The Visual Computer
https://doi.org/10.1007/s00371-024-03641-7

REVIEW

Innovative AI techniques for photorealistic 3D clothed human reconstruction from monocular images or videos: a survey

Shuo Yang[1] · Xiaoling Gu[1] · Zhenzhong Kuang[1] · Feiwei Qin[1] · Zizhao Wu[1]

Accepted: 3 September 2024

Abstract

The reconstruction of high-quality 3D clothed humans from monocular images or videos has gained popularity in recent years due to its significant practical applications. While several surveys have addressed the reconstruction of full-body parametric human models from images or videos, this survey specifically delves into the challenges and methodologies of reconstructing 3D clothed humans. It covers both pose-dependent and dynamic approaches to clothed human reconstruction. Regarding pose-dependent clothed human reconstruction from monocular images, we investigate methodologies that employ regression models trained on high-quality 3D scans to estimate human geometry with clothing. Additionally, we explore research leveraging texture priors within large-scale diffusion models to enhance the inference of human appearance in occluded or unseen areas. In terms of dynamic clothed human reconstruction from monocular and sparse multi-view videos, we analyze human modeling techniques utilizing neural radiance fields and 3D Gaussian representations, which employ deformation fields to capture human movements across frames. Furthermore, we provide an overview of the datasets and commonly used quantitative evaluation metrics in these studies. Finally, we conclude by discussing open issues and proposing future research directions in the realistic reconstruction of clothed humans, emphasizing areas that warrant additional investigation.

Keywords Clothed human reconstruction · NeRF · 3D Gaussian splatting · SMPL

1 Introduction

The field of 3D computer vision and computer graphics has long recognized the importance of research efforts in 3D human reconstruction. 3D human reconstruction aims to estimate the geometry and appearance of humans. This area of study has considerable applications across various domains such as virtual reality/augmented reality (VR/AR) [1], online meetings [2], virtual fitting [3], and gaming [4]. However, generating high-fidelity 3D human models typically requires complex multi-view systems, time-consuming offline processing, or manual design using software tools [5, 6]. These methods often pose challenges, particularly for individuals lacking specialized expertise, thereby rendering the process cost-prohibitive. In recent years, researchers have increasingly turned to data-driven and deep learning approaches to enhance various types of clothed human representations [7–11], making the creation of 3D clothed human models more accessible for widespread application.

Photorealistic reconstruction of clothed humans from monocular images and videos presents numerous challenges, primarily due to the complexity of human poses and the diversity of clothing in real-world scenarios. Firstly, varied human representations involve different degrees of compromise concerning storage capacity, expressiveness, manipulability, and compatibility with existing tools. For instance, parametric human models excel at efficiently representing the human body in diverse poses and shapes, providing a high level of operational flexibility [12, 13]. However, they often lack the capability to accurately depict surface details of the human body and the various clothing styles worn on it [14, 15].

✉ Xiaoling Gu
guxl@hdu.edu.cn

Shuo Yang
yangshuo@hdu.edu.cn

Zhenzhong Kuang
zzkuang@hdu.edu.cn

Feiwei Qin
qinfeiwei@hdu.edu.cn

Zizhao Wu
wuzizhao@hdu.edu.cn

[1] HangZhou DianZi University, Hangzhou 310018, Zhejiang Province, China

Published online: 26 September 2024

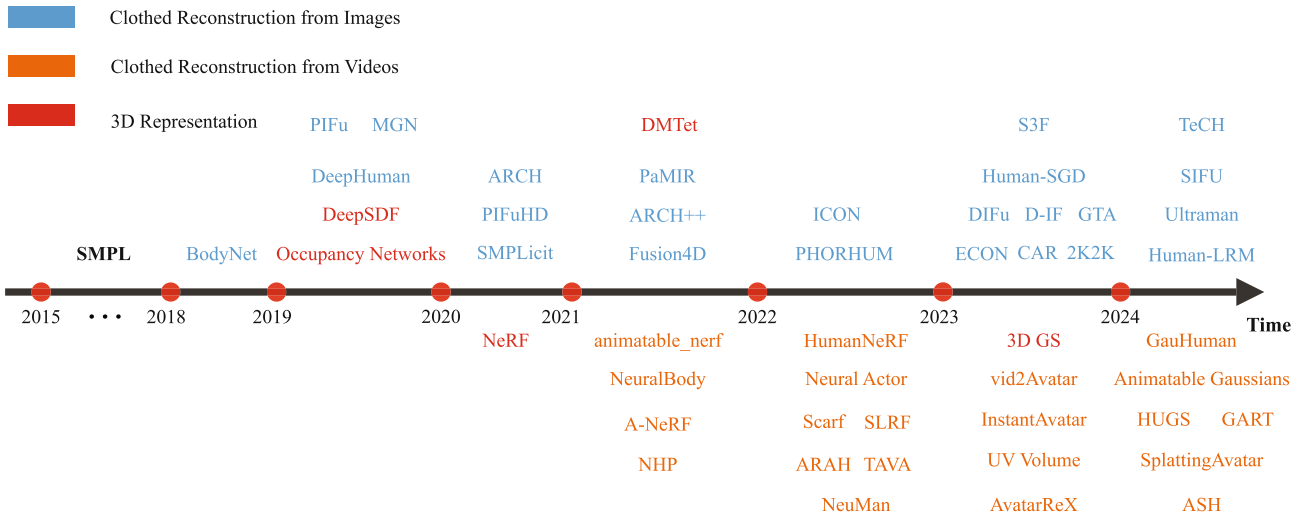

Fig. 1 Representative works on clothed human reconstruction from monocular images or videos over time. Blue text indicates studies focusing on monocular images, orange text denotes studies using video, and red text highlights research on 3D representation

Secondly, models trained on a restricted set of 3D human scans frequently display incomplete geometric surfaces and discontinuities in the reconstructed body parts and clothing, particularly when inferring human shapes with complex poses and voluminous garments. This issue primarily stems from the overfitting of trained models to the particular distributions of human poses and clothing found within the training set of 3D scans [7–9, 16]. Lastly, achieving dynamic reconstruction of clothed humans for a specific subject not only demands consistent representation across various poses in the video but also necessitates the reconstruction of human motion. Typically, this motion reconstruction involves modeling deformation fields learned from videos, which rely on linear blend skinning (LBS) transformations of the skeletal structure [10, 11]. However, accurately reconstructing non-rigid deformations, like clothing, is often challenging and prone to producing artifacts.

1.1 Scope

In this survey, we concentrate on the reconstruction of high-quality 3D clothed humans from monocular images and videos. Figure 1 outlines the timeline of two grouped approaches: *clothed human reconstruction from images* and *clothed human reconstruction from videos*. Note that this survey excludes research focusing on the reconstruction of specific human body parts, such as hands [17–19] and head avatar [20–22], as well as human parametric model estimation [13, 23, 24], and traditional reconstruction methods using multi-view sensor camera systems [5, 6]. Similarly, research on 3D human reconstruction involving depth information inputs [25, 26] will not be discussed.

Related survey Prior to our investigation, several surveys have explored the reconstruction of 3D humans from images or videos [27–29]. Tian et al. [27] provide an overview of research on estimating human shape and pose from monocular images. Sun et al. [29] present a concise overview of prevalent implicit neural representation techniques employed in reconstructing human bodies, hands, and heads. Chen et al. [28] provide a brief overview of traditional reconstruction pipelines, regression-based models, and optimization-based methods for realistic clothed human reconstruction. In contrast, our survey focuses specifically on methodologies for reconstructing 3D clothed humans from monocular images and videos in practical scenarios. We explore a broader range of human representation techniques, including deformable meshes and dual depth maps, which are not covered by Chen et al. [28]. Additionally, we cover recent advancements in computer vision relevant to human reconstruction, such as leveraging diffusion model priors and exploring 3D Gaussian representations. Our review also includes an analysis of datasets and quantitative evaluation metrics essential for training and assessing these reconstruction techniques.

1.2 Organization

In Sect. 2, we discuss conventional explicit and implicit representations of 3D humans. Section 3 focuses on clothed human reconstruction from images, employing regression models trained on 3D human scans and texture inference in unobserved regions using diffusion priors. In Sect. 4, we investigate dynamic human reconstruction from monocular videos using NeRF and 3D Gaussian representations. Additionally, this survey offers insights into commonly used datasets for training and evaluation, accompanied by a dis-

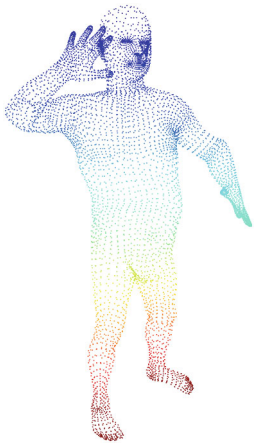

Fig. 2 Example of a point cloud representation of a human

Fig. 3 Example of a mesh representation of a human. The mesh is from the THuman2.0 Dataset [25]

cussion of quantitative evaluation metrics in Sect. 5. Lastly, Sect. 6 summaries this survey and discusses several directions deserving further exploration in the field of clothed human reconstruction.

2 Human representations

Current methods of 3D human representation are primarily categorized into explicit and implicit representations. Explicit representations directly encode geometric information about the human body and include techniques such as point clouds [30, 31], meshes [32, 33], the SMPL model [12], voxels [34, 35], depth maps [36, 37], and 3D Gaussians [38]. Implicit representations encode geometry indirectly by using functions that describe the spatial relationship to the surface. These methods determine whether points are inside or outside the surface, or how far they are from it, through representations such as signed distance functions (SDF) [39] and occupancy fields [40]. Additionally, methods like NeRF [41] use neural networks to implicitly encode both geometry and texture. These advancements enhance the accuracy and fidelity of human modeling, addressing various challenges in capturing and rendering complex human forms.

2.1 Explicit representations

Point clouds A point cloud is a collection of points defined by Cartesian coordinates in Euclidean space [42]. In addition to spatial position coordinates, each point may possess supplementary attributes, such as normal vectors or color for surfaces, as well as opacity or density for volumes (Fig. 2). A point endowed with shape and shading attributes is typically referred to as a surface element or surfel, which approximates a patch of the surface [43]. SCALE [30] and Ma et al. [31] use point cloud representations to explore the modeling and animation of clothed humans.

Mesh A mesh is a common representation of an object's surface, defined by vertices and faces that determine the connectivity of these vertices (Fig. 3). To texture a mesh, the UV texture map is widely used. This technique involves "unwrapping" the 3D surface of the mesh onto a 2D plane, where each vertex is mapped to a corresponding point on the 2D texture image. Parametric human models [12, 44–46] are particularly popular for representing human body meshes and are extensively utilized in studies of human shape and pose estimation.

SMPL The SMPL model [12] is a skinned vertex-based representation of the human body shape and pose, developed from a large dataset of 3D human scans (Fig. 4). It decomposes body shape into two distinct blend shapes: an identity-dependent blend shape and a pose-dependent blend shape. The blend shape is represented as a vertex offset vector corresponding to the template mesh. A standard blend skinning function is applied to the corrective template mesh to obtain the deformed vertices. The SMPL model $M(\beta, \theta; \Phi)$ can be described as follows:

$$M(\beta, \theta; \Phi) = W(T_{\mathrm{P}}(\beta, \theta), J(\beta), \theta, \mathcal{W}) \tag{1}$$

$$T_{\mathrm{P}}(\beta, \theta) = \bar{\mathbf{T}} + B_{\mathrm{S}}(\beta) + B_{\mathrm{P}}(\theta) \tag{2}$$

$$\Phi = \left\{\bar{\mathbf{T}}, \mathcal{W}, \mathcal{S}, \mathcal{J}, \mathcal{P}\right\} \tag{3}$$

where Φ represents the parameters of the SMPL model, $\bar{\mathbf{T}}$ denotes the positions of the template mesh vertices in the zero pose and $T_{\mathrm{P}}(\beta, \theta)$ represents the positions of each vertex in the mesh after incorporating shape and pose offsets. $B_{\mathrm{S}}(\beta)$ and $B_{\mathrm{P}}(\theta)$ correspond to the shape blend shape and pose blend shape, which are vectors of vertices representing offsets from the template. β and θ correspond to linear coefficients representing different body shapes and poses, respectively. $\mathcal{S}$ and $\mathcal{P}$ represent the orthogonal principal components of the shape and pose offsets, respectively. $J(\beta)$ denotes a function used to calculate the positions of the body joints, based on shape. $\mathcal{J}$ represents a matrix that transforms rest vertices into rest joints. $\mathcal{W}$ represents a set of blend skinning weights, and W denotes the standard blend skinning

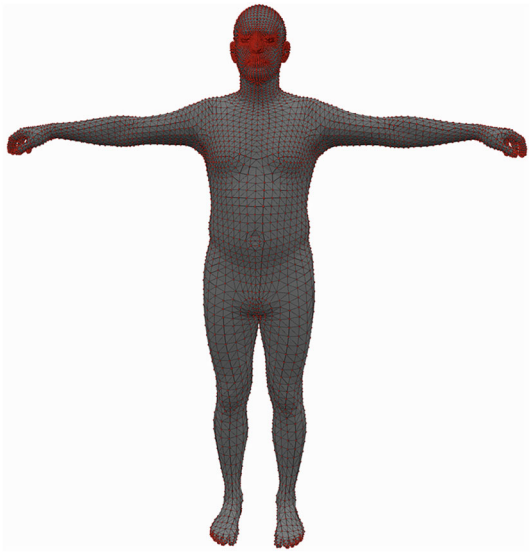

Fig. 4 Example of a SMPL model [12]

function, with linear blend skinning being the most commonly used.

The SMPL model is compatible with existing rendering engines and has been employed in numerous studies investigating human-centric vision, including shape and pose estimation [13, 23, 24], human surface reconstruction [9, 16, 47], and generation [48–54]. Subsequently, FLAME [55] developed a parametric model of the face, while MANO [56] defined a hand model. SMPL-X [57] extended this approach to jointly model the human body, face, and hands. The SMPL model, derived from minimal clothing 3D scans, cannot accurately represent garments on the human body. Extending this parametric representation to encompass clothed humans involves introducing a vertex offset term [58–60], or garments layer [33, 61–65].

Voxel Similar to pixels in 2D images, voxels are the fundamental units for representing 3D objects [42]. In voxel-based modeling for clothed human reconstruction, spatial information is represented using cubic grids. Various studies have explored 3D voxel representations [34, 35, 66] and probabilistic visual hulls [67] to model human body shapes.

Depth maps Depth maps represent the distance from a camera to the surface of objects in a scene, with each pixel indicating a specific depth value. Dual depth maps use two separate depth maps, one for the front surface and one for the back surface, to capture the full 3D structure of an object, such as a human body. By combining these maps, the object's 3D geometry can be approximated. For instance, several studies have explored reconstructing clothed humans using dual depth maps [36, 37].

Fig. 5 Visualization of the 3D Gaussians [38] multi-view reconstruction results of the DNA-rendering dataset [91] using the SIBR tool [92]

3D Gaussians In previous research on human performance capture, 3D Gaussians have been employed for modeling the human body. This approach has been employed in various studies, such as those by Stoll et al. [68] and Robertini et al. [69]. The 3D Gaussian representation models the human body volumetrically and is typically associated with a kinematic skeletal model. Unlike the 3D Gaussians used in 3D scene modeling, those used for human performance capture are often fixed in number and relatively low in quantity. The optimization of mean and covariance is achieved through the similarity between projected 3D Gaussians and 2D image Gaussians.

The model presented in Kerbl et al. [38] employs a set of discrete 3D Gaussians, which can be rendered to an image using an efficient rasterization algorithm. This representation is similar to point clouds and is referred to as "splatting" in the rendering process (Fig. 5). Each 3D Gaussian possesses several learnable attributes and can be described as follows:

$$\mathcal{G} = (\mu, \Sigma, \alpha, C) \tag{4}$$

$$\Sigma = \mathbf{RSS^T R^T} \tag{5}$$

The position of the Gaussian center is denoted by μ, while the covariance matrix Σ is decomposed into a scaling matrix $\mathbf{S}$ and a rotation matrix $\mathbf{R}$. The scaling matrix $\mathbf{S}$ is represented as a 3D vector $\mathbf{s}$, while the rotation matrix $\mathbf{R}$ is represented as a quaternion $\mathbf{q}$. The opacity value is represented by α, and directional color C is represented via spherical harmonics (SH).

3D Gaussian splatting, as a more memory-efficient representation with faster training and inference speeds, has widespread applications [70–72] in general dynamic scene modeling [73–76], object surface extraction [77–79], 3D object generation [80–82], and clothed human reconstruction [83–88] and generation [89, 90].

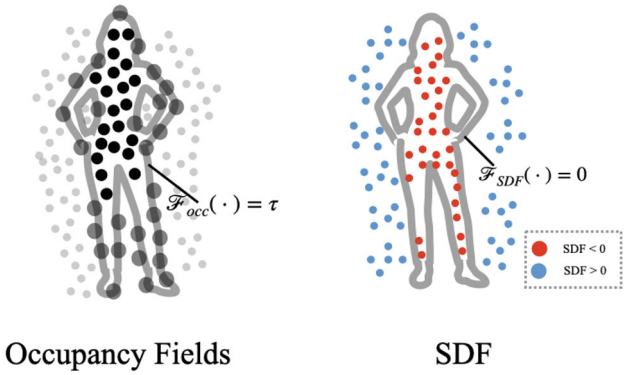

Fig. 6 Illustrative explanations of occupancy fields and SDF

2.2 Implicit representations

Occupancy fields and SDF Occupancy fields represent a 3D object by predicting whether a given point in space is inside or outside the object. This is done by assigning an occupancy value to each point, typically ranging between 0 and 1. SDF represent 3D surfaces by predicting the shortest distance from any point in space to the object's surface. The distance is signed, meaning it has a positive value if the point is outside the object, a negative value if the point is inside, and zero if the point is exactly on the surface. The occupancy value and SDF value of query points in space can be computed using neural networks, allowing for the representation of intricate surface details of objects [34, 39]. The classic marching cubes algorithm [93] can be employed to extract meshes from occupancy fields and SDF (Fig. 6). Occupancy fields and SDF have extensive applications in the reconstruction of clothed human from from single images [7, 8, 94–96].

NeRF The NeRF [41] stands out as a prominent implicit volume representation of scenes. It characterizes a static scene using a continuous 5D function that furnishes volume density and directional emitted radiance at any spatial point. This function is approximated by a multilayer perceptron (MLP), which predicts volume density and view-dependent RGB color conditioned on a 5D coordinate input (x, y, z, θ, ϕ). Formally, the static scene can be defined as:

$$F_\Theta : (\gamma(\mathbf{x}), \gamma(\mathbf{d})) \mapsto (\mathbf{c}, \sigma) \tag{6}$$

Here, F represents the NeRF, and Θ denotes its parameters. $\mathbf{x}$ represents the spatial coordinates of the query point, and $\mathbf{d}$ denotes the direction of the query point. $\gamma(\cdot)$ is the position encoding function. c represents the emitted color, and σ represents the volume density. NeRF, as a highly efficient and flexible 3D representation, has been widely applied in the reconstruction [97, 98] and generation [48, 49] of 3D clothed humans.

2.3 Advantages and limitations

The various 3D human representation methods each have their own advantages and disadvantages concerning intuitiveness, expressive power, storage, and computation. Choosing the most suitable method often involves balancing trade-offs between speed and accuracy based on specific requirements. Point clouds and voxels are more intuitive but may require higher storage for detailed representations. Meshes provide a clear topological structure, making them suitable for representing complex structures and widely applicable. However, they can require high resolution and lead to higher storage costs. SMPL models are efficient and flexible for human representation but cannot capture detailed surface features, such as clothing. 3D Gaussians offer faster reconstruction and better rendering quality but may face challenges in extracting high-quality geometric surfaces. Implicit representations, such as occupancy fields, SDF, and NeRF, offer greater expressive power and flexibility. However, they often require dense sampling, which results in higher computational costs.

3 Clothed human reconstruction from images

The pipeline for reconstructing clothed humans from monocular images input, as illustrated in Fig. 7, consists of three primary stages: (1) extracting feature from the input image; (2) predicting the geometry and texture of the clothed human; (3) computing loss using ground-truth data from 3D scans of clothed humans to optimize the model parameters. Table 1 summarizes extracted features, reconstruction output, and loss functions in these approaches. The reconstruction process is formulated as follows:

$$\mathcal{F}(\mathcal{E}_{2D}, \mathcal{E}_{3D}, \mathcal{E}_o, \mathbf{x}) \mapsto s \tag{7}$$

Here, $\mathcal{E}$ denotes the feature extractor responsible for deriving information on human geometry and appearance from the image. $\mathbf{x}$ usually represents the query points or vertices in the SMPL model. Specifically, $\mathcal{E}_{2D}$ represents the 2D features of the human in the image, while $\mathcal{E}_{3D}$ pertains to the spatial features associated with human geometry. $\mathcal{E}_o$ denotes additional feature information. The function $\mathcal{F}$ encapsulates the modeling of geometry and appearance of clothed humans based on these extracted features. The variable s represents the estimated geometry and predicted appearance. Section 3.1 discusses the geometric reconstruction of clothed human under different 3D representations, while Sect. 3.2 focuses on the estimation of the textures for the reconstructed human.

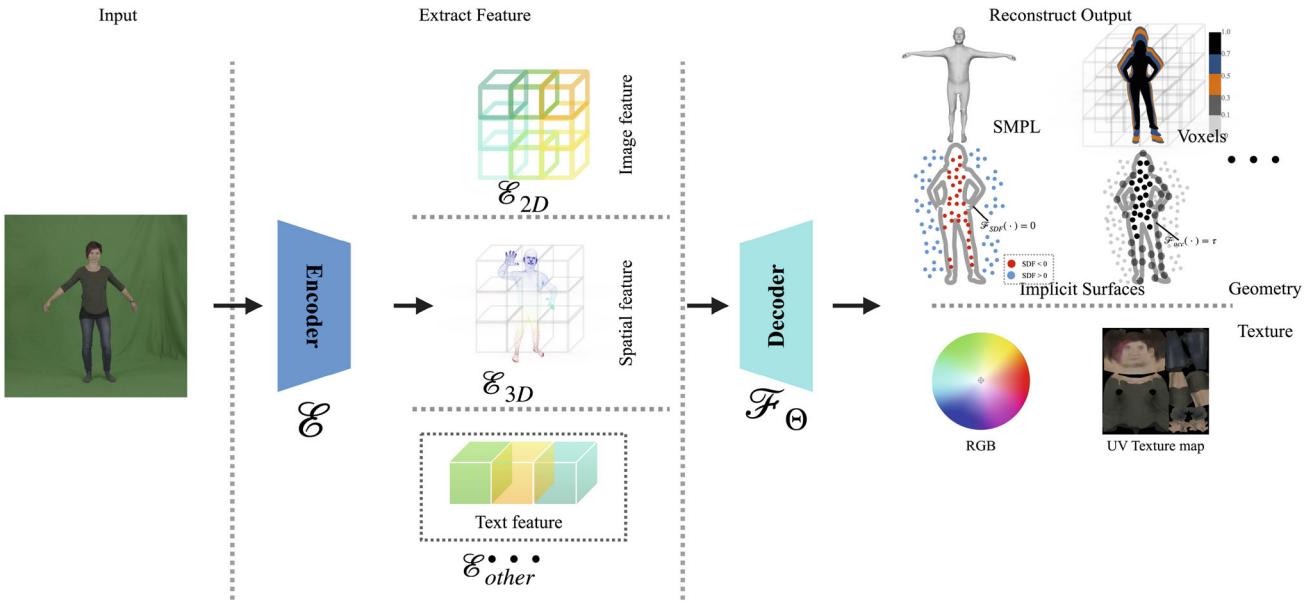

Fig. 7 An illustrative framework outlines the reconstruction of clothed humans from images. The process is divided into three primary components: feature extraction, reconstruction output, and optimization. Initially, the feature extractor encodes crucial information to depict the geometry and appearance of humans, including image features, spatial features, and textual descriptions. Subsequently, the decoder predicts diverse 3D representations of clothed humans conditioned on these features extracted from the image. The reconstruction outputs consist of geometry and appearance branches. Various loss functions are defined based on 3D ground-truth scan data, and regularization terms are employed to optimize and constrain model training. The input image and UV texture image source from the People-Snapshot dataset [99]

3.1 Geometry

SMPL models with clothing offsets The SMPL model inherently characterizes humans in minimal attire across diverse shapes and poses, thus limiting its capacity to delineate intricate geometry details inherent in complex clothing. Extending this parametric representation to encompass clothed humans involves introducing a vertex offset term [58–60], or garments layer [61–65]. Alldieck et al. [15] propose Tex2Shape, which employs a strategy entailing the prediction of normal maps and vertex displacement maps, both contingent upon a partial UV texture map, thereby facilitating the refinement of fine geometry details in human models. The methodology commences by generating a partial UV texture map derived from a single human input image, subsequently completing it by inpainting with the full normal map and vertex displacement map, utilizing a pixel2pixel network [109]. The ground-truth full UV maps are rendered from 3D human scans. Additionally, Lazova et al. [59] propose a technique for modeling clothed humans via predicted vertex displacement maps, conditioned on estimated partial segmentation, culminating in the generation of a complete UV texture map from the partial texture. Bhatnagar et al. [61] introduce multi-garment network (MGN), which delineates several garment templates as supplementary offset terms of the SMPL model, prognosticating both the separable garment templates and underlying human shapes in input images. The inference model is trained with 3D vertex loss and 2D segmentation loss. Corona et al. [64] propose SMPLicit, which harnesses a generative model to forecast the clothing layer atop the human body, conditioned on a set of latent variables encapsulating the garment's cut and style. This model optimizes the fit of each garment to the image by minimizing a loss function between the detected garment's pose and its projected points on the semantic segmentation map. Moon et al. [110] devise two regressors to estimate body shape and pose, cloth existence score, and latent code, advocating a densepose-based loss function amalgamating clothing segmentations and densepose to ensure alignment of the estimated clothing layer with the clothing segmentations (Fig. 8). Zhu et al. [101] register predefined categorized garment templates to the human in the image based on the estimated clothing human shape, boundary, and clothing semantics.

Voxels Varol et al. [34] and similarly, Zheng et al. [35] endeavor to construct reconstructed clothed human using explicit occupancy voxel volumes. Varol et al. [34] propose BodyNet, a volumetric representation of human shape based on fixed-resolution voxel grids (Fig. 9). BodyNet is trained to estimate 3D human shapes directly from single images by minimizing the binary cross-entropy loss between ground-truth and predicted occupancy values for each grid cell. Similarly, Zheng et al. [35] introduce DeepHuman, which constructs a semantic volume by leveraging the SMPL

Table 1 This examination provides a comprehensive review of extracted features, reconstruction outputs, and loss functions across various representative methods for reconstructing clothed humans from monocular RGB images

Methods	Feature			Output		Loss
	2D image	3D spatial	Others	Geometry	Appearance	
Tex2Shape [15]	–	partial UV	–	Offsets	–	MS-SSIM & GAN [100]
ReEF [101]	Bound	Cloth shape	–	Templates	–	–
BodyNet [34]	Pose	Pose	–	Voxel grids	Pose	–
ECON [16]	–	Normal and depth	–	D-BiNI [102]	–	BiNI & Depth
2K2K [103]	–	Normal and depth	–	Depths	–	Depth & SSIM
PIFu [7]	Pixel-aligned	z value	–	Occupancy	RGB	Recon
PIFuHD [8]	Pixel-aligned	3D embed and normal	–	Occupancy	–	Recon
PaMIR [47]	Pixel-aligned	Voxelized SMPL	–	Occupancy	RGB	Recon
ICON [9]	–	Norm and SMLP SDF	–	Occupancy	–	Recon & SMPL and Normal
ARCH [104]	Pixel-aligned	SemS & SemDF	–	Occupancy	RGB	Recon
ARCH++ [105]	Pixel-aligned	Spatial-align	–	Occupancy	RGB	Recon
CAR [106]	Pixel-aligned	Normal and SDF	–	SDF	–	Recon
PHORHUM [94]	Pixel-aligned	–	Illum	SDF	Albedo	Recon
S3F [95]	Pixel-aligned	3D	Illum	SDF	Albedo	Recon
TeCH [50]	Human embed	Normal	Text	DMTet	Color	SDS [107]
Human-SGD [51]	Mask	Normal	–	–	UV texture	Image
Ultraman [108]	–	Depth	Text	–	Mult-view img	–

The symbol "–" indicates content that may not be explicitly mentioned in the original papers or may be scattered across different sections. In the loss function: "Recon" stands for "reconstruction," referring to the loss between the network-predicted occupancy or SDF values at query points and the actual occupancy or SDF values. "Normal" typically denotes normal maps or normal loss. "Embed" is an abbreviation for "embedding"

Fig. 8 Examples of clothed human reconstruction based on SMPL model with offsets. The reconstruction results source from Moon et al. [110]

Fig. 9 Examples of clothed human reconstruction based on voxels. The reconstruction results source from BodyNet [34]

Fig. 10 Examples of clothed human reconstruction based on dual depth maps. The reconstruction results source from ECON [16]

Fig. 11 Examples of clothed human reconstruction based on deformable meshes. The reconstruction results source from TeCH [50]

model derived from input images and integrating multi-scale image features through volumetric feature transformation. This occupancy volume is further refined using a normal refinement model that enhances the geometric details of the reconstructed human body.

Dual depth maps Gabeur et al. [111] delineate methodologies for estimating visible and hidden depth maps from a single input image. Subsequently, full-body 3D point clouds are derived from these two depth maps, akin to aligning two halves of a mold. The surface mesh is subsequently obtained through Poisson surface reconstruction [112]. ECON [16] approximates front and back partial surfaces via depth-aware silhouette-consistent bilateral normal integration (d-BiNI) optimization [102] (Fig. 10). Guided by clothed human normal maps proposed in ICON [9] and body depth maps rendered from estimated SMPL models, this method employs SMPL-X [57] guided IF-Nets [113] to inpaint missing geometry of partial surfaces, thereby obtaining sided and occluded triangles. The final watertight mesh is synthesized via screened Poisson reconstruction [114], amalgamating the two d-BiNI surfaces, inpainted surfaces, and faces or hands cropped from the estimated SMPL-X model. 2K2K [103] initially prognosticates low-resolution depth maps, delineating the global structure of the human, alongside high-resolution normal maps representing finer details. These high-resolution maps are conditioned on the normal maps. The proposed methodology involves a part-wise image-to-normal network, predicting the front and back normals of the human subject within the image based on the subject's joints. The final mesh is crafted from the predicted high-resolution depth maps using screened Poisson surface reconstruction [114].

Deformable meshes Recently, DefTet [115] and DMTet [116] have proposed the use of deformable tetrahedron meshes for object shape reconstruction. These approaches offer enhanced flexibility in modeling complex geometries, representing a significant advancement in mesh-based reconstruction techniques. These studies integrate the DMTet representation with diffusion models [117–119], optimizing the reconstructed DMTet representation of the clothed human through score distillation sampling (SDS) loss [107].

TeCH [50] adopts a hybrid 3D representation based on DMTet [116] and refines it via multi-view score distillation sampling loss and reconstruction loss (Fig. 11). The diffusion model, a personalized fine-tuned text-to-image model [119], is employed to model certain indescribable appearances conditioned on descriptive text prompts generated by the BLIP model [120]. Puzzleavatar [121] reconstructs the shape and appearance of the human in an image collection by composing multiple assets into a T-posed, textured tetrahedral body mesh via score distillation sampling. Multiple human-related assets, such as garments, accessories, faces,

and hair, are decomposed from an image collection, with each asset linked to unique learned tokens by a personalized T2I model, PuzzleBooth. ConTex-Human [122] firstly leverages view-aware 2D diffusion model Zero-1-to-3 [123] to perform score distillation sampling for optimizing a human NeRF. The NeRF representation of the human is then converted into a DMTet mesh and further optimized using front and back normal maps, estimated during the back-view synthesis stage, through a visibility-aware patch consistency loss.

Occupancy fields and SDF PIFu [7] constructs a 3D occupancy field using pixel-aligned features from a single image to predict occupancy probability at sample points, enabling human surface and texture generation from monocular input. GeoPIFu [124] introduces latent voxel features derived from an input image, augmented by a 3D U-Net incorporation, serving as supplementary input for the implicit surface function to alleviate geometry ambiguities associated with query points. PIFuHD [8] with a coarse-to-fine framework to achieve high-resolution 3D reconstructions (Fig. 12). The coarse stage generates a 3D geometry embedding from a downsampled input image and normal maps, while an image-to-image module [125] predicts front and back normals, improving the reconstruction of geometric details. PaMIR [47] voxelizes the estimated SMPL model and employs a 3D encoder to extract voxel-aligned features. Furthermore, a depth-ambiguity-aware reconstruction loss rectifies discrepancies between the predicted model and ground-truth along the z-axis.

ICON [9] leverages the SMPL model to guide normal map estimation for both the human body and clothing, using local features derived from the cloth-body normal map and SMPL model to regress the clothed human surface. D-IF [126] builds on ICON by extracting 7D local features and predicting the distance distribution between points and the surface, combining sampled occupancy with residual offsets calculated by an MLP. Cao et al. [127] introduce the self-evolved signed distance field (SeSDF) module, which refines SDF from the SMPL-X model using pixel-aligned features, 3D features from SMPL-X, and distance encoding to enhance clothed human model accuracy. Song et al. [128] propose DIFu, which generates a back-side image using a hallucinator, constructs depth volumes from dual depth maps, and combines voxel-aligned with pixel-aligned features for human occupancy prediction. Zhang et al. [129] develop the global-correlated 3D-decoupling transformer to extract decoupled triplane features, integrating spatial locality and human structural priors through a feature-mixing query strategy.

The aforementioned studies focus on modeling pose-dependent human modeling. Moreover, several investigations have aimed to generate reconstructed surfaces of pose-independent clothed human from images depicting bodies in arbitrary poses. ARCH [104] introduces a semantic deformation field to map query points from posed space to canonical space and a semantic space wherein each point is associated with a spatial feature. Subsequently, it computes the occupancy, normal, and color of the human in canonical space based on spatial and pixel-aligned features. ARCH++ [105] incorporates a geometry encoder to extract spatial features of human geometry and pose from input images utilizing the SMPL model. It then optimizes the occupancy estimator of the human in both canonical and posed space. Liao et al. [106] initially estimate a coarse SDF of the human in canonical pose, relying on wrapped query points and a normal vector. Subsequently, an SDFNet is developed to refine the coarse human geometry extracted from the canonical SDF field and deformed into canonical space. Moreover, a meta HyperNetwork is defined to initialize the refinement SDFNet based on the estimated SMPL model.

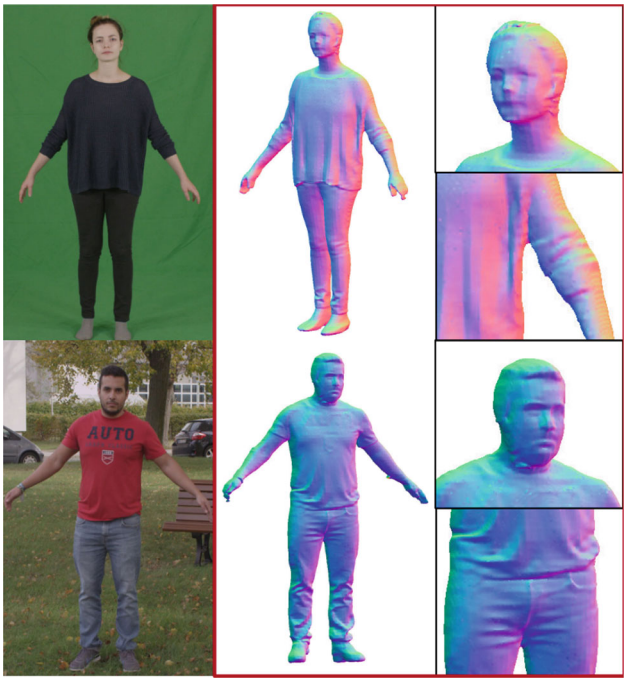

Fig. 12 Examples of clothed human reconstruction based on Implicit surfaces. The reconstruction results source from PIFuHD [8]

NeRFs Some studies utilize NeRF as a representation for human reconstruction, training a generalized NeRF model that predicts images from single input images under given viewpoints or even novel poses, and can extract meshes from them. MoNoNHR [130] is a novel NeRF-based architecture that robustly renders free-viewpoint images of human from a monocular image input. It consists of an image feature backbone, a mesh inpainter, a geometry branch, and a texture branch. However, the reconstructed human NeRF model cannot be animated by novel pose. SHERF [97] builds the generalizable human NeRF model which can synthe-

size novel views and poses of human from a single image input. The sample points are transformed to the canonical space through inverse LBS deformation. The RGB values and density are generated by the feature fusion transformer and NeRF decoder based on the 3D-aware global, point-level, and pixel-aligned features. ELICIT [98] can create free-viewpoint motion videos from a single image by constructing an animatable NeRF representation. It introduces 3D geometry prior and visual semantic prior to assist in the estimation of the human shape and full-body clothing from a single image. Human-LRM [131] models geometry and appearance color within a predicted triplane NeRF by LRM [132], rather than relying on human geometry priors from the SMPL model. Subsequently, a diffusion model is employed to generate high-quality novel view images based on the coarse rendered image from the triplane NeRF. The final geometry and appearance of the human are acquired via a multi-view reconstruction model, conditioned on consistent multi-view images predicted in the preceding stage. The Human-LRM model necessitates an extensive dataset comprising multi-view images and 3D human scans.

3.2 Appearance

In clothed human reconstruction, the appearance of the human is typically represented by RGB values [7, 97, 104, 130], UV texture maps [51, 59, 133–135], and generated multi-view images [108]. The fidelity of reconstruction in unobserved regions is of paramount importance for achieving realistic clothed human reconstruction. However, due to the limitations of single front-view image inputs or inferred back-view images, the appearance in the posterior sections of reconstructed models often appears excessively smooth and blurred. With the advent of advancements in the 2D diffusion model [117–119], there has been a surge in research focusing on optimizing 3D human representations by integrating geometry and appearance priors within these models. Certain methodologies [50, 51, 108, 131, 136] have introduced personalized pretrained text-to-image models [117–119] to guide the appearance inference of unobserved areas.

RGB values Similar to human shape estimation, the appearance of a human is defined by the RGB values of query points based on pixel-aligned features and reconstructed human geometry embeddings [7, 97, 104, 130]. Natsume et al. [137] align the frontal and back views of a person spatially by sharing the same contour and many visual features, building a front-to-back synthesis network to infer the back-view image of a human. PHORHUM [94] anticipates the albedo colors of corresponding surface points. The ultimate shaded color is disentangled into an albedo color and surface point shading, which is predicted by a shading network contingent on the surface normal vector and scene illumination embedding. S3F [95] projects the points sampled around the estimated GHUM model [45] onto the 2D feature map derived from the input image, thereby engendering structured 3D features. Subsequently, the geometry, albedo color, and final shaded color are computed based on PHORHUM [94]. Sengupta et al. [138] estimate the back-view image and front/back albedo of the human in the single image input based on a diffusion probabilistic model.

UV texture maps The 2D UV texture map is an efficient representation of human appearance. Several studies [51, 59, 133–135, 139–141] have derived the appearance of reconstructed clothed humans by inferring UV texture maps from single image inputs. Lazova et al. [59] predict a complete map in the UV space using an image-to-image translation network based on estimated partial textures and garment segmentation. Texformer [133] proposes a transformer-based framework to estimate UV texture maps from single image inputs, overcoming the limitations of existing methods that solely rely on convolutional neural networks. It also introduces a part-style loss to enhance the high fidelity of reconstructed colors and reduce artifacts. HUman-SGD [51] constructs a support set comprising an input image and an inferred back-view image. It amalgamates visible pixels of novel viewpoints from images within the support set and inpaints the appearance in unobserved areas via a diffusion model, conditional on the normal map and silhouette map derived from estimated clothed human geometry. Subsequently, the synthesized novel view image is assimilated into the support set until all requisite viewpoint images are generated. Finally, the images within the support set are amalgamated via inverse rendering to procure an optimized UV texture map. DINAR [134] first uses UV maps to sample the partial RGB textures of the input image based on the estimated SMPL-X model and then converts the input image into a neural texture using a StyleGAN2 generator. It inpaints the concatenation of the neural texture and partial texture to generate a complete texture using a diffusion model. SHERT [135] obtains a partial texture map from the input image based on the reconstructed semantic mesh and camera parameters. It employs a ControlNet model [118] to generate complete human textures, conditioned on the partial texture and prompts of the texture descriptions with an additional partial UV mask.

3.3 Advantages and limitations

By combining extensive high-quality 3D clothed human scan data with efficient and flexible human representations, current monocular image-based methods can quickly reconstruct relatively high-quality 3D human models from monocular images input. However, several factors still constrain further improvements in reconstruction quality:

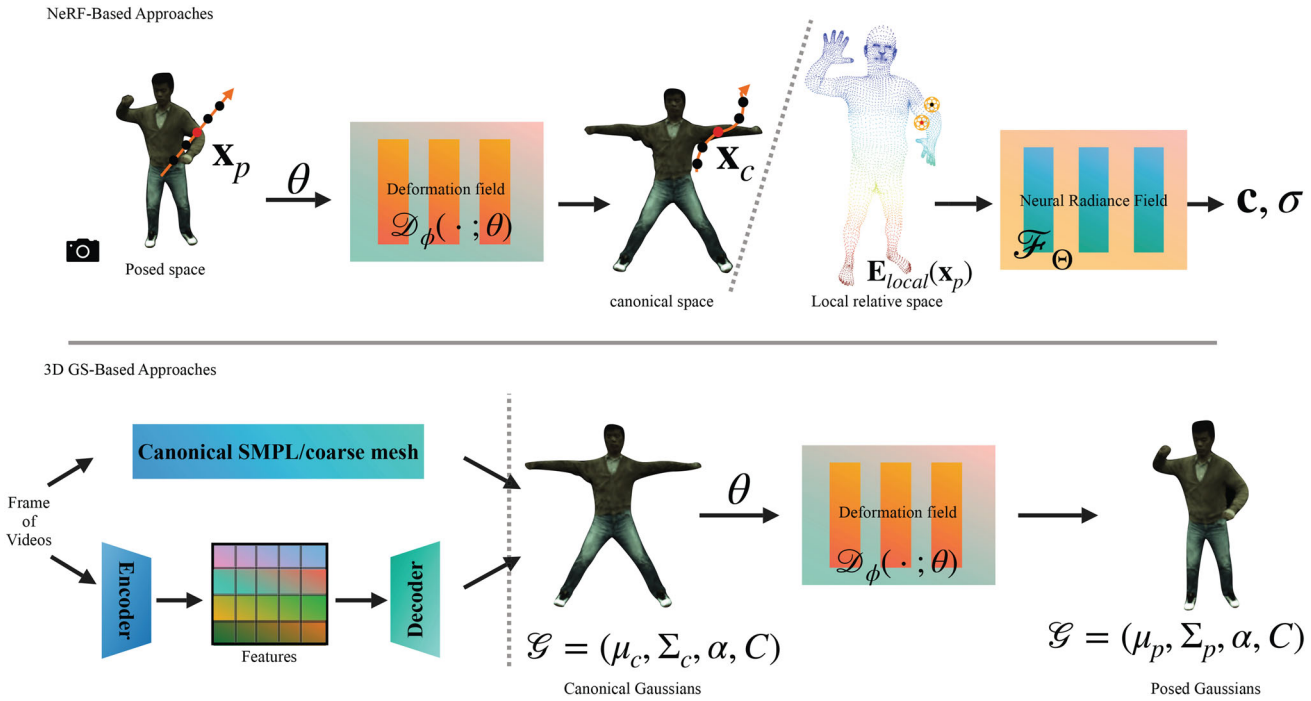

Fig. 13 The framework for dynamic clothed human reconstruction from video sequences comprises three essential components: the 3D human representation, the deformation field, and the optimization procedure. The 3D human representation can be broadly categorized into implicit NeRF and explicit 3D Gaussians. The formulation of the deformation field depends on the chosen 3D human representation. In NeRF representation, two types of transformations govern sample coordinate alterations: global world space transformations and local correspondence embedding. The input to NeRF typically includes optimizable latent codes encoding pose-dependent information, frame-specific intricacies, and relevant features. Conversely, in the realm of 3D Gaussians, canonical representations undergo deformation, resulting in posed 3D Gaussians based on human pose parameters. These canonical 3D Gaussians can originate from the SMPL model or be generated from auxiliary maps, such as UV position maps or triplane maps. Images used in this figure source from ZJU-MoCap [142, 143] dataset

- Dataset bias: Existing datasets for training clothed human scans often have significant gaps in the variety and distribution of human poses and clothing compared to real-world scenarios.
- SMPL parameter estimation: Accurately estimating SMPL parameters from monocular images remains a challenging step in reconstruction, limiting further advancements in human model quality.
- Model reasoning: Monocular images typically capture only forward-facing information. While incorporating human structural information can enhance geometric completeness, accurately inferring texture, especially in occluded regions, continues to be a difficult problem in monocular human reconstruction.

4 Clothed human reconstruction from videos

Departing from clothed human reconstruction from images, dynamic human reconstruction requires not only accurate rendering of human movements within each frame but also consistency in reconstructed human representations across different frames.

Some studies [144–146] construct a rigged clothed human template of the actor featured in the video and capture the human performance by optimizing the skeleton parameters of the template fitting to the detected information from each frame, such as joint positions. Alldieck et al. [99, 147] transform the silhouette cones corresponding to dynamic human silhouettes at each frame to obtain a visual hull in canonical space via unposing silhouette camera rays. It back-projects the image color from several frames to all visible vertices of SMPL model to generate a full texture map. SelfRecon [148] combines implicit SDF and explicit SMPL+D representations to reconstruct space–time coherent clothed human shape from a monocular self-rotating human video. This hybrid representation is defined in the canonical space and a deformation field is used to transform the canonical human shape conditional on estimated human pose parameters from each frame to posed human shape in corresponding frame.

Figure 13 illustrates the general pipeline for reconstructing dynamical clothed human from monocular videos, comprising three components: human representation, deformation

field, and optimization process. NeRF and 3D Gaussian serve as efficient and flexible 3D representations, defining the moving human in videos implicitly and explicitly, respectively, and are widely applied in human reconstruction from videos. Despite their differing reconstruction processes, Sect. 4.1 discusses human reconstruction using NeRF representations from monocular videos, while Sect. 4.2 discusses dynamic human reconstruction based on 3D Gaussian representations. The primary role of the deformation field is to transform the reconstructed human representation to a posed shape based on provided human pose parameters. Table 2 summarizes these three components across various studies.

4.1 NeRF-based approaches

As shown in the upper part of Fig. 13, in dynamic human reconstruction based on NeRF representations, the deformation field is used to transform the pose-dependent human NeRF representation to the corresponding pose. NeRF provides radiance values for each query point in space, implicitly modeling the clothed human. The deformation field converts the Cartesian coordinates of the sampling points along the camera ray in the posed space into the pose-dependent coordinates in the reference space. This reference space defines the human NeRF in various poses. The reference space is primarily divided into two categories: the first is typically the canonical space in an unposed state, and the second is a combination of local spaces relative to different human body parts. The design of the deformation field primarily comprises two main variants. One involves the conversion of the global reference coordinate space from the posed space to the canonical space. The other entails the translation of the global space into the local relative space. Moreover, certain investigations [143, 160] have amalgamated these two variants, wherein the initial step involves transforming each local component (latent code or node) into the global space, followed by the computation of local coordinate deformation for sample points corresponding to the local reference. This process can be mathematically described as follows:

$$\mathcal{F}_\Theta(\mathcal{D}_\Phi(\mathbf{x};\theta),\mathbf{d}) \mapsto (\mathbf{c},\sigma) \tag{8}$$

Here, $\mathcal{F}_\Theta$ denotes the general human NeRF parameterized by Θ. The sample point in posed space is denoted by $\mathbf{x}$, the human pose parameters by θ, and the camera view direction by $\mathbf{d}$. The symbol $\mathcal{D}_\Phi$ represents the deformation fields and Φ is the parameters of the deformation field.

Human NeRFs The representation of dynamic human subjects draws upon NeRF and its variants. In studies such as those by Wang et al. [155], Guo et al. [170], and Liu et al. [163], researchers aim to construct two distinct neural networks: one for predicting opacity and the other for predicting color. This bifurcation serves the purpose of disentangling the geometry and appearance aspects of the canonical human model. Within the geometry branch, the opacity of sample points is contingent upon their SDF value, thereby facilitating the extraction of a clothed human mesh from this geometry network.

Fig. 14 Examples of clothed human reconstruction based on NeRF. The leftmost image represents the ground truth. The middle two columns display the rendering results of the reconstructed human from different camera perspectives. The rightmost three columns show the rendering results after transformation to new poses. The reconstruction results source from InstantAvatar [157]

Approaches like Scarf [159] and Delta [171] employ a hybrid representation to disentangle the human body from garments. Here, the upsampled SMPL-X model, augmented with an offset term, represents the human body, while a NeRF model portrays the clothing. Additionally, mesh-integrated volume rendering is introduced, predicated upon the intersection between rays and the human body mesh. In Wang et al. [172], the disentanglement of human body and clothing is achieved through the utilization of a double-layer NeRF, capturing the movements of both body and clothing in canonical space. A physical simulation loss is further employed to preserve physically plausible clothing deformation.

MonoHuman [154] incorporates a forward correspondence search module to expedite feature search across different frames, guided by forward deformation, thereby facilitating novel view synthesis. SLRF [160] employs a set of structured local radiance fields to depict the human body, with each local radiance field corresponding to a predefined node based on the SMPL model. Neural Actor [163] integrates 2D texture maps defined on the SMPL model as latent variables to capture the dynamic deformation and appearance of the human body. GP-NeRF [173] proposes a geometry-guided multi-view feature integration approach to refine the coarse geometry prior, leveraging the estimated SMPL model and pixel-aligned features extracted from images.

Some methodologies [157, 174, 175] adopt a human model based on InstantNGP [158] to reduce training costs (Fig. 14). The UV volume introduced by Chen et al. [164] is devised to predict the density and texture coordinates of query points, along with a pose-dependent neural texture stack (NTS) to encode appearance information. The final color of query points is determined based on UV coordinates and texture embedding interpolated from NTS. Kwon et al. [176] initiate by generating a time-augmented skeletal representation of human body motion, which is then fused with pixel-aligned features of sample points at each timestep

Table 2 A comparative analysis of reconstructing dynamic clothed humans from sparse multi-view and monocular video data

Methods	Viewpoints	Representation	Deformation fields		Loss
			Global	Local relative	
SelfRecon [148]	Monocular	Mesh and SDF	Forward	–	Exp & imp & cons
VideoAvatar [99]	Monocular	SMPL+ offset	Inverse	–	Data & reg
Neural body [143]	Sparse multi-view	NeRF	Forward	Latent code	Rendering
Animatable_NeRF [149]	Sparse multi-view	NeRF	Dual	×	Rendering & nsf
Anim-NeRF [150]	Monocular	NeRF	Inverse	×	Rendering & pose
HumanNeRF [10]	Monocular	NeRF	Inverse	–	Rendering & LPIPS [151]
TAVA [152]	Sparse multi-view	NeRF	Inverse	–	Rendering & pose reg
NeuMan [153]	Monocular	NeRF	Inverse	–	Rendering & LPIPS
MonoHuman [154]	Monocular	NeRF	Dual	–	Rendering & LPIPS
ARAH [155]	Spare multi-view	SDF & NeRF	Inverse	–	Rendering & Eik reg [156]
InstantAvatar [157]	monocular	InstantNGP [158]	Inverse	–	Rendering & Occupancy reg
Scarf [159]	Monocular	NeRF & SMPL-X [57]	Inverse	–	Rendering & clothing & body
SLRF [160]	One & several	NeRF	Forward	Structural local embed	Rendering & trans & embed
A-NeRF [161]	One & several	NeRF	–	Skeleton-relative embed	Rendering & pose
Xu et al. [162]	Sparse multi-view	NeRF	–	Surface-aligend embed	Rendering
Neural actor [163]	Spare multi-view	NeRF	Inverse	barycentric coordinates	Rendering
UV volume [164]	Sparse multi-view	3D UV volume & 2D texture			Rendering & VGG & UV
SplattingAvatar [84]	Monocular	3D Gaussians	Forward	Mesh embed	Rendering & LPIPS
AnimatableGaussian [165]	Sparse multi-view	3D Gaussians	Forward	×	Rendering & LPIPS
GART [166]	Monocular	3D Gaussians	Forward	×	Rendering & SSIM
HUGS [167]	Monocular	3D Gaussians	Forward	×	Rendering & SSIM & VGG
GauHuman [11]	Monocular	3D Gaussians	Forward	×	Rendering & SSIM & LPIPS
GaussianAvatar [168]	Monocular	3D Gaussians	Forward	×	Rendering & SSIM & LPIPS
ASH [169]	Sparse multi-view	3D Gaussians	Forward	UV mapping	Rendering & SSIM

The table offers a concise overview of the representation, deformation field, and loss function utilized in various representative methods. "Exp" and "Imp" are abbreviations for "explicit" and "implicit," respectively, representing a combination of multiple loss terms used in SelfRecon [148]. "Cons" is short for "Consistency," referring to a regularization term. "Reg" is an abbreviation for "Regularization," and "Embed" is short for "embedding"

to obtain the fused query feature. GHuNeRF [177] constructs a 3D feature volume from target SMPL vertex-size features through visibility-aware feature aggregation, further enhancing this volume with temporally aligned features.

Global coordinates deformation In the realm of global coordinates deformation, the human representation takes the form of NeRF or its derivatives, all of which are delineated within the canonical space. Sample points are extracted within the posed space, contingent upon the camera view. To compute their radiance concerning the canonical NeRF, a wrapping procedure is requisite to transition them into the canonical space, necessitating a Cartesian coordinate transformation within the sampling world space. This transformation conventionally adopts an inverse LBS transformation premised on the human skeletal structure [150, 153, 163, 170, 178, 179]:

$$\mathbf{x}_p = \left(\sum_{i=1}^{n} w^{c \to p}(\mathbf{x_c})_i \, B_i(\theta) \right) \mathbf{x}_c \tag{9}$$

$$\mathbf{x}_c = \left(\sum_{i=1}^{n} w^{p \to c}(\mathbf{x_p})_i \, B_i(\theta) \right)^{-1} \mathbf{x}_p \tag{10}$$

The former operation signifies the forward deformation, which translates a point from the canonical space to points in the posed space contingent upon the pose parameters, while the latter represents the inverse operation. The forward and inverse skinning weights of points corresponding to human joint i, denoted by $w_i^{c \to p}$ and $w_i^{p \to c}$, respectively, are expected to be equivalent. The parameter n signifies the number of joints, and the vector $B_1(\theta), \ldots, B_n(\theta)$ encapsulates the bone transformations corresponding to the human pose θ. For sample points, their inverse skinning weights can be approximated either by the nearest points on the SMPL surface [163] or by the average of the skinning weights of several nearest SMPL surface vertices [150, 159, 170, 171].

HumanNeRF [10] devises an optimizable explicit volume representation to prognosticate the skinning weights of its rigid deformation. Such methodologies aptly approximate the skinning weights of the inverse LBS transformation. Peng et al. [149] and Yu et al. [154] introduce auxiliary forward deformation meshes to guarantee parity between the skinning weights of forward and inverse deformations. Meanwhile, Tava [152] and InstantAvatar [157] fashion deformations via a forward skinning weight optimization algorithm introduced in SNARF [179]. ARAH [155] pioneers a neural network for predicting the forward skinning weights of points in space and proffers a novel joint root finding algorithm to pinpoint the corresponding canonical point of intersection of camera rays and the posed SDF iso-surface. AnimatableNeRF [149] avails a consistency loss between the blend weights of the posed-to-canonical and canonical-to-posed transformations to optimize the canonical neural blend weight field and novel pose latent code. Typically, a neural network is enlisted to predict the offset stemming from nonrigid pose-dependent deformation [10, 149, 152, 153, 163]. The pose parameters estimated per frame by off-the-shelf tools may lack precision and could undergo refinement during the optimization process [10, 150]. In Zhi et al. [180], query point coordinates are morphed from the posed space to the canonical space through barycentric mapping. Initially, the canonical normal vector is computed, and subsequently, it is retrogressed into the posed space to derive the normal vector of the query points. Conversely, ActorsNeRF [181] adopts a coarse-to-fine approach, harnessing a pretrained coarse-level canonical model derived from multiple monocular video sequences to capture a general coarse shape. Additionally, an instance-level canonical model is leveraged to encapsulate human specifics emanating from distinct human movements.

Local relative coordinates An alternative modality for mapping Cartesian coordinates of sample points from posed space to canonical space involves aligning the global Cartesian coordinates of sample points with local coordinates corresponding to distinct human body parts [155, 160, 161, 174]. In this paradigm, the global human form is typically segmented into various local parts predicated on the human skeletal structure.

A-nerf [161] introduces a skeleton-relative encoding mechanism to process the positional and directional coordinates of each query point, alongside a cutoff operation to attenuate the influence of extraneous bones. The NARF model [182] conceptualizes the human as a composite assembly of several movable rigid bone parts, whereby the radiance field of a 3D position is correlated with the most pertinent bone. In SLRF [160], every sample point is tethered to a set of predefined nodes, with each node undergoing deformation via forward skinning and a learnable residual corresponding to garment movements. Subsequently, the coordinates of sample points are transposed to local correspondences, aligning with each bone part [182] or node [160]. Furthermore, SLRF [160] advances a conditional variational auto-encoder (cVAE) to compute node-associated residual motions and dynamic detail embeddings for novel poses. SelfNeRF [174] introduces a surface-relative representation predicated on the observation that the k-nearest vertices on the human model surface of sample points remain constant during human movement. In Te et al. [183], a query embedding predicated on the nearest projected vertex of the query point on the posed SMPL mesh, the k-nearest adjacent vertices of the projected vertex in the canonical SMPL mesh, and the Euclidean distance between these adjacent vertices and sample points is proposed. Meanwhile, Xu et al. [162] propose a surface-aligned representation of query points, entailing scattered projection points on the canonical human mesh surface

of query points and the signed distance between the query points and projected points. Su et al. [184] forge a volume feature for each body part founded on a graph neural network (GNN), wherein local features of sample points corresponding to each body part are amalgamated into a global feature, subsequently utilized to compute the opacity and color of the human. The surface projection method, delineated in Zhang et al. [185], involves projecting sample points onto the estimated SMPL model surface, with a deformation net predicting the nonlinear offset of projected points on the surface. This process transmutes the coordinates of sample points in posed space to a locally deformed coordinate system within the neural deformable field space via surface projection and offset calibration.

Additionally, research has been dedicated to amalgamating global coordinate transformations of query points with local spatial embedding. In such hybrid deformation fields, the global deformation of each local reference space typically embodies forward deformation. Each local reference space is posited in the canonical space with a base human pose. Neural body [143] delineates a set of structural latent codes serving as the conditional feature of NeRF input predicated on the SMPL model, while SLRF [160] introduces several nodes to construct a scene embedding of query points within a local space on the SMPL model.

4.2 3D Gaussians-based approaches

Animatable human models can also be constructed using a set of 3D Gaussians [38] augmented with an elaborate deformation field. These 3D Gaussians represent the fundamental elements of the human, explicitly modeling it. Instead of transforming sample points, the deformation field is employed to transform the 3D Gaussians. Defined in canonical space, the Gaussians undergo forward-oriented deformation, which proves to be more accurate, especially for sample points at the intersection of multiple body parts. The integration of human representation priors, such as the SMPL model, into the initialization and optimization process of the 3D Gaussians is more straightforward.

Figure 13 illustrates the general pipeline for animatable human reconstruction from videos based on a 3D Gaussians representation. This pipeline comprises three main stages: 3D Gaussians generation, forward deformation field generation, and optimization. There are two distinct approaches for generating the canonical Gaussians. One approach initializes the 3D Gaussians directly from the SMPL model or estimated canonical human mesh and then optimizes each attribute during the training process. The other approach involves predicting auxiliary feature maps such as UV position maps, 2D Gaussian maps, and triplane feature maps and subsequently deriving Gaussian attributes from these maps. These approaches are not mutually exclusive and can be combined, with some Gaussian attributes being predicted by auxiliary networks.

Initialization In the original research on 3D Gaussians Splatting [38], the initial sparse point cloud is derived from structure-from-motion point sets. A well-executed initialization can significantly expedite the optimization process and enhance the quality of the output [11, 38, 186]. However, this approach necessitates the use of multi-view images, which is often infeasible due to the typical input format of monocular videos. Consequently, in these studies, the 3D Gaussians is typically initialized with a human representation prior. The SMPL model [12, 57] represents human shape and pose in a simple and efficient manner and is frequently employed as a reliable prior for human representation in the field of animatable human modeling. A straightforward initialization strategy is to utilize the vertices in the SMPL mesh surface as the initial point clouds [11, 167, 187–189].

However, the SMPL model does not account for the geometry of complex garments or the appearance of humans. Consequently, the point clouds for initialization lack information about the representation of clothed humans. In Jung et al. [190], the authors utilize the center of each SMPL template mesh face as the initial point clouds, along with the parent index and surface normal of the face as additional features. In lieu of directly utilizing the SMPL model, some research endeavors employ a mesh estimated by off-the-shelf tools as the initialization point clouds. Compared to the SMPL model, the estimated mesh exhibits enhanced accuracy in terms of geometry and appearance information for the representation of clothed humans. SplatArmor [191] initially recovers a coarse per-face color SMPL + D mesh and initializes the Gaussians with the optimized mesh. In addition to the SMPL vertices, Li et al. [186] also employ the ECON model [16] to generate a clothed, viewpoint-only textured mesh from a selected frame. These meshes are then deformed to the canonical space, where they are fused with color to initialize the human Gaussians in the canonical space.

Human 3D Gaussians In several studies, such as [167, 169], spherical harmonics or their residuals are predicted by neural networks. Instead of directly using spherical harmonics to represent the view-dependent color of the Gaussians, some approaches, Hu et al. [168] and Moreau et al. [187], use optimizable RGB values as color. In other works, such as Jena et al. [191] and Qian et al. [188], a MLP with additional pose-dependent features and other attributes is employed to compute the color of the Gaussians. Kocabas et al. [167] decodes the position offset and attributes of the Gaussians using three MLPs, which take the position of the Gaussians and features obtained from interpolating at a feature tree plane as input. The Gaussians are attached to the 2D UV map, and additional expressive features can be derived

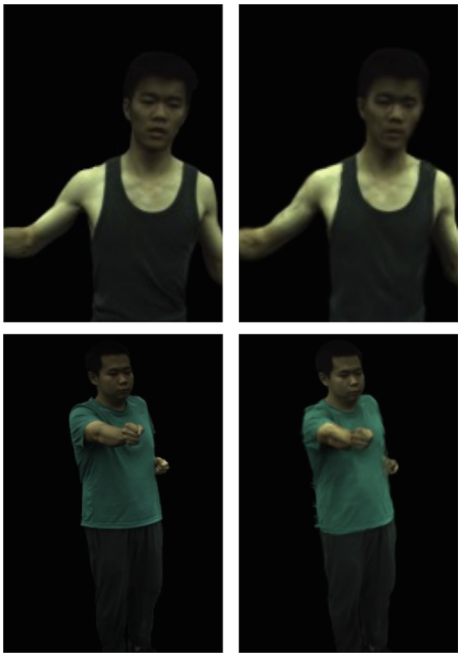

Fig. 15 Examples of clothed human reconstruction based on 3D Gaussians. The left column presents the ground truth, while the right column shows the rendering results of the reconstructed human from the corresponding view. The reconstruction results source from GauHuman [11]

from the 2D UV space via a 2D convolutional neural network (CNN) [192], enhancing the representation capability of the Gaussians. In Li et al. [165], a 3D Gaussian is generated from front and back pose-dependent Gaussian maps, which are produced by StyleUnet on position maps obtained from a canonical template via orthogonal projection. Zhu et al. [169], a texel-based 2D parameterization of the 3D Gaussians is proposed, utilizing the texel of the template mesh to store the parameters of the Gaussians. The parameters are calculated from two 2D CNN decoders based on the motion-aware textures map rendered from the template mesh and the deformation network.

Deformation fields The deformation field is a fundamental aspect of animatable human modeling, capturing human movements observed in videos and controlling the human representation to alter its appearance based on specific pose parameters. The deformation field is primarily applied to the mean and covariance of the canonical 3D Gaussians distribution and can be described as follows:

$$\mu_{\mathrm{p}} = \mathcal{D}_{\Phi}(\mu_{\mathrm{c}}; \theta) \tag{11}$$

$$\Sigma_{\mathrm{p}} = \mathcal{D}_{\Phi}(\Sigma_{\mathrm{c}}; \theta) \tag{12}$$

Here, the positions of the 3D Gaussians centers, denoted by μ_{p} and μ_{c}, represent the means of the 3D Gaussians in the posed space and canonical space, respectively. The covariance matrices of the 3D Gaussians, denoted by Σ_{p} and Σ_{c}, represent the covariance in posed and canonical spaces, respectively. The pose parameters of the human, denoted by θ, represent the transformation of the human in the video frame. As previously mentioned, the human 3D Gaussians is defined in canonical space. However, before rendering, it is necessary to transform these Gaussians to the target pose. This differs from the approach typically employed by NeRF-based systems, which use inverse deformation to warp the ray in posed space to canonical space. The deformation field is specifically designed for this purpose and is critical to the performance of human representation and rendering. The deformation field consists of two parts: rigid deformation and nonrigid pose-dependent deformation. The deformation of the covariance matrix is mainly focused on the rotation of the Gaussians' rotational attributes.

The predominant form of rigid deformation is the LBS transformation, which relies on the skeleton structure of the SMPL model [165, 168, 186, 189, 190]. For a more nuanced representation of rigid deformation, the blend weights of LBS often incorporate additional learnable weights, known as residuals associated with each Gaussians [11, 166, 188] (Fig. 15). These residuals, derived from Gaussians shape functions, are employed to refine the blend weights [167, 187]. SplatArmor [191] defines a weight function based on the k-nearest neighbor SMPL vertices of a point in the entire space to extend the blend weights transformation. Li et al. [186], the closest triangular facet of each Gaussians on the SMPL model is identified based on Cartesian distance anchor in both canonical and posed space. Subsequently, the Gaussians rotation transformation matrix is calculated and applied to the rotation and spherical harmonics direction of the Gaussians. The nonrigid deformation of the pose-dependent Gaussians maps is utilized to extract the Gaussians [165]. ASH [169] describes the wrapping of the canonical Gaussians stored in the texel to a posed Gaussians via UV mapping and dual quaternion skinning. To generalize novel poses, the approach proposed in Li et al. [165] utilizes principal component analysis to project a novel driving pose signal into the distribution of observed training poses.

Optimization The optimization process for Gaussians attributes typically follows a series of steps, including rendering loss, split, clone, and prune operations, consistent with the 3D Gaussians optimization process. GauHuman [11] proposes the utilization of Kullback–Leibler (KL) divergence of 3D Gaussians to guide the split and clone processes. Moreover, a KL-based merge operation is employed to amalgamate redundant 3D Gaussians. Human masks acquired via off-the-shelf tools are commonly used to regularize the Gaussians distribution, thus preventing overfitting to the background in Jena et al. [191]. To mitigate the error resulting from inaccurate pose parameter estimation, a pose parame-

ter refinement module is typically integrated into the model training pipeline [11, 165, 168, 188, 189]. The final blend weights are expected to exhibit spatial smoothness [166] and closely resemble the blend weights of the nearest SMPL vertices [167, 187]. In Li et al. [189], it is demonstrated that [75] utilizes local constraint information to regularize the deformation of Gaussians between canonical and posed space, thereby minimizing visual artifacts. 3DGS-Avatar [188] proposes an as-isometric-as-possible regularization to constrain neighboring 3D Gaussians centers, ensuring a consistent distance preservation after deformation and reducing noise generation.

4.3 Advantages and limitations

Human reconstruction from videos using NeRF and 3D Gaussian representations leverages preprocessed video datasets that are widely accessible and applicable across various contexts. These methods offer high flexibility, enabling adaptive adjustments to human parameters in the video and generating high-quality renderings based on camera parameters. The reconstructed human models typically generalize well to new poses. However, extracting mesh surfaces from these methods can be complex and often requires additional sophisticated regularization techniques. Additionally, limitations in input viewpoints, typically restricted to monocular perspectives, can result in the loss of surface detail, such as clothing wrinkles, during rendering. Currently, deformation fields representing human motion in videos are mainly based on SMPL model movements, and accurately capturing the motion of clothing, particularly loose garments, remains a challenging task.

5 Evaluation

This section provides an overview of datasets commonly employed for both training and evaluating clothed human reconstruction, along with the frequently used quantitative evaluation metrics.

5.1 Datasets

Table 3 provides a comprehensive overview of the most commonly employed high-quality clothed human scan datasets for training 3D human reconstruction models and human movement datasets for evaluating dynamic human reconstruction.

The following datasets have been proposed for use in 3D pose estimation algorithms: Human3.6M [193], CMU Panoptic Studio dataset [194], 3D Poses in the Wild (3DPW) dataset [196], MPI-INF-3DHP [195], and AIST++ dataset [197]. Human3.6M [193] comprises 3.6 million accurate 3D human poses with moderately realistic clothing, captured in an indoor marker-based motion capture system. The CMU Panoptic Studio dataset [194] is captured by a massively multi-view system and includes multiple people engaging in social games. The 3DPW [196] records challenging movements in outdoor scenes captured by a handheld camera phone, including some multiple person movements. The MPI-INF-3DHP [195] is captured in a markerless multi-camera green screen studio and augmented by replacing masked regions. AIST++ dataset [197] is constructed from the AIST Dance Video DB dataset [198]. It is a large-scale video dataset of street dances, employing multiple fixed camera angles to capture the dance movements of subjects.

MonoPerfCap [144], DeepCap [146], and DynaCap [205] have been proposed for the evaluation of human performance capture. MonoPerfCap [144] contains human motions captured in various settings with both daily and challenging movements, along with sections providing accurate surface ground-truth. DeepCap [146] contains 13 human movement sequences in indoor and outdoor scenes, encompassing various types of motions, with the ground-truth 3D joint positions. DynaCap [205] is a multi-view human motion dataset that provides 3D scans of rigged skeletons.

Multi-view Neural Human Rendering (NHR) [142], ZJU-MoCap [142, 143], and Neural Actors [163] datasets are representative of a common multi-view human movement dataset in which actors perform complex motions with daily clothing. The THuman4.0 dataset [160] contains three multi-view human motion sequences that are used to evaluate the method proposed for animatable human avatar reconstruction in SFRF [160]. The People-Snapshot dataset [99] captures the rotation of actors in an A-pose under a static camera in a variety of backgrounds. Each actor is depicted wearing daily attire. These datasets are frequently utilized in the evaluation of novel view synthesis of human movements.

Recently, there has been a proliferation of higher quality and larger-scale datasets aimed at evaluating approaches to novel view and novel pose synthesis of human movements. ActorsHQ [199] stands out as a high-fidelity dataset of human motion captured by 160 synchronized cameras at 12MP resolution. Alongside raw RGB images, it offers 3D meshes at every frame. HuMMan [200] emerges as a large-scale, multi-view human movements dataset comprising 1,000 human subjects captured in 400,000 sequences and 60 million frames, employing 10 synchronized RGB-D cameras. The GeneBody-1.0 [201], DNA-Rendering [91], and MVHumanNet [202] datasets are noteworthy for their comprehensive coverage of human movements. These datasets encompass a diverse range of human types, body shapes, ages, and genders. GeneBody-1.0 [201] and DNA-Rendering [91] datasets present actors in a variety of clothing types, materials, and textures, including everyday attire and specific professional scenarios such as theatrical costumes. MVHumanNet

Table 3 Summary of commonly used datasets for training and evaluating human reconstruction models

Dataset	#Sub	#Outfits	#Motions	#Views	#Frames	Videos	Scans
Human3.6M [193]	11	11	17	4	3.6M	✓	✓
CMU panoptic [194]	8	8	5	480	–	✓	×
MPI-INF-3DHP [195]	8	8	16	14	1.3M	✓	×
3DPW [196]	7	18	60	1	51K	✓	✓
AIST++ [197, 198]	30	30	–	9	10.1M	✓	×
People-snapeshot [99]	11	24	1	1	–	✓	×
ZJU-MoCap [142, 143]	9	9	–	21	–	✓	×
NHR [142]	3	3	5	80	–	✓	✓
Neural actors [163]	4	4	–	79–86	87886	✓	×
ActorsHQ [199]	8	8	52	160	40K	✓	✓
HuMMan [200]	1000	1000	500	10	60M	✓	✓
GeneBody-1.0 [201]	50	100	61	48	2.95M	✓	×
DNA-rendering [91]	500	1500	1187	60	67.5M	✓	×
MVHumanNet [202]	4500	9000	500	48	645.1M	✓	×
BUFF [203]	6	12	–	–	13.6K	✓	✓
CAPE [60]	11	88	–	–	80K	✓	✓
THuman2.0 [25]	500	500	–	–	–	×	✓
THuman3.0 [204]	154	154	–	–	–	×	✓
2K2K [103]	2050	2050	–	–	–	×	✓

The columns "#Sub" and "#Outfits" denote the number of actors and total garments, respectively. "#Motions" indicates the observed human movement types, and "#Views" represents the number of cameras used. "–" signifies that the attribute is not specified, while ✓ indicates the presence of the specified data type and × indicates the absence

[202], on the other hand, offers a wide array of everyday clothing styles and colors found in real-world settings. It defines an action library that represents a broad spectrum of human actions, including both daily-life activities and professional actions such as sports activities. DNA-Rendering [91] dataset also incorporates human-object interactivity, capturing instances where human motions interact with objects of varying sizes.

Bodies Under Flowing Fashion (BUFF) [203] and CAPE [60] provide several 3D scans sequence of actors performing simple actions. The BUFF dataset [203] also provides the texture of scans. CAPE [60] separates the clothing from the body, thereby providing an accurate ground-truth body shape under clothing. THuman2.0 dataset [25] and 2K2K dataset [103] represent two of the largest publicly available datasets of clothed human scans captured by a multi-view DSLR camera system. In addition to the original scans, they also provide comprehensive label information, including 3D pose, texture map, and SMPL model parameters. THuman3.0 dataset [204] is a collection of human-garment combinations, with each combination comprising multiple scans. HUMBI [206, 207] is a large multi-view image dataset of human body expressions with natural clothing. In addition to the raw images, the dataset provides processed point clouds of the entire body, different body parts, and garments. There are also some large-scale commercial photorealistic 3D clothed human scans, such as RenderPeople [208].

5.2 Evaluation metrics

Evaluation metrics for clothed human reconstruction The final output of the clothed human reconstruction from single image is the mesh which can be extracted from the parametric human model, depth maps and implicit functions. Common quantitative evaluation metrics of geometry reconstruction include point-to-surface Euclidean distance (P2S), chamfer distance, and normal reprojection error [7]. P2S represents the average distance from the vertices on the reconstructed surface to the ground-truth, while the chamfer distance calculates the distance between the reconstructed surface and the ground-truth surface. The normal reprojection error measures the L2 error between two normal maps rendered from the reconstructed surface and the ground-truth surface at the input viewpoint, assessing the fineness of the reconstructed local details and projection consistency from the input image.

Table 4 summarizes the metrics associated with typical human reconstruction approaches as discussed in Sect. 3.

Evaluation metrics for dynamic human reconstruction In the study of dynamic human reconstruction using NeRF and 3D Gaussian representations, the quality of the reconstruc-

Table 4 Quantitative evaluation of typical methods of clothed human reconstruction from single images

Methods	RenderPeople [208]			BUFF [203]			CAPE [60]			Thuman2.0 [25]		
	Normal↓	P2S↓	Chamfer↓	Normal↓	P2S↓	Chamfer↓	Normal↓	P2S↓	Chamfer↓	Normal↓	P2S↓	Chamfer↓
PIFu [7]	0.084	1.52	1.50	0.0928	1.15	1.14	–	–	–	–	–	–
PIFuHD [8]	0.107	1.37	1.43	0.134	1.63	1.75	–	–	–	–	–	–
ARCH [104]	0.038	0.74	0.85	0.04	0.82	0.87	–	–	–	–	–	–
ARCH++ [105]	0.030	0.50	0.61	0.61	0.03	0.64	–	–	–	–	–	–
ICON [9]	–	–	–	–	–	–	0.066	1.065	1.142	–	–	–
ECON [16]	0.0478	1.458	1.342	–	–	–	0.0367	0.917	0.926	–	–	–
CAR [106]	0.0871	1.4147	1.5142	–	–	–	–	–	–	–	–	–
2K2K [103]	0.0364	1.12	0.92	–	–	–	–	–	–	0.0665	1.19	1.21
DIFu [128]	–	–	–	0.138	3.375	3.318	–	–	–	0.119	2.992	2.952
GTA [129]	–	–	–	–	–	–	0.035	0.763	0.763	0.055	0.862	0.814
TeCH [50]	–	–	–	–	–	–	0.0306	0.6962	0.7416	0.0642	1.2715	1.2364

Table 5 Quantitative comparison of animatable clothed human reconstruction from videos in novel view rendering

Methods	Human 3.6M [193]			People-Snapshot [99]			ZJU-MoCap [143]			Train	Infer	Devices
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓			
Neural Body [143]	–	–	–	–	–	–	28.10	0.944	–	14h	–	4 × 2080Ti
Animatable_NeRF [149]	23.00	0.890	–	–	–	–	27.10	0.949	–	12h	–	4 × 2080Ti
Anim-NeRF [150]	–	–	–	28.89	0.968	0.0206	–	–	–	13h	–	2 × 3090
HumanNeRF [10]	–	–	–	–	–	–	30.24	0.968	0.0317	72h	–	4 × 2080Ti
TAVA [152]	–	–	–	–	–	–	33.11	0.981	–	–	–	
MonoHuman [154]	–	–	–	–	–	–	30.26	0.969	0.0392	70h	–	1 × V100
ARAH [155]	24.79	0.918	–	–	–	–	28.51	0.948	0.0813	36h	–	4 × 2080Ti
SLRF [160]	–	–	–	23.95	0.9115	–	29.55	0.955	–	25h	5 s	1× 3090
Xu et al. [162]	24.28	0.909	–	–	–	–	28.27	0.945	–	14h	–	1 × V100
UV volume [164]	25.33	0.913	0.096	–	–	–	27.96	0.935	0.072	–	19.46 ms	1 × A100
GART [166]	–	–	–	28.37	0.97	0.0465	31.76	0.976	0.034	30 s	150 FPS	1 × 3080
GauHuman [11]	–	–	–	–	–	–	31.34	0.965	0.0351	1 m	189 FPS	1 × 3090
GEA [85]	–	–	–	–	–	–	32.14	0.977	0.02613	3 m	–	1 × A30
Human101 [186]	–	–	–	–	–	–	31.29	0.964	0.0394	100 s	104 FPS	1 × 3090
3DGS-Avatar [188]	–	–	–	32.06	0.966	0.0186	30.90	0.971	0.02696	45 m	0.5h	1 × 3090

The metrics for training time and inference time are reported only for the original paper referenced

Table 6 Quantitative comparison of animatable clothed human reconstruction from video in novel pose rendering

Methods	Human 3.6M [193]			ZJU-MoCap [143]			Inference time ↓
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	
Animatable_NeRF [149]	22.55	0.880	–	23.16	0.893	–	30 m
TAVA [152]	–	–	–	32.02	0.975	–	–
ARAH [155]	23.42	0.896	–	24.63	0.911	0.107	10–20 s
SLRF [160]	–	–	–	25.17	0.916	–	
Xu et al. [162]	23.25	0.892	–	24.42	0.902	–	–
UV volume [164]	25.04	0.874	0.141	23.69	0.910	0.104	68.23 ms

tion is typically assessed by evaluating the similarity between novel view renderings and real images. Commonly employed metrics include peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) [209], and learned perceptual image patch similarity (LPIPS) [151]. PSNR is a widely adopted metric that evaluates the quality of reconstructed images by comparing the logarithmic difference between the maximum possible pixel value and the mean squared error of the reconstructed and reference images. SSIM [209] measures the similarity between two images by assessing luminance, contrast, and structural information, thus providing a perceptual metric that reflects the characteristics of the human visual system. Higher values of PSNR and SSIM indicate a greater similarity between the two images. LPIPS [151] evaluates the perceptual similarity between images using deep neural network features, offering a more accurate representation of human visual perception compared to traditional metrics. Additionally, the training time and frame rate (FPS) during inference are crucial metrics for evaluating the performance of these scene modeling approaches.

Table 5 presents a quantitative comparison of common dynamic human reconstruction methods on three widely used evaluation datasets: Human 3.6M [193], People-Snapshot [99], and ZJU-MoCap [143]. The results in the table are averaged over multiple individual actor's assessments. Most papers only report qualitative results of reconstructed human rendered in novel poses. Table 6 provides quantitative evaluation results from several studies on reconstructed human driven by novel poses.

6 Conclusion and future directions

This survey summarizes the research on high-quality clothed human reconstruction from monocular images and video inputs over the past five years. It provides an overview of common 3D human representations and discusses the reconstruction of clothed human geometry and texture from monocular image inputs under different representations. Furthermore, the survey introduces studies on dynamic human reconstruction from monocular videos. Additionally, it summarizes the commonly used datasets for training and evaluation, along with quantitative analysis results of some representative methods.

The following three issues are critical for achieving high-quality reconstruction of clothed humans from monocular images and deserve further thorough investigation.

Improving detail in avatar reconstruction Current research predominantly focuses on the holistic representation of clothed human avatars. However, this approach often leads to insufficient geometric detail and overly smooth results, particularly in areas such as the hands and face. Integrating the study of hand and face reconstruction with full-body reconstruction can help address the challenges posed by complex human poses and enhance local details [16, 210, 211]. This more refined prior knowledge can be utilized to achieve a more realistic reconstruction.

Separating body and clothing layers In prevailing methodologies of clothed human reconstruction, the body and clothing of the subject are frequently treated as a single layer. This holistic representation limits applications such as clothing editing and virtual try-on. A more effective approach is to treat them as two separate layers: the outer clothing and the human body underneath, with each layer being reconstructed independently [159, 212, 213]. However, as discussed in Sect. 3.1, the parametric or generated clothing layer often lacks realism, making it challenging to independently infer the realistic underlying body and clothing geometry.

Occlusions and incomplete inputs In real world, occlusions caused by other objects, self-occlusions from different body parts, and incomplete visibility of the human part in input images pose significant challenges for accurate and complete human reconstruction. Similar to inferring textures in occluded areas, the possible reconstructions of the human in occluded or incomplete regions are not unique. Recent studies have attempted to reconstruct human meshes [214, 215], NeRF [216], and 3D Gaussian [217] representations from these complex environments. These studies on occlu-

sions and incomplete inputs are of significant importance for applying clothed human reconstruction in real-life scenarios.

Acknowledgements This work was supported by the National Science Foundation of China (No. 62471168, 61802100 and 62372147). This work was also supported by the Zhejiang Provincial Natural Science Foundation of China (No. LDT23F02025F02, No.LY21F020019 and No. LY22F020028) and the Open Project Program of the State Key Laboratory of CAD&CG (No. A2314, No. A2304 and A2306), Zhejiang University. This work was also partially supported by Aeronautical Science Foundation of China (No. 2022Z0710T5001).

Author Contributions SY and XG wrote the main manuscript text and SY prepared all figures and tables. All authors reviewed the manuscript.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

1. Salagean, A., Crellin, E., Parsons, M., Cosker, D., Fraser, D.S.: Meeting your virtual twin: effects of photorealism and personalization on embodiment, self-identification and perception of self-avatars in virtual reality. In: CHI, pp. 499–149916 (2023). https://doi.org/10.1145/3544548.3581182
2. Panda, P., Nicholas, M.J., González-Franco, M., Inkpen, K., Ofek, E., Cutler, R., Hinckley, K., Lanier, J.: AllTogether: effect of avatars in mixed-modality conferencing environments. In: CHI-WORK, pp. 8–1810 (2022). https://doi.org/10.1145/3533406.3539658
3. Manfredi, G., Gilio, G., Baldi, V., Youssef, H., Erra, U.: VICO-DR: a collaborative virtual dressing room for image consulting. J. Imaging **9**(4), 76 (2023). https://doi.org/10.3390/JIMAGING9040076
4. Szolin, K., Kuss, D.J., Nuyens, F.M., Griffiths, M.D.: Exploring the user-avatar relationship in videogames: a systematic review of the Proteus effect. Hum. Comput. Interact. **38**(5–6), 374–399 (2023). https://doi.org/10.1080/07370024.2022.2103419
5. Guo, K., Lincoln, P., Davidson, P.L., Busch, J., Yu, X., Whalen, M., Harvey, G., Orts-Escolano, S., Pandey, R., Dourgarian, J., Tang, D., Tkach, A., Kowdle, A., Cooper, E., Dou, M., Fanello, S.R., Fyffe, G., Rhemann, C., Taylor, J., Debevec, P.E., Izadi, S.: The relightables: volumetric performance capture of humans with realistic relighting. ACM Trans. Graph. **38**(6), 217–121719 (2019). https://doi.org/10.1145/3355089.3356571
6. Collet, A., Chuang, M., Sweeney, P., Gillett, D., Evseev, D., Calabrese, D., Hoppe, H., Kirk, A.G., Sullivan, S.: High-quality streamable free-viewpoint video. ACM Trans. Graph. **34**(4), 69–16913 (2015). https://doi.org/10.1145/2766945
7. Saito, S., Huang, Z., Natsume, R., Morishima, S., Li, H., Kanazawa, A.: PIFu: pixel-aligned implicit function for high-resolution clothed human digitization. In: ICCV, pp. 2304–2314 (2019). https://doi.org/10.1109/ICCV.2019.00239
8. Saito, S., Simon, T., Saragih, J.M., Joo, H.: PIFuHD: multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In: CVPR, pp. 81–90 (2020). https://doi.org/10.1109/CVPR42600.2020.00016
9. Xiu, Y., Yang, J., Tzionas, D., Black, M.J.: ICON: implicit clothed humans obtained from normals. In: CVPR, pp. 13286–13296 (2022). https://doi.org/10.1109/CVPR52688.2022.01294
10. Weng, C., Curless, B., Srinivasan, P.P., Barron, J.T., Kemelmacher-Shlizerman, I.: HumanNeRF: free-viewpoint rendering of moving people from monocular video. In: CVPR, pp. 16189–16199 (2022). https://doi.org/10.1109/CVPR52688.2022.01573
11. Hu, S., Liu, Z.: GauHuman: articulated Gaussian splatting from monocular human videos. In: CVPR (2024)
12. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: a skinned multi-person linear model. ACM Trans. Graph. **34**(6), 248–124816 (2015). https://doi.org/10.1145/2816795.2818013
13. Feng, Y., Choutas, V., Bolkart, T., Tzionas, D., Black, M.J.: Collaborative regression of expressive bodies using moderation. In: 3DV, pp. 792–804 (2021). https://doi.org/10.1109/3DV53792.2021.00088
14. Alldieck, T., Magnor, M.A., Bhatnagar, B.L., Theobalt, C., Pons-Moll, G.: Learning to reconstruct people in clothing from a single RGB camera. In: CVPR, pp. 1175–1186 (2019). https://doi.org/10.1109/CVPR.2019.00127
15. Alldieck, T., Pons-Moll, G., Theobalt, C., Magnor, M.A.: Tex2Shape: detailed full human body geometry from a single image. In: ICCV, pp. 2293–2303 (2019). https://doi.org/10.1109/ICCV.2019.00238
16. Xiu, Y., Yang, J., Cao, X., Tzionas, D., Black, M.J.: ECON: explicit clothed humans optimized via normal integration. In: CVPR, pp. 512–523 (2023). https://doi.org/10.1109/CVPR52729.2023.00057
17. Corona, E., Hodan, T., Vo, M., Moreno-Noguer, F., Sweeney, C., Newcombe, R.A., Ma, L.: LISA: learning implicit shape and appearance of hands. In: CVPR, pp. 20501–20511 (2022). https://doi.org/10.1109/CVPR52688.2022.01988
18. Chen, X., Wang, B., Shum, H.: Hand Avatar: free-pose hand animation and rendering from monocular video. In: CVPR, pp. 8683–8693 (2023). https://doi.org/10.1109/CVPR52729.2023.00839
19. Chen, Z., Moon, G., Guo, K., Cao, C., Pidhorskyi, S., Simon, T., Joshi, R., Dong, Y., Xu, Y., Pires, B., Wen, H., Evans, L., Peng, B., Buffalini, J., Trimble, A., McPhail, K., Schoeller, M., Yu, S.-I., Romero, J., Zollhöfer, M., Sheikh, Y., Liu, Z., Saito, S.: URHand: universal relightable hands. In: CVPR (2024)
20. Saito, S., Schwartz, G., Simon, T., Li, J., Nam, G.: Relightable Gaussian codec avatars. In: CVPR (2024)
21. Bi, S., Lombardi, S., Saito, S., Simon, T., Wei, S., McPhail, K., Ramamoorthi, R., Sheikh, Y., Saragih, J.M.: Deep relightable appearance models for animatable faces. ACM Trans. Graph. **40**(4), 89–18915 (2021). https://doi.org/10.1145/3450626.3459829
22. Li, X., Sheng, B., Li, P., Kim, J., Feng, D.D.: Voxelized facial reconstruction using deep neural network. In: CGI, pp. 1–4 (2018). https://doi.org/10.1145/3208159.3208170
23. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P.V., Romero, J., Black, M.J.: Keep It SMPL: automatic estimation of 3D human pose and shape from a single image. In: ECCV, pp. 561–578 (2016). https://doi.org/10.1007/978-3-319-46454-1_34
24. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR, pp. 7122–7131 (2018).https://doi.org/10.1109/CVPR.2018.00744
25. 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: CVPR, pp. 5746–5756 (2021). https://doi.org/10.1109/CVPR46437.2021.00569
26. Wang, L., Zhao, X., Yu, T., Wang, S., Liu, Y.: NormalGAN: learning detailed 3D human from a single RGB-D image. In: ECCV,

vol. 12365, pp. 430–446 (2020). https://doi.org/10.1007/978-3-030-58565-5_26
27. Tian, Y., Zhang, H., Liu, Y., Wang, L.: Recovering 3D human mesh from monocular images: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **45**(12), 15406–15425 (2023). https://doi.org/10.1109/TPAMI.2023.3298850
28. Chen, L., Peng, S., Zhou, X.: Towards efficient and photorealistic 3D human reconstruction: a brief survey. Vis. Inform. **5**(4), 11–19 (2021). https://doi.org/10.1016/J.VISINF.2021.10.003
29. 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: 2022 14th International Conference on Signal Processing Systems (ICSPS), pp. 818–827. IEEE (2022)
30. Ma, Q., Saito, S., Yang, J., Tang, S., Black, M.J.: SCALE: modeling clothed humans with a surface codec of articulated local elements. In: CVPR, pp. 16082–16093 (2021). https://doi.org/10.1109/CVPR46437.2021.01582
31. Ma, Q., Yang, J., Tang, S., Black, M.J.: The power of points for modeling humans in clothing. In: ICCV, pp. 10954–10964 (2021). https://doi.org/10.1109/ICCV48922.2021.01079
32. Manfredi, G., Capece, N., Erra, U., Gilio, G., Baldi, V., Domenico, S.G.D.: TryItOn: a virtual dressing room with motion tracking and physically based garment simulation. In: XR, vol. 13445, pp. 63–76 (2022). https://doi.org/10.1007/978-3-031-15546-8_5
33. Fan, T., Yang, B., Bao, C., Wang, L., Zhang, G., Cui, Z.: HybridAvatar: efficient mesh-based human avatar generation from few-shot monocular images with implicit mesh displacement. In: IEEE International Symposium on Mixed and Augmented Reality Adjunct, ISMAR 2023, Sydney, Australia, October 16–20, 2023, pp. 371–376 (2023).https://doi.org/10.1109/ISMAR-ADJUNCT60411.2023.00080
34. Varol, G., Ceylan, D., Russell, B.C., Yang, J., Yumer, E., Laptev, I., Schmid, C.: BodyNet: volumetric inference of 3D human body shapes. In: ECCV, pp. 20–38 (2018). https://doi.org/10.1007/978-3-030-01234-2_2
35. Zheng, Z., Yu, T., Wei, Y., Dai, Q., Liu, Y.: DeepHuman: 3D human reconstruction from a single image. In: ICCV, pp. 7738–7748 (2019).https://doi.org/10.1109/ICCV.2019.00783
36. Tang, S., Tan, F., Cheng, K., Li, Z., Zhu, S., Tan, P.: A neural network for detailed human depth estimation from a single image. In: ICCV, pp. 7749–7758 (2019). https://doi.org/10.1109/ICCV.2019.00784
37. Smith, D., Loper, M., Hu, X., Mavroidis, P., Romero, J.: FACSIMILE: fast and accurate scans from an image in less than a second. In: ICCV, pp. 5329–5338 (2019). https://doi.org/10.1109/ICCV.2019.00543
38. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3D Gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–113914 (2023). https://doi.org/10.1145/3592433
39. Park, J.J., Florence, P.R., Straub, J., Newcombe, R.A., Lovegrove, S.: DeepSDF: learning continuous signed distance functions for shape representation. In: CVPR, pp. 165–174 (2019). https://doi.org/10.1109/CVPR.2019.00025
40. Mescheder, L.M., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: learning 3D reconstruction in function space. In: CVPR, pp. 4460–4470 (2019). https://doi.org/10.1109/CVPR.2019.00459
41. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: representing scenes as neural radiance fields for view synthesis. In: ECCV, pp. 405–421 (2020). https://doi.org/10.1007/978-3-030-58452-8_24
42. Tewari, A., Thies, J., Mildenhall, B., Srinivasan, P.P., Tretschk, E., Wang, Y., Lassner, C., Sitzmann, V., Martin-Brualla, R., Lombardi, S., Simon, T., Theobalt, C., Nießner, M., Barron, J.T., Wetzstein, G., Zollhöfer, M., Golyanik, V.: Advances in neural rendering. Comput. Graph. Forum **41**(2), 703–735 (2022). https://doi.org/10.1111/CGF.14507
43. Pfister, H., Zwicker, M., Baar, J., Gross, M.H.: Surfels: surface elements as rendering primitives. In: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH, pp. 335–342 (2000). https://doi.org/10.1145/344779.344936
44. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: SCAPE: shape completion and animation of people. ACM Trans. Graph. **24**(3), 408–416 (2005). https://doi.org/10.1145/1073204.1073207
45. Xu, H., Bazavan, E.G., Zanfir, A., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: GHUM & GHUML: generative 3D human shape and articulated pose models. In: CVPR, pp. 6183–6192 (2020). https://doi.org/10.1109/CVPR42600.2020.00622
46. Osman, A.A.A., Bolkart, T., Black, M.J.: STAR: sparse trained articulated human body regressor. In: ECCV, vol. 12351, pp. 598–613 (2020). https://doi.org/10.1007/978-3-030-58539-6_36
47. Zheng, Z., Yu, T., Liu, Y., Dai, Q.: PaMIR: parametric model-conditioned implicit representation for image-based human reconstruction. IEEE Trans. Pattern Anal. Mach. Intell. **44**(6), 3170–3184 (2022). https://doi.org/10.1109/TPAMI.2021.3050505
48. Hong, F., Chen, Z., Lan, Y., Pan, L., Liu, Z.: EVA3D: compositional 3D human generation from 2D image collections. In: ICLR (2023)
49. Dong, Z., Chen, X., Yang, J., Black, M.J., Hilliges, O., Geiger, A.: AG3D: learning to generate 3D avatars from 2D image collections. In: ICCV, pp. 14870–14881 (2023). https://doi.org/10.1109/ICCV51070.2023.01370
50. Huang, Y., Yi, H., Xiu, Y., Liao, T., Tang, J., Cai, D., Thies, J.: TeCH: text-guided reconstruction of lifelike clothed humans. In: 3DV (2024)
51. Albahar, B., Saito, S., Tseng, H., Kim, C., Kopf, J., Huang, J.: Single-image 3D human digitization with shape-guided diffusion. In: SIGGRAPH Asia 2023 Conference Papers, pp. 62–16211 (2023). https://doi.org/10.1145/3610548.3618153
52. Yao, J., Chen, J., Niu, L., Sheng, B.: Scene-aware human pose generation using transformer. In: MM, pp. 2847–2855 (2023). https://doi.org/10.1145/3581783.3612439
53. Kamel, A., Liu, B., Li, P., Sheng, B.: An investigation of 3D human pose estimation for learning Tai Chi: a human factor perspective. Int. J. Hum. Comput. Interact. **35**(4–5), 427–439 (2019). https://doi.org/10.1080/10447318.2018.1543081
54. Kamel, A., Sheng, B., Li, P., Kim, J., Feng, D.D.: Efficient body motion quantification and similarity evaluation using 3-D joints skeleton coordinates. IEEE Trans. Syst. Man Cybern. Syst. **51**(5), 2774–2788 (2021). https://doi.org/10.1109/TSMC.2019.2916896
55. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4D scans. ACM Trans. Graph. **36**(6), 194–119417 (2017). https://doi.org/10.1145/3130800.3130813
56. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: modeling and capturing hands and bodies together. ACM Trans. Graph. **36**(6), 245–124517 (2017). https://doi.org/10.1145/3130800.3130883
57. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: CVPR, pp. 10975–10985 (2019). https://doi.org/10.1109/CVPR.2019.01123
58. Zhu, H., Zuo, X., Wang, S., Cao, X., Yang, R.: Detailed human shape estimation from a single image by hierarchical mesh deformation. In: CVPR, pp. 4491–4500 (2019). https://doi.org/10.1109/CVPR.2019.00462

59. Lazova, V., Insafutdinov, E., Pons-Moll, G.: 360-Degree textures of people in clothing from a single image. In: 3DV, pp. 643–653 (2019). https://doi.org/10.1109/3DV.2019.00076
60. 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: CVPR, pp. 6468–6477 (2020). https://doi.org/10.1109/CVPR42600.2020.00650
61. Bhatnagar, B.L., Tiwari, G., Theobalt, C., Pons-Moll, G.: Multi-Garment Net: learning to dress 3D people from images. In: ICCV, pp. 5419–5429 (2019). https://doi.org/10.1109/ICCV.2019.00552
62. Jiang, B., Zhang, J., Hong, Y., Luo, J., Liu, L., Bao, H.: BCNet: learning body and cloth shape from a single image. In: ECCV, vol. 12365, pp. 18–35 (2020). https://doi.org/10.1007/978-3-030-58565-5_2
63. Patel, C., Liao, Z., Pons-Moll, G.: TailorNet: predicting clothing in 3D as a function of human pose, shape and garment style. In: CVPR, pp. 7363–7373 (2020).https://doi.org/10.1109/CVPR42600.2020.00739
64. Corona, E., Pumarola, A., Alenyà, G., Pons-Moll, G., Moreno-Noguer, F.: SMPLicit: topology-aware generative model for clothed people. In: CVPR, pp. 11875–11885 (2021). https://doi.org/10.1109/CVPR46437.2021.01170
65. Luigi, L.D., Li, R., Guillard, B., Salzmann, M., Fua, P.: DrapeNet: garment generation and self-supervised draping. In: CVPR, pp. 1451–1460 (2023). https://doi.org/10.1109/CVPR52729.2023.00146
66. Mikić, I., Trivedi, M., Hunter, E., Cosman, P.: Human body model acquisition and tracking using voxel data. Int. J. Comput. Vis. **53**, 199–223 (2003)
67. Gilbert, A., Volino, M., Collomosse, J.P., Hilton, A.: Volumetric performance capture from minimal camera viewpoints. In: ECCV, vol. 11215, pp. 591–607 (2018). https://doi.org/10.1007/978-3-030-01252-6_35
68. Stoll, C., Hasler, N., Gall, J., Seidel, H., Theobalt, C.: Fast articulated motion tracking using a sums of Gaussians body model. In: ICCV, pp. 951–958 (2011).https://doi.org/10.1109/ICCV.2011.6126338
69. Robertini, N., Casas, D., Rhodin, H., Seidel, H., Theobalt, C.: Model-based outdoor performance capture. In: 3DV, pp. 166–175 (2016). https://doi.org/10.1109/3DV.2016.25
70. Chen, G., Wang, W.: A survey on 3D Gaussian splatting (2024). arXiv preprint arXiv:2401.03890
71. Bai, S., Li, J.: Progress and prospects in 3D generative AI: a technical overview including 3D human (2024). arXiv preprint arXiv:2401.02620
72. Wu, T., Yuan, Y.-J., Zhang, L.-X., Yang, J., Cao, Y.-P., Yan, L.-Q., Gao, L.: Recent advances in 3D Gaussian Splatting. Comput. Vis. Media (2024). https://doi.org/10.1007/s41095-024-0436-y
73. Xu, Z., Peng, S., Lin, H., He, G., Sun, J., Shen, Y., Bao, H., Zhou, X.: 4K4D: real-time 4D view synthesis at 4K resolution. In: CVPR (2024)
74. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Xinggang, W.: 4D Gaussian splatting for real-time dynamic scene rendering. In: CVPR (2024)
75. Luiten, J., Kopanas, G., Leibe, B., Ramanan, D.: Dynamic 3D Gaussians: tracking by persistent dynamic view synthesis. In: 3DV (2024)
76. Yang, Z., Gao, X., Zhou, W., Jiao, S., Zhang, Y., Jin, X.: Deformable 3D Gaussians for high-fidelity monocular dynamic scene reconstruction. In: CVPR (2024)
77. Huang, B., Yu, Z., Chen, A., Geiger, A., Gao, S.: 2D Gaussian splatting for geometrically accurate radiance fields. In: ACM SIGGRAPH 2024 Conference Papers, SIGGRAPH 2024, Denver, CO, USA, 27 July 2024–1 August 2024, pp. 32 (2024). https://doi.org/10.1145/3641519.3657428
78. Guédon, A., Lepetit, V.: Sugar: Surface-aligned gaussian splatting for efficient 3D mesh reconstruction and high-quality mesh rendering. In: CVPR (2024)
79. Chen, H., Li, C., Lee, G.H.: NeuSG: neural implicit surface reconstruction with 3D Gaussian splatting guidance (2023). arXiv preprint arXiv:2312.00846
80. Chen, Z., Wang, F., Liu, H.: Text-to-3D using Gaussian splatting (2023). arXiv preprint arXiv:2309.16585
81. Li, X., Wang, H., Tseng, K.-K.: GaussianDiffusion: 3D Gaussian splatting for denoising diffusion probabilistic models with structured noise (2023). arXiv preprint arXiv:2311.11221
82. Tang, J., Ren, J., Zhou, H., Liu, Z., Zeng, G.: DreamGaussian: generative Gaussian splatting for efficient 3D content creation (2023). arXiv preprint arXiv:2309.16653
83. Zielonka, W., Bagautdinov, T., Saito, S., Zollhöfer, M., Thies, J., Romero, J.: Drivable 3D Gaussian avatars (2023). arXiv preprint arXiv:2311.13404
84. Shao, Z., Wang, Z., Li, Z., Wang, D., Lin, X., Zhang, Y., Fan, M., Wang, Z.: SplattingAvatar: realistic real-time human avatars with mesh-embedded Gaussian splatting. In: CVPR (2024)
85. Liu, X., Wu, C., Liu, J., Liu, X., Zhao, C., Feng, H., Ding, E., Wang, J.: GVA: reconstructing Vivid 3D Gaussian avatars from monocular videos. Arxiv (2024)
86. Svitov, D., Morerio, P., Agapito, L., Del Bue, A.: HAHA: highly articulated Gaussian human avatars with textured mesh prior (2024). arXiv preprint arXiv:2404.01053
87. Wen, J., Zhao, X., Ren, Z., Schwing, A., Wang, S.: GoMAvatar: efficient animatable human modeling from monocular video using Gaussians-on-mesh. In: CVPR (2024)
88. Jiang, Y., Liao, Q., Li, X., Ma, L., Zhang, Q., Zhang, C., Lu, Z., Shan, Y.: UV Gaussians: joint learning of mesh deformation and gaussian textures for human avatar modeling (2024). arXiv preprint arXiv:2403.11589
89. Liu, X., Zhan, X., Tang, J., Shan, Y., Zeng, G., Lin, D., Liu, X., Liu, Z.: HumanGaussian: text-driven 3D human generation with Gaussian splatting. In: CVPR (2024)
90. Abdal, R., Yifan, W., Shi, Z., Xu, Y., Po, R., Kuang, Z., Chen, Q., Yeung, D.-Y., Wetzstein, G.: Gaussian shell maps for efficient 3D human generation. In: CVPR (2024)
91. Cheng, W., Chen, R., Fan, S., Yin, W., Chen, K., Cai, Z., Wang, J., Gao, Y., Yu, Z., Lin, Z., Ren, D., Yang, L., Liu, Z., Loy, C.C., Qian, C., Wu, W., Lin, D., Dai, B., Lin, K.: DNA-rendering: a diverse neural actor repository for high-fidelity human-centric rendering. In: ICCV, pp. 19925–19936 (2023). https://doi.org/10.1109/ICCV51070.2023.01829
92. Bonopera, S., Hedman, P., Esnault, J., Prakash, S., Rodriguez, S., Thonat, T., Benadel, M., Chaurasia, G., Philip, J., Drettakis, G.: SIBR: a system for image based rendering (2020). https://sibr.gitlabpages.inria.fr/
93. Lorensen, W.E., Cline, H.E.: Marching cubes: a high resolution 3D surface construction algorithm. In: Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1987, pp. 163–169 (1987). https://doi.org/10.1145/37401.37422
94. Alldieck, T., Zanfir, M., Sminchisescu, C.: Photorealistic monocular 3D reconstruction of humans wearing clothing. In: CVPR, pp. 1496–1505 (2022). https://doi.org/10.1109/CVPR52688.2022.00156
95. Corona, E., Zanfir, M., Alldieck, T., Bazavan, E.G., Zanfir, A., Sminchisescu, C.: Structured 3D features for reconstructing controllable avatars. In: CVPR, pp. 16954–16964 (2023). https://doi.org/10.1109/CVPR52729.2023.01626
96. Lin, L., Zhu, J.: Topology-preserved human reconstruction with details. Vis. Comput. **39**(8), 3609–3619 (2023). https://doi.org/10.1007/S00371-023-02957-0

97. Hu, S., Hong, F., Pan, L., Mei, H., Yang, L., Liu, Z.: SHERF: generalizable human nerf from a single image. In: ICCV, pp. 9318–9330 (2023). https://doi.org/10.1109/ICCV51070.2023.00858
98. Huang, Y., Yi, H., Liu, W., Wang, H., Wu, B., Wang, W., Lin, B., Zhang, D., Cai, D.: One-shot implicit animatable avatars with model-based priors. In: ICCV, pp. 8940–8951 (2023). https://doi.org/10.1109/ICCV51070.2023.00824
99. Alldieck, T., Magnor, M.A., Xu, W., Theobalt, C., Pons-Moll, G.: Video based reconstruction of 3D people models. In: CVPR, pp. 8387–8397 (2018). https://doi.org/10.1109/CVPR.2018.00875
100. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Commun. ACM **63**(11), 139–144 (2020). https://doi.org/10.1145/3422622
101. Zhu, H., Qiu, L., Qiu, Y., Han, X.: Registering explicit to implicit: towards high-fidelity garment mesh reconstruction from single images. In: CVPR, pp. 3835–3844 (2022). https://doi.org/10.1109/CVPR52688.2022.00382
102. Cao, X., Santo, H., Shi, B., Okura, F., Matsushita, Y.: Bilateral normal integration. In: ECCV **13661**, 552–567 (2022). https://doi.org/10.1007/978-3-031-19769-7_32
103. Han, S., Park, M., Yoon, J.H., Kang, J., Park, Y., Jeon, H.: High-fidelity 3D human digitization from single 2K resolution images. In: CVPR, pp. 12869–12879 (2023).https://doi.org/10.1109/CVPR52729.2023.01237
104. Huang, Z., Xu, Y., Lassner, C., Li, H., Tung, T.: ARCH: animatable reconstruction of clothed humans. In: CVPR, pp. 3090–3099 (2020). https://doi.org/10.1109/CVPR42600.2020.00316
105. He, T., Xu, Y., Saito, S., Soatto, S., Tung, T.: ARCH++: animation-ready clothed human reconstruction revisited. In: ICCV, pp. 11026–11036 (2021). https://doi.org/10.1109/ICCV48922.2021.01086
106. Liao, T., Zhang, X., Xiu, Y., Yi, H., Liu, X., Qi, G., Zhang, Y., Wang, X., Zhu, X., Lei, Z.: High-fidelity clothed avatar reconstruction from a single image. In: CVPR, pp. 8662–8672 (2023). https://doi.org/10.1109/CVPR52729.2023.00837
107. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: DreamFusion: text-to-3D using 2D diffusion. In: ICLR (2023)
108. Chen, M., Chen, J., Ye, X., Gao, H.-a., Chen, X., Fan, Z., Zhao, H.: Ultraman: single image 3D human reconstruction with ultra speed and detail. arXiv preprint arXiv:2403.12028 (2024)
109. Isola, P., Zhu, J., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR, pp. 5967–5976 (2017). https://doi.org/10.1109/CVPR.2017.632
110. Moon, G., Nam, H., Shiratori, T., Lee, K.M.: 3D clothed human reconstruction in the wild. In: ECCV, vol. 13662, pp. 184–200 (2022). https://doi.org/10.1007/978-3-031-20086-1_11
111. Gabeur, V., Franco, J., Martin, X., Schmid, C., Rogez, G.: Moulding humans: non-parametric 3D human shape estimation from single images. In: ICCV, pp. 2232–2241 (2019).https://doi.org/10.1109/ICCV.2019.00232
112. Kazhdan, M.M., Bolitho, M., Hoppe, H.: Poisson surface reconstruction. In: Proceedings of the Fourth Eurographics Symposium on Geometry Processing, Cagliari, Sardinia, Italy, June 26–28, 2006. ACM International Conference Proceeding Series, vol. 256, pp. 61–70 (2006). https://doi.org/10.2312/SGP/SGP06/061-070
113. Chibane, J., Alldieck, T., Pons-Moll, G.: Implicit functions in feature space for 3D shape reconstruction and completion. In: CVPR, pp. 6968–6979 (2020). https://doi.org/10.1109/CVPR42600.2020.00700
114. Kazhdan, M.M., Hoppe, H.: Screened Poisson surface reconstruction. ACM Trans. Graph. **32**(3), 29–12913 (2013). https://doi.org/10.1145/2487228.2487237
115. Gao, J., Chen, W., Xiang, T., Jacobson, A., McGuire, M., Fidler, S.: Learning deformable tetrahedral meshes for 3D reconstruction. In: NeurIPS (2020)
116. Shen, T., Gao, J., Yin, K., Liu, M., Fidler, S.: Deep marching tetrahedra: a hybrid representation for high-resolution 3D shape synthesis. In: NeurIPS, pp. 6087–6101 (2021)
117. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR, pp. 10674–10685 (2022). https://doi.org/10.1109/CVPR52688.2022.01042
118. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV, pp. 3813–3824 (2023). https://doi.org/10.1109/ICCV51070.2023.00355
119. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: DreamBooth: fine tuning text-to-image diffusion models for subject-driven generation. In: CVPR, pp. 22500–22510 (2023). https://doi.org/10.1109/CVPR52729.2023.02155
120. Li, J., Li, D., Xiong, C., Hoi, S.C.H.: BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML, vol. 162, pp. 12888–12900 (2022)
121. Xiu, Y., Ye, Y., Liu, Z., Tzionas, D., Black, M.J.: PuzzleAvatar: assembling 3D avatars from personal albums (2024). arXiv preprint arXiv:2405.14869
122. Gao, X., Li, X., Zhang, C., Zhang, Q., Cao, Y., Shan, Y., Quan, L.: ConTex-Human: free-view rendering of human from a single image with texture-consistent synthesis (2023). arXiv preprint arXiv:2311.17123
123. Liu, R., Wu, R., Hoorick, B.V., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3D object. In: ICCV, pp. 9264–9275 (2023). https://doi.org/10.1109/ICCV51070.2023.00853
124. He, T., Collomosse, J.P., Jin, H., Soatto, S.: Geo-PIFu: geometry and pixel aligned implicit functions for single-view human reconstruction. In: NeurIPS (2020)
125. Wang, T., Liu, M., Zhu, J., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional GANs. In: CVPR, pp. 8798–8807 (2018). https://doi.org/10.1109/CVPR.2018.00917
126. Yang, X., Luo, Y., Xiu, Y., Wang, W., Xu, H., Fan, Z.: D-IF: uncertainty-aware human digitization via implicit distribution field. In: ICCV, pp. 9088–9098 (2023). https://doi.org/10.1109/ICCV51070.2023.00837
127. Cao, Y., Han, K., Wong, K.K.: SeSDF: self-evolved signed distance field for implicit 3D clothed human reconstruction. In: CVPR, pp. 4647–4657 (2023). https://doi.org/10.1109/CVPR52729.2023.00451
128. Song, D., Lee, H., Seo, J., Cho, D.: DIFu: depth-guided implicit function for clothed human reconstruction. In: CVPR, pp. 8738–8747 (2023). https://doi.org/10.1109/CVPR52729.2023.00844
129. Zhang, Z., Sun, L., Yang, Z., Chen, L., Yang, Y.: Global-correlated 3D-decoupling transformer for clothed avatar reconstruction. In: NeurIPS (2023)
130. Choi, H., Moon, G., Armando, M., Leroy, V., Lee, K.M., Rogez, G.: MonoNHR: monocular neural human renderer. In: 3DV, pp. 242–251 (2022). https://doi.org/10.1109/3DV57658.2022.00036
131. Weng, Z., Liu, J., Tan, H., Xu, Z., Zhou, Y., Yeung-Levy, S., Yang, J.: Single-view 3D human digitalization with large reconstruction models. arXiv preprint arXiv:2401.12175 (2024)
132. Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., Liu, F., Sunkavalli, K., Bui, T., Tan, H.: LRM: large reconstruction model for single image to 3D (2023). arXiv preprint arXiv:2311.04400
133. Xu, X., Loy, C.C.: 3D human texture estimation from a single image with transformers. In: ICCV, pp. 13829–13838 (2021). https://doi.org/10.1109/ICCV48922.2021.01359
134. Svitov, D., Gudkov, D., Bashirov, R., Lempitsky, V.: DINAR: diffusion inpainting of neural textures for one-shot human avatars. In: ICCV, pp. 7039–7049 (2023). https://doi.org/10.1109/ICCV51070.2023.00650

135. Zhan, X., Yang, J., Li, Y., Guo, J., Guo, Y., Wang, W.: Semantic human mesh reconstruction with textures (2024). arXiv preprint arXiv:2403.02561
136. Zhang, J., Li, X., Zhang, Q., Cao, Y., Shan, Y., Liao, J.: HumanRef: single image to 3D human generation via reference-guided diffusion. arXiv preprint arXiv:2311.16961 (2023)
137. Natsume, R., Saito, S., Huang, Z., Chen, W., Ma, C., Li, H., Morishima, S.: SiCloPe: silhouette-based clothed people. In: CVPR, pp. 4480–4490 (2019). https://doi.org/10.1109/CVPR.2019.00461
138. Sengupta, A., Alldieck, T., Kolotouros, N., Corona, E., Zanfir, A., Sminchisescu, C.: DiffHuman: probabilistic photorealistic 3D reconstruction of humans (2024). arXiv preprint arXiv:2404.00485
139. Wang, J., Zhong, Y., Li, Y., Zhang, C., Wei, Y.: Re-identification supervised texture generation. In: CVPR, pp. 11846–11856 (2019). https://doi.org/10.1109/CVPR.2019.01212
140. Xu, X., Chen, H., Moreno-Noguer, F., Jeni, L.A., Torre, F.D.: 3D human pose, shape and texture from low-resolution images and videos. IEEE Trans. Pattern Anal. Mach. Intell. **44**(9), 4490–4504 (2022). https://doi.org/10.1109/TPAMI.2021.3070002
141. Altindis, S.F., Meric, A., Dalva, Y., Gudukbay, U., Dundar, A.: Refining 3D human texture estimation from a single image (2023). arXiv preprint arXiv:2303.03471
142. Fang, Q., Shuai, Q., Dong, J., Bao, H., Zhou, X.: Reconstructing 3D human pose by watching humans in the mirror. In: CVPR, pp. 12814–12823 (2021). https://doi.org/10.1109/CVPR46437.2021.01262
143. 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: CVPR, pp. 9054–9063 (2021). https://doi.org/10.1109/CVPR46437.2021.00894
144. Xu, W., Chatterjee, A., Zollhöfer, M., Rhodin, H., Mehta, D., Seidel, H., Theobalt, C.: MonoPerfCap: human performance capture from monocular video. ACM Trans. Graph. **37**(2), 27 (2018). https://doi.org/10.1145/3181973
145. Habermann, M., Xu, W., Zollhöfer, M., Pons-Moll, G., Theobalt, C.: LiveCap: real-time human performance capture from monocular video. ACM Trans. Graph. **38**(2), 14–11417 (2019). https://doi.org/10.1145/3311970
146. Habermann, M., Xu, W., Zollhöfer, M., Pons-Moll, G., Theobalt, C.: DeepCap: monocular human performance capture using weak supervision. In: CVPR, pp. 5051–5062 (2020)
147. Alldieck, T., Magnor, M.A., Xu, W., Theobalt, C., Pons-Moll, G.: Detailed human avatars from monocular video. In: 3DV, pp. 98–109 (2018). https://doi.org/10.1109/3DV.2018.00022
148. Jiang, B., Hong, Y., Bao, H., Zhang, J.: SelfRecon: self reconstruction your digital avatar from monocular video. In: CVPR, pp. 5595–5605 (2022). https://doi.org/10.1109/CVPR52688.2022.00552
149. Peng, S., Dong, J., Wang, Q., Zhang, S., Shuai, Q., Zhou, X., Bao, H.: Animatable neural radiance fields for modeling dynamic human bodies. In: ICCV, pp. 14294–14303 (2021). https://doi.org/10.1109/ICCV48922.2021.01405
150. Chen, J., Zhang, Y., Kang, D., Zhe, X., Bao, L., Jia, X., Lu, H.: Animatable neural radiance fields from monocular RGB videos (2021). arXiv preprint arXiv:2106.13629
151. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR, pp. 586–595 (2018). https://doi.org/10.1109/CVPR.2018.00068
152. Li, R., Tanke, J., Vo, M., Zollhöfer, M., Gall, J., Kanazawa, A., Lassner, C.: TAVA: template-free animatable volumetric actors. In: ECCV, vol. 13692, pp. 419–436 (2022). https://doi.org/10.1007/978-3-031-19824-3_25
153. Jiang, W., Yi, K.M., Samei, G., Tuzel, O., Ranjan, A.: NeuMan: Neural human radiance field from a single video. In: ECCV, vol. 13692, pp. 402–418 (2022).https://doi.org/10.1007/978-3-031-19824-3_24
154. Yu, Z., Cheng, W., Liu, X., Wu, W., Lin, K.: MonoHuman: animatable human neural field from monocular video. In: CVPR, pp. 16943–16953 (2023).https://doi.org/10.1109/CVPR52729.2023.01625
155. Wang, S., Schwarz, K., Geiger, A., Tang, S.: ARAH: animatable volume rendering of articulated human SDFs. In: ECCV, vol. 13692, pp. 1–19 (2022). https://doi.org/10.1007/978-3-031-19824-3_1
156. Gropp, A., Yariv, L., Haim, N., Atzmon, M., Lipman, Y.: Implicit geometric regularization for learning shapes. In: ICML. Proceedings of Machine Learning Research, vol. 119, pp. 3789–3799 (2020)
157. Jiang, T., Chen, X., Song, J., Hilliges, O.: InstantAvatar: learning avatars from monocular video in 60 seconds. In: CVPR, pp. 16922–16932 (2023).https://doi.org/10.1109/CVPR52729.2023.01623
158. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution Hash encoding. ACM Trans. Graph. **41**(4), 102–110215 (2022). https://doi.org/10.1145/3528223.3530127
159. Feng, Y., Yang, J., Pollefeys, M., Black, M.J., Bolkart, T.: Capturing and animation of body and clothing from monocular video. In: SIGGRAPH Asia 2022 Conference Papers, pp. 45–1459 (2022). https://doi.org/10.1145/3550469.3555423
160. Zheng, Z., Huang, H., Yu, T., Zhang, H., Guo, Y., Liu, Y.: Structured local radiance fields for human avatar modeling. In: CVPR, pp. 15872–15882 (2022). https://doi.org/10.1109/CVPR52688.2022.01543
161. Su, S., Yu, F., Zollhöfer, M., Rhodin, H.: A-NeRF: articulated neural radiance fields for learning human shape, appearance, and pose. In: NeurIPS, pp. 12278–12291 (2021)
162. Xu, T., Fujita, Y., Matsumoto, E.: Surface-aligned neural radiance fields for controllable 3D human synthesis. In: CVPR, pp. 15862–15871 (2022). https://doi.org/10.1109/CVPR52688.2022.01542
163. Liu, L., Habermann, M., Rudnev, V., Sarkar, K., Gu, J., Theobalt, C.: Neural actor: neural free-view synthesis of human actors with pose control. ACM Trans. Graph. **40**(6), 219–121916 (2021). https://doi.org/10.1145/3478513.3480528
164. Chen, Y., Wang, X., Chen, X., Zhang, Q., Li, X., Guo, Y., Wang, J., Wang, F.: UV volumes for real-time rendering of editable free-view human performance. In: CVPR, pp. 16621–16631 (2023). https://doi.org/10.1109/CVPR52729.2023.01595
165. Li, Z., Zheng, Z., Wang, L., Liu, Y.: Animatable Gaussians: learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In: CVPR (2024)
166. Lei, J., Wang, Y., Pavlakos, G., Liu, L., Daniilidis, K.: GART: Gaussian articulated template models. In: CVPR (2024)
167. Kocabas, M., Chang, J.-H.R., Gabriel, J., Tuzel, O., Ranjan, A.: HUGS: human Gaussian splats. In: CVPR (2024)
168. Hu, L., Zhang, H., Zhang, Y., Zhou, B., Liu, B., Zhang, S., Nie, L.: GaussianAvatar: towards realistic human avatar modeling from a single video via animatable 3D Gaussians. In: CVPR (2024)
169. Pang, H., Zhu, H., Kortylewski, A., Theobalt, C., Habermann, M.: ASH: animatable Gaussian splats for efficient and photoreal human rendering. In: CVPR (2024)
170. Guo, C., Jiang, T., Chen, X., Song, J., Hilliges, O.: Vid2Avatar: 3D avatar reconstruction from videos in the wild via self-supervised scene decomposition. In: CVPR, pp. 12858–12868 (2023). https://doi.org/10.1109/CVPR52729.2023.01236
171. Feng, Y., Liu, W., Bolkart, T., Yang, J., Pollefeys, M., Black, M.J.: Learning disentangled avatars with hybrid 3D representations. arXiv (2023)

172. Wang, K., Zhang, G., Cong, S., Yang, J.: Clothed human performance capture with a double-layer neural radiance fields. In: CVPR, pp. 21098–21107 (2023). https://doi.org/10.1109/CVPR52729.2023.02021
173. Chen, M., Zhang, J., Xu, X., Liu, L., Cai, Y., Feng, J., Yan, S.: Geometry-guided progressive nerf for generalizable and efficient neural human rendering. In: ECCV, vol. 13683, pp. 222–239 (2022). https://doi.org/10.1007/978-3-031-20050-2_14
174. Peng, B., Hu, J., Zhou, J., Zhang, J.: SelfNeRF: fast training NeRF for human from monocular self-rotating video (2022). arXiv preprint arXiv:2210.01651
175. Geng, C., Peng, S., Xu, Z., Bao, H., Zhou, X.: Learning neural volumetric representations of dynamic humans in minutes. In: CVPR, pp. 8759–8770 (2023).https://doi.org/10.1109/CVPR52729.2023.00846
176. Kwon, Y., Kim, D., Ceylan, D., Fuchs, H.: Neural human performer: learning generalizable radiance fields for human performance rendering. In: NeurIPS, pp. 24741–24752 (2021)
177. Li, C., Lin, J., Lee, G.H.: GHuNeRF: generalizable human NeRF from a monocular video (2023). arXiv preprint arXiv:2308.16576
178. Chen, X., Zheng, Y., Black, M.J., Hilliges, O., Geiger, A.: SNARF: differentiable forward skinning for animating non-rigid neural implicit shapes. In: ICCV, pp. 11574–11584 (2021). https://doi.org/10.1109/ICCV48922.2021.01139
179. Chen, X., Jiang, T., Song, J., Rietmann, M., Geiger, A., Black, M.J., Hilliges, O.: Fast-SNARF: a fast deformer for articulated neural fields. IEEE Trans. Pattern Anal. Mach. Intell. **45**(10), 11796–11809 (2023). https://doi.org/10.1109/TPAMI.2023.3271569
180. Zhi, Y., Qian, S., Yan, X., Gao, S.: Dual-space NeRF: learning animatable avatars and scene lighting in separate spaces. In: 3DV, pp. 1–10 (2022). https://doi.org/10.1109/3DV57658.2022.00048
181. Mu, J., Sang, S., Vasconcelos, N., Wang, X.: ActorsNeRF: animatable few-shot human rendering with generalizable NeRFs. In: ICCV, pp. 18345–18355 (2023). https://doi.org/10.1109/ICCV51070.2023.01686
182. Noguchi, A., Sun, X., Lin, S., Harada, T.: Neural articulated radiance field. In: ICCV, pp. 5742–5752 (2021). https://doi.org/10.1109/ICCV48922.2021.00571
183. Te, G., Li, X., Li, X., Wang, J., Hu, W., Lu, Y.: Neural capture of animatable 3D human from monocular video. In: ECCV, vol. 13666, pp. 275–291 (2022). https://doi.org/10.1007/978-3-031-20068-7_16
184. Su, S., Bagautdinov, T.M., Rhodin, H.: DANBO: disentangled articulated neural body representations via graph neural networks. In: ECCV, vol. 13662, pp. 107–124 (2022).https://doi.org/10.1007/978-3-031-20086-1_7
185. Zhang, R., Chen, J.: NDF: neural deformable fields for dynamic human modelling. In: ECCV, vol. 13692, pp. 37–52 (2022).https://doi.org/10.1007/978-3-031-19824-3_3
186. Li, M., Tao, J., Yang, Z., Yang, Y.: Human101: Training 100+FPS human Gaussians in 100s from 1 view (2023). arXiv preprint arXiv:2312.15258
187. Moreau, A., Song, J., Dhamo, H., Shaw, R., Zhou, Y., Pérez-Pellitero, E.: Human Gaussian splatting: real-time rendering of animatable avatars (2023). arXiv preprint arXiv:2311.17113
188. Qian, Z., Wang, S., Mihajlovic, M., Geiger, A., Tang, S.: 3DGS-Avatar: animatable avatars via deformable 3D Gaussian splatting. In: CVPR (2024)
189. Li, M., Yao, S., Xie, Z., Chen, K., Jiang, Y.-G.: GaussianBody: clothed human reconstruction via 3D Gaussian splatting (2024). arXiv preprint arXiv:2401.09720
190. Jung, H., Brasch, N., Song, J., Perez-Pellitero, E., Zhou, Y., Li, Z., Navab, N., Busam, B.: Deformable 3D Gaussian splatting for animatable human avatars (2023). arXiv preprint arXiv:2312.15059
191. Jena, R., Iyer, G.S., Choudhary, S., Smith, B., Chaudhari, P., Gee, J.: SplatArmor: articulated Gaussian splatting for animatable humans from monocular RGB videos (2023). arXiv preprint arXiv:2311.10812
192. Kamel, A., Sheng, B., Yang, P., Li, P., Shen, R., Feng, D.D.: Deep convolutional neural networks for human action recognition using depth maps and postures. IEEE Trans. Syst. Man Cybern. Syst. **49**(9), 1806–1819 (2019). https://doi.org/10.1109/TSMC.2018.2850149
193. 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. **36**(7), 1325–1339 (2014). https://doi.org/10.1109/TPAMI.2013.248
194. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B.C., Matthews, I.A., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: a massively multiview system for social motion capture. In: ICCV, pp. 3334–3342 (2015). https://doi.org/10.1109/ICCV.2015.381
195. 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: 3DV, pp. 506–516 (2017). https://doi.org/10.1109/3DV.2017.00064
196. 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: ECCV, vol. 11214, pp. 614–631 (2018). https://doi.org/10.1007/978-3-030-01249-6_37
197. Tsuchida, S., Fukayama, S., Hamasaki, M., Goto, M.: AIST dance video database: multi-genre, multi-dancer, and multi-camera database for dance information processing. In: ISMIR, pp. 501–510 (2019)
198. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: AI choreographer: music conditioned 3D dance generation with AIST++. In: ICCV, pp. 13381–13392 (2021).https://doi.org/10.1109/ICCV48922.2021.01315
199. Isik, M., Rünz, M., Georgopoulos, M., Khakhulin, T., Starck, J., Agapito, L., Nießner, M.: HumanRF: high-fidelity neural radiance fields for humans in motion. ACM Trans. Graph. **42**(4), 160–116012 (2023). https://doi.org/10.1145/3592415
200. Cai, Z., Ren, D., Zeng, A., Lin, Z., Yu, T., Wang, W., Fan, X., Gao, Y., Yu, Y., Pan, L., Hong, F., Zhang, M., Loy, C.C., Yang, L., Liu, Z.: HuMMan: multi-modal 4D human dataset for versatile sensing and modeling. In: ECCV, vol. 13667, pp. 557–577 (2022). https://doi.org/10.1007/978-3-031-20071-7_33
201. Cheng, W., Xu, S., Piao, J., Qian, C., Wu, W., Lin, K.-Y., Li, H.: Generalizable neural performer: learning robust radiance fields for human novel view synthesis (2022). arXiv preprint arXiv:2204.11798
202. Xiong, Z., Li, C., Liu, K., Liao, H., Hu, J., Zhu, J., Ning, S., Qiu, L., Wang, C., Wang, S., et al.: MVHumanNet: a large-scale dataset of multi-view daily dressing human captures (2023). arXiv preprint arXiv:2312.02963
203. Zhang, C., Pujades, S., Black, M.J., Pons-Moll, G.: Detailed, accurate, human shape estimation from clothed 3D scan sequences. In: CVPR, pp. 5484–5493 (2017).https://doi.org/10.1109/CVPR.2017.582
204. Su, Z., Yu, T., Wang, Y., Liu, Y.: DeepCloth: neural garment representation for shape and style editing. IEEE Trans. Pattern Anal. Mach. Intell. **45**(2), 1581–1593 (2023). https://doi.org/10.1109/TPAMI.2022.3168569
205. Habermann, M., Liu, L., Xu, W., Zollhöfer, M., Pons-Moll, G., Theobalt, C.: Real-time deep dynamic characters. ACM Trans. Graph. **40**(4), 94–19416 (2021). https://doi.org/10.1145/3450626.3459749
206. 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 expres-

sions. In: CVPR, pp. 2987–2997 (2020). https://doi.org/10.1109/CVPR42600.2020.00306
207. Yoon, J.S., Yu, Z., Park, J., Park, H.S.: HUMBI: a large multiview dataset of human body expressions and benchmark challenge. IEEE Trans. Pattern Anal. Mach. Intell. **45**(1), 623–640 (2023). https://doi.org/10.1109/TPAMI.2021.3138762
208. Over 4,000 Scanned 3D People Models. https://renderpeople.com/
209. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861
210. Zheng, Z., Zhao, X., Zhang, H., Liu, B., Liu, Y.: AvatarReX: real-time expressive full-body avatars. ACM Trans. Graph. **42**(4), 158–115819 (2023). https://doi.org/10.1145/3592101
211. Dong, J., Fang, Q., Guo, Y., Peng, S., Shuai, Q., Zhou, X., Bao, H.: TotalSelfScan: learning full-body avatars from self-portrait videos of faces, hands, and bodies. In: NeurIPS (2022)
212. Yu, T., Zheng, Z., Guo, K., Zhao, J., Dai, Q., Li, H., Pons-Moll, G., Liu, Y.: DoubleFusion: real-time capture of human performances with inner body shapes from a single depth sensor. In: CVPR, pp. 7287–7296 (2018). https://doi.org/10.1109/CVPR.2018.00761
213. Lin, S., Li, Z., Su, Z., Zheng, Z., Zhang, H., Liu, Y.: LayGA: layered Gaussian avatars for animatable clothing transfer (2024). arXiv preprint arXiv:2405.07319
214. Khirodkar, R., Tripathi, S., Kitani, K.: Occluded human mesh recovery. In: CVPR, pp. 1705–1715 (2022). https://doi.org/10.1109/CVPR52688.2022.00176
215. Wang, J., Yoon, J.S., Wang, T.Y., Singh, K.K., Neumann, U.: Complete 3D human reconstruction from a single incomplete image. In: CVPR, pp. 8748–8758 (2023). https://doi.org/10.1109/CVPR52729.2023.00845
216. Xiang, T., Sun, A., Wu, J., Adeli, E., Fei-Fei, L.: Rendering Humans from object-occluded monocular videos. In: ICCV, pp. 3216–3227 (2023). https://doi.org/10.1109/ICCV51070.2023.00300
217. Ye, J., Zhang, Z., Jiang, Y., Liao, Q., Yang, W., Lu, Z.: OccGaussian: 3D Gaussian splatting for occluded human rendering (2024). arXiv preprint arXiv:2404.08449

Shuo Yang is currently a master candidate at Hangzhou Dianzi University. His research interests include computer vision and computer graphics.

Xiaoling Gu received the Ph.D. degree in computer science from Zhejiang University in 2017. Now she is working at School of Computer Science and Technology, Hangzhou Dianzi University. Her current research interests are in the areas of computer vision, machine learning, vision and language, and fashion data analysis. She has published several top-tier conference and journal papers, e.g. SIGIR, ACM Multimedia, IEEE Transaction on Multimedia, etc.

Zhenzhong Kuang received Ph.D. degree from the China University of Petroleum, China, in 2017. He is currently works as an associate professor with the School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China. His research interests include 2D/3D visual recognition, privacy protection, multimedia analysis, and machine learning.

Feiwei Qin received the Ph.D. degree in computer science and technology from Zhejiang University, Hangzhou, China, in 2014. He is currently a Professor with the School of Computer Science and Technology, Hangzhou Dianzi University. His research interests include artificial intelligence, machine learning, image processing, and computer-aided design.

Zizhao Wu is currently an Associate Professor with the Faculty of Digital Media Technology, Hangzhou Dianzi University. He received the Ph.D. degree from the Department of Computer Science and Technology, Zhejiang University, in 2013. His main research interests include computer vision and computer graphics.