting using lifted optimization. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Part I 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 687–703.
95. Guan, P.; Weiss, A.; Balan, A.O.; Black, M.J. Estimating human shape and pose from a single image. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 1381–1388.
96. Zhu, H.; Su, H.; Wang, P.; Cao, X.; Yang, R. View extrapolation of human body from a single image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4450–4459.
97. Smith, B.M.; Chari, V.; Agrawal, A.; Rehg, J.M.; Sever, R. Towards accurate 3D human body reconstruction from silhouettes. In Proceedings of the 2019 International Conference on 3D Vision (3DV), Quebec City, QC, Canada, 16–19 September 2019; pp. 279–288.
98. Robinette, K.M.; Blackwell, S.; Daanen, H.; Boehmer, M.; Fleming, S.; Brill, T.; Hoeferlin, D.; Burnsides, D. Civilian American and European Surface Anthropometry Resource (CAESAR), Final Report, Volume I: Summary. Sytronics Inc Dayton Oh. 2002. Available online: https://www.humanics-es.com/CAESARvol1.pdf (accessed on 20 October 2023).
99. Beacco, A.; Gallego, J.; Slater, M. Automatic 3D character reconstruction from frontal and lateral monocular 2d rgb views. In Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 2785–2789.
100. Kanazawa, A.; Zhang, J.Y.; Felsen, P.; Malik, J. Learning 3d human dynamics from video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5614–5623.
101. Mahmood, N.; Ghorbani, N.; Troje, N.F.; Pons-Moll, G.; Black, M.J. AMASS: Archive of motion capture as surface shapes. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5442–5451.
102. Sarbolandi, H.; Lefloch, D.; Kolb, A. Kinect range sensing: Structured-light versus Time-of-Flight Kinect. *Comput. Vis. Image Underst.* **2015**, *139*, 1–20.
103. Petkova, R.; Poulkov, V.; Manolova, A.; Tonchev, K. Challenges in Implementing Low-Latency Holographic-Type Communication Systems. *Sensors* **2022**, *22*, 9617.