Recovering 3D Human Mesh from Monocular Images: A Survey

Yating Tian, Hongwen Zhang, Yebin Liu, *Member, IEEE*, and Limin Wang, *Member, IEEE*

Abstract—Estimating human pose and shape from monocular images is a long-standing problem in computer vision. Since the release of statistical body models, 3D human mesh recovery has been drawing broader attention. With the same goal of obtaining well-aligned and physically plausible mesh results, two paradigms have been developed to overcome challenges in the 2D-to-3D lifting process: i) an optimization-based paradigm, where different data terms and regularization terms are exploited as optimization objectives; and ii) a regression-based paradigm, where deep learning techniques are embraced to solve the problem in an end-to-end fashion. Meanwhile, continuous efforts are devoted to improving the quality of 3D mesh labels for a wide range of datasets. Though remarkable progress has been achieved in the past decade, the task is still challenging due to flexible body motions, diverse appearances, complex environments, and insufficient in-the-wild annotations. To the best of our knowledge, this is the first survey that focuses on the task of monocular 3D human mesh recovery. We start with the introduction of body models and then elaborate recovery frameworks and training objectives by providing in-depth analyses of their strengths and weaknesses. We also summarize datasets, evaluation metrics, and benchmark results. Open issues and future directions are discussed in the end, hoping to motivate researchers and facilitate their research in this area. A regularly updated project page can be found at https://github.com/tinatiansjz/hmr-survey.

Index Terms—3D human mesh recovery, 3D from monocular images, gestures and pose, deep learning, literature survey.

✦

1 INTRODUCTION

Understanding humans from monocular images is one of the fundamental tasks in computer vision. Over the past two decades, the research community has focused on predicting 2D contents such as keypoints [1], [2], [3], silhouettes [4], and part segmentations [5] from RGB images. With these advances, researchers further seek to estimate human pose in 3D space [6], [7], [8], [9], [10], [11], [12]. Although simple movements can be represented relatively clearly by 2D contents or a few sparse 3D joints, complex human behaviors require more descriptions of the human body with a finer granularity. Moreover, it is critical to reason about body shape, contact, gesture, and expression since we interact with the world using our surface skin instead of unobserved joints.

In recent years, the community has shifted its interests towards 3D mesh recovery of human bodies [13], [14], [15], [16], [17], [18], [19], [20], [21] along with expressive face and hands [22], [23], [24], [25], [26]. This trend is inseparable from the success of statistical human models. As shown in Fig. 1, since the release of the SMPL model [27] in 2015 and the SMPL-X model [22] in 2019, they have gained increasing interest as their annual citations grow rapidly year by year. The recovery of human body meshes plays a key role in facilitating the downstream tasks such as clothed human reconstruction [28], [29], [30], [31], [32], [33], [34], rendering [35], [36], and avatar modeling [37], [38], [39],

- *Yating Tian and Limin Wang are with the Department of Computer Science and Technology, Nanjing University, China. Email: yatingtian@smail.nju.edu.cn; lmwang@nju.edu.cn (Corresponding author: Limin Wang)*
- *Hongwen Zhang and Yebin Liu are with the Department of Automation, Tsinghua University, China. Email: zhanghongwen@mail.tsinghua.edu.cn; liuyebin@mail.tsinghua.edu.cn (Yating Tian and Hongwen Zhang are co-first authors.)*
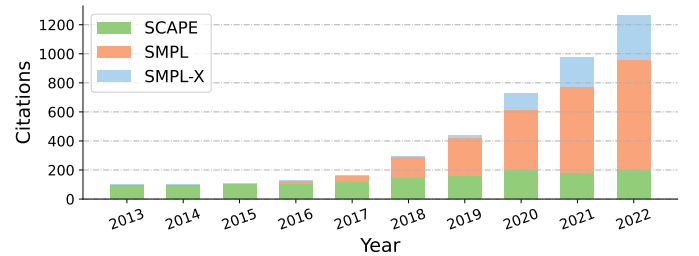

Fig. 1. The annual citations of three representative 3D statistical human models, i.e., SCAPE [41], SMPL [27], and SMPL-X [22].

[40]. It is also involved in widespread applications such as VR/AR content creation, virtual try-on, and computer-assistant coaching, as depicted in Fig. 2.

Recovering 3D human mesh from monocular images is quite challenging, owing to the issues such as inherent ambiguities in lifting 2D observations to 3D space, flexible body kinematic structures, complex intersections with the environments, and insufficient annotated 3D data. To address these issues, two different paradigms have been investigated in this field for the recovery of well-aligned and physically plausible results. Following the optimization-based paradigm [13], [15], [42], methods explicitly fit body models to 2D observations in an iterative manner. Various data terms and regularization terms are explored as optimization objectives. Alternatively, the regression-based paradigm [16], [17], [18], [43], [44] takes advantage of the powerful nonlinear mapping capability of neural networks and directly predicts model parameters from raw image pixels. Different network architectures and regression targets are designed to achieve better performances. Meanwhile, significant efforts have also been devoted to creating various datasets to facilitate the research of this task. Despite the remarkable progress achieved in recent years, the research
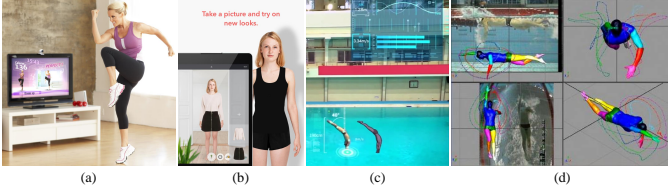
Fig. 2. Real-world applications of human mesh recovery: (a) a video game for fitness ☑; (b) virtual try-on ☑; (c) a 3D+AI coaching system for diving ☑; (d) dynamic simulations during swimming ☑.

community still faces challenges toward the ultimate goal of robust, accurate, and efficient human mesh recovery.

1.1 Scope

This survey mainly focuses on approaches to monocular 3D human mesh recovery (a.k.a. 3D human pose and shape estimation) in the deep learning era. Single RGB images and monocular RGB videos (or "monocular images" to refer to them collectively) as input are considered. In addition to single-person recovery from monocular images, we also take multi-person recovery into account. As for the reconstruction target, statistical human models are used to estimate body shape under clothing. RGBD and multi-view inputs are beneficial to resolve ambiguities, but they are not in the scope of this review. We simply ignore the modeling of clothes, which is a step towards photorealism. We refer readers to [45] for clothed human reconstruction. We also do not cover work on neural rendering [35], [46] that focuses on the appearance modeling instead of geometry. This survey is also complementary to existing survey papers focusing on 2D/3D human pose estimation [47], [48], [49].

1.2 Organization

The rest of the survey is organized as follows. In Section 2, we give a brief introduction of the development history of human models and provide detailed information on the SMPL model [27], the most widely used template for human reasoning. Section 3 describes approaches to body recovery and whole-body recovery with hands and face. Methods are categorized into an optimization-based paradigm or a regression-based paradigm. In Section 4 and 5, we sort out novel modules that help to deal with videos or multi-person recovery. However, results may be physically unreasonable and suffer from visual defects if we merely supervise the human body with regular data terms. Thus, in Section 7, we discuss the strategies used to enhance physical plausibility by involving realistic camera models, contact constraints, and human priors. The commonly used datasets and evaluation criteria, along with the benchmark leaderboard, are summarized in Sections 8 and 9. Finally, we draw conclusions and point out worthwhile future directions in Section 10.

2 HUMAN MODELING

The human body can be abstracted as a stick figure [50], simply marking the keypoints in body, hands, and face and connecting them with sticks, as shown in Fig. 3(a). However, we interact with the world through surface contacts and facial expressions, which requires the modeling of both

body pose and shape. In early work [51], [52], a wide variety of geometric primitives have been studied to approximate body shapes. Later, inspired by the breakthrough [53] in face modeling, researchers derive body shape constraints from 3D scanned data and create body models [27], [41] from a statistical viewpoint. Based on modeling details, we classify the modeling process into two classes: methods that represent the human body with *geometric primitives*, and methods that use subject-specific *body scanned data* to build a statistical 3D mesh model.

2.1 Geometric Primitives

Body modeling starts by manipulating a bunch of geometric primitives, including planar rectangles [52], cylinders [51], [54], [55], [56], [57], [58], and epllisoids [59], as shown in Fig. 3(b). Nevatia et al. [51] use generalized cylinders to fit range data. Marr et al. [54] propose a general, compositional 3D shape representation. Pentlan et al. [60] attempt to track a jumping man using a model with spring-like connections between body parts. Later, more sophisticated primitives were proposed, such as superquadric ellipsoids [61], [62], [63], metaballs [64] and customized graphical model [59], [65]. By then, human body models were hand-crafted, unrealistic, and tended to be brittle.

2.2 Statistical Modeling

Compared to primitives-based models, full-body 3D scans offer more detailed measurements of the body surface, but the modeling process is much more complicated. To convert a dense point cloud and a triangulated mesh from 3D scans to a watertight and animatable 3D human body mesh, three main pre-processing steps are taken [66]: (i) *template mesh registration*: fit a template mesh to the 3D point cloud to deal with holes that the triangulated mesh contains; (ii) *skeleton fitting*: determine the number of joints and the location and axis orientations of rotations for each joint; (iii) *skinning*: bind every vertex in the surface to the skeleton for animation.

2.2.1 Body Modeling

Statistical body modeling refers to learning a statistical body model by exploiting an extensive collection of 3D body scans and simply ignoring hand articulation or facial expression. There has been a lot of research [41], [67], [68], [69], [70], [71], [72], [73], [74], [75] on learning highly realistic human body models from scanning data like CAESAR [76]. Among them, SCAPE [41] and SMPL [27] are two representative models that factor body deformations into identity-dependent and pose-dependent shape deformations.

SCAPE [41] is a deformable human body model that represents the individual shape and the pose-dependent shape via triangle deformations. During processing, Anguelov et al. combine static scans of several people with the scans of a single person in various poses. SCAPE is one of the most successful human models. Many models [68], [69], [70], [71], [72], [77] are built upon SCAPE. The stitched puppet model [77] combines the realism of statistical models with the advantages of part-based representations. Dyna [72], an extension of SCAPE, relates soft-tissue deformations to

motion and body shape and enables itself to produce a wide range of realistic soft-tissue motions.

SMPL [27] is a vertex-based linear model depicting minimally-clothed humans in natural poses, which is currently the most widely used human body model in the research community. It is compatible with existing rendering engines. Like SCAPE [41], SMPL factors deformations into shape and pose deformations. Two basic sets of parameters control pose deformation $\boldsymbol{\theta}$ and shape variation $\boldsymbol{\beta}$, respectively. The pose parameters $\boldsymbol{\theta} = [\boldsymbol{w}_0^T, ..., \boldsymbol{w}_K^T]^T$ are defined by a standard skeletal rig at $K = 23$ joints, where $\boldsymbol{w}_k \in \mathbb{R}^3$ denotes the relative rotation of part k w.r.t. its parent in the kinematic tree and $\boldsymbol{w}_0$ refers to the root orientation. The shape parameters $\boldsymbol{\beta} \in \mathbb{R}^m$ are coefficients of the top-m principal components in a low-dimensional shape space after principal component analysis (PCA). SMPL can be represented as a function $M(\cdot)$ mapping pose parameters $\boldsymbol{\theta}$ and shape parameters $\boldsymbol{\beta}$ to a triangulated mesh with $N = 6890$ vertices. It is formulated as an additive model in the vertex space. Specifically, a posed human body instance can be obtained as follows:

$$M(\boldsymbol{\beta}, \boldsymbol{\theta}) = W(T(\boldsymbol{\beta}, \boldsymbol{\theta}), J(\boldsymbol{\beta}), \boldsymbol{\theta}; \mathcal{W}), \quad (1)$$

$$T(\boldsymbol{\beta}, \boldsymbol{\theta}) = \bar{\mathbf{T}} + B_s(\boldsymbol{\beta}) + B_p(\boldsymbol{\theta}), \quad (2)$$

where a rest pose $T(\boldsymbol{\beta}, \boldsymbol{\theta})$ is first generated by learning corrective blend shapes, i.e., pose-dependent deformations $B_P(\boldsymbol{\theta}) : \mathbb{R}^{|\boldsymbol{\theta}|} \mapsto \mathbb{R}^{3N}$ and shape-dependent deformations $B_S(\boldsymbol{\beta}) : \mathbb{R}^{|\boldsymbol{\beta}|} \mapsto \mathbb{R}^{3N}$, in order to deal with standard LBS artifacts [78]. A blend shape is a vector of displacements in the mean template shape $\bar{\mathbf{T}}$. Secondly, linear blend skinning function W with a set of blend weights $\mathcal{W} \in \mathbb{R}^{N \times K}$ and the pose parameters $\boldsymbol{\theta}$ allow to pose the T-shape mesh $T(\boldsymbol{\beta}, \boldsymbol{\theta})$ based on its skeleton joints locations $J(\boldsymbol{\beta}) : \mathbb{R}^{|\boldsymbol{\beta}|} \mapsto \mathbb{R}^{3K}$. Moreover, SMPL can be extended to capture soft-tissue dynamics [72]. Dynamic deformations of the resulting DMPL model is parameterized by coefficients $\boldsymbol{\delta}$.

The SMPL family has been growing. FLAME face model [79], MANO hands model [80], and SMIL infant body model [81] have been proposed, which are overall a linear blend skinning with shape and pose blend-shapes. Despite the success in the application, SMPL still has its limitation. First, its global blend shapes capture spurious long-range correlations and result in non-local deformation artifacts. Second, SMPL ignores correlations between body shape and pose-dependent shape deformation. In addition, SMPL relies on a linear PCA subspace to represent soft-tissue deformations, struggling to reproduce highly nonlinear deformations. Many researchers seek an improvement for descriptive capability [82], [83]. STAR [83] is a drop-in replacement for SMPL. It factorizes pose-dependent deformation into a set of sparse and spatially local pose-corrective blend-shape functions. SoftSMPL [82] defines a highly efficient nonlinear subspace to encode tissue deformations, compared to the linear descriptors [27], [72]. Recently, learning-based solutions are also explored to represent the body model in implicit [84], [85], [86], [87] or explicit [88] manners.

2.2.2 Whole Body Modeling

Recently, much progress has been made in modeling the human body together with hands [80], or with hands and
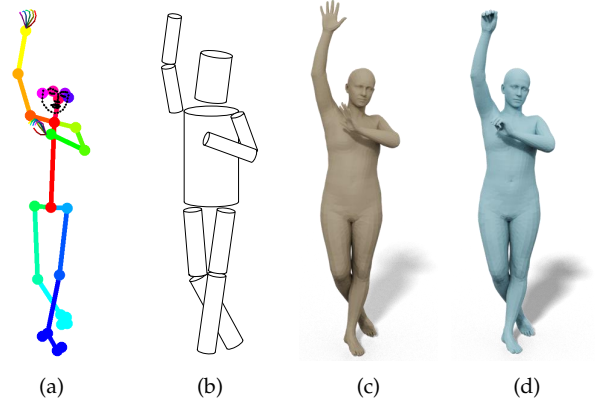

Fig. 3. Typical 2D and 3D human models representing the same posing human. (a) 2D skeletons [1], formed from the keypoints of body, hands and face; (b) a cylindrical body model; (c) SMPL [27]; (d) SMPL-X [22].

face [22], [89]. Romero et al. [80] attach MANO to SMPL to obtain a new articulated model (SMPL+H) with hands and body interaction. Frankenstein Model [89] combines a simplified version of SMPL [27] with an artist-designed hand rig and the FaceWarehouse face model [90]. These disparate models are integrated together, resulting in a model that is slightly out of proportion. A simpler parameterized model, Adam, is also introduced, which is more capable of body motion capture. Pavlakos et al. [22] learn a new, holistic model named SMPL-X that jointly models the human body, face, and hands. They extend SMPL with the FLAME head model [79] and the MANO hand model [80] and then register this combined model to CAESAR [76] scans to curate for quality. SMPL-X has several parameters representing the body, hand, and face. Initially, there are 75 rotational parameters for the global rotation and {body, eyes, jaw} poses; 24 low-dimensional PCA coefficient or 90 rotational parameters for hand poses; 10 for the body shape and 10 for the facial expressions. Following SMPL-X, SUPR [91] is proposed recently for more expressive and accurate modeling of head, hand, and foot. Besides a series of linear models based on SMPL, some attempts are devoted to different modeling strategies. For example, GHUM and GHUML shape spaces [92] are based on variational auto-encoders (VAE), which are nonlinear. All the model parameters, including shape spaces, pose-space deformation correctives, skeleton joint center predictors, and blend skinning functions, are trained end-to-end in a single consistent learning loop.

3 HUMAN MESH RECOVERY

Since the release of statistical body models, researchers have used them to estimate the shape and pose from monocular images. Balan et al. [93] pioneer in estimating the parameters of SCAPE [41] from images. Nowadays, SMPL [27] has been prevailing in academia for 3D body shape recovery. The credit goes to SMPL's open-source nature and its fast-developing community around it: the ground-truth acquisition methods [13], [94], datasets with extended SMPL annotations [42], [95], [96], [97], [98], and milestone works [16], [44], [99]. This section will sort out articles on human mesh recovery based on predefined body
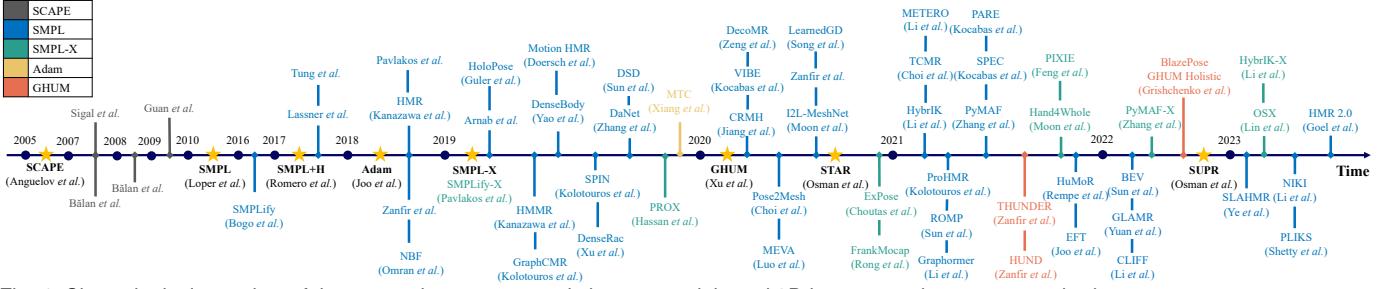
Fig. 4. Chronological overview of the most relevant parametric human models and 3D human mesh recovery methods.

models [22], [27], [41]. Body models capture the shape and pose variability but do not account for clothing or hair. Thus, to put it more precisely, approaches estimate the shape and pose of the body under clothing or in tight clothing. In Fig. 4, we demonstrate some representative methods. We categorize them based on the human models they adopt. In Table 2, we provide a summary of two paradigms, i.e., optimization and regression methods, for the goal of better alignment and physical plausibility.

According to the level of detail in the reconstruction, related approaches are categorized into body recovery (Section 3.1) and whole-body recovery (Section 3.2) with expressive hands and face. In each case, we further divide them into two paradigms. *Optimization-based* or *fitting-based* approaches explicitly fit a parametric human model to 2D observations in an iterative manner. On the contrary, *regression-based* methods make use of a deep neural network to regress the representation from image pixels directly.

3.1 Body Recovery

Algorithms dealing with body recovery are expected to yield a mesh that reflects the body pose and shape, without considering the detailed recovery of hands and face.

3.1.1 Optimization-based Paradigm

Optimization-based approaches attempt to estimate a 3D body mesh consistent with 2D image observations. The objective function typically contains two parts: data terms and regularization terms. Data terms are the measure of alignment between 2D cues and the re-projection of a mesh. To obtain physically plausible body mesh, it is important to introduce regularization terms to favor probable poses over improbable ones. Before deep learning became all the rage, optimization-based approaches were the leading paradigm for model-based human reconstruction. In the early work, silhouette cues are crucial in fitting a 3D body model, SCAPE [41] in most cases, to images. The objective function penalizes pixels in non-overlapping regions [93], [100], [101], [102], [103]. Some literature also requires manually clicked 2D keypoints [101] or correspondences [102] for a rough fit or camera estimation as initialization.

With the advances in 2D detection in the deep learning era, Bogo et al. [13] proposed SMPLify that iteratively fits the SMPL model [27] to detected 2D keypoints of an unconstrained image. They adopt an off-the-shelf 2D pose Convolutional Network (ConvNet) to detect the keypoints and perform gradient-based optimization. The objective function is the sum of a joint-based data term and several regularization terms, including an interpenetration error

term, two pose priors, and a shape prior. Specifically, the data term penalizes the distance between detected 2D joints and the projected SMPL joints. The pose priors consist of a penalty on unnatural rotations of elbows and knees, and a mixture of Gaussians trained on CMU marker data [104]. The shape prior is a quadratic penalty on the shape coefficients estimated via PCA. The interpenetration error term exploits capsule approximations and penalizes the capsule intersections. The 3D pose generated by SMPLify is relatively well-aligned. However, the shape remains highly unconstrained since the connection length between two keypoints is the only indicator that can be used to estimate the body shape. To further add constraints, instead of relying solely on one geometric term, [15], [42], [101], [105], [106] combine multiple cues for optimization, including 2D keypoints, silhouettes, and segmentations. For example, [15] leverage a multi-task neural network that estimates multiple cues to guide a joint multi-person optimization under constraints. In the refinement stage of HoloPose [106], the FCN-based estimates of DensePose [107], 2D, and 3D keypoints drive the regressed 3D models to better align with image evidence.

Moreover, deep learning techniques can be embedded into the gradient-based optimization process as a powerful tool to enhance robustness and plausibility [21], [108]. Given 2D keypoint annotations, Exemplar Fine-Tuning (EFT) [21] leverages a fully-trained 3D pose regressor and carries out optimization in the neighborhood of the pre-trained parameters. After the fitting is completed for one sample, the regressor's parameters are re-initialized for a new round. EFT optimizes all body parts without any external regularization terms since the pre-trained regressor implicitly embodies a strong prior. Song et al. [108] resort to neural networks to generate the parameter update rule. Current parameters, target 2D joints, and the gradient are passed into the network to get the updated term for the next iteration.

Besides, inverse kinematics has also been studied. Forward kinematics (FK) computes the positions of each body joint from specified joint rotations. Conversely, inverse kinematics (IK) calculates body joint rotations that match the given body joints or vertices. Iqbal et al. [109] calculate rotations for every joint accordingly based on the number of children. They follow SMPLify to refine the pose and estimate the shape. Differential IK module in [110] relies on a set of kinematics prior knowledge to infer 3D rotations from estimated 3D skeletons. HybrIK [111] decompose relative rotations into twist and swing. An adaptive IK algorithm is designed to recover swing angles. The shape and twist angles are learned in a regression-based manner.

Li et al. [112] later propose NIKI [112]. It combines the FK and IK processes using an invertible neural network to explicitly decouple errors from plausible poses. PLIKS [113] approximate the rotations based on the UV position map inputs X_{uvd} and then solve IK from 2D pixel-aligned vertex inputs X_{uv}.

3.1.2 Regression-based Paradigm

Regression-based methods take advantage of thriving deep learning techniques to process pixels directly. Here, we take a step further by breaking the networks apart and going through the similarities and differences. We examine the output types to represent a human mesh, and their motivations and setbacks. Then, we talk about the intermediate representations embedded in the networks as well as various ways to supervise in 2D and 3D space. Finally, we elaborate on the network architectures, which reflect researchers' observations and insights into this task. Fig. 5 summarizes a typical pipeline of regression-based methods. Moreover, Fig. 6 provides an illustration of various output types and intermediate representations in the regression networks. We also summarize representative regression-based methods in Table 1.

Output Type. Outputs are mainly divided into two groups: parametric outputs and non-parametric outputs.

Parametric Output. The majority of image-based human mesh recovery methods [16], [17], [18], [19], [20], [25], [44], [111], [114], [115], [116], [117], [118], [119], [120] choose to regress the parameters of the parametric models directly. They are also categorized as "model-based" approaches. Since this representation is embedded in a latent space, it is highly abstract. Networks simply need to output a low-dimensional vector, which corresponds to a body with a specific pose and shape.

Pose parameters θ contain the angle-axis representation of relative rotations of body joints plus the root orientation. Intuitively, networks can directly regress a vector corresponding to joint rotations in axis angle [16], [17], [121], [122], [123], [124]. HoloPose [106] choose to use Euler angles as the regression target alternatively. However, as demonstrated in [125], axis angle and Euler angle representations are discontinuous in the three-dimensional Euclidean space. To overcome the discontinuity, rotation matrices are adopted as the learning objective [18], [114], [126]. Learning rotation matrices is beneficial in avoiding discontinuity, but the trade-off is increasing representational redundancy and consequently dimensionality. Recently, there has been a growing trend to use a 6D representation [125], which is continuous in space, more compact than a matrix, and thus considered more suitable for deep learning [19], [20], [25], [44], [99], [116], [127], [128], [129]. To alleviate the error accumulation issue, SGRE [130] directly estimates the global rotation of each joint instead of relative rotations.

Non-parametric Output. The key to the model-based paradigm may be a stumbling block. The template serves as a strong structure prior to handling severe occlusions or ambiguities and generating likely results. In the meantime, it gets stuck in the predefined embedded space, making it harder to align with 2D cues. Researchers seek to relax this heavy reliance on the parameter space while still retaining the topology. Instead of predicting the template's

parameters, some methods directly regress non-parametric body shapes in the form of voxels [131], [132] or 3D positions of mesh vertices [133], [134], [135]. Among them, BodyNet [131] predicts a volumetric representation and then fits a SMPL model. Kolotouros et al. [133] pioneer in 3D mesh vertex coordinates regression. [135], [136] choose to predict 3D coordinates of mesh vertices and body joints in parallel. Luan et al. [137] build up a non-rigid transformation with the guidance of a concise 3D target pose and apply it to every vertex to correct the results from HMR [16]. To model uncertainty and maintain the spatial relationship between pixels in images, I2L-MeshNet [134] uses lixel (line+pixel)-based 1D heatmap for dense mesh vertex localization. It's a memory-efficient version of voxel-based 3D heatmaps. Recently released state-of-the-arts [133], [134], [138] show that evaluation on non-parametric results generally outperforms model-based ones due to their flexibility.

Apart from generating the locations of each vertex in 3D space, [139], [140], [141] utilize UV map and turn the vertex inference problem into an image-to-image translation task, which fits well with the characteristic of convolutional layers. UV maps is a pixel-to-surface dense correspondence map, which are often used for texture rendering. By storing the vertex coordinates as the (R, G, B) color values in the UV map, the UV position map is obtained and used as a suitable regression objective for fully convolutional networks. In practice, [139], [141] leverage the default UV map provided by the SMPL model. [140] propose a new UV map to maintain neighboring relations on the original mesh surface.

Probabilistic Output. The above-mentioned are deterministic and uni-modal regression models, typically yielding a single estimate for one input. Due to reconstruction ambiguity, we can also design a network to produce a set of plausible poses or a probabilistic distribution. Biggs et al. [142] learn a multi-hypothesis neural network to generate multiple sets of parameters that are plausible estimates and consistent with the ambiguous views. Sengupta et al. [143] assume simple multivariate Gaussian distributions over SMPL pose parameters θ and let the network to predict $\mu_\theta(I)$ and $\delta_\theta(I)$. ProHMR [144] models a conditional probability distribution $p(\theta|I)$ using Conditional Normalizing Flow, which is more powerful and expressive than Gaussian distributions. Sengupta et al. [145] estimate a hierarchical matrix-Fisher distribution over the relative 3D rotation matrix of each joint. This probability density function is conditioned on the parent joint along with the body's kinematic tree structure. The shape is still based on a Gaussian distribution. Fang et al. [146] propose to learn probability distributions for human joint rotations by leveraging the learned analytical posterior probability. Sengupta et al. [147] improve the consistency and diversity of predictions by modeling the ancestor-conditioned per-body-part pose distributions in an autoregressive manner.

Intermediate/Proxy Representation. Instead of directly lifting a raw RGB image to a 3D pose, plenty of approaches introduce intermediate representation into network architectures. Intermediate representations are the outputs of generic human analysis ConvNets. The benefits of involving
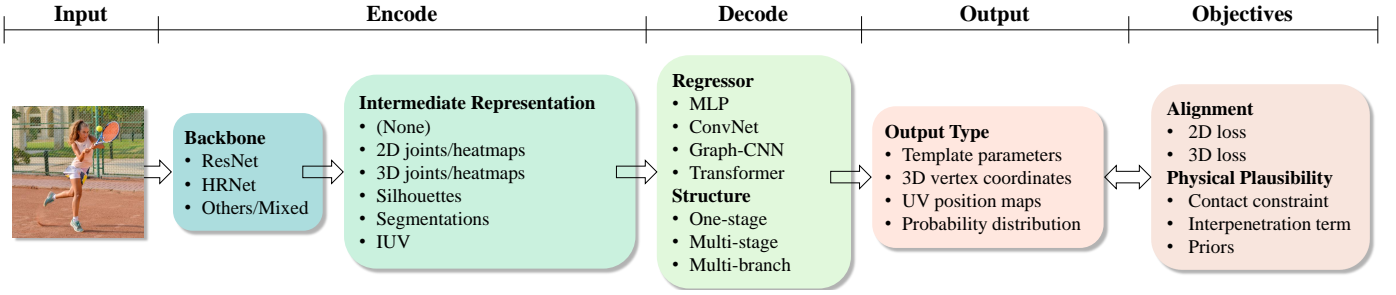
Input	Encode	Decode	Output	Objectives

Backbone
- ResNet
- HRNet
- Others/Mixed

Intermediate Representation
- (None)
- 2D joints/heatmaps
- 3D joints/heatmaps
- Silhouettes
- Segmentations
- IUV

Regressor
- MLP
- ConvNet
- Graph-CNN
- Transformer

Structure
- One-stage
- Multi-stage
- Multi-branch

Output Type
- Template parameters
- 3D vertex coordinates
- UV position maps
- Probability distribution

Alignment
- 2D loss
- 3D loss

Physical Plausibility
- Contact constraint
- Interpenetration term
- Priors

Fig. 5. The pipeline of regression-based methods for human mesh recovery.

2D/3D cues in the intermediate stage can be summarized by two words: "simplification" and "guidance".

Intermediate representations can be viewed as a simplification over RGB inputs, ignoring illumination, clothing, or background clutter, which do not necessarily correlate with human pose and shape. Intermediate estimates take the place of RGB images to be the actual input to the regression network. In this case, they are also referred to as "proxy representation", such as silhouettes [17], [116], [143], segmentations [18], [118], [148], [149], 2D heatmaps [17], [116], [118], [128], [143], [150], 2D keypoint coordinates [151], optical flow [150], [152], IUV [122], [126], [140], 3D keypoint coordinates [111], [128], [151], and surface markers [138]. The introduction of proxy representation distinctly contributes to overcoming data scarcity. The initial stage processes the RGB inputs to proxy representations. However, we can involve synthetic instances in the following stages to make a difference in performance. Compared to the synthesis of raw RGB images, proxy inputs lead to a smaller synthetic-to-real domain gap which is more readily bridged by data augmentation [17], [116], [122], [150], [153].

On the other hand, intermediate representations guide toward finer information for accurate prediction. 2D keypoint coordinates can be used to obtain part-based information to represent local body structure that is invariant to global image deformation. [106], [126] use the pose-guided pooling around keypoints to extract image features and partial IUV, and then adopt a multi-branch framework for individual part-based prediction. Besides explicit extraction or cropping, features can also be "purified" implicitly. Tung et al. [43] concatenate the RGB image and corresponding 2D heatmaps and feed to the network. Sun et al. [123] use the detected 2D keypoint coordinates and employ bilinear transformation to disentangle the skeleton from the rest details. Hand4Whole [25] calculates 3D positional pose from 3D heatmaps and interpolates on the image feature map to obtain joint-level features. 3DCrowdNet [154] concatenates image features and 2D heatmaps along the channel dimension, which will be further processed to output a 2D human pose-guided feature with high activation on a target person. PARE [20] predicts part attention masks to model the likelihood of a pixel belonging to a particular joint. The attention masks and image feature maps are fused to aggregate information from attended regions.

Supervision. Supervision signals are categorized based on the dimension of space where they play a role. 3D supervision matches the task better as the output is defined in 3D space. We can supervise the pose parameters θ in the form of axis angle [16], [117], [124], [155], rotation

matrix [17], [18], [114], [115], [122], [123], [126], [156], or a 6D representation [16], [17], [18]. Once mesh vertices are obtained, we can compute 3D joints using a pre-trained linear regressor and penalize the distance between regressed 3D joints and ground truth [16], [17], [18]. Given predicted mesh vertices and the corresponding ground truth vertices, we can also supervise the network with an additional 3D per-vertex loss [17], [116], [133], [136]. Though 3D joints and vertices are fully determined by the parameters, the redundancy leads to more stable training and better performance empirically, as each supervision signal has a different granularity [116]. In the approaches that directly regress vertices [134], [151], surface normal loss and surface edge loss are included to improve surface smoothness and details.

When 3D annotations are unavailable, we can train networks in a weakly-supervised or unsupervised manner. The strategy of reprojection-and-compare or render-and-compare has been extensively studied to transform 3D outputs to a 2D plane for supervision. The most common form of 2D supervision is 2D joints. Predicting camera parameters allows us to obtain corresponding 2D joints through reprojection and measure the displacement between ground truth and estimated 2D joints. Results can also benefit from the supervision for silhouettes [43], [110], segmentations [118], [122], [157] and dense correspondences [19], [106], [126], [140]. As pointed out in [156], dense correspondences such as IUV map [107] are effective substitutes for 3D annotations.

Network Architecture. Generally speaking, network architectures follow an encoder-decoder paradigm. The encoder is a convolutional backbone that extracts features of input images, while the decoder, or regressor, takes image features as input and outputs regressed results. Therefore, the core issue is how to design a powerful encoder and an efficient decoder to capture more information from the input and parse it adequately to boost performance. We review the existing network architectures, summarize design strategies, and organize them into three main categories:

a. One-stage frameworks that predict the pose and shape from a RGB image in a single path. No intermediate modalities are generated.

b. Multi-stage frameworks that break down the estimation into a series of sub-tasks, then leverage intermediate cues to generate final 3D outputs.

c. Multi-branch frameworks that predict pose and shape, or each body part independently in different branches after feature disentanglement.

One-stage Frameworks. In a one-stage framework, convolutional backbones like ResNet [158] and HRNet [159] are
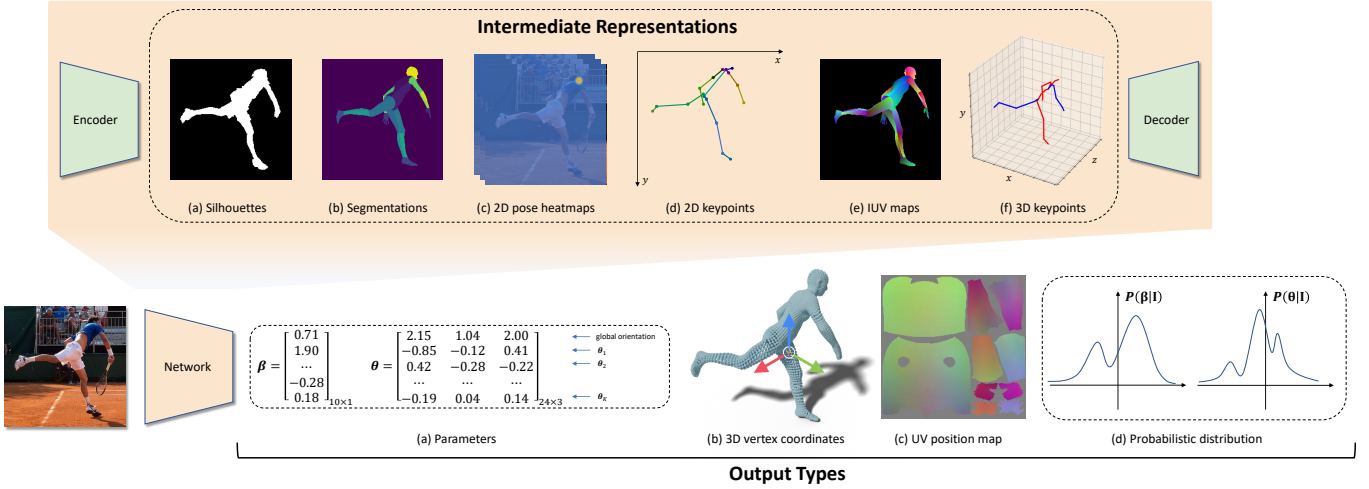
Fig. 6. Illustration of various output types and intermediate representations in the regression networks. We investigate four output types: (a) parametric output; (b) 3D coordinates of the mesh vertices; (c) UV position maps; (d) probability distribution over pose and/or shape parameters. Intermediate representations adopted in the multi-stage frameworks include (a) silhouettes; (b) segmentations; (c) 2D pose heatmaps; (d) 2D keypoint coordinates; (e) IUV maps; (d) 3D keypoint coordinates, which can serve as a simplification of the inputs or a guidance.

employed as an encoder to generate a global feature [16], [44], [133] or spatial feature [19], [139]. As for the decoders, the Iterative Error Feedback (IEF) loop in HMR [16] reduces the prediction risk compared to regressing in one go. However, it reuses the same global feature during iteration, making the regressor hardly perceive spatial information. PyMAF [19] proposes a mesh alignment feedback that leverages mesh-aligned evidence sampled from spatial feature maps to correct parameters in each loop. HUND [149] utilizes multiple RNN layers, with shared parameters and internal memory, to optimize the result stage by stage. GraphCMR [133] attaches the global feature to vertices and employs a Graph-CNN to parse neighborhood vertex-vertex interactions and then regress the 3D coordinates of each vertex. The decoder in [139], [140] comprises up-sampling and convolutional layers to generate UV position maps. In [144], a global feature vector is fed to a conditional normalizing flow to decode the probability distribution over pose parameters. ImpHMR [160] introduces the neural feature fields and learns the 3D shape and pose with volume-rendered features. HMR 2.0 [161] uses ViT [162] as the image encoder and a standard transformer decoder with multi-head self-attention to make predictions. METRO [136] and Graphormer [135] leverage an encoder-based transformer as a decoder to model non-local intersections among mesh vertices and joints, which complements convolutional operations. Following METRO [136], the follow-up works improve the architecture from the aspects of reducing computational cost [163], [164], [165], leveraging pixel-aligned features [19] in their architectures [166], [167], or fusing 2D and 3D features [168].

Multi-stage Architecture. Existing methods have also investigated breaking down the process into multiple sub-tasks. The intermediate results gradually get close to the final representation. An intermediate estimate provides a new starting point, which alleviates the reconstruction difficulty. A direct strategy is regressing body model parameters on top of intermediate predictions, including 2D/3D joints [17], [25], [118], [143], [151], sihouettes [17], [143], semantic parts [18], [118], [149], and IUV [122], [126].

Multi-branch Architecture. Pose parameters represent rel-

ative rotations of local body parts. Shape parameters, however, reflect the holistic body figure. Given the above observation, researchers seek to disentangle global shape features and local part-specific features, resulting in a multi-branch architecture. Pavlakos [17] design a two-branch network. One branch takes 2D pose heatmaps as input to regress the pose, and the other processes the silhouettes to yield the shape. HoloPose [106] pools convolutional features around each keypoint. The pooled local features are sent to a series of linear layers to infer the votes for putative joint angles. DaNet [126] decomposes the prediction task into one global stream and multiple local streams. A global IUV map is produced for the camera and shape prediction. A set of partial IUV maps are estimated based on joint-centric RoI pooling for independent predictions of each joint. HKMR [124] expresses the pose as a concatenation of six individual chains and estimate pose parameters on each kinematic chain with a network. Kocabas et al. [20] use part attention maps to aggregate 3D body features. After obtaining the final feature, they use separate linear layers to predict each SMPL joint rotation.

Regression with More Accurate Shape. Most of the regression-based methods focus on the accuracy of the poses and overlook the inaccurate shapes. This issue becomes critical when the inputs contains humans with extreme shapes since the results of typical regression-based methods are close to mean shapes. To predict more accurate body shapes, Sengupta et al. [116] leverage synthetic training data to overcome the lack of shape diversity in prevalent datasets. SHAPY [169] improves body shape estimation by exploiting the data labeled with anthropometric measurements and linguistic shape attributes. Ma et al. [170] propose to use virtual markers, which are learned from large-scale MoCap data, as intermediate representations for better capture of body shapes.

3.2 Whole Body Recovery with Hands and Face

To comprehensively understand human behavior, we need to further capture facial expressions and hand gestures along with body poses. A straightforward way to get there is

Table 1
Summary of representative regression-based methods for human mesh recovery.

Frame-based	Single Person	Output Type	**a) Template paramters:** [16], [17], [18], [19], [20], [21], [25], [44], [106], [110], [111], [115], [116], [117], [118], [119], [122], [124], [126], [127], [142], [148], [155], [171]
			b) 3D vertex coordinates: GraphCMR [133], Pose2Mesh [151], I2L-MeshNet [134], PC-HMR [137], METRO [136], Graphormer [135]
			c) Voxels: BodyNet [131], DeepHuman [132]
			d) UV position maps: DenseBody [139], DecoMR [140], Zhang et al. [141], PLIKS [113]
			e) Probabilistic outputs: Biggs et al. [142], Sengupta et al. [143], [145], ProHMR [144]
			f) Whole-body: SMPLify-X [22], ExPose [23], PIXIE [24], Hand4Whole [25], PyMAF-X [172]
		Intermediate/ Proxy Representation	**a) Silhouettes:** Pavlakos et al. [17], STRAPS [116], Skeleton2Mesh [110]
			b) Segmentations: NBF [18], Rueegg et al. [148], STRAPS [116], Zanfir et al. [118], HUND [149]
			c) 2D pose heatmaps: Tung et al. [43], Pavlakos et al. [17], STRAPS [116], Zanfir et al. [118], HUND [149], Sengupta et al. [145]
			d) 2D keypoint coordinates: HoloPose [106], Pose2Mesh [151], Skeleton2Mesh [110]
			e) IUV maps: DenseRac [122], DaNet [126], DecoMR [140], Wang et al. [173]
			f) 3D keypoint coordinates: I2L-MeshNet [134], Pose2Mesh [151], HybrIK [111], Hand4Whole [25], Skeleton2Mesh [110], Wang et al. [173], NIKI [112]
			g) Markers / Dense vertices: THUNDR [138], Ma et al. [170], PLIKS [113]
		Network Architecture	**a) One-stage:** HMR [16], GraphCMR [133], DenseBody [139], SPIN [44], PyMAF [19], METRO [135], Graphormer [135], ProHMR [144], CLIFF [120], HMR 2.0 [161]
			b) Multi-stage: Pavlakos et al. [17], NBF [18], DenseRac [122], DaNet [126], Zanfir et al. [118], Pose2Mesh [151], STRAPS [116], I2L-MeshNet [134], DecoMR [140], Zhang et al. [141], PARE [20], THUNDR [138], HUND [149], Skeleton2Mesh [110]
			c) Multi-branch: Pavlakos [17], HoloPose [106], DaNet [126], HKMR [124], PARE [20]
	Multiple Person		**a) Top-down:** Jiang et al. [174], 3DCrowdNet [154], Ugrinovic et al. [175], REMIPS [176], Cha et al. [177], OCHMR [178]
			b) Bottom-up: MubyNet [179], ROMP [180], BEV [181], PSVT [182]
Temporal	Single Person		Tung et al. [43], HMMR [121], Doersch et al. [150], Pavlakos et al. [155], Arnab et al. [183], DSD-SATN [123], VIBE [99], MEVA [127], TCMR [129], Lee et al. [184], MAED [185], SimPoE [186], DTS-VIBE [152], MPS-Net [187]
	Multiple Person		XNect [11], HMAR [188], GLAMR [189]

by performing individual reconstruction of the body, hands, and face from images and stitching them together. However, such a strategy leads to unrealistic and unnatural results. To overcome this, the community has introduced expressive human models [22], [89] for a unified reconstruction.

3.2.1 Individual Reconstruction of Hands and Face

We start with the individual methods of hand and face reconstruction. These methods can be directly combined with the body reconstruction methods to achieve a naive whole-body recovery.

Hands Reconstruction. There are also considerable efforts devoted to 3D hand pose prediction from monocular images [199], [200], [201], [202], [203], [204], [205], [206]. Based on the outputs, these methods can be grouped into two categories, i.e., methods for 3D joints prediction [203], [205], [206], [207], methods producing statistical mesh models [199], [200], [202], [204], [208], [209], [210], [211]. Since the release of the two-hand dataset InterHand2.6M [212], there have been considerable efforts devoted to reconstructing interacting hands from monocular images. Similar to the body or hand mesh recovery methods, existing approaches to two-hand reconstruction have also explored different intermediate representations [213], [214], [215], refinement strategies [215], [216], graph convolution networks [215], the implicit representation [217], the attention mechanism [215], [218], and strategies to handle in-the-wild inputs [219].

We believe these advances in integrating hand reconstruction could also provide helpful insights and solutions for integrating human mesh recovery and whole-body mesh recovery. For a thorough review of the recent advances in 3D hand pose and shape estimation, please refer to [220], [221].

Face Reconstruction. To tackle the monocular 3D face reconstruction problem, existing solutions also follow the optimization-based [222], [223], [224] and regression-based strategies [225], [226], [227], [228]. Recent state-of-the-art methods [228], [229], [230], [231], [232], [233], [234] typically render face images with estimated lighting, albedo, and geometry of the face model using a differentiable renderer [235], [236] and compare the synthetic images with the inputs. Such an analysis-by-synthesis strategy facilitates the demand for in-the-wild images and helps to recover geometric details. Moreover, recent progress [230], [232], [237] also exploits face recognition [238] to obtain more accurate facial reconstruction results. For a complete overview of recent face reconstruction methods, please refer to [239].

3.2.2 Unified Reconstruction

After unified 3D human models [22], [89], [92] are developed to account for the limitations in expressiveness, whole human body recovery methods have been proposed accordingly to estimate body posture, facial expression together with hand gestures as a whole.

Table 2
Summary of optimization/regression objectives for better alignment and physical plausibility. For each term, representative methods are listed.

Alignment	2D Loss	**a) 2D keypoints**: [13], [16], [17], [18], [19], [20], [21], [22], [23], [25], [43], [44], [99], [105], [118], [119], [121], [122], [126], [136], [180]
		b) Silhouettes: Balan et al. [93], [100], Sigal et al. [190], Guan et al. [101], Tung et al. [43], Pavlakos et al. [17], Zhang et al. [141], Skeleton2Mesh [110]
		c) Segmentations: Lassner et al. [42], Zanfir et al. [118], PARE [20]
		d) UV Position Map: DenseBody [139], DecoMR [140], Zhang et al. [141]
		e) IUV: HoloPose [106], Rong et al. [156], DenseRac [122], DaNet [126], PyMAF [19]
	3D Loss	**a) Parameters**: [16], [17], [18], [19], [20], [21], [22], [23], [25], [25], [44], [99], [118], [119], [121], [126], [180]
		b) 3D keypoints: [16], [17], [18], [19], [20], [21], [23], [25], [99], [111], [116], [119], [122], [126], [134], [135], [136], [180]
		c) Per-vertex: [17], [25], [115], [116], [118], [126], [133], [134], [135], [136]
		d) Surface (edge/normal): Pose2Mesh [151], I2L-MeshNet [134]
Physical Plausibility	Contact/ Interpenetration	SMPLify [13], Zanfir et al. [15], PROX [191], SMPLify-X [22], MotioNet [192], Jiang et al. [174], Ugrinovic et al. [175], LEMO [193], Müller et al. [194], Rempe et al. [195], [196]
	Pose Prior	**a) Handcrafted prior**: Lee etal [50], Akhter et al. [197], SMPlify [13], SMPLify-X [22]
		b) GMM: SMPLify [13], MTC [105]
		c) MoE: Sigal et al. [190], HoloPose [106]
		d) GAN: HMR [16], DenseRac [122], Jiang et al. [174], BMP [198]
		e) VAE: SMPLify-X [22], HKMR [124], GHUM [92]
		f) Normalizing Flows: Biggs et al. [142], Zanfir et al. [118], GHUM [92], Fan et al. [171], ProHMR [144]
	Motion Prior	HMMR [121], VIBE [99], MEVA [127], HuMoR [196], LEMO [193], SimPoE [186], GLAMR [189]

Optimization-based Paradigm. Similar to human body recovery, optimization-based methods [22], [92], [105], [240], [241] for whole-body recovery detect reliable 2D cues using pre-trained detectors and fit the parametric model to these observations. Xiang et al. [105] train a ConvNet to predict joint confidence maps and Part Orientation Fields (POF) for the body, hands, face, and feet. They iteratively optimize the objective function to fit the Adam model [89] to data terms. To fit SMPL-X to a single RGB image, Pavlakos et al. [22] present SMPLify-X that follows SMPLify [13] by first detecting 2D features [1], [242] corresponding to the face, hands, and feet and optimizing the model parameters afterward. SMPLify-X makes several improvements, including a better-performing pose prior based on a variational auto-encoder (VAE), self-collision penalty terms, and an updated interpenetration term. Xu et al. [92] set anatomical joint angle limits and optimize GNUM parameters using a joint reprojection term and a semantic body-part alignment term. Like body-only recovery, optimization-based methods tend to be slow and sensitive to initialization.

Regression-based Paradigm. Leveraging expressive human models and paired data, the community has also resorted to an end-to-end training fashion for whole-body reconstruction. Among existing solutions, the *divide-and-conquer* strategy [23], [24], [25], [149], [172], [243], [244] is commonly used to break the reconstruction problem down into its parts where the estimation of the bodies, hands, faces is conducted separately with part-specific models. The final expressive 3D human mesh is obtained by forwarding the outputs of each branch to the body template layer. For example, ExPose [23] directly predicts hands, face, and body parameters in the SMPL-X format and utilizes the body estimation to localize the face and hands regions and crop them from the high-resolution inputs for refinement. It learns part-specific knowledge from existing face-

and hand-only datasets to improve performance. Zhou et al. [128] is a real-time method that captures body, hands, and face with competitive accuracy by exploiting the inter-part relationship between body and hands. SMPL+H [80] and 3DMM [53] are used to represent the body+hands and face. Hand4Whole [25] obtains the joint-level features from feature maps, and regresses the 3D body/hand joint rotations from them. PIXIE [24] estimates the confidence of part-specific features and fuses the face-body and hand-body features weighted according to moderators. The fused features are fed to the independent regressors for robust regression. In addition, fine facial details, i.e., geometry, albedo, and illumination, are predicted in [24], [128]. Sun et al. [245] predict hands, and face parameters based on detected whole-body 2D keypoints, making it feasible to take advantage of synthetic data during training. To resolve conflicts and merge the results from all sub-networks, PyMAF-X [172] proposes an adaptive integration with elbow-twist compensation. HybrIK-X [244] recalculates the rotations of the parents of the conflict joints. OSX [246] proposes a transformer-based one-stage method to capture the connections of body parts. SMPLer-X [247] investigates scaling up the sizes of models and data for whole-body recovery. RoboSMPLX [248] improves the localization and feature extraction of body parts for a more robust recovery of whole-body models. SGNify [249] improves the 3D hand poses by leveraging linguistic priors as constraints for more natural whole-body mesh recovery from sign-language videos. Despite the recent progress, recovering the whole-body model with plausible hand gestures remains challenging, especially in the cases of interacting hands, occlusions, and motion blur.

4 MULTI-PERSON RECOVERY

In order to recover 3D human mesh from crowded scenes, we categorize the mainstream methods into two classes based on the design strategy: 1) the top-down strategy and 2) the bottom-up strategy.

The top-down design reduces the multi-person recovery task to the single-person setting. Cropped single-person images from off-the-shelf detectors [250], [251] are the actual input to the network. In this way, we get to adopt all kinds of single-person regression methods mentioned above. However, truncations, person-person occlusions, and human-scene intersections are ubiquitous in multi-person scenes, which impede the network from perceiving holistic information about a target person. As pointed out in [180], when two people overlap each other badly, it lacks sufficient context for a network to distinguish the regression target from similar patches. Thus, 3DCrowdNet [154] takes advantage of robust 2D pose outputs to produce a pose-guided feature that disentangles the target person's feature from others. 3D joint coordinates and SMPL parameters are later derived from the 2D pose-guided feature. Zanfir et al. [15] fit a parametric human model for every person based on 2D and 3D observations provided by a multi-task deep neural network. They jointly perform multi-person optimization over all people in the scene, including collision and ground-plane constraints. Zanfir et al. [179] identify and score different body joint connections, and assemble limbs into skeletons. The feature volume and its identified skeleton are mapped into a shape and pose parameter pair for each person. Note that it is still a multi-stage pipeline, and the last module operates in a top-down manner. Jiang et al. [174] explore a R-CNN-based architecture for detection and estimation for all people in the image. To encourage reconstruction in the depth order and avoid overlapping, they incorporate a depth ordering-aware loss and an interpenetration loss during training. REMIPS [176] creates a sequence of spatial feature tokens and person tokens based on the detected bounding boxes. The tokens are fed to a transformer encoder to make predictions. In [177], a relation-aware transformer takes every person's image feature and 3D mesh as input to refine the multi-person predictions.

The top-down paradigm has been criticized for repeated feature extraction and limited receptive field within the bounding box. These drawbacks make it harder to speed up and perform robustly in occlusion and truncation cases. Instead of designing a multi-stage pipeline, the bottom-up paradigm preserves a holistic view and provides simple one-stage solutions that are computationally efficient. Single-shot methods [180], [198] exploit point-based representation to represent instances by a single point at their center. Using multiple heads, they simultaneously predict an instance localization heatmap and a mesh parameter map. ROMP [180] constructs a repulsion field to push apart body centers that are too close. BMP [198] improves the inter-instance ordinal depth loss and adopts a keypoint-aware augmentation strategy during training. Crowd3D [252] proposes a framework to reconstruct the body model and global locations of hundreds of people from a single large-scene image. PSVT [182] is an end-to-end multiperson 3D human pose and shape estimation framework with the proposed progressive video Transformer.

To deal with the human-human interactions, Zanfir et al. [15] introduce a collision constraint to prevent the human models from overlapping. Parallelepipeds are fitted to each person at first. If the far-range check fails, the authors adaptively represent the limbs by a series of spheres and calculate the distances based on centers and radius as the volume occupancy loss. REMIPS [176] employs an interaction-contact loss based on the contact signature and the distance at the facet level. Jiang et al. [174] deploy an adapted Signed Distance Field (SDF) to the multi-person scene that takes positive inside each human and zero outside. Based on this, they compute an interpenetration loss for every vertex in every human model. OCHMR [178] uses a global centermap and a subject-specific local centermap to encode the spatial context for each person, which serves as a conditioned input to normalize intermediate features. Besides interpenetration, depth order incorrection often occurs in rendering multiple persons. The authors also propose a depth ordering-aware loss based on the segmentation and depth maps.

5 RECOVERY FROM MONOCULAR VIDEOS

Human mesh recovery from monocular videos is a step forward in understanding human behavior. Image-based methods process each frame independently. The reconstruction results are prone to suffer from occlusions and motion jitters due to the lack of temporal constraints. For this reason, a good design for videos should exploit the full potential in feature encoding to enhance consistency in spatial and temporal dimensions.

Temporal encoding functions are typically represented as convolutional and recurrent networks. For example, Doersch et al. [150] extract features from a combination of optical flow and 2D heatmaps via a single-frame ConvNet followed by an LSTM. In typical methods [99], [121], [123], [127], [129], [184], [187], [253], a pre-trained backbone like ResNet-50 [158] is used to process raw images to generate static features. After that, Kanazawa et al. [121] adopt a 1D fully convolutional model as a temporal encoder. Follow-up works [99], [127], [129], [184] employ bidirectional GRUs to incorporate the information from all frames and get temporally correlated features. Besides, TCMR [129] applies two more GRUs to forecast additional temporal features for the current target pose from the past and future frames. Lee et al. [184] consider the uncertainty-aware embedding and include optical flow information. Wei et al. [187] extend the non-local operation [254] to recalibrate the range of attention in a motion sequence. Lately, there has been a trend to adopt the multi-head self-attention (MHA) module [255] for long-term sequence dependency modeling [123], [155]. Wan et al. [185] modify the original MHA to perform spatial and temporal encoding simultaneously. GLoT [256] proposes a transformer to decouple the long-term and short-term modeling of temporal motions. HMR-ViT [257] takes into account both temporal and kinematic information by leveraging temporal-kinematic features in a vision transformer. To handle out-of-domain video inputs, Guan et al. [258] propose a bilevel online adaptation with temporal

constraints, while Nam et al. [259] propose a cyclic test-time adaptation strategy.

On the other hand, different decoding strategies and optimization objectives have been proposed to reduce jitters and improve smoothness. The decoder in [99], [121], [127], [129] iteratively refines the parameters for each frame based on HMR [16]. HMMR [121] includes a dynamics predictor to predict the change of pose parameters in a time step. MEVA [127] learns a human motion subspace via variational autoencoder (VAE) to generate coarse but smooth motions. Finer motions are later retrieved as residuals. TCMR [129] passes integrated features, features from past frames, and features from future frames to a shared regressor. All three outputs are supervised with the ground truth of the current frame.

Apart from the architecture, different supervision strategies have also been explored in existing solutions. Tung et al. [43] compute a motion re-projection error between the predicted 2D body flow and estimated 2D optical flow field in two consecutive frames. Zanfir et al. [15] design a velocity prior, assuming that the displacement of pose angles and translation in two adjacent frames should be constants. Sun et al. [123] first shuffle the frames and adopt an adversarial training strategy to recover the correct temporal order. Arnab et al. [183] adopt a temporal error on 3D joints, camera parameters, and 2D keypoint re-projection to penalize the changes between two consecutive frames. MTC [105] defines a texture consistency term based on the flow mapping and enforces a smoothness constraint for the depth of 3D joint locations. Tripathi et al. [260] use a sliding window to penalize 3D joints of the same frames before and after the window strides. Wan et al. [185] use a series of learnable linear regressors to decode joint rotations in a hierarchical order. Some objective terms are predefined empirically or learned from large motion capture datasets [98], [104]. We treat them as "motion priors", which are of great importance and will be discussed thoroughly in § 7.4.

There has been a movement to predict in the world coordinate system by combining camera motion, multi-person tracking, and reconstruction into one system [161], [189], [261], [262]. GLAMR [189] recovers human meshes in a consistent global coordinate system after extracting motions in the local coordinate system, infilling missing poses, predicting global trajectories, and jointly optimizing camera poses and global motions. It deals with monocular videos that are recorded with dynamic cameras. D&D [263] proposes an inertial force control module to improve the dynamic motions estimated from videos with moving cameras. SLAHMR [261] first initialize relative camera motion with SfM, and people tracks with PHALP. These are fed to a joint optimization system to solve the camera scale, the ground plane, and each person's trajectory in the world coordinate system. TRACE [262] predicts a motion offset map, a world motion map to reason about human trajectories in camera coordinates and world coordinates, respectively. A memory unit is used to predict the tracked identities. 4DHumans [161] takes HMR 2.0 as the backbone and adapts the PHALP tracker [188] with a vanilla transformer to track people as well as predict future pose parameters.

6 HUMAN-SCENE INTERACTIONS

Human-Scene interactions are ubiquitous. However, given monocular inputs, most works perceive 3D humans in isolation from the surroundings. Considering 3D humans, scenes, and interacting objects as a whole and inferring the spatial arrangement and contacts help us understand interactive scenarios better. This section discusses several works that reason 3D humans together with scenes from monocular RGB images. Pioneering attempts [264], [265], [266], [267] infer 3D humans and objects separately, which is prone to be visually unrealistic. To encourage plausibility, various objective functions are proposed over interactions, collisions, and contacts to optimize modules in the scene jointly, which we will introduce in Section 7 with more details.

HolisticMesh [264] imposes human-scene losses at the joint optimization stage, including human-scene collision, human-object contact, and ground support. PHOSA [265] optimizes spatial arrangement using instance-level and part-level interaction losses, a scale loss, and an ordinal depth loss. But both of these two methods [264], [265] depend on pre-defined candidate contact vertices or pairs to constrain interaction, which limits the generalization to diverse scenes. CHORE [266] first learns to predict several 3D neural fields that are more robust than plain 2D evidence. The predicted neural fields serve as stronger 3D terms to provide constraints in the optimization process of SMPL and the object template. MOVER [267] optimizes a plausible scene by refining the camera orientation, object layout, interactions, and ground plane based on expected contacts and 2D segmentations. There are also scene-aware approaches [268], [269] to recovering plausible human motions in a pre-scanned 3D scene.

7 PHYSICAL PLAUSIBILITY

Existing methods can produce 3D shape and pose well-aligned to 2D joints but still suffer from visual artifacts, such as ground penetration, foot skating, and body leaning. Only supervising the human body is insufficient to get a consistent result. To obtain a physically plausible reconstruction, realistic camera models, contact constraints, and shape/pose priors should be taken into account.

7.1 Camera Model

Due to the lack of camera information in uncalibrated images, it is virtually impossible to find the exact intrinsic and extrinsic parameters of a perspective camera from monocular images. For simplicity, a weak-perspective camera model with a large constant focal length is widely used to approximate perspective cameras, in which only scale $s \in \mathbb{R}$ and camera translation $t \in \mathbb{R}^2$ along the x, y-axis are unknown and need to be predicted [16], [19], [44], [116], [118]. The scale parameter is further converted into camera translation along the z-axis. However, this simplification does not completely tally with the real-world data. Focal length impacts the field of view (fov), depth of field, and the sense of perspective. Also, the effect of camera rotation, such as a significant pitch, can not be entirely counteracted by the rotations along the body kinematic chain. Besides,

by merely processing cropped images, we fail to know the actual location in the original image, causing difficulty in a real-world application.

There are methods [119], [189], [270] that recover human meshes based on original images and estimate in the world coordinate system. Kissos et al. [270] and CLIFF [120] approximate a realistic focal based on the width and height of the original image and convert the camera translation parameter to calculate the reprojection loss in the full image instead of the cropped one. SPEC [119] computes the camera intrinsics and rotation by predicting the pitch, roll, and vertical field of view (vfov). GLAMR [189] adopts a dynamic camera in global coordinates and jointly optimizes the camera poses and global motions to match the video evidence. For more robust pose estimation in the real world, Cho et al. [271] and Zolly [272] also take the perspective distortion issue into account.

7.2 Contact Constraint

The primary purpose of contact constraints is to encourage proper contacts and penalize erroneous interpenetration.

We start with the human-scene contact. [15], [175] fit a ground plane to the selected 3D ankle positions of all people in a frame, and use the estimated normal vector and a reference point fixed in the plane to penalize the ankle joints away from the plane. In [192], [195], [196], [269], the human-scene contact status is predicted to improve the plausibility. Specifically, Rempe et al. [195] design a physics-based trajectory optimization that takes the predicted foot contacts from 2D poses as input and outputs 6D center-of-mass motion, feet positions, and contact forces. The physics-based models are also used to enable full-body contacts [273] or trajectory optimization [274]. Shi et al. [192] supervise the network to infer a binary label indicating whether the foot is in contact with the ground and encourage the velocity of foot positions to be zero in contact. HuMoR [196] generates a contact probability for each joint. The contact probability output gives weight to an environment regularizer to ensure consistency in joint positions and joint heights among frames. Xie et al. [275] exert contact forces on the feet at 4 different points and design a contact loss to penalize violation of Signorini conditions. Going beyond the flat ground, [191], [193] delve deep into the vertex representation and perform scene reconstruction as the first step. PROX [191] penalizes the contact candidate vertices of the body far away from the nearest 3D scene mesh vertices. The contact term only considers body-scene proximity and thus fails to prevent the foot-ground skating problem. To address this issue, LEMO [193] decomposes the velocity of contacted vertices and regularizes the component along the scene normal to be non-negative, and the component tangential to the scene to be small to prevent sliding. Huang et al. [276] propose to train a motion distribution prior with a physics simulator and introduce an interaction constraint based on signed distance fields to enforce ground contact modeling. IPMAN [277] defines a stability loss based on the estimated Center of Pressure (CoP) and Center of Mass (CoM), and a ground contact loss based on the vertices' height.

Apart from the human-human interactions elaborated in Section 4, some self-contacts exist between body parts. To vividly model the hands touching the body and contact between other body parts, Müller et al. [194] compute an approximated surface-to-surface distance to detect self-contact. They adapt SMPLify-X [22] by adding self-contact-related objectives, and one of them encourages every vertex in the self-contact pairs to be in contact. Similarly, Fieraru et al. [278] detect self-contact and design losses to enforce the constraint explicitly. On the other hand, to avoid self-collision and penetration of several body parts, Bogo et al. [13] approximate the body parts using an ensemble of capsules and penalize the intersections between the incompatible capsules. Although the approximation is computationally efficient, it lacks details. [22], [191] leverage Bounding Volume Hierarchies (BVH) [279] to detect a list of colliding body triangles for a more accurate collision penalizer. Müller et al. [194] design a term to push the vertices inside the mesh to the surface. ProxyCap [280] introduces a contact-aware neural motion descent module such that the network can be aware of foot-ground contact and the misalignment with 2D observations.

7.3 Pose Prior and Shape Prior

The inherent ambiguity in lifting 2D observations to 3D space gives rise to the need for prior knowledge. Priors favor plausible predictions and rule out impossible ones, helping to restrict the outputs to a feasible distribution. Besides, priors play an indispensable role when 3D labels are not available. Existing shape and pose priors are set heuristically by handcrafted designs or learned by generative models. Classic generative models like the Gaussian Mixture Model (GMM), and Mixture of Experts (MoE) are used to discover patterns and correlations in data. Compared to traditional machine learning methods, deep generative models such as Generative Adversarial Network (GAN) [281], Variational Autoencoder (VAE) [282], Normalizing Flows [283] are better qualified for prior modeling, especially when large-scale training data is available. Priors can be treated as loss terms and added to objective functions in the training or iterative optimization processes. The decoder or generator of a generative prior can also be integrated into a regression network as a human mesh regressor.

Handcrafted Prior. Priors can be designed empirically to achieve a certain direct-viewing effect. For example, known limb lengths are adopted [50] and pose-dependent joint angle limits are explored [197]. A pose prior in [13], [22] is represented as the sum of the exponentials to penalize the unnatural bending in elbows and knees for the exponential value would soar if the rotations violate natural constraints. As for shape prior, Bogo et al. [13] compute the shape prior quadratically with the squared singular values estimated via PCA. A simple L_2 shape prior is adopted in [22], [24], [44], [118], assuming β should stay near the neutral zero vector.

Gaussian Mixture Model (GMM). Bogo et al. [13] study the multi-model nature of the pose by fitting a mixture of 8 Gaussian distributions to a collection of reasonable pose parameters. Xiang et al. [105] compute a Gaussian distribution for the pose parameters as a whole.

K-Means. Güler et al. [106] obtain K representative angle values for each body joint after applying K-Means. The prediction outputs are restricted within the convex hull

of the rotation clusters. Rong et al. [284] build a prototypical memory using K-Means to store multiple sets of mean parameters for regression initialization.

Generative Adversarial Network (GAN). Researchers first resort to GAN [281] to obtain adversarial priors. The discriminator is forced to distinguish between candidates produced by the network and real data [16], [122], [174], [198]. For instance, Kanazawa et al. [16] assign a discriminator for shape and pose independently, and further train an adversarial prior for each joint. Similarly, DenseRac [122] trains the discriminator with millions of synthetic samples to learn an admissible manifold of IUV representation. Davydov et al. [285] define a generator and a discriminator with the same architecture as the decoder in VPoser [22] and the discriminator in HMR [16], respectively. After training, the GAN-based pose prior can be used in the optimization process to optimize a latent vector z in the latent space. It can also serve as a drop-in human mesh regressor.

Variational Autoencoder (VAE). In a VAE [282], the encoder compresses the data $\mathbf{x}$ into a latent distribution $P(\mathbf{z} \mid \mathbf{x})$. A latent variable $\mathbf{z}$ is sampled from $P(\mathbf{z})$, typically $\mathcal{N}(0, I)$. The decoder reconstructs $\hat{\mathbf{x}}$ given the hidden vector $\mathbf{z}$. Pavlakos et al. [22] propose a VAE-based pose prior, VPoser, to learn a regularized latent distribution of human poses. To employ VPoser in the optimization, the pose parameters are encoded to a latent variable $\mathbf{z}$, and a quadratic penalty is applied to $\mathbf{z}$. A similar strategy is used in Georgakis et al. [124] to obtain plausible poses. Besides, the body and facial deformation in the GHUM/GHUML models [92] is also based on the latent space in VAE.

Normalizing Flows. Normalizing flows [283] are powerful in distribution approximation and efficient in derivation calculation. Zanfir et al. [118] introduce normalizing flows to model 3D human pose. They cascade multiple Real-NVP steps [286] to build a model that embodies a flow-based prior for weakly-supervised training. Inspired by this, Biggs et al. [142] also adopt the Real-NVP architecture. Fan et al. [171] design a normalizing flow using fully-connected layers. The GHUM/GHUML models [92] rely on normalizing flows to represent skeleton kinematics. The authors also train a kinematic prior for hands and body based on normalizing flows. ProHMR [144] acts as an image-based pose prior to the fitting process, predicting the distribution of plausible poses given an input image. This distribution is modeled using conditional normalizing flows.

Diffusion Model. The Diffusion model [287] is a generative model based on stochastic processes. Recently, there have been several approaches [288], [289], [290] applying diffusion models in the task of human mesh recovery. Thanks to the probabilistic nature of diffusion models, these approaches can produce multiple hypotheses to handle the ambiguity in cases of occlusions.

7.4 Motion Prior

Motions can be predicted to some extent since they have some patterns in nature. Simply penalizing the velocity or acceleration of each joint will degrade motion naturalness. Instead, priors based on recurrent models [99] and autoencoders [193], [196], [291] have larger temporal receptive fields to learn motion patterns. VIBE [99] contains a motion discriminator and MPoser. The motion discriminator consists of multiple GRU layers to identify plausibility. MPoser, an extension of VPoser [22] to temporal sequences, is based on sequential VAE. Inspired by VIBE, He et al. [292] generate marker-based motion maps as input to a discriminator to obtain an adversarial motion prior. In HuMoR [196], the probability distribution of possible state transitions is formulated by a conditional variational autoencoder (CVAE). This dynamic prior is later used for robust test-time optimization. LEMO [193] smooths the motion in the latent space of a convolutional autoencoder to reduce the pose jitters. GLAMR [189] contains a CVAE-based generative motion infiller to infill missing poses. SimPoE [186] resorts to reinforcement learning and introduces a simulation-based motion modeling approach. HM-VAE [293] contains skeleton-based convolution, pooling, and unpooling operations. With the learned HM-VAE, one can refine noisy motion sequences by first projecting into the latent space and then decoding back. Xu et al. [294] exploit sequence-based and segment-based frequencies to compress input motions adaptively. The pretrained motion prior can be embedded into VIBE [99] in a video-to-mesh regression task.

8 DATASETS

In this section, we focus on the commonly-used datasets. First, we introduce the acquisition of human mesh annotations. Then, we give brief descriptions of the commonly used datasets.

8.1 The Acquisition of Mesh Annotations

Obtaining samples paired with 3D mesh labels is not easy. The most precise image-label pairs are generated by **rendering** 3D body models [97] or human scans [296] to images. The lack of realism remains a major issue in these synthetic images. In order to collect real images and obtain corresponding 3D labels, marker/sensor-based [95], [96] and marker-less [14], [301] MoCap systems are deployed to capture body motions. **Marker/sensor-based systems** attach reflective markers or Inertial Measurement Units (IMU) to the subjects' bodies and track them over time. These 3D sparse point sets are later processed by MoSh [94] to fit a body mesh. **Marker-less systems** capture person images from multiple cameras, where 2D cues are further fitted to the body mesh by exploiting multi-view geometry. The MoCap data is generally limited to constrained environments and lacks the diversity of subjects and actions.

To obtain mesh annotations for in-the-wild images, researchers fit the body model to image evidence to generate **pseudo-3D labels** in semi-automatic [13], [42] or full-automatic [21], [44], [120], [315] manners. For instance, SPIN [44] combines the fitting and regression process in a loop. The regressed outputs serve as better initialization for optimization. EFT [21] finetunes a pretrained SPIN network to 2D joint coordinates for each sample. But as pointed out in [315], this may lead to overfitting, especially when the input image is partially invisible. To overcome this, NeuralAnnot [315] is trained on a mixture of 3D datasets and the target 2D in-the-wild dataset. It is optimized for

Table 3
Comparison of datasets involved in network training and evaluation. Each dataset provides images with paired 3D (pseudo) ground truth.

Type	Dataset	# Frames	# Scenes	# Subjects	# Subjects Per Frame	In-the-wild	Mesh Type	Mesh Annotation Source
Rendered Datasets	SURREAL [97]	6.5M	2,607	145	1	-	SMPL	[97]
	GTA-Human [295]	1.4M	-	> 600	1	-	SMPL	[295]
	AGORA [296]	17K	> 350	4,240	$5 \sim 15$	-	SMPL-X	[296]
	THUman2.0 [297]	-	-	~ 200	1	-	SMPL-X	[297]
	MultiHuman [30]	-	-	~ 50	$1 \sim 3$	-	SMPL-X	[30]
	HSPACE [298]	1M	100	100	avg. 5	-	GHUM	[298]
	BEDLAM [299]	380K	103	-	$1 \sim 10$	-	SMPL-X	[299]
Marker/Sensor-based MoCap	HumanEva [58]	80K	1	4	1	-	-	-
	Human3.6M [95]	3.6M	1	11	1	-	SMPL	[16], [97], [134]
	Total Capture [300]	~ 1.9M	1	5	1	-	-	-
	3DPW [96]	> 51K	60	7	$1 \sim 2$	✓	SMPL	[96]
Markerless Multiview MoCap	CMU Panoptic [301]	1.5M	1	40	$3 \sim 8$	-	-	-
	MPI-INF-3DHP [302]	> 1.3M	1	8	1	-	SMPL	[44]
	MuCo-3DHP [303]	200K	1	8	$1 \sim 4$	-	-	-
	MuPoTs-3D [303]	> 8K	20	8	3	✓	-	-
	MannequinChallenge [304]	24,428	567	742	5	✓	SMPL	[305]
	3DOH50K [141]	51,600	1	-	1	-	SMPL	[141]
	Mirrored-Human [306]	1.8M	> 200	> 200	≥ 1	✓	SMPL	[306]
	MTC [105]	834K	1	40	1	-	-	-
	EHF [22]	100	1	1	1	-	SMPL-X	[22]
	HUMBI [307]	17.3M	1	772	1	-	SMPL	[307]
	ZJU-MoCap [35]	-	1	9	1	-	SMPL-X	[308]
Datasets with Pseudo-3D Labels	LSP [309]	2,000	-	-	1	✓	SMPL	[21], [42], [44]
	LSP-Extended [310]	10,000	-	-	1	✓	SMPL	[21], [44]
	MSCOCO [311]	38K	-	-	≥ 1	✓	SMPL	[21], [44]
	MPII [312]	24,920	3,913	> 40k	≥ 1	✓	SMPL	[21], [42], [44]
	UP-3D [42]	8,515	-	-	1	✓	SMPL	[21], [44]
	PoseTrack [313]	66,374	550	550	> 1	✓	SMPL	[21]
	SSP-3D [116]	311	62	62	1	✓	SMPL	[116]
	OCHuman [314]	4,731	-	8110	> 1	✓	SMPL	[21]
	MTP [194]	3,731	-	148	1	✓	SMPL-X	[194]
	Ubody [246]	> 1,050K	-	-	≥ 1	✓	SMPL-X	[246]

entire samples. CLIFF [120] trains an annotator with the information from the original frames instead of the cropped ones. Thus, the CLIFF annotator produces more accurate labels, especially the global rotations. Even though pseudo-labels for in-the-wild datasets are not as accurate as MoCap data, they still remarkably improve the generalization of regression-based methods thanks to their scale and diversity.

8.2 Datasets

Datasets involved in training and evaluation can be categorized into four groups based on data and label acquisition strategies, i.e., rendered datasets, marker/sensor-based MoCap datasets, marker-less MoCap datasets, and datasets with pseudo-3D labels. Table 3 summarizes some key information about these datasets.

8.2.1 Rendered Datasets

SURREAL. *Synthetic hUmans foR REAL tasks* [97] is a large-scale synthetic human body dataset. Bodies are created with the SMPL body model and driven by 3D MoCap motions. Textures are rendered with random attributes on the background images. The dataset contains ground truth depth maps, optical flow, surface normals, human part segmentations, and 2D/3D joint locations.

GTA-Human [295] is a large-scale 3D human dataset with diverse subjects, actions, and scenarios. The dataset is generated with the GTA-V game engine. There are 20K video sequences with SMPL annotations in this dataset.

AGORA. *Avatars in Geography Optimized for Regression Analysis* dataset [296] is a recently released synthetic dataset with high realism and accurate SMPL/SMPL-X models fitted to 3D scans. Over 4,000 photorealistic textured human scans, including some children's scans, are positioned in panoramic scenes. AGORA has become a popular benchmark for SMPL and SMPL-X estimation from monocular images.

THUman2.0 [297] contains 500 high-quality human scans with different clothing and poses captured by a 128 DSLR camera dome system. The dataset provides the 3D scan model with the corresponding texture map and fitted SMPL-X model for each scan. The person images can be generated from any viewpoint using the rendering strategy mentioned in PIFu [316] and PaMIR [29].

MultiHuman [30] consists of 453 high-quality 3D human scans with raw scan meshes, texture maps, and the fitted SMPL-X models. Each scan contains 1-3 persons under occluded or interactive scenes. Images can be synthesized in the same way as THUman2.0.

HSPACE [298] relies on a corpus of 100 human scans. After fitting the scans with GHUM mesh [92], the authors

Table 4

Evaluation of the body recovery methods on 3DPW [96] and Human3.6M [95] datasets. The comparison is not completely fair considering the factors of backbones, output types, the quality of pseudo labels, dataset selection, and training strategy. ♯ represents approaches dealing with the multi-person 3D mesh recovery task. † denotes the approaches using training data from 3DPW. Please refer to § 9.2 for a comprehensive discussion.

	Method	Publication	Output Type	Pseudo-GT (2D Datasets)	3DPW			Human3.6M	
					MPJPE	PA-MPJPE	PVE	MPJPE	PA-MPJPE
Frame-based	Pavlakos et al. [17]	CVPR'18	Parameters	[42]	-	-	-	-	75.9
	HMR [16]	CVPR'18	Parameters	-	130.0	76.7	-	88.0	56.8
	NBF [18]	3DV'18	Parameters	[42]	-	-	-	-	59.9
	GraphCMR [133]	CVPR'19	Vertices	[42]	-	70.2	-	-	50.1
	HoloPose [106]	CVPR'19	Parameters	-	-	-	-	60.3	46.5
	DenseRac [122]	ICCV'19	Parameters	-	-	-	-	76.8	48.0
	SPIN [44]	ICCV'21	Parameters	[44]	96.9	59.2	135.1	62.5	41.1
	Jiang et al. [174]♯	CVPR'20	Parameters	-	-	-	-	-	52.7
	Zhang et al. [141]	CVPR'20	Position map	-	-	-	-	-	41.7
	DecoMR [140]	CVPR'20	Position map	[42]	-	61.7	-	-	39.3
	Zanfir et al. [118]	ECCV'20	Parameters	-	90.0	57.1	-	-	-
	LearnedGD [108]	ECCV'20	Parameters	-	-	55.9	-	-	56.4
	Pose2Mesh [151]	ECCV'20	Vertices	[44]	89.2	58.9	-	64.9	47.0
	HKMR [124]	ECCV'20	Parameters	-	-	-	-	59.6	43.2
	I2L-MeshNet [134]	ECCV'20	Vertices	[134]	93.2	57.7	-	55.7	41.1
	DaNet [114]	MM'19	Parameters	[42]	85.5	54.8	110.8	-	40.5
	Hand4Whole [25]	CVPRW'22	Parameters	[315]	86.6	54.4	-	71.0	47.4
	HybrIK [111]	CVPR'21	Parameters	-	80.0	48.8	94.5	-	-
	METRO [136]	CVPR'21	Vertices	[42], [44]	77.1	47.9	88.2	54.0	36.7
	Sengupta et al. [143]	CVPR'21	Probabilistic	[42]	-	61.0	-	-	-
	BMP [198]♯	CVPR'21	Parameters	-	104.1	63.8	119.3	-	51.3
	HUND [149]	CVPR'21	Parameters	-	81.4	57.5	-	69.5	52.6
	EFT [21]	3DV'21	Parameters	[21]	-	54.2	-	-	43.7
	ProHMR [144]	ICCV'21	Probabilistic	[44]	-	59.8	-	-	41.2
	DSR [157]	ICCV'21	Parameters	[21]	91.7	54.1	105.8	60.9	40.3
	ROMP [180]♯	ICCV'21	Parameters	[44]	89.3	53.5	103.1	-	-
	Graphormer [135]	ICCV'21	Vertices	-	74.7	45.6	87.7	51.2	34.5
	THUNDR [138]	ICCV'21	Parameters	-	74.8	51.5	88.0	55.0	39.8
	PyMAF [19]	ICCV'21	Parameters	[44]	92.8	58.9	110.1	57.7	40.5
	SPEC [119]	ICCV'21	Parameters	[21]	-	53.2	-	-	-
	PARE [20] †	ICCV'21	Parameters	[21]	74.5	46.5	88.6	-	-
	BEV [181]♯	CVPR'22	Parameters	[21]	78.5	46.9	92.3	-	-
	CLIFF [120] †	ECCV'22	Parameters	[120]	69.0	43.0	81.2	47.1	32.7
	FastMETRO [163] †	ECCV'22	Vertices	-	73.5	44.6	84.1	52.2	33.7
	Cha [177]♯	ECCV'22	Parameters	-	66.0	39.0	76.3		
	PyMAF-X [172] †	TPAMI'23	Parameters	[21]	74.2	45.3	87.0	-	-
	POTTER [164] †	CVPR'23	Parameters	-	75.0	44.8	87.4	56.5	35.1
	ProPose [146] †	CVPR'23	Parameters	-	68.3	40.6	79.4	45.7	29.1
	NIKI [112] †	CVPR'23	Parameters	-	71.3	40.6	86.6	-	-
	PLIKS [113] †	CVPR'23	Parameters	[21]	60.5	38.5	73.3	47.0	34.5
	PointHMR [166] †	CVPR'23	Vertices	-	73.9	44.9	85.5	48.3	32.9
	ImpHMR [160] †	CVPR'23	Parameters	-	74.3	45.4	87.1	-	-
	VirtualMarker [170] †	CVPR'23	Parameters	-	67.5	41.3	77.9	47.3	32.0
Temporal	HMMR [121]	CVPR'19	Parameters	-	116.5	72.6	139.3	-	56.9
	Arnab et al. [183]	CVPR'19	Parameters	-	-	72.2	-	77.8	54.3
	DSD-SATN [123]	ICCV'19	Parameters	-	-	69.5	-	59.1	42.4
	Sim2Real [150]	NeurIPS'19	Parameters	[42]	-	74.7	-	-	-
	VIBE [99] †	CVPR'20	Parameters	-	82.9	51.9	99.1	65.6	41.4
	MEVA [127] †	ACCV'20	Parameters	-	86.9	54.7	-	-	-
	TCMR [129] †	CVPR'21	Parameters	-	95.0	55.8	111.3	62.3	41.1
	Lee et al. [184] †	ICCV'21	Parameters	-	92.8	52.2	106.1	58.4	38.4
	MAED [185] †	ICCV'21	Parameters	-	79.1	45.7	92.6	56.4	38.7
	MPS-Net [187] †	CVPR'22	Parameters	-	91.6	54.0	109.6	69.4	47.4
	GLoT [256] †	CVPR'23	Parameters	-	80.7	50.6	96.3	67.0	46.3
	PSVT [182]♯ †	CVPR'23	Parameters	-	73.1	43.5	84.0	-	-

Table 5
Evaluation of the whole-body recovery methods on the AGORA [296] dataset. **FB**, **B**, **F**, **LH/RH** denote evaluation results on the full-body, body, face, left-hand/righ-hand, respectively.

Method	Train on AGORA	NMVE		NMJE		MVE				MPJPE			
		FB	B	FB	B	FB	B	F	LH/RH	FB	B	F	LH/RH9
SMPLify-X [22]	-	333.1	263.3	326.5	256.5	236.5	187	48.9	48.3/51.4	231.8	182.1	52.9	46.5/49.6
ExPose [23]	✗	265.0	184.8	263.3	183.4	217.3	151.5	51.1	74.9/71.3	215.9	150.4	55.2	72.5/68.8
FrankMocap [243]	✗		207.8		204.0		168.3		54.7/55.7		165.2		52.3/53.1
PIXIE [24]	✗	233.9	173.4	230.9	171.1	191.8	142.2	50.2	49.5/49.0	189.3	140.3	54.5	46.4/46.0
Hand4Whole [25]	✓	144.1	96.0	141.1	92.7	135.5	90.2	41.6	46.3/48.1	132.6	87.1	46.1	44.3/46.2
PyMAF-X [172]	✓	141.2	94.4	140.0	93.5	125.7	84.0	35.0	44.6/45.6	124.6	83.2	37.9	42.5/43.7
OSX [246]	✓	130.6	85.3	127.6	83.3	122.8	80.2	36.2	45.4/46.1	119.9	78.3	37.9	43.0/43.9
HybrIK-X [244]	✓	120.5	73.7	115.7	72.3	112.1	68.5	37.0	46.7/47.0	107.6	67.2	38.5	41.2/41.4

Table 6
Evaluation of the whole-body recovery methods on the EHF [22] dataset.

Methods	Body Model	PA-V2V			PA-MPJPE	
		All	Hands	Face	Body	Hands
MTC [105]	Adam	67.2	-	-	107.8	16.7
SMPLify-X [22]	SMPL-X	65.3	12.3	6.3	87.6	12.9
ExPose [23]	SMPL-X	54.5	12.8	5.8	62.8	13.1
FrankMocap [243]	SMPL-X	57.5	12.6	-	62.3	12.9
PIXIE [24]	SMPL-X	55.0	11.1	4.6	61.5	11.6
Hand4Whole [25]	SMPL-X	50.3	10.8	5.8	60.4	10.8
PyMAF-X [172]	SMPL-X	50.2	10.2	5.5	52.8	10.3

augment them with 16 different shape parameters. Human meshes are placed in 100 synthetic environments and are animated with over 100 motion snippets.

BEDLAM [299] is a synthetic dataset aiming to increase the scale and realism by expanding the diversity of body poses, shapes, skin tones, hair, and clothing. Moreover, the clothing is more realistic clothing as they are simulated on the moving bodies using commercial clothing physics simulation.

SynBody [317] is a large-scale synthetic dataset comprising 1.2M images with corresponding 3D annotations. It covers 1,187 actions in various viewpoints, 10,000 body models, and 26,960 video clips with 2.7M SMPL/SMPLX annotations.

8.2.2 Marker/Sensor-based MoCap

HumanEva [58] includes HumanEva-I and HumanEva-II. The two datasets are captured in a multi-camera MoCap system. Reflective markers are attached to subjects to record 4 subjects performing 6 actions in HumanEva-I and 2 subjects performing 1 action in HumanEva-II. Both datasets contain synchronized video from multiple camera views and associated 3D pose ground truth.

Human3.6M [95] is a benchmark dataset for 3D pose estimation. It consists of 3.6 million video frames captured against indoor backgrounds from 4 viewpoints. 5 female and 6 male subjects perform 15 actions, with reflective markers attached to their body. The extended SMPL model annotations are generated by [16], [97] after applying MoSh [94] to sparse marker data. Alternatively, Moon et al. [134] apply SMPLify-X [22] to the ground truth 3D joints to get the label.

Total Capture [300] has fully synchronized video, IMU, and Vicon labeling for about 1.7M frames. There were 4 male and

1 female subjects participated, each performing five actions, repeated 3 times.

3DPW. *3D Poses in the Wild Dataset* [96] is captured in challenging outdoor scenes. The dataset includes over 51,000 frames for 7 actors in 18 clothing styles. A hand-held smartphone camera records 1 or 2 IMU-equipped actors performing rich activities. This dataset provides accurate mesh ground truth annotations by fitting the SMPL model to the raw ground-truth markers using a similar method to [94].

AMASS. *Archive of Motion Capture as Surface Shapes* [98] is a large and varied human motion dataset that spans over 300 subjects and contains more than 40 hours of motion data for over 110K motions. It unifies 15 marker-based MoCap datasets, including CMU MoCap [104] and PosePrior [197]. The SMPL model is used to represent motions via the proposed method MoSh++. Given credit for its sufficient richness, AMASS is widely adopted to learn human motion prior and assess the rationality of predicted poses or sequences of motions.

8.2.3 Marker-less Multi-view MoCap

CMU Panoptic [301] is a large-scale multi-person dataset captured by 480 synchronized cameras in the Panoptic Studio. For each session, 3 to 8 participants are asked to play various games together to get involved in social interactions. 1.5M frames with ground truth 3D skeletons from 65 sequences are currently available.

MPI-INF-3DHP [302] is a single-person 3D pose dataset collected in a multi-camera green screen studio. The system is equipped with 14 cameras and records 8 subjects in total. Each subject features 2 sets of clothing and performs 8 activities. Ground truth 3D pose annotations are available, but some noise exists. The authors further propose MuCo-3DHP [303] as data augmentation. It is built on the person masks in MPI-INF-3DHP. 1 to 4 subjects are pasted to real-world background images, resulting in 200K images that cover a range of inter-person overlap and activity scenarios.

MuPoTs-3D. *Multiperson Pose Test Set in 3D* [303] is a multi-person dataset for evaluation. It consists of more than 8000 frames covering 5 indoor and 15 outdoor settings. The ground-truth 3D poses are captured in a multi-view marker-less motion capture system.

MannequinChallenge [304] contains videos in which multiple people freeze in the pose, and the camera moves around to film the static scenes. The dataset originally provides estimated camera poses and dense depth. Leroy et al. [305]

further extend the annotations with 3D keypoints locations and visibility information using a SMPL-based approach.

3DOH50K. *3D Occlusion Human 50K* dataset [141] is captured in indoor scenes with 6 viewpoints. It contains more than 51,600 images, most of which are human activities in occlusion scenarios. The authors adapt SMPLify-X in a multi-view strategy to get the SMPL mesh ground truth.

Mirrored-Human [306] consists of videos from the Internet, in which we can see a person and the person's image in a mirror. The mirror reflection provides an additional view to resolve the depth ambiguity. The dataset provides 2D keypoints and pseudo-ground truth SMPL annotations generated by an optimization-based framework.

MTC. *Monocular Total Capture* dataset [105] is captured by the Panoptic Studio [301] with 31 HD cameras in a multi-view setup. The dataset contains about 834K body images and 111K hand images, representing a wide range of motions in the body and hand of multiple subjects.

EHF. *Expressive Hands and Faces* dataset [22] contains 100 samples for evaluation. Following [80], the SMPL-X model is aligned to original 4D scans. With special attention paid to hand poses and facial expressions, mesh annotations of the selected samples are of good alignment quality.

HUMBI [307] is a large multi-view dataset for human body expressions with natural clothing. 107 synchronized HD cameras are employed to capture 772 distinctive subjects. The subjects at the capture stage are asked to perform a series of gaze, face, hand, and body expression tasks. Each frame contains up to 4 representations: multi-view images, 3D keypoints, 3D mesh, and appearance maps. Basel Face Model [318], MANO [80], and SMPL [27] are adopted for face, hands, and body reconstruction, respectively.

ZJU-MoCap [35] consists of 9 dynamic human sequences captured by 21 synchronized cameras in a multi-view setup. The sequences have a length between 60 to 300 frames, in which actors do complex movements like twirling and kicking. The SMPL-X annotations are also available after iteratively optimizing the human model to align with the multi-view observations.

8.2.4 Datasets with Human-Scene Interactions

There are several datasets for investigating the task of human/hand mesh recovery with human-object interactions. PROX [191] uses a single Kinect camera to capture 20 subjects interacting with the indoor scenes. The dataset provides 12 indoor scene meshes and 100K RGB-D frames with pseudo SMPL-X labels. BEHAVE [319] captures dynamic human-object interactions using 4 Kinects in natural environments. The dataset contains multi-view RGB-D sequences and corresponding human models, objects, and contact annotations. It has 10.7k frames for training and 4.5k frames for testing. GRAB [320] uses a marker-based capture system to capture 10 subjects interacting with 51 everyday objects. The SMPL-X model is fitted to Mocap markers to present body pose, shape, facial expression, and hand gestures. However, the dataset does not have corresponding RGB(-D) frames. RICH [321] contains multiview outdoor/indoor high-resolution video sequences, ground-truth 3D human bodies, 3D body scans, and high-resolution 3D scene scans. SLOPER4D [322] is a scene-aware dataset

collected in urban environments, consisting of 15 sequences, more than 100K LiDAR frames, 300k video frames, and 500K IMU-based motion frames.

8.2.5 Datasets with Pseudo 3D Labels

2D pose datasets are known for their richness and diversity in subjects, poses, and scenes, but lack 3D pose or mesh annotations. Researchers have explored algorithms to generate pseudo-ground truth in an automatic or semi-automatic manner. **LSP** [309], **LSP-Extended** [310], **MSCOCO** [311], **MPII** [312], **PoseTrack** [313], **OCHuman** [314] are in-the-wild 2D human pose estimation datasets. Their labels are fitted in an optimization process [42] or with the help of regression networks [21], [44].

SSP-3D [116] is collected from the Sports-1M video dataset [323]. SSP-3D comprises 311 in-the-wild images of 62 tightly-clothed sportspersons with a diverse range of body shapes and corresponding pseudo-ground truth SMPL shape and pose labels.

MTP. *Mimic The Pose* dataset [194] contains 3,731 images corresponding to 1,653 SMPL-X meshes. 3D meshes exhibit self-contact, and images are collected after asking participants to mimic the poses and contacts. Since the presented pose, shape, and gender are not aligned perfectly, the authors further adapt SMPLify-X [22] to refine the original meshes.

UBody. *Upper-Body* dataset [246] mainly focuses on representing upper bodies. It contains a series of close-up shots of humans with rich hand gestures and facial expressions in 15 real-life scenarios. The dataset has 2D annotations and high-quality 3D pseudo-GT SMPL-X fits.

Motion-X [324] is a large-scale motion dataset comprising 15.6M 3D whole-body pose annotations in the form of SMPL-X. It consists of 81.1K motion sequences of in-the-wild scenes and provides corresponding semantic labels and pose descriptions.

9 EVALUATION

In this section, we discuss the evaluation metrics and the benchmark results from multiple perspectives.

9.1 Metrics

MPJPE. *Mean Per Joint Position Error* measures the average Euclidean distance between predicted 3D joints and ground truth after root matching. It is defined in the local space. Recently, SPEC [119] introduces W-MPJPE that computes 3D joints error in the world coordinates. The authors believe it can better reflect performance in real-world applications.

PA-MPJPE. *Procrustes-aligned MPJPE* denotes MPJPE after rigid alignment of the predicted pose and ground truth. Procrustes Analysis removes the effects of translation, rotation, and scale. Thus, PA-MPJPE concerns the reconstructed 3D mesh/pose itself. It is also referred to as the reconstruction error.

PVE/V2V. *Mean Per-vertex Error* or *Vertex-to-Vertex* is defined as the average point-to-point Euclidean distance between predicted mesh vertices and ground truth mesh vertices. Similar to MPJPEC, W-PVE, a variant of PVE, is proposed [119] to calculate in the world space.

MPJAE. *Mean Per Joint Angle Error* represents the orientation deviation between predicted 3D joints and ground truth, which is measured in $SO(3)$ using the geodesic distance.

PA-MPJAE. *Procrustes-aligned MPJAE* is calculated according to MPJAE after executing Procrustes Analysis to align predicted poses with ground truth.

9.2 Benchmark Leaderboards

The quantitative comparison of 3D body mesh recovery on Human3.6M [95] and 3DPW [96] are illustrated in Table 4. With the researchers' persistent efforts, the performance has been improving each year. However, the deployment and evaluation standards for comparison are not fully consistent. Different combinations of backbones, output types, pseudo labels, datasets, training strategies, and evaluation protocols would lead to a fluctuation in values. SPIN [44] establishes an evaluation protocol that is widely adopted by the follow-ups in the table. In 3DPW, most approaches follow *Protocol 2* and use the test set for evaluation without any fine-tuning on the training set. But the strategy is different in [118], [119], [127], [135], [136], [143], [149], [184] in which 3DPW train set is used during training. [138], [149] use the GHUM model [92] to represent the pose and shape, while others adopt the SMPL model [27]. In general, ResNet-50 [158] serves as a generic convolutional backbone to extract features from images, except that [135], [136] use HRNet [159] and multi-stage pipelines [17], [18], [114], [141], [150] have multiple convolutional modules. For the methods that yield non-parametric outputs, metrics will degrade in general after the outputs are converted to parameters with an additional parameter regression module [133], [134].

There are much fewer algorithms that deal with full-body mesh recovery with face and hands, compared to body-only mesh recovery. Table 5 and Table 6 show the performances of the full-body recovery task on the AGORA dataset [296] and the EHF dataset [22], respectively. Results on body-only, face, and hands are also included in these tables for comprehensive evaluations. Since AGORA does not provide the ground-truth labels for its test set, the performances are calculated after uploading the results to the official evaluation platform [296]. Comparing the results of **FB** and **B** in Table 5 and Table 6, we can observe that this task is very challenging as the reconstruction error becomes much higher when taking face and hands into consideration.

10 CONCLUSION AND FUTURE DIRECTIONS

In this survey, we provide a thorough overview of 3D human mesh recovery methods in the past decade. The categorization is based on design paradigm, reconstruction granularity, and application scenarios. We also give special considerations for physical plausibility, including camera models, contact constraints, and human priors. In the experiment section, we introduce relevant datasets, evaluation metrics and provide performance comparison. Next, we highlight a few promising future directions, hoping to promote advances in this field.

Under Heavy Occlusions. In real-world scenarios, occlusions are ubiquitous. People often appear partially or heavily occluded due to self-overlapping, close-range interaction with other people, or occlusion of scene objects. Even though the occlusion has been extensively studied for years [20], [114], [115], [141], robustness and stability are still need to be improved. Besides, the visual evidence may be insufficient to identify a 3D reconstruction uniquely, recover several plausible reconstructions [142] or a pose distribution [144], [145] for one input is worthwhile.

Stable Reconstruction from Videos. Motion jitters, i.e., irregular movement and variation across frames, remain a severe issue in existing regression-based temporal-based methods [99], [127], [129]. The visual performance is largely influenced by motion jitters. Jitters are slight when much of the body is observable, while severe jitters occur in those frames with heavy occlusion or in a complex context. To improve temporal smoothness, we need to deal with long-term motion jitters. There is a trend to perform pose refinement after primary estimation using low-pass filters or learning-based refinement networks [325].

Reconstruction with Scene Constraints. Standard methods perform 3D human pose estimation without explicitly considering the scene. This may lead to inter-penetration with the 3D scene. Most methods ignore the scene constraint during estimation. In the methods that aim to reconstruct physically consistent results, scenes are typically assumed as flat floors [195], [196], [275] for simplicity. [191], [326], [327], [328] are among the first to go beyond flat floors and resolve human pose and shape from static 3D scenes. Further work may take scene mesh into consideration to better capture the motion of humans interacting with a real static 3D scene.

Beyond Fully Supervised Learning. Building 3D human mesh datasets is time-consuming and of high cost. A MoCap system needs to be set up beforehand. After capturing, the cleaning and annotation process of raw 3D data is highly demanding. Besides, 3D datasets lack diversity in human motion and background, but 2D datasets are far more substantial. In light of this, it is promising to make use of the abundant unlabeled data to train a network in an unsupervised fashion. Recent unsupervised 3D pose estimation [329], [330] has achieved exciting performance. Compared to this, unsupervised [110], [260] or self-supervised [115], [117], [331] human mesh recovery is much more difficult due to richer reconstruction information.

Grouped Person Reconstruction. In public scenes, people often walk, talk, or work together in groups as family members, teammates, etc. An interesting future direction is reconstructing a group of people over space and time, which reveals the relationships and activities in the target group. Moreover, when considering person matching across different cameras or long-range temporal sequences, the relationship of individuals within a group provides a more stable context that can be exploited to handle occlusions or detection failures. This task can also be combined with person tracking [188] and re-identification [332], [333] for more robust reconstruction in crowded scenarios.

Whole-body Human Mesh Recovery. There is a trend to utilize a unified framework to regress the body, hands, and face parameters of expressive human models [22], [92]. Compared with body-only mesh recovery, there are much fewer methods to deal with whole-body mesh recovery [23], [24], [172], [243], [244]. One major challenge is that the

whole-body datasets are rather scarce for training. Separate body/hand/face-only datasets are typically used to compensate for the incompleteness of whole-body data. This brings challenges to the consistent recovery of body poses and hand gestures. Moreover, the occlusions, motion blur, depth ambiguity, and interaction of the hand regions also impose great challenges to the monocular whole-body mesh recovery with plausible hand poses.

Detailed Shape Reconstruction with Clothing. Parametric models like SMPL and SMPL-X can only represent minimally clothed humans. The research community needs to exploit other representations with more flexibility to go beyond the representation power of parametric models. In existing work, meshes [334], [335], [336], point clouds [38], [337], [338], [339], and implicit fields [86], [316], [340], [341] have been used to model the detailed deformation of clothing. Though these methods can produce reasonable results, their reconstructed surfaces tend to be over-smoothed and not robust to novel poses. These issues can be alleviated by incorporating different types of representations [29], [342], [343], [344], [345] to leverage the modeling power of different representations.

ACKNOWLEDGMENTS

REFERENCES

[1] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *TPAMI*, vol. 43, no. 1, pp. 172–186, 2019.

[2] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *ICCV*, 2017, pp. 2334–2343.

[3] S. Kreiss, L. Bertoni, and A. Alahi, "OpenPifPaf: Composite fields for semantic keypoint detection and spatio-temporal association," *TITS*, vol. 23, no. 8, pp. 13 498–13 511, 2021.

[4] Q. Chen, T. Ge, Y. Xu, Z. Zhang, X. Yang, and K. Gai, "Semantic human matting," in *ACM MM*, 2018, pp. 618–626.

[5] J. Zhao, J. Li, Y. Cheng, T. Sim, S. Yan, and J. Feng, "Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing," in *ACM MM*, 2018, pp. 792–800.

[6] K. Grauman, G. Shakhnarovich, and T. Darrell, "Inferring 3D structure with a statistical image-based shape model," in *ICCV*, 2003, pp. 641–648.

[7] A. Agarwal and B. Triggs, "Recovering 3D human pose from monocular images," *TPAMI*, vol. 28, no. 1, pp. 44–58, 2005.

[8] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3D human pose estimation," in *ICCV*, 2017, pp. 2659–2668.

[9] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Coarse-to-fine volumetric prediction for single-image 3D human pose," in *CVPR*, 2017, pp. 1263–1272.

[10] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, "Integral human pose regression," in *ECCV*, 2018, pp. 536–553.

[11] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, M. Elgharib, P. Fua, H.-P. Seidel, H. Rhodin, G. Pons-Moll, and C. Theobalt, "XNect: Real-time multi-person 3D motion capture with a single RGB camera," *TOG*, vol. 39, no. 4, pp. 82–1, 2020.

[12] P. Weinzaepfel, R. Brégier, H. Combaluzier, V. Leroy, and G. Rogez, "DOPE: Distillation of part experts for whole-body 3D pose estimation in the wild," in *ECCV*, 2020, pp. 380–397.

[13] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image," in *ECCV*. Springer, 2016, pp. 561–578.

[14] Y. Huang, F. Bogo, C. Lassner, A. Kanazawa, P. V. Gehler, J. Romero, I. Akhter, and M. J. Black, "Towards accurate markerless human shape and pose estimation over time," in *3DV*. IEEE, 2017, pp. 421–430.

[15] A. Zanfir, E. Marinoiu, and C. Sminchisescu, "Monocular 3D pose and shape estimation of multiple people in natural scenes - the importance of multiple scene constraints," in *CVPR*, 2018, pp. 2148–2157.

[16] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *CVPR*, 2018, pp. 7122–7131.

[17] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis, "Learning to estimate 3D human pose and shape from a single color image," in *CVPR*, 2018, pp. 459–468.

[18] M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, and B. Schiele, "Neural Body Fitting: Unifying deep learning and model-based human pose and shape estimation," in *3DV*. IEEE, 2018, pp. 484–494.

[19] H. Zhang, Y. Tian, X. Zhou, W. Ouyang, Y. Liu, L. Wang, and Z. Sun, "PyMAF: 3D human pose and shape regression with pyramidal mesh alignment feedback loop," in *ICCV*, 2021.

[20] M. Kocabas, C.-H. P. Huang, O. Hilliges, and M. J. Black, "PARE: Part attention regressor for 3D human body estimation," in *ICCV*, 2021, pp. 11 127–11 137.

[21] H. Joo, N. Neverova, and A. Vedaldi, "Exemplar fine-tuning for 3D human pose fitting towards in-the-wild 3D human pose estimation," in *3DV*, 2021, pp. 42–52.

[22] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3D hands, face, and body from a single image," in *CVPR*, 2019, pp. 10 975–10 985.

[23] V. Choutas, G. Pavlakos, T. Bolkart, D. Tzionas, and M. J. Black, "Monocular expressive body regression through body-driven attention," in *ECCV*. Springer, 2020, pp. 20–40.

[24] Y. Feng, V. Choutas, T. Bolkart, D. Tzionas, and M. J. Black, "Collaborative regression of expressive bodies using moderation," in *3DV*, 2021.

[25] G. Moon, H. Choi, and K. M. Lee, "Accurate 3D hand pose estimation for whole-body 3D human mesh estimation," in *CVPRW*, 2022, pp. 2308–2317.

[26] Y. Zhang, Z. Li, L. An, M. Li, T. Yu, and Y. Liu, "Lightweight multi-person total motion capture using sparse multi-view cameras," in *ICCV*, 2021, pp. 5560–5569.

[27] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *TOG*, vol. 34, no. 6, pp. 1–16, 2015.

[28] T. Yu, Z. Zheng, K. Guo, J. Zhao, Q. Dai, H. Li, G. Pons-Moll, and Y. Liu, "DoubleFusion: Real-time capture of human performances with inner body shapes from a single depth sensor," in *CVPR*, 2018, pp. 7287–7296.

[29] Z. Zheng, T. Yu, Y. Liu, and Q. Dai, "PaMIR: Parametric model-conditioned implicit representation for image-based human reconstruction," *TPAMI*, 2021.

[30] Y. Zheng, R. Shao, Y. Zhang, T. Yu, Z. Zheng, Q. Dai, and Y. Liu, "DeepMultiCap: Performance capture of multiple characters using sparse multiview cameras," in *ICCV*, 2021, pp. 6239–6249.

[31] K. Li, H. Wen, Q. Feng, Y. Zhang, X. Li, J. Huang, C. Yuan, Y.-K. Lai, and Y. Liu, "Image-guided human reconstruction via multi-scale graph transformation networks," *TIP*, vol. 30, pp. 5239–5251, 2021.

[32] Q. Feng, Y. Liu, Y.-K. Lai, J. Yang, and K. Li, "FOF: Learning fourier occupancy field for monocular real-time human reconstruction," in *NeurIPS*, 2022.

[33] Y. Xiu, J. Yang, D. Tzionas, and M. J. Black, "ICON: implicit clothed humans obtained from normals," in *CVPR*, 2022, pp. 13 286–13 296.

[34] Y. Xiu, J. Yang, X. Cao, D. Tzionas, and M. J. Black, "ECON: Explicit clothed humans optimized via normal integration," in *CVPR*, 2023, pp. 512–523.

[35] S. Peng, Y. Zhang, Y. Xu, Q. Wang, Q. Shuai, H. Bao, and X. Zhou, "Neural Body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans," in *CVPR*, 2021, pp. 9054–9063.

[36] T. Hu, T. Yu, Z. Zheng, H. Zhang, Y. Liu, and M. Zwicker, "HVTR: Hybrid volumetric-textural rendering for human avatars," in *3DV*. IEEE, 2022, pp. 197–208.

[37] Z. Huang, Y. Xu, C. Lassner, H. Li, and T. Tung, "ARCH: Animatable reconstruction of clothed humans," in *CVPR*, 2020, pp. 3093–3102.

[38] Q. Ma, J. Yang, S. Tang, and M. J. Black, "The power of points for modeling humans in clothing," in *ICCV*, 2021, pp. 10974–10984.

[39] Z. Zheng, H. Huang, T. Yu, H. Zhang, Y. Guo, and Y. Liu, "Structured local radiance fields for human avatar modeling," in *CVPR*, 2022, pp. 15893–15903.

[40] Z. Zheng, X. Zhao, H. Zhang, B. Liu, and Y. Liu, "AvatarReX: Real-time expressive full-body avatars," *ACM TOG*, vol. 42, no. 4, 2023.

[41] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis, "SCAPE: shape completion and animation of people," *TOG*, vol. 24, pp. 408–416, 2005.

[42] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler, "Unite the people: Closing the loop between 3D and 2D human representations," in *CVPR*, 2017, pp. 6050–6059.

[43] H.-Y. F. Tung, H.-W. Tung, E. Yumer, and K. Fragkiadaki, "Self-supervised learning of motion capture," *NeurIPS*, pp. 5236–5246, 2017.

[44] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis, "Learning to reconstruct 3D human pose and shape via model-fitting in the loop," in *ICCV*, 2019, pp. 2252–2261.

[45] L. Chen, S. Peng, and X. Zhou, "Towards efficient and photorealistic 3D human reconstruction: A brief survey," *Visual Informatics*, vol. 5, no. 4, pp. 11–19, 2021.

[46] A. Tewari, O. Fried, J. Thies, V. Sitzmann, S. Lombardi, K. Sunkavalli, R. Martin-Brualla, T. Simon, J. Saragih, M. Nießner *et al.*, "State of the art on neural rendering," in *CGF*, vol. 39. Wiley Online Library, 2020, pp. 701–727.

[47] Y. Chen, Y. Tian, and M. He, "Monocular human pose estimation: A survey of deep learning-based methods," *CVIU*, vol. 192, p. 102897, 2020.

[48] C. Zheng, W. Wu, T. Yang, S. Zhu, C. Chen, R. Liu, J. Shen, N. Kehtarnavaz, and M. Shah, "Deep learning-based human pose estimation: A survey," *arXiv preprint arXiv:2012.13392*, 2020.

[49] W. Liu, Q. Bao, Y. Sun, and T. Mei, "Recent advances of monocular 2D and 3D human pose estimation: A deep learning perspective," *ACM Computing Surveys*, vol. 55, no. 4, pp. 1–41, 2022.

[50] H.-J. Lee and Z. Chen, "Determination of 3D human body postures from a single view," *Computer Vision, Graphics, and Image Processing*, vol. 30, no. 2, pp. 148–168, 1985.

[51] R. Nevatia and T. O. Binford, "Description and recognition of curved objects," *Artificial intelligence*, vol. 8, no. 1, pp. 77–98, 1977.

[52] S. X. Ju, M. J. Black, and Y. Yacoob, "Cardboard people: A parameterized model of articulated image motion," in *FG*. IEEE, 1996, pp. 38–44.

[53] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *SIGGRAPH*, 1999, pp. 187–194.

[54] D. Marr and H. K. Nishihara, "Representation and recognition of the spatial organization of three-dimensional shapes," *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 200, no. 1140, pp. 269–294, 1978.

[55] K. Rohr, "Towards model-based recognition of human movements in image sequences," *CVGIP: Image understanding*, vol. 59, no. 1, pp. 94–115, 1994.

[56] S. Wachter and H.-H. Nagel, "Tracking of persons in monocular image sequences," *CVIU*, vol. 74, no. 3, pp. 174–192, 1999.

[57] H. Sidenbladh, M. J. Black, and D. J. Fleet, "Stochastic tracking of 3D human figures using 2D image motion," in *ECCV*. Springer, 2000, pp. 702–718.

[58] L. Sigal, A. O. Balan, and M. J. Black, "HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *IJCV*, vol. 87, no. 1-2, p. 4, 2010.

[59] M. Wang, F. Qiu, W. Liu, C. Qian, X. Zhou, and L. Ma, "Monocular human pose and shape reconstruction using part differentiable rendering," in *CGF*, vol. 39. Wiley Online Library, 2020, pp. 351–362.

[60] A. Pentland and B. Horowitz, "Recovery of nonrigid motion and structure," *TPAMI*, vol. 13, no. 07, pp. 730–742, 1991.

[61] D. Metaxas and D. Terzopoulos, "Shape and nonrigid motion estimation through physics-based synthesis," *TPAMI*, vol. 15, no. 6, pp. 580–591, 1993.

[62] D. M. Gavrila, *Vision-based 3-D tracking of humans in action*. University of Maryland, College Park, 1996.

[63] C. Sminchisescu and B. Triggs, "Estimating articulated human motion with covariance scaled sampling," *International Journal of Robotics Research*, vol. 22, no. 6, pp. 371–391, 2003.

[64] R. Plänkers and P. Fua, "Tracking and modeling people in video sequences," *CVIU*, vol. 81, no. 3, pp. 285–302, 2001.

[65] L. Kakadiaris and D. Metaxas, "Model-based estimation of 3D human motion," *TPAMI*, vol. 22, no. 12, pp. 1453–1459, 2000.

[66] G. Pons-Moll and B. Rosenhahn, "Model-based pose estimation," *Visual Analysis of Humans*, pp. 139–170, 2011.

[67] B. Allen, B. Curless, and Z. Popović, "The space of human body shapes: reconstruction and parameterization from range scans," *TOG*, vol. 22, no. 3, pp. 587–594, 2003.

[68] N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H.-P. Seidel, "A statistical model of human pose and body shape," in *CGF*, vol. 28. Wiley Online Library, 2009, pp. 337–346.

[69] Y. Chen, Z. Liu, and Z. Zhang, "Tensor-based human body modeling," in *CVPR*, 2013, pp. 105–112.

[70] O. Freifeld and M. J. Black, "Lie bodies: A manifold representation of 3D human shape," in *ECCV*. Springer, 2012, pp. 1–14.

[71] D. A. Hirshberg, M. Loper, E. Rachlin, and M. J. Black, "Coregistration: Simultaneous alignment and modeling of articulated 3D shape," in *ECCV*. Springer, 2012, pp. 242–255.

[72] G. Pons-Moll, J. Romero, N. Mahmood, and M. J. Black, "Dyna: A model of dynamic human shape in motion," *TOG*, vol. 34, no. 4, pp. 1–14, 2015.

[73] B. Allen, B. Curless, Z. Popović, and A. Hertzmann, "Learning a correlated model of identity and pose-dependent body shape variation for real-time synthesis," in *SCA*. ACM, 2006, pp. 147–156.

[74] N. Hasler, T. Thormählen, B. Rosenhahn, and H.-P. Seidel, "Learning skeletons for shape and pose," in *I3D*, 2010, pp. 23–30.

[75] H. Wang, R. A. Güler, I. Kokkinos, G. Papandreou, and S. Zafeiriou, "BLSM: A bone-level skinned model of the human mesh," in *ECCV*. Springer, 2020, pp. 1–17.

[76] K. M. Robinette, S. Blackwell, H. Daanen, M. Boehmer, S. Fleming, T. Brill, D. Hoeferlin, and D. Burnsides, "Civilian American and European Surface Anthropometry Resource (CAESAR) final report," US Air Force Research Laboratory, Tech. Rep. AFRL-HE-WP-TR-2002-0169, 2002.

[77] S. Zuffi and M. J. Black, "The stitched puppet: A graphical model of 3D human shape and pose," in *CVPR*, 2015, pp. 3537–3546.

[78] A. Mohr and M. Gleicher, "Building efficient, accurate character skins from examples," *TOG*, vol. 22, no. 3, pp. 562–568, 2003.

[79] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 2D scans," *TOG*, vol. 36, no. 6, pp. 194–1, 2017.

[80] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *TOG*, vol. 36, no. 6, pp. 1–17, 2017.

[81] N. Hesse, S. Pujades, J. Romero, M. J. Black, C. Bodensteiner, M. Arens, U. G. Hofmann, U. Tacke, M. Hadders-Algra, R. Weinberger *et al.*, "Learning an infant body model from RGB-D data for accurate full body motion analysis," in *MICCAI*. Springer, 2018, pp. 792–800.

[82] I. Santesteban, E. Garces, M. A. Otaduy, and D. Casas, "SoftSMPL: Data-driven modeling of nonlinear soft-tissue dynamics for parametric humans," in *CGF*, vol. 39. Wiley Online Library, 2020, pp. 65–75.

[83] A. A. Osman, T. Bolkart, and M. J. Black, "STAR: Sparse trained articulated human body regressor," in *ECCV*. Springer, 2020, pp. 598–613.

[84] B. Deng, J. P. Lewis, T. Jeruzalski, G. Pons-Moll, G. Hinton, M. Norouzi, and A. Tagliasacchi, "NASA: Neural articulated shape approximation," in *ECCV*. Springer, 2020, pp. 612–628.

[85] M. Mihajlovic, Y. Zhang, M. J. Black, and S. Tang, "LEAP: Learning articulated occupancy of people," in *CVPR*, 2021, pp. 10461–10471.

[86] X. Chen, Y. Zheng, M. J. Black, O. Hilliges, and A. Geiger, "SNARF: Differentiable forward skinning for animating nonrigid neural implicit shapes," in *ICCV*, 2021, pp. 11594–11604.

[87] M. Mihajlovic, S. Saito, A. Bansal, M. Zollhoefer, and S. Tang, "COAP: Compositional articulated occupancy of people," in *CVPR*, 2022, pp. 13201–13210.

[88] X. Sun, Q. Feng, X. Li, J. Zhang, Y.-K. Lai, J. Yang, and K. Li, "Learning semantic-aware disentangled representation for flexible 3D human body editing," in *CVPR*, 2023.

[89] H. Joo, T. Simon, and Y. Sheikh, "Total Capture: A 3D deformation model for tracking faces, hands, and bodies," in *CVPR*, 2018, pp. 8320–8329.

[90] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "FaceWarehouse: a 3D facial expression database for visual computing," *TVCG*, vol. 20, no. 3, pp. 413–425, 2013.

[91] A. A. Osman, T. Bolkart, D. Tzionas, and M. J. Black, "SUPR: A sparse unified part-based human representation," in *ECCV*. Springer, 2022, pp. 568–585.

[92] H. Xu, E. G. Bazavan, A. Zanfir, W. T. Freeman, R. Sukthankar, and C. Sminchisescu, "GHUM & GHUML: Generative 3D human shape and articulated pose models," in *CVPR*, 2020, pp. 6184–6193.

[93] A. O. Balan, L. Sigal, M. J. Black, J. E. Davis, and H. W. Haussecker, "Detailed human shape and pose from images," in *CVPR*. IEEE, 2007, pp. 1–8.

[94] M. Loper, N. Mahmood, and M. J. Black, "MoSh: Motion and shape capture from sparse markers," *TOG*, vol. 33, no. 6, pp. 1–13, 2014.

[95] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments," *TPAMI*, vol. 36, no. 7, pp. 1325–1339, 2014.

[96] T. von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll, "Recovering accurate 3D human pose in the wild using IMUs and a moving camera," in *ECCV*, 2018, pp. 601–617.

[97] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid, "Learning from synthetic humans," in *CVPR*, 2017, pp. 109–117.

[98] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "AMASS: Archive of motion capture as surface shapes," in *ICCV*, 2019, pp. 5442–5451.

[99] M. Kocabas, N. Athanasiou, and M. J. Black, "VIBE: Video inference for human body pose and shape estimation," in *CVPR*, 2020, pp. 5253–5263.

[100] A. O. Bălan and M. J. Black, "The naked truth: Estimating body shape under clothing," in *ECCV*. Springer, 2008, pp. 15–29.

[101] P. Guan, A. Weiss, A. O. Balan, and M. J. Black, "Estimating human shape and pose from a single image," in *ICCV*. IEEE, 2009, pp. 1381–1388.

[102] N. Hasler, H. Ackermann, B. Rosenhahn, T. Thormählen, and H.-P. Seidel, "Multilinear pose and body shape estimation of dressed subjects from image sets," in *CVPR*. IEEE, 2010, pp. 1823–1830.

[103] S. Zhou, H. Fu, L. Liu, D. Cohen-Or, and X. Han, "Parametric reshaping of human bodies in images," *TOG*, vol. 29, no. 4, pp. 1–10, 2010.

[104] "Carnegie mellon university - cmu graphics lab - motion capture library," http://mocap.cs.cmu.edu/, 2010.

[105] D. Xiang, H. Joo, and Y. Sheikh, "Monocular total capture: Posing face, body, and hands in the wild," in *CVPR*, 2019, pp. 10 965–10 974.

[106] R. A. Güler and I. Kokkinos, "HoloPose: Holistic 3D human reconstruction in-the-wild," in *CVPR*, 2019, pp. 10 884–10 894.

[107] R. A. Güler, N. Neverova, and I. Kokkinos, "DensePose: Dense human pose estimation in the wild," in *CVPR*, 2018, pp. 7297–7306.

[108] J. Song, X. Chen, and O. Hilliges, "Human body model fitting by learned gradient descent," in *ECCV*. Springer, 2020, pp. 744–760.

[109] U. Iqbal, K. Xie, Y. Guo, J. Kautz, and P. Molchanov, "KAMA: 3D keypoint aware body mesh articulation," in *3DV*. IEEE, 2021, pp. 689–699.

[110] Z. Yu, J. Wang, J. Xu, B. Ni, C. Zhao, M. Wang, and W. Zhang, "Skeleton2Mesh: Kinematics prior injected unsupervised human mesh recovery," in *ICCV*, 2021, pp. 8619–8629.

[111] J. Li, C. Xu, Z. Chen, S. Bian, L. Yang, and C. Lu, "HybrIK: A hybrid analytical-neural inverse kinematics solution for 3D human pose and shape estimation," in *CVPR*, 2021, pp. 3383–3393.

[112] J. Li, S. Bian, Q. Liu, J. Tang, F. Wang, and C. Lu, "NIKI: Neural inverse kinematics with invertible neural networks for 3D human pose and shape estimation," in *CVPR*, 2023, pp. 12 933–12 942.

[113] K. Shetty, A. Birkhold, S. Jaganathan, N. Strobel, M. Kowarschik, A. Maier, and B. Egger, "PLIKS: A pseudo-linear inverse kinematic solver for 3D human body estimation," in *CVPR*, 2023, pp. 574–584.

[114] H. Zhang, J. Cao, G. Lu, W. Ouyang, and Z. Sun, "DaNet: Decompose-and-aggregate network for 3D human shape and pose estimation," in *ACM MM*, 2019, pp. 935–944.

[115] C. Rockwell and D. F. Fouhey, "Full-body awareness from partial observations," in *ECCV*. Springer, 2020, pp. 522–539.

[116] A. Sengupta, I. Budvytis, and R. Cipolla, "Synthetic training for accurate 3D human pose and shape estimation in the wild," in *BMVC*, September 2020.

[117] X. Xu, H. Chen, F. Moreno-Noguer, L. A. Jeni, and F. De la Torre, "3D human shape and pose from a single low-resolution image with self-supervised learning," in *ECCV*. Springer, 2020, pp. 284–300.

[118] A. Zanfir, E. G. Bazavan, H. Xu, W. T. Freeman, R. Sukthankar, and C. Sminchisescu, "Weakly supervised 3D human pose and shape reconstruction with normalizing flows," in *ECCV*. Springer, 2020, pp. 465–481.

[119] M. Kocabas, C.-H. P. Huang, J. Tesch, L. Muller, O. Hilliges, and M. J. Black, "SPEC: Seeing people in the wild with an estimated camera," in *ICCV*, 2021, pp. 11 035–11 045.

[120] Z. Li, J. Liu, Z. Zhang, S. Xu, and Y. Yan, "CLIFF: Carrying location information in full frames into human pose and shape estimation," in *ECCV*. Springer, 2022, pp. 590–606.

[121] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik, "Learning 3D human dynamics from video," in *CVPR*, 2019, pp. 5614–5623.

[122] Y. Xu, S.-C. Zhu, and T. Tung, "DenseRaC: Joint 3D pose and shape estimation by dense render-and-compare," in *ICCV*, 2019, pp. 7760–7770.

[123] Y. Sun, Y. Ye, W. Liu, W. Gao, Y. Fu, and T. Mei, "Human mesh recovery from monocular images via a skeleton-disentangled representation," in *ICCV*, 2019, pp. 5349–5358.

[124] G. Georgakis, R. Li, S. Karanam, T. Chen, J. Košecká, and Z. Wu, "Hierarchical kinematic human mesh recovery," in *ECCV*. Springer, 2020, pp. 768–784.

[125] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, "On the continuity of rotation representations in neural networks," in *CVPR*, 2019, pp. 5745–5753.

[126] H. Zhang, J. Cao, G. Lu, W. Ouyang, and Z. Sun, "Learning 3D human shape and pose from dense body parts," *TPAMI*, 2020.

[127] Z. Luo, S. A. Golestaneh, and K. M. Kitani, "3D human motion estimation via motion compression and refinement," in *ACCV*, 2020.

[128] Y. Zhou, M. Habermann, I. Habibie, A. Tewari, C. Theobalt, and F. Xu, "Monocular real-time full body capture with inter-part correlations," in *CVPR*, 2021, pp. 4811–4822.

[129] H. Choi, G. Moon, J. Y. Chang, and K. M. Lee, "Beyond static features for temporally consistent 3D human pose and shape from a video," in *CVPR*, 2021, pp. 1964–1973.

[130] D. Wang and S. Zhang, "3D human mesh recovery with sequentially global rotation estimation," in *ICCV*, 2023, pp. 14 953–14 962.

[131] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid, "BodyNet: Volumetric inference of 3D human body shapes," in *ECCV*, 2018, pp. 20–36.

[132] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu, "DeepHuman: 3D human reconstruction from a single image," in *ICCV*, 2019, pp. 7738–7748.

[133] N. Kolotouros, G. Pavlakos, and K. Daniilidis, "Convolutional mesh regression for single-image human shape reconstruction," in *CVPR*, 2019, pp. 4501–4510.

[134] G. Moon and K. M. Lee, "I2L-MeshNet: Image-to-lixel prediction network for accurate 3D human pose and mesh estimation from a single RGB image," in *ECCV*. Springer, 2020, pp. 752–768.

[135] K. Lin, L. Wang, and Z. Liu, "Mesh Graphormer," in *ICCV*, 2021.

[136] ——, "End-to-end human pose and mesh reconstruction with transformers," in *CVPR*, 2021, pp. 1954–1963.

[137] T. Luan, Y. Wang, J. Zhang, Z. Wang, Z. Zhou, and Y. Qiao, "PC-HMR: Pose calibration for 3D human mesh recovery from 2D images/videos," in *AAAI*, vol. 35, no. 3, 2021, pp. 2269–2276.

[138] M. Zanfir, A. Zanfir, E. G. Bazavan, W. T. Freeman, R. Sukthankar, and C. Sminchisescu, "THUNDR: Transformer-based 3D human reconstruction with markers," in *CVPR*, 2021.

[139] P. Yao, Z. Fang, F. Wu, Y. Feng, and J. Li, "DenseBody: Directly regressing dense 3D human pose and shape from a single color image," *arXiv preprint arXiv:1903.10153*, 2019.

[140] W. Zeng, W. Ouyang, P. Luo, W. Liu, and X. Wang, "3D human mesh regression with dense correspondence," in *CVPR*, 2020, pp. 7054–7063.

[141] T. Zhang, B. Huang, and Y. Wang, "Object-occluded human shape and pose estimation from a single color image," in *CVPR*, 2020, pp. 7376–7385.

[142] B. Biggs, D. Novotny, S. Ehrhardt, H. Joo, B. Graham, and A. Vedaldi, "3D multi-bodies: Fitting sets of plausible 3D human models to ambiguous image data," *NeurIPS*, vol. 33, 2020.

[143] A. Sengupta, I. Budvytis, and R. Cipolla, "Probabilistic 3D human shape and pose estimation from multiple unconstrained images in the wild," in *CVPR*, 2021, pp. 16 094–16 104.

[144] N. Kolotouros, G. Pavlakos, D. Jayaraman, and K. Daniilidis, "Probabilistic modeling for human mesh recovery," in *ICCV*, 2021, pp. 11 605–11 614.

[145] A. Sengupta, I. Budvytis, and R. Cipolla, "Hierarchical kinematic probability distributions for 3D human shape and pose estimation from images in the wild," in *ICCV*, 2021, pp. 11 219–11 229.

[146] Q. Fang, K. Chen, Y. Fan, Q. Shuai, J. Li, and W. Zhang, "Learning analytical posterior probability for human mesh recovery," in *CVPR*, 2023, pp. 8781–8791.

[147] A. Sengupta, I. Budvytis, and R. Cipolla, "HuManiFlow: Ancestor-conditioned normalising flows on SO (3) manifolds for human pose and shape distribution estimation," in *CVPR*, 2023, pp. 4779–4789.

[148] N. Rueegg, C. Lassner, M. Black, and K. Schindler, "Chained representation cycling: Learning to estimate 3D human pose and shape by cycling between representations," in *AAAI*, vol. 34, 2020, pp. 5561–5569.

[149] A. Zanfir, E. G. Bazavan, M. Zanfir, W. T. Freeman, R. Sukthankar, and C. Sminchisescu, "Neural descent for visual 3D human pose and shape," in *CVPR*, 2021, pp. 14 484–14 493.

[150] C. Doersch and A. Zisserman, "Sim2real transfer learning for 3D human pose estimation: motion to the rescue," *NeurIPS*, vol. 32, pp. 12 949–12 961, 2019.

[151] H. Choi, G. Moon, and K. M. Lee, "Pose2Mesh: Graph convolutional network for 3D human pose and mesh recovery from a 2D human pose," in *ECCV*, 2020.

[152] Z. Li, B. Xu, H. Huang, C. Lu, and Y. Guo, "Deep two-stream video inference for human body pose and shape estimation," in *WACV*, 2022, pp. 430–439.

[153] X. Gong, M. Zheng, B. Planche, S. Karanam, T. Chen, D. Doermann, and Z. Wu, "Self-supervised human mesh recovery with cross-representation alignment," in *ECCV*. Springer, 2022, pp. 212–230.

[154] H. Choi, G. Moon, J. Park, and K. M. Lee, "Learning to estimate robust 3D human mesh from in-the-wild crowded scenes," in *CVPR*, 2022.

[155] G. Pavlakos, J. Malik, and A. Kanazawa, "Human mesh recovery from multiple shots," in *CVPR*, 2022, pp. 1485–1495.

[156] Y. Rong, Z. Liu, C. Li, K. Cao, and C. C. Loy, "Delving deep into hybrid annotations for 3D human recovery in the wild," in *ICCV*, 2019, pp. 5340–5348.

[157] S. K. Dwivedi, N. Athanasiou, M. Kocabas, and M. J. Black, "Learning to regress bodies from images using differentiable semantic rendering," in *ICCV*, 2021, pp. 11 250–11 259.

[158] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[159] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, X. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *TPAMI*, 2020.

[160] H. Cho, Y. Cho, J. Ahn, and J. Kim, "Implicit 3D human mesh recovery using consistency with pose and shape from unseen-view," in *CVPR*, 2023, pp. 21 148–21 158.

[161] S. Goel, G. Pavlakos, J. Rajasegaran, A. Kanazawa, and J. Malik, "Humans in 4D: Reconstructing and tracking humans with transformers," in *ICCV*, 2023.

[162] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[163] J. Cho, K. Youwang, and T.-H. Oh, "Cross-attention of disentangled modalities for 3D human mesh recovery with transformers," in *ECCV*. Springer, 2022, pp. 342–359.

[164] C. Zheng, X. Liu, G.-J. Qi, and C. Chen, "POTTER: Pooling attention transformer for efficient human mesh recovery," in *CVPR*, 2023, pp. 1611–1620.

[165] Z. Dou, Q. Wu, C. Lin, Z. Cao, Q. Wu, W. Wan, T. Komura, and W. Wang, "TORE: Token reduction for efficient human mesh recovery with transformer," in *ICCV*, 2023, pp. 15 143–15 155.

[166] J. Kim, M.-G. Gwon, H. Park, H. Kwon, G.-M. Um, and W. Kim, "Sampling is matter: Point-guided 3D human mesh reconstruction," in *CVPR*, 2023, pp. 12 880–12 889.

[167] Y. Yoshiyasu, "Deformable mesh transformer for 3D human mesh recovery," in *CVPR*, 2023, pp. 17 006–17 015.

[168] J. Li, Z. Yang, X. Wang, J. Ma, C. Zhou, and Y. Yang, "JOTR: 3D joint contrastive learning with transformers for occluded human mesh recovery," in *ICCV*, 2023, pp. 9110–9121.

[169] V. Choutas, L. Müller, C.-H. P. Huang, S. Tang, D. Tzionas, and M. J. Black, "Accurate 3D body shape regression using metric and semantic attributes," in *CVPR*, 2022, pp. 2718–2728.

[170] X. Ma, J. Su, C. Wang, W. Zhu, and Y. Wang, "3D human mesh estimation from virtual markers," in *CVPR*, 2023, pp. 534–543.

[171] T. Fan, K. V. Alwala, D. Xiang, W. Xu, T. Murphey, and M. Mukadam, "Revitalizing optimization for 3D human pose and shape estimation: A sparse constrained formulation," in *ICCV*, 2021.

[172] H. Zhang, Y. Tian, Y. Zhang, M. Li, L. An, Z. Sun, and Y. Liu, "PyMAF-X: Towards well-aligned full-body model regression from monocular images," *TPAMI*, 2023.

[173] Z. Wang, J. Yang, and C. Fowlkes, "The best of both worlds: combining model-based and nonparametric approaches for 3D human body estimation," in *CVPRW*, 2022, pp. 2318–2327.

[174] W. Jiang, N. Kolotouros, G. Pavlakos, X. Zhou, and K. Daniilidis, "Coherent reconstruction of multiple humans from a single image," in *CVPR*, 2020, pp. 5579–5588.

[175] N. Ugrinovic, A. Ruiz, A. Agudo, A. Sanfeliu, and F. Moreno-Noguer, "Body size and depth disambiguation in multi-person reconstruction from single images," in *3DV*. IEEE, 2021, pp. 53–63.

[176] M. Fieraru, M. Zanfir, T. Szente, E. Bazavan, V. Olaru, and C. Sminchisescu, "REMIPS: Physically consistent 3D reconstruction of multiple interacting people under weak supervision," *NeurIPS*, vol. 34, pp. 19 385–19 397, 2021.

[177] J. Cha, M. Saqlain, G. Kim, M. Shin, and S. Baek, "Multi-person 3D pose and shape estimation via inverse kinematics and refinement," in *ECCV*. Springer, 2022, pp. 660–677.

[178] R. Khirodkar, S. Tripathi, and K. Kitani, "Occluded human mesh recovery," in *CVPR*, 2022, pp. 1715–1725.

[179] A. Zanfir, E. Marinoiu, M. Zanfir, A.-I. Popa, and C. Sminchisescu, "Deep network for the integrated 3D sensing of multiple people in natural images," in *NeurIPS*, vol. 31, 2018, pp. 8410–8419.

[180] Y. Sun, Q. Bao, W. Liu, Y. Fu, M. J. Black, and T. Mei, "Monocular, one-stage, regression of multiple 3D people," in *ICCV*, 2021, pp. 11 179–11 188.

[181] Y. Sun, W. Liu, Q. Bao, Y. Fu, T. Mei, and M. J. Black, "Putting people in their place: Monocular regression of 3D people in depth," in *CVPR*, 2021.

[182] Z. Qiu, Q. Yang, J. Wang, H. Feng, J. Han, E. Ding, C. Xu, D. Fu, and J. Wang, "PSVT: End-to-end multi-person 3D pose and shape estimation with progressive video transformers," in *CVPR*, 2023, pp. 21 254–21 263.

[183] A. Arnab, C. Doersch, and A. Zisserman, "Exploiting temporal context for 3D human pose estimation in the wild," in *CVPR*, 2019, pp. 3395–3404.

[184] G.-H. Lee and S.-W. Lee, "Uncertainty-aware human mesh recovery from video by learning part-based 3D dynamics," in *ICCV*, 2021, pp. 12 375–12 384.

[185] Z. Wan, Z. Li, M. Tian, J. Liu, S. Yi, and H. Li, "Encoder-decoder with multi-level attention for 3D human shape and pose estimation," in *ICCV*, 2021, pp. 13 033–13 042.

[186] Y. Yuan, S.-E. Wei, T. Simon, K. Kitani, and J. Saragih, "SimPoE: Simulated character control for 3D human pose estimation," in *CVPR*, 2021, pp. 7159–7169.

[187] W.-L. Wei, J.-C. Lin, T.-L. Liu, and H.-Y. M. Liao, "Capturing humans in motion: temporal-attentive 3D human pose and shape estimation from monocular video," in *CVPR*, 2022, pp. 13 211–13 220.

[188] J. Rajasegaran, G. Pavlakos, A. Kanazawa, and J. Malik, "Tracking people with 3D representations," in *NeurIPS*, 2021.

[189] Y. Yuan, U. Iqbal, P. Molchanov, K. Kitani, and J. Kautz, "GLAMR: Global occlusion-aware human mesh recovery with dynamic cameras," in *CVPR*, 2022, pp. 11 038–11 049.

[190] L. Sigal, A. Balan, and M. Black, "Combined discriminative and generative articulated pose and non-rigid shape estimation," *NeurIPS*, vol. 20, pp. 1337–1344, 2007.

[191] M. Hassan, V. Choutas, D. Tzionas, and M. J. Black, "Resolving 3D human pose ambiguities with 3D scene constraints," in *ICCV*, 2019, pp. 2282–2292.

[192] M. Shi, K. Aberman, A. Aristidou, T. Komura, D. Lischinski, D. Cohen-Or, and B. Chen, "MotioNet: 3D human motion reconstruction from monocular video with skeleton consistency," *TOG*, vol. 40, no. 1, pp. 1–15, 2020.

[193] S. Zhang, Y. Zhang, F. Bogo, M. Pollefeys, and S. Tang, "Learning motion priors for 2D human body capture in 3D scenes," in *ICCV*, 2021, pp. 11343–11353.

[194] L. Müller, A. A. Osman, S. Tang, C.-H. P. Huang, and M. J. Black, "On self-contact and human pose," in *CVPR*, 2021, pp. 9990–9999.

[195] D. Rempe, L. J. Guibas, A. Hertzmann, B. Russell, R. Villegas, and J. Yang, "Contact and human dynamics from monocular video," in *ECCV*. Springer, 2020, pp. 71–87.

[196] D. Rempe, T. Birdal, A. Hertzmann, J. Yang, S. Sridhar, and L. J. Guibas, "HuMoR: 3D human motion model for robust pose estimation," in *ICCV*, 2021.

[197] I. Akhter and M. J. Black, "Pose-conditioned joint angle limits for 3D human pose reconstruction," in *CVPR*, 2015, pp. 1446–1455.

[198] J. Zhang, D. Yu, J. H. Liew, X. Nie, and J. Feng, "Body meshes as points," in *CVPR*, 2021, pp. 546–556.

[199] S. Baek, K. I. Kim, and T.-K. Kim, "Pushing the envelope for RGB-based dense 3D hand pose estimation via neural rendering," in *CVPR*, 2019, pp. 1067–1076.

[200] A. Boukhayma, R. d. Bem, and P. H. Torr, "3D hand shape and pose from images in the wild," in *CVPR*, 2019, pp. 10843–10852.

[201] S. Hampali, M. Rad, M. Oberweger, and V. Lepetit, "HOnnotate: A method for 3D annotation of hand and object poses," in *CVPR*, 2020, pp. 3193–3203.

[202] Y. Hasson, G. Varol, D. Tzionas, I. Kalevatykh, M. J. Black, I. Laptev, and C. Schmid, "Learning joint reconstruction of hands and manipulated objects," in *CVPR*, 2019, pp. 11807–11816.

[203] U. Iqbal, P. Molchanov, T. Breuel, J. Gall, and J. Kautz, "Hand pose estimation via latent 2.5D heatmap regression," in *ECCV*, 2018, pp. 125–143.

[204] D. Kulon, H. Wang, R. A. Güler, M. M. Bronstein, and S. Zafeiriou, "Single image 3D hand reconstruction with mesh convolutions," in *BMVC*, 2019.

[205] F. Mueller, F. Bernard, O. Sotnychenko, D. Mehta, S. Sridhar, D. Casas, and C. Theobalt, "GANerated hands for real-time 3D hand tracking from monocular RGB," in *CVPR*, 2018, pp. 49–59.

[206] B. Tekin, F. Bogo, and M. Pollefeys, "H+O: Unified egocentric recognition of 3D hand-object poses and interactions," in *CVPR*, 2019, pp. 4511–4520.

[207] C. Zimmermann and T. Brox, "Learning to estimate 3D hand pose from single RGB images," in *ICCV*, 2017, pp. 4913–4921.

[208] X. Zhang, Q. Li, H. Mo, W. Zhang, and W. Zheng, "End-to-end hand mesh recovery from a monocular RGB image," in *ICCV*, 2019, pp. 2354–2364.

[209] L. Ge, Z. Ren, Y. Li, Z. Xue, Y. Wang, J. Cai, and J. Yuan, "3D hand shape and pose estimation from a single RGB image," in *CVPR*, 2019, pp. 10833–10842.

[210] D. Kulon, R. A. Güler, I. Kokkinos, M. M. Bronstein, and S. Zafeiriou, "Weakly-supervised mesh-convolutional hand reconstruction in the wild," in *CVPR*, 2020, pp. 4989–4999.

[211] J. Park, Y. Oh, G. Moon, H. Choi, and K. M. Lee, "HandOccNet: Occlusion-robust 3D hand mesh estimation network," in *CVPR*, 2022, pp. 1496–1505.

[212] G. Moon, S.-I. Yu, H. Wen, T. Shiratori, and K. M. Lee, "InterHand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image," in *ECCV*. Springer, 2020, pp. 548–564.

[213] J. Wang, F. Mueller, F. Bernard, S. Sorli, O. Sotnychenko, N. Qian, M. A. Otaduy, D. Casas, and C. Theobalt, "RGB2Hands: real-time tracking of 3D hand interactions from monocular RGB video," *ACM TOG*, vol. 39, no. 6, pp. 1–16, 2020.

[214] B. Zhang, Y. Wang, X. Deng, Y. Zhang, P. Tan, C. Ma, and H. Wang, "Interacting two-hand 3D pose and shape reconstruction from single color image," in *ICCV*, 2021, pp. 11354–11363.

[215] M. Li, H. An, H. Zhang, L. Wu, F. Chen, T. Yu, and Y. Liu, "Interacting attention graph for single image two-hand reconstruction," in *CVPR*, 2022.

[216] C. Wang, F. Zhu, and S. Wen, "MeMaHand: Exploiting mesh-mano interaction for single image two-hand reconstruction," in *CVPR*, 2023, pp. 564–573.

[217] J. Lee, M. Sung, H. Choi, and T.-K. Kim, "Im2Hands: Learning attentive implicit representation of interacting two-hand shapes," in *CVPR*, 2023, pp. 21169–21178.

[218] Z. Yu, S. Huang, C. Fang, T. P. Breckon, and J. Wang, "ACR: Attention collaboration-based regressor for arbitrary two-hand reconstruction," in *CVPR*, 2023, pp. 12955–12964.

[219] G. Moon, "Bringing inputs to shared domains for 3D interacting hands recovery in the wild," in *CVPR*, 2023, pp. 17028–17037.

[220] T. Chatzis, A. Stergioulas, D. Konstantinidis, K. Dimitropoulos, and P. Daras, "A comprehensive study on deep learning-based 3D hand pose estimation methods," *Applied Sciences*, vol. 10, no. 19, p. 6850, 2020.

[221] L. Huang, B. Zhang, Z. Guo, Y. Xiao, Z. Cao, and J. Yuan, "Survey on depth and RGB image-based 3D hand shape and pose estimation," *Virtual Reality & Intelligent Hardware*, vol. 3, no. 3, pp. 207–234, 2021.

[222] O. Aldrian and W. A. Smith, "Inverse rendering of faces with a 3D morphable model," *TPAMI*, vol. 35, no. 5, pp. 1080–1093, 2013.

[223] T. Vetter and V. Blanz, "Estimating coloured 3D face models from single images: An example based approach," in *ECCV*, 1998, pp. 499–513.

[224] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2Face: Real-time face capture and reenactment of RGB videos," in *CVPR*, 2016, pp. 2387–2395.

[225] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, "Joint 3D face reconstruction and dense alignment with position map regression network," in *ECCV*, 2018, pp. 557–574.

[226] A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos, "Large pose 3D face reconstruction from a single image via direct volumetric CNN regression," in *ICCV*, 2017, pp. 1031–1039.

[227] S. Sanyal, T. Bolkart, H. Feng, and M. J. Black, "Learning to regress 3D face shape and expression from an image without 3D supervision," in *CVPR*, 2019, pp. 7763–7772.

[228] A. Tewari, M. Zollhöfer, P. Garrido, F. Bernard, H. Kim, P. Pérez, and C. Theobalt, "Self-supervised multi-level face model learning for monocular reconstruction at over 250 Hz," in *CVPR*, 2018, pp. 2549–2559.

[229] A. Tewari, M. Zollhöfer, H. Kim, P. Garrido, F. Bernard, P. Perez, and C. Theobalt, "MoFA: model-based deep convolutional face autoencoder for unsupervised monocular reconstruction," in *ICCV*, 2017, pp. 3735–3744.

[230] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong, "Accurate 3D face reconstruction with weakly-supervised learning: From single image to image set," in *CVPRW*, 2019, pp. 285–295.

[231] L. Tran, F. Liu, and X. Liu, "Towards high-fidelity nonlinear 3D face morphable model," in *CVPR*, 2019, pp. 1126–1135.

[232] Y. Feng, H. Feng, M. J. Black, and T. Bolkart, "Learning an animatable detailed 3D face model from in-the-wild images," *TOG*, vol. 40, no. 4, pp. 88:1–88:13, 2021.

[233] L. Wang, Z. Chen, T. Yu, C. Ma, L. Li, and Y. Liu, "FaceVerse: a fine-grained and detail-controllable 3D face morphable model from a hybrid dataset," in *CVPR*, 2022.

[234] W. Zielonka, T. Bolkart, and J. Thies, "Towards metrical reconstruction of human faces," in *ECCV*. Springer, 2022, pp. 250–269.

[235] M. M. Loper and M. J. Black, "OpenDR: An approximate differentiable renderer," in *ECCV*, 2014, pp. 154–169.

[236] N. Ravi, J. Reizenstein, D. Novotny, T. Gordon, W.-Y. Lo, J. Johnson, and G. Gkioxari, "Accelerating 3D deep learning with PyTorch3D," *arXiv preprint arXiv:2007.08501*, 2020.

[237] K. Genova, F. Cole, A. Maschinot, A. Sarna, D. Vlasic, and W. T. Freeman, "Unsupervised training for 3D morphable model regression," in *CVPR*, 2018, pp. 8377–8386.

[238] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *FG*, 2018, pp. 67–74.

[239] B. Egger, W. A. P. Smith, A. Tewari, S. Wuhrer, M. Zollhoefer, T. Beeler, F. Bernard, T. Bolkart, A. Kortylewski, S. Romdhani, C. Theobalt, V. Blanz, and T. Vetter, "3D morphable face models - past, present and future," *TOG*, vol. 39, no. 5, pp. 157:1–157:38, 2020.

[240] H. Yi, H. Liang, Y. Liu, Q. Cao, Y. Wen, T. Bolkart, D. Tao, and M. J. Black, "Generating holistic 3D human motion from speech," in *CVPR*, 2023, pp. 469–480.

[241] N. Zioulis and J. F. O'Brien, "KBody: Towards general, robust, and aligned monocular whole-body estimation," in *CVPRW*, 2023, pp. 6214–6224.

[242] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *CVPR*, 2017, pp. 4645–4653.

[243] Y. Rong, T. Shiratori, and H. Joo, "FrankMocap: A monocular 3D whole-body pose estimation system via regression and integration," in *ICCVW*, 2021.

[244] J. Li, S. Bian, C. Xu, Z. Chen, L. Yang, and C. Lu, "HybrIK-X: Hybrid analytical-neural inverse kinematics for whole-body mesh recovery," *arXiv preprint arXiv:2304.05690*, 2023.

[245] Y. Sun, T. Huang, Q. Bao, W. Liu, G. Wenpeng, and Y. Fu, "Learning monocular mesh recovery of multiple body parts via synthesis," in *ICASSP*, 2022.

[246] J. Lin, A. Zeng, H. Wang, L. Zhang, and Y. Li, "One-stage 3D whole-body mesh recovery with component aware transformer," in *CVPR*, 2023, pp. 21 159–21 168.

[247] Z. Cai, W. Yin, A. Zeng, C. Wei, Q. Sun, Y. Wang, H. E. Pang, H. Mei, M. Zhang, L. Zhang, C. C. Loy, L. Yang, and Z. Liu, "SMPLer-X: Scaling up expressive human pose and shape estimation," in *NeurIPS Datasets and Benchmarks*, 2023.

[248] H. E. Pang, Z. Cai, L. Yang, T. Qingyi, W. Zhonghua, T. Zhang, and Z. Liu, "Towards robust and expressive whole-body human pose and shape estimation," *NeurIPS*, 2023.

[249] M.-P. Forte, P. Kulits, C.-H. P. Huang, V. Choutas, D. Tzionas, K. J. Kuchenbecker, and M. J. Black, "Reconstructing signing avatars from video using linguistic priors," in *CVPR*, 2023, pp. 12 791–12 801.

[250] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[251] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *ICCV*, 2017, pp. 2961–2969.

[252] H. Wen, J. Huang, H. Cui, H. Lin, Y.-K. Lai, L. Fang, and K. Li, "Crowd3D: Towards hundreds of people reconstruction from a single image," in *CVPR*, 2023, pp. 8937–8946.

[253] B. Zhang, K. Ma, S. Wu, and Z. Yuan, "Two-stage co-segmentation network based on discriminative representation for recovering human mesh from videos," in *CVPR*, 2023, pp. 5662–5670.

[254] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *CVPR*, 2018, pp. 7794–7803.

[255] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[256] X. Shen, Z. Yang, X. Wang, J. Ma, C. Zhou, and Y. Yang, "Global-to-local modeling for video-based 3D human pose and shape estimation," in *CVPR*, 2023, pp. 8887–8896.

[257] H. Cho, J. Ahn, Y. Cho, and J. Kim, "Video inference for human mesh recovery with vision transformer," in *FG*, 2023, pp. 1–6.

[258] S. Guan, J. Xu, Y. Wang, B. Ni, and X. Yang, "Bilevel online adaptation for out-of-domain human mesh reconstruction," in *CVPR*, 2021, pp. 10 472–10 481.

[259] H. Nam, D. S. Jung, Y. Oh, and K. M. Lee, "Cyclic test-time adaptation on monocular video for 3D human mesh reconstruction," in *ICCV*, 2023, pp. 14 829–14 839.

[260] S. Tripathi, S. Ranade, A. Tyagi, and A. Agrawal, "PoseNet3D: Learning temporally consistent 3D human pose via knowledge distillation," in *3DV*. IEEE, 2020, pp. 311–321.

[261] V. Ye, G. Pavlakos, J. Malik, and A. Kanazawa, "Decoupling human and camera motion from videos in the wild," in *CVPR*, 2023, pp. 21 222–21 232.

[262] Y. Sun, Q. Bao, W. Liu, T. Mei, and M. J. Black, "TRACE: 5D temporal regression of avatars with dynamic cameras in 3D environments," in *CVPR*, 2023, pp. 8856–8866.

[263] J. Li, S. Bian, C. Xu, G. Liu, G. Yu, and C. Lu, "D &D: Learning human dynamics from dynamic camera," in *ECCV*. Springer, 2022, pp. 479–496.

[264] Z. Weng and S. Yeung, "Holistic 3D human and scene mesh estimation from single view images," in *CVPR*, 2021, pp. 334–343.

[265] J. Y. Zhang, S. Pepose, H. Joo, D. Ramanan, J. Malik, and A. Kanazawa, "Perceiving 3D human-object spatial arrangements from a single image in the wild," in *ECCV*. Springer, 2020, pp. 34–51.

[266] X. Xie, B. L. Bhatnagar, and G. Pons-Moll, "CHORE: Contact, human and object reconstruction from a single RGB image," in *ECCV*. Springer, 2022, pp. 125–145.

[267] H. Yi, C.-H. P. Huang, D. Tzionas, M. Kocabas, M. Hassan, S. Tang, J. Thies, and M. J. Black, "Human-aware object placement for visual environment reconstruction," in *CVPR*, 2022, pp. 3959–3970.

[268] Z. Luo, S. Iwase, Y. Yuan, and K. M. Kitani, "Embodied scene-aware human pose estimation," in *NeurIPS*, 2022.

[269] Z. Shen, Z. Cen, S. Peng, Q. Shuai, H. Bao, and X. Zhou, "Learning human mesh recovery in 3D scenes," in *CVPR*, 2023, pp. 17 038–17 047.

[270] I. Kissos, L. Fritz, M. Goldman, O. Meir, E. Oks, and M. Kliger, "Beyond weak perspective for monocular 3D human pose estimation," in *ECCV*. Springer, 2020, pp. 541–554.

[271] H. Cho, Y. Cho, J. Yu, and J. Kim, "Camera distortion-aware 3D human pose estimation in video with optimization-based meta-learning," in *ICCV*, 2021, pp. 11 169–11 178.

[272] W. Wang, Y. Ge, H. Mei, Z. Cai, Q. Sun, Y. Wang, C. Shen, L. Yang, and T. Komura, "Zolly: Zoom focal length correctly for perspective-distorted human mesh reconstruction," in *ICCV*, 2023, pp. 3925–3935.

[273] E. Gärtner, M. Andriluka, E. Coumans, and C. Sminchisescu, "Differentiable dynamics for articulated 3D human motion reconstruction," in *CVPR*, 2022, pp. 13 190–13 200.

[274] E. Gärtner, M. Andriluka, H. Xu, and C. Sminchisescu, "Trajectory optimization for physics-based reconstruction of 3D human pose from monocular video," in *CVPR*, 2022, pp. 13 106–13 115.

[275] K. Xie, T. Wang, U. Iqbal, Y. Guo, S. Fidler, and F. Shkurti, "Physics-based human motion estimation and synthesis from videos," in *ICCV*, 2021, pp. 11 532–11 541.

[276] B. Huang, L. Pan, Y. Yang, J. Ju, and Y. Wang, "Neural MoCon: Neural motion control for physically plausible human motion capture," in *CVPR*, 2022, pp. 6417–6426.

[277] S. Tripathi, L. Müller, C.-H. P. Huang, O. Taheri, M. J. Black, and D. Tzionas, "3D human pose estimation via intuitive physics," in *CVPR*, 2023, pp. 4713–4725.

[278] M. Fieraru, M. Zanfir, E. Oneata, A.-I. Popa, V. Olaru, and C. Sminchisescu, "Learning complex 3D human self-contact," in *AAAI*, 2021.

[279] M. Teschner, S. Kimmerle, B. Heidelberger, G. Zachmann, L. Raghupathi, A. Fuhrmann, M.-P. Cani, F. Faure, N. Magnenat-Thalmann, W. Strasser *et al.*, "Collision detection for deformable objects," in *CGF*, vol. 24, no. 1. Wiley Online Library, 2005, pp. 61–81.

[280] Y. Zhang, H. Zhang, L. Hu, J. Zhang, H. Yi, S. Zhang, and Y. Liu, "ProxyCap: Real-time monocular full-body capture in world space via human-centric proxy-to-motion learning," *arXiv preprint arXiv:2307.01200*, 2023.

[281] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *NeurIPS*, vol. 27, 2014.

[282] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *ICLR*, 2014.

[283] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *International conference on machine learning*. PMLR, 2015, pp. 1530–1538.

[284] Y. Rong, Z. Liu, and C. C. Loy, "Chasing the tail in monocular 3D human reconstruction with prototype memory," *TIP*, vol. 31, pp. 2907–2919, 2022.

[285] A. Davydov, A. Remizova, V. Constantin, S. Honari, M. Salzmann, and P. Fua, "Adversarial parametric pose prior," in *CVPR*, 2022, pp. 10 997–11 005.

[286] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real NVP," in *ICLR*, 2017.

[287] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *NeurIPS*, vol. 33, pp. 6840–6851, 2020.

[288] L. G. Foo, J. Gong, H. Rahmani, and J. Liu, "Distribution-aligned diffusion for human mesh recovery," in *ICCV*, 2023, pp. 9221–9232.

[289] H. Cho and J. Kim, "Generative approach for probabilistic human mesh recovery using diffusion models," in *ICCV Workshops*, 2023, pp. 4183–4188.

[290] S. Zhang, Q. Ma, Y. Zhang, S. Aliakbarian, D. Cosker, and S. Tang, "Probabilistic human mesh recovery in 3d scenes from egocentric views," in *ICCV*, 2023, pp. 7989–8000.

[291] M. Kaufmann, E. Aksan, J. Song, F. Pece, R. Ziegler, and O. Hilliges, "Convolutional autoencoders for human motion infilling," in *3DV*. IEEE, 2020, pp. 918–927.

[292] Y. He, A. Pang, X. Chen, H. Liang, M. Wu, Y. Ma, and L. Xu, "ChallenCap: Monocular 3D capture of challenging human performances using multi-modal references," in *CVPR*, 2021, pp. 11 400–11 411.

[293] J. Li, R. Villegas, D. Ceylan, J. Yang, Z. Kuang, H. Li, and Y. Zhao, "Task-generic hierarchical human motion prior using VAEs," in *3DV*. IEEE, 2021, pp. 771–781.

[294] J. Xu, M. Wang, J. Gong, W. Liu, C. Qian, Y. Xie, and L. Ma, "Exploring versatile prior for human motion via motion frequency guidance," in *3DV*. IEEE, 2021, pp. 606–616.

[295] Z. Cai, M. Zhang, J. Ren, C. Wei, D. Ren, J. Li, Z. Lin, H. Zhao, S. Yi, L. Yang *et al.*, "Playing for 3D human recovery," *arXiv preprint arXiv:2110.07588*, 2021.

[296] P. Patel, C.-H. P. Huang, J. Tesch, D. T. Hoffmann, S. Tripathi, and M. J. Black, "AGORA: Avatars in geography optimized for regression analysis," in *CVPR*, 2021, pp. 13 468–13 478.

[297] T. Yu, Z. Zheng, K. Guo, P. Liu, Q. Dai, and Y. Liu, "Function4D: Real-time human volumetric capture from very sparse consumer RGBD sensors," in *CVPR*, 2021, pp. 5746–5756.

[298] E. G. Bazavan, A. Zanfir, M. Zanfir, W. T. Freeman, R. Sukthankar, and C. Sminchisescu, "HSPACE: Synthetic parametric humans animated in complex environments," *arXiv preprint arXiv:2112.12867*, 2021.

[299] M. J. Black, P. Patel, J. Tesch, and J. Yang, "BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion," in *CVPR*, 2023, pp. 8726–8737.

[300] M. Trumble, A. Gilbert, C. Malleson, A. Hilton, and J. Collomosse, "Total Capture: 3D human pose estimation fusing video and inertial sensors," in *BMVC*, 2017, pp. 1–13.

[301] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh, "Panoptic studio: A massively multiview system for social motion capture," in *ICCV*, 2015, pp. 3334–3342.

[302] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt, "Monocular 3D human pose estimation in the wild using improved CNN supervision," in *3DV*. IEEE, 2017, pp. 506–516.

[303] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, S. Sridhar, G. Pons-Moll, and C. Theobalt, "Single-shot multi-person 3D pose estimation from monocular RGB," in *3DV*. IEEE, 2018, pp. 120–130.

[304] Z. Li, T. Dekel, F. Cole, R. Tucker, N. Snavely, C. Liu, and W. T. Freeman, "Learning the depths of moving people by watching frozen people," in *CVPR*, 2019, pp. 4521–4530.

[305] V. Leroy, P. Weinzaepfel, R. Brégier, H. Combaluzier, and G. Rogez, "SMPLy benchmarking 3D human pose estimation in the wild," in *3DV*. IEEE, 2020, pp. 301–310.

[306] Q. Fang, Q. Shuai, J. Dong, H. Bao, and X. Zhou, "Reconstructing 3D human pose by watching humans in the mirror," in *CVPR*, 2021, pp. 12 814–12 823.

[307] Z. Yu, J. S. Yoon, I. K. Lee, P. Venkatesh, J. Park, J. Yu, and H. S. Park, "HUMBI: A large multiview dataset of human body expressions," in *CVPR*, 2020, pp. 2990–3000.

[308] zju3dv, "EasyMoCap - make human motion capture easier." GitHub, 2021. [Online]. Available: https://github.com/zju3dv/EasyMocap

[309] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *BMVC*, 2010, pp. 12.1–12.11.

[310] ——, "Learning effective human pose estimation from inaccurate annotation," in *CVPR*. IEEE, 2011, pp. 1465–1472.

[311] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *ECCV*. Springer, 2014, pp. 740–755.

[312] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *CVPR*, 2014, pp. 3686–3693.

[313] M. Andriluka, U. Iqbal, E. Insafutdinov, L. Pishchulin, A. Milan, J. Gall, and B. Schiele, "PoseTrack: A benchmark for human pose estimation and tracking," in *CVPR*, 2018, pp. 5167–5176.

[314] S.-H. Zhang, R. Li, X. Dong, P. Rosin, Z. Cai, X. Han, D. Yang, H. Huang, and S.-M. Hu, "Pose2Seg: Detection free human instance segmentation," in *CVPR*, 2019, pp. 889–898.

[315] G. Moon, H. Choi, and K. M. Lee, "NeuralAnnot: Neural annotator for 3D human mesh training sets," in *CVPRW*, 2022, pp. 2299–2307.

[316] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li, "PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization," in *ICCV*, 2019, pp. 2304–2314.

[317] Z. Yang, Z. Cai, H. Mei, S. Liu, Z. Chen, W. Xiao, Y. Wei, Z. Qing, C. Wei, B. Dai, W. Wu, C. Qian, D. Lin, Z. Liu, and L. Yang, "SynBody: Synthetic dataset with layered human models for 3D human perception and modeling," in *ICCV*, 2023, pp. 20 282– 20 292.

[318] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3D face model for pose and illumination invariant face recognition," in *AVSS*. Ieee, 2009, pp. 296–301.

[319] B. L. Bhatnagar, X. Xie, I. A. Petrov, C. Sminchisescu, C. Theobalt, and G. Pons-Moll, "BEHAVE: Dataset and method for tracking human object interactions," in *CVPR*, 2022, pp. 15 935–15 946.

[320] O. Taheri, N. Ghorbani, M. J. Black, and D. Tzionas, "GRAB: A dataset of whole-body human grasping of objects," in *ECCV*. Springer, 2020, pp. 581–600.

[321] C.-H. P. Huang, H. Yi, M. Höschle, M. Safroshkin, T. Alexiadis, S. Polikovsky, D. Scharstein, and M. J. Black, "Capturing and inferring dense full-body human-scene contact," in *CVPR*, 2022, pp. 13 274–13 285.

[322] Y. Dai, Y. Lin, X. Lin, C. Wen, L. Xu, H. Yi, S. Shen, Y. Ma, and C. Wang, "SLOPER4D: A scene-aware dataset for global 4D human pose estimation in urban environments," in *CVPR*, 2023, pp. 682–692.

[323] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *CVPR*, 2014, pp. 1725–1732.

[324] J. Lin, A. Zeng, S. Lu, Y. Cai, R. Zhang, H. Wang, and L. Zhang, "Motion-X: A large-scale 3D expressive whole-body human motion dataset," *NeurIPS*, 2023.

[325] A. Zeng, L. Yang, X. Ju, J. Li, J. Wang, and Q. Xu, "SmoothNet: A plug-and-play network for refining human poses in videos," in *ECCV*. Springer, 2022, pp. 625–642.

[326] S. Zhang, Y. Zhang, Q. Ma, M. J. Black, and S. Tang, "PLACE: Proximity learning of articulation and contact in 3D environments," in *3DV*. IEEE, 2020, pp. 642–651.

[327] Y. Zhang, M. Hassan, H. Neumann, M. J. Black, and S. Tang, "Generating 3D people in scenes without people," in *CVPR*, 2020, pp. 6194–6204.

[328] M. Liu, D. Yang, Y. Zhang, Z. Cui, J. M. Rehg, and S. Tang, "4D human body capture from egocentric video via 3D scene grounding," in *3DV*. IEEE, 2021, pp. 930–939.

[329] C.-H. Chen, A. Tyagi, A. Agrawal, D. Drover, S. Stojanov, and J. M. Rehg, "Unsupervised 3D pose estimation with geometric self-supervision," in *CVPR*, 2019, pp. 5714–5724.

[330] H. Rhodin, M. Salzmann, and P. Fua, "Unsupervised geometry-aware representation for 3D human pose estimation," in *ECCV*, 2018, pp. 750–767.

[331] J. N. Kundu, M. Rakesh, V. Jampani, R. M. Venkatesh, and R. Venkatesh Babu, "Appearance consensus driven self-supervised human mesh recovery," in *ECCV*. Springer, 2020, pp. 794–812.

[332] W.-S. Zheng, S. Gong, and T. Xiang, "Associating groups of people," in *BMVC*, vol. 2, no. 6, 2009, pp. 1–11.

[333] G. Lisanti, N. Martinel, A. Del Bimbo, and G. Luca Foresti, "Group re-identification via unsupervised transfer of sparse features encoding," in *ICCV*, 2017, pp. 2449–2458.

[334] W. Xu, A. Chatterjee, M. Zollhöfer, H. Rhodin, D. Mehta, H.-P. Seidel, and C. Theobalt, "MonoPerfCap: Human performance capture from monocular video," *TOG*, vol. 37, no. 2, pp. 1–15, 2018.

[335] Q. Ma, J. Yang, A. Ranjan, S. Pujades, G. Pons-Moll, S. Tang, and M. J. Black, "Learning to dress 3D people in generative clothing," in *CVPR*, 2020, pp. 6468–6477.

[336] H. Zhu, X. Zuo, S. Wang, X. Cao, and R. Yang, "Detailed human shape estimation from a single image by hierarchical mesh deformation," in *CVPR*, 2019, pp. 4491–4500.

[337] Q. Ma, S. Saito, J. Yang, S. Tang, and M. J. Black, "SCALE: Modeling clothed humans with a surface codec of articulated local elements," in *CVPR*, 2021, pp. 16 082–16 093.

[338] S. Lin, H. Zhang, Z. Zheng, R. Shao, and Y. Liu, "Learning implicit templates for point-based clothed human modeling," in *ECCV*. Springer, 2022, pp. 210–228.

[339] H. Zhang, S. Lin, R. Shao, Y. Zhang, Z. Zheng, H. Huang, Y. Guo, and Y. Liu, "CloSET: Modeling clothed humans on continuous surface with explicit template decomposition," in *CVPR*, 2023, pp. 501–511.

[340] Z. Li, T. Yu, C. Pan, Z. Zheng, and Y. Liu, "Robust 3D self-portraits in seconds," in *CVPR*, 2020, pp. 1344–1353.

[341] T. Alldieck, M. Zanfir, and C. Sminchisescu, "Photorealistic monocular 3D reconstruction of humans wearing clothing," in *CVPR*, 2022, pp. 1506–1515.

[342] B. L. Bhatnagar, C. Sminchisescu, C. Theobalt, and G. Pons-Moll, "Combining implicit function learning and parametric models for 3D human reconstruction," in *ECCV*. Springer, 2020, pp. 311–329.

[343] R. Shao, H. Zhang, H. Zhang, M. Chen, Y. Cao, T. Yu, and Y. Liu, "DoubleField: Bridging the neural surface and radiance fields for high-fidelity human reconstruction and rendering," in *CVPR*, 2022.

[344] Y. Feng, J. Yang, M. Pollefeys, M. J. Black, and T. Bolkart, "SCARF: Capturing and animation of body and clothing from monocular video," in *SIGGRAPH Asia Conference Papers*, 2022, p. 9.

[345] G. Moon, H. Nam, T. Shiratori, and K. M. Lee, "3D clothed human reconstruction in the wild," in *ECCV*. Springer, 2022, pp. 184–200.