Parametric Human Body Representation in Human Pose Estimation & Action Recognition: A Survey

Martin Purgat
Practicum 2024W
TU Wien
e12333396@student.tuwien.ac.at

Abstract

Human body modeling has long been a central topic in computer vision, graphics, and simulation and has experienced a significant advancements in recent years. Application areas range from film and video games to virtual avatars and human motion capture. Parametric models such as SMPL in particular have become fundamental components for pose estimation and animation. Implicit models, by contrast, fill gaps in high-detail, flexible representations and support advanced reconstructions from limited image data. Research increasingly combines both paradigms: a parametric core provides plausibility and is easy to animate, while implicit components capture detailed geometry or clothing. Open challenges, such as modeling hair, realistic soft tissue, or physically plausible dynamics, continue to drive the field. The paper further explores the current state of the art in parametric human body modeling, their use in human pose estimation, action and activity recognition. Besides the human pose estimation, where the SMPL-based models achieve state of the art results, the models are being increasingly employed in action recognition and broader activity understanding and generating synthetic data.

1. Introduction

Human body modeling and 3D pose estimation have long been central topics in both computer vision and graphics, enabling analysis of human motion and appearance across a broad spectrum of applications. Early work typically employed geometric or polygonal models, yet the need for standardized, data-driven solutions consequently led to the development of parametric approaches. Such models can represent variability in body shape and pose typically by linear combination of learned so-called blend-shape functions. Seminal works include SCAPE (9) and SMPL (101), both of which pioneered data-driven decompositions that factor identity-dependent shapes from pose-induced deformations. By regressing a small number of low-dimensional parameters, these parametric approaches can generate high-fidelity human meshes suitable for motion capture, animation, biomechanics, and beyond.

A parallel line of research focuses on human pose estimation, extracting the body's configuration from images or videos, ranging from the classic optimization-based SMPLify (19) to more recent, regression-oriented techniques (71; 80). These methods reconstruct a coherent 3D human mesh by minimizing reprojection errors of 2D keypoints or silhouettes while enforcing prior knowledge of plausible body shapes. Nowadays, algorithms are able to operate on single or multi-view videos, incorporate probabilistic and adversarial frameworks, and model full-body articulations (body, hands, and face) in a unified parameter space.

Beyond academia, the impact of these methods spans virtual reality (VR) and telepresence systems, where holographic and immersive communication is enabled (29; 123), gaming and film industries, which frequently rely on realistic avatars and motion capture pipelines (14), robotics to understand human motion in dynamic environments (42), and healthcare, for instance in patient tracking and movement analysis (59).

Importantly, the parametric human body models are increasingly used to support action recognition and broader activity understanding. In particular, the pose and shape priors encoded by models like SMPL can serve as stable, viewpoint-invariant descriptors for classifying human behaviors (138). By representing each individual's motion in a shared 3D space, tasks such as group activity recognition, multi-person tracking, and long-term motion forecasting all benefit from improved geometric consistency. Aditionally, the readiness of mesh representation allows for practical visualization and further analysis of interaction with the environment. Overall, the synergy among these three fields promises

a new level of realism and robustness in digital human analysis, positioning the 3D body models at the core of future interactive and intelligent systems.

Contributions of this review. The survey provides a structured overview of the main principles and recent developments in parametric human body modeling respectively in Section 4 and Section 5. We first introduce a basic taxonomy of human body models, including explicit and implicit models, and in Section 6 further discuss how parametric models serve as components in the state of the art human pose estimation. In Section 7, we turn to methods that leverage parametric representations for action recognition. Finally, in Section 8, we outline emerging research directions and practical challenges, emphasizing the need for larger datasets, real-time systems, and multi-modal integration to fully unlock the capabilities of 3D human body models in today's world.

2. Related Work

Over the past two decades, numerous surveys and reviews described the evolution of human body modeling and pose estimation from purely geometric or kinematic representations to modern, data-driven parametric methods. Early reviews, such as the one by Aggarwal and Cai (2), provide an overview of classic approaches that rely on manual feature extraction, part-based models, and low-level motion cues. Subsequent works refined this perspective by further segmenting body modeling into appearance, viewpoint, and spatial relations modules, as in Perez-Sala *et al.* (122), and by emphasizing part-parsing strategies for 2D pose inference (100).

Parametric Models and 3D Mesh Recovery. Among surveys focusing on parametric human body modeling, Cheng *et al.* (29) and Petkova *et al.* (123) provide particularly comprehensive accounts of how the introduction of statistical shape representations revolutionized the field. Their emphasis lies in data-driven methods that capture shape-and-pose variations through principal component analysis, such as SCAPE (9), or SMPL (101), not omitting the advancements in 3D scanning pipelines and mesh morphing techniques. Additionally, Hu (59) reviews how parametric human modeling can address subject-specific anatomy in biomechanics and safety simulations. More recently, Garcia-D'Urso *et al.* (48) broadened the discussion by contrasting classical parametric pipelines with newer neural implicit representations, focusing on the hybrid direction.

Advancements in 3D Pose Estimation. Alongside shape modeling, deep 3D pose estimation has matured into a robust research domain, as exemplified by the survey of Tian *et al.* (155) and Wang *et al.* (167). These works trace the evolution from optimization-based fitting of SMPL-like models (e.g., SMPLify (19)) to end-to-end regression methods (e.g., HMR (71), SPIN (80)), while contrasting single-person pipelines with multi-person or multi-view approaches. They also underscore the emergence of adversarial and temporal priors to enhance plausibility and reduce jitter, especially in monocular video settings.

Action Recognition and Synthetic Data Generation. The link between parametric modeling and action recognition has been discussed in surveys focusing on broader motion analysis (2) or purely skeleton-based methods (100). Recent work has shown that SMPL-based or SMPL-X-based reconstructions can mitigate viewpoint changes and self-occlusions in action recognition pipelines, and that generating synthetic datasets via parametric avatars can substantially augment training data.

Compared to these prior reviews, our work places a stronger emphasis on how parametric models are employed in action and activity recognition. While existing surveys do mention pose recovery pipelines, they typically provide either high-level coverage of linear modeling (29), or emphasize photorealistic reconstruction pipelines (123). Furthermore, our survey also covers more recent model refinements and identifies how these richer representations facilitate tasks beyond static 3D reconstruction, including data generation for training advanced recognition models.

3. Methodology

To conduct the review, we followed a structured methodology that combined elements of rapid review and scoping review spanning the following three areas:

1. Human body models, with an emphasis on parametric models,
2. Human pose estimation using parametric models,
3. The role of parametric human body models in action and activity recognition.

Each subreview followed the same protocol. The methodology was informed by the scoping review principles (157), involving steps to identify, screen, and extract data from the relevant literature.

3.1. Objectives and Review Questions

The objective was to aggregate and map the state of the art in parametric human body modeling, to summarize developments in human pose estimation using these

models, and to explore their application in action and activity recognition. These questions were addressed:

1. **Human Body Models:** What are the main classes of parametric body models, and what are their defining characteristics?

2. **Human Pose Estimation:** In what ways have parametric body models been employed for pose estimation, and how have these approaches evolved?

3. **Action and Activity Recognition:** How are parametric models contributing to action or activity recognition tasks, and what approaches have emerged?

3.2. Eligibility Criteria

All three subreviews applied the following inclusion criteria:

- **Population.** Any study or report involving human body models, pose estimation, or action/activity recognition contexts.

- **Intervention/Method.** Approaches that propose, refine, or apply parametric or hybrid (parametric–implicit) human body models for relevant tasks.

- **Outcomes.** Description of the model structure, algorithmic innovation, or empirical performance measures for pose estimation, action recognition, or related tasks.

- **Publication Type.** Peer-reviewed journal articles, conference proceedings, research reports, or doctoral theses.

- **Time Frame.** Published between 2000 and 2025 with a bias towards recent works and number of citations.

- **Language.** Limited to English.

Exclusion criteria consisted of:

- Non-human modeling (e.g., animal shape models) unrelated to parametric human representations,

- Publications addressing only 2D pose or methods not involving parametric 3D representations,

- Opinion pieces, editorials, or abstracts lacking sufficient methodological detail.

3.3. Literature Search Strategy

A systematic search was conducted across the following electronic databases: *IEEE Xplore*, *ACM Digital Library*, *Web of Science*, *arXiv*, *Google Scholar*.
The search was performed in December 2024 and February 2025 for human body models and pose estimation/action recognition, respectively. The strategy used a combination of controlled vocabulary (where applicable) and free-text terms. Key terms for parametric body modeling included:

```
``parametric AND (human OR
body) AND (model OR shape
OR mesh)'',    ``(SCAPE OR SMPL
OR SMPL-X) AND shape'',    ``3D
body representation AND blend
shapes AND modeling'',
```

while pose estimation keywords included:

```
``3D pose estimation AND
parametric model'',    ``human
mesh recovery AND SMPL'',
```

and activity recognition terms included:

```
``(action OR activity)
recognition AND (parametric
human model OR SMPL OR
SMPL-X)''.
```

Additional synonyms and variations were used to account for differences in field terminology. To capture gray literature, conference websites (e.g., CVPR, ICCV, ECCV, SIGGRAPH) were manually searched, and references of included studies were screened. The full search strings as well as full list of searched literature are available upon request.

3.4. Study Selection

All retrieved records were imported into a reference management software, and duplicates were removed.

3.5. Data Extraction

A standardized extraction form was used for each of the three subreviews.

- Publication details (title, authors, year, source)

- Description of the model or method (e.g., SMPL, SCAPE, hybrid parametric-implicit)

- Data sources (image datasets, motion capture systems, synthetic generation)

- Methodological approach (pose estimation pipeline, inference strategies, optimization frameworks)

- Performance metrics (pose accuracy, reconstruction error, action classification accuracy, etc.)
- Qualitative findings related to usability, limitations, or integration into additional tasks

Discrepancies were reconciled by re-checking source articles. The extracted data for each subreview were tabulated for further analysis. The information contained in the Tab. 3 was extracted using LLM Deepseek (33), by querying the model with a paper to fill in the table columns with the relevant information and providing proof of the information in the given paper.

3.6. Quality Assessment

Because the objective was to present a scoping overview of methods and applications, no formal risk-of-bias instrument was applied. Studies were not excluded based on design quality. However, studies lacking minimum methodological transparency (e.g., no technical detail, unclear evaluation protocol) were excluded during the full-text screening stage.

3.7. Data Synthesis and Analysis

The analysis was conducted separately for each of the three subreviews. Extracted information was aggregated in narrative summaries. For the quantitative results (e.g., pose estimation errors, action recognition accuracy), the reported values were compiled in comparative tables, segmented by approach or dataset. Wherever possible, trends in the use of parametric models, frequently combined methods, and typical evaluation protocols were noted. Cross-cutting themes were identified based on repeated patterns:

- Common parametric modeling backbones,
- Hybrid methods coupling parametric priors with implicit or volumetric representations,
- Integration of parametric models in end-to-end learning pipelines,
- Use of parametric shape parameters for action classification features.

3.8. Reporting

Results are presented as three separate sections (Sections 4, 6, and 7) corresponding to each subreview. Studies that addressed multiple review questions appear in more than one subreview. No meta-analysis was conducted since the identified literature featured heterogeneous outcomes and methodological designs. All included studies are cited, and descriptive tables are used to summarize core findings.

3.9. Limitations of the Method

The approach was adapted from rapid and scoping review principles. The chosen date restriction and language limit might exclude relevant but older or non-English works. The lack of a meta-analysis may limit quantitative synthesis. Nonetheless, the procedure was deemed sufficient to map the main trends, typical methodologies, and open research directions in parametric body modeling and its applications.

4. Human Body Models

The proposed high-level taxonomy naturally emerged over time as the field of human body modeling matured. The classification is based on the underlying representation of the human body, distinguishing between explicit and implicit models. Recent research often times combines the two paradigms to achieve high-fidelity reconstructions with detailed surface geometry and plausible articulation.

Parametric Models. Parametric models encode the human body as a low-dimensional set of shape and pose parameters. One representative instance is SCAPE (9), which learns separate components for body shape and pose-induced deformation by aligning a shared template to multiple scans. Another model, SMPL (101), adopts a linear blend skinning framework and introduces additive blend shapes for pose and shape. Variants of SMPL, such as SMPL-X (116), include additional components for hands and expressive faces. Further refinements incorporate localized deformation bases (112; 113) or biomechanical elements (75) for more consistent articulation. These models typically rely on principal component analysis or learned subspaces. Many works extend this concept to specialized settings, including garment-aware shape estimation (52; 110), body reshaping (13; 31), or large-scale face-and-head modeling (12).

Implicit Models. Implicit models represent the human body through functions in continuous three-dimensional space. These functions typically map a point to signed distance (6; 98) or occupancy (107; 121). They allow high-resolution reconstructions and flexible topology. Some methods learn volumetric fields directly from single images, for instance by predicting an occupancy grid (137). Others incorporate multi-part strategies (130) or single-view depth priors (149), while volume rendering approaches can further generate appearance. These models typically extract surfaces as level sets via marching cubes.

Hybrid Models. Hybrid models couple a parametric body representation with an implicit component.

They exploit parametric body priors and augment them with continuous functions for capturing surface details, clothes, and hair. Some approaches learn offsets around a parametric mesh (15; 195), while others refine normal maps and partial surface reconstructions using implicit functions (34; 172). Point-based methods can further combine parametric rigs with learned displacements on a continuous surface (96; 191). In each case, the parametric body imposes a global shape prior, and the implicit or volumetric layer accounts for local geometry. Additional examples include motion or skeleton-driven deformations (86; 114) and layered volume representations (175).

Other Approaches. Some other methods do not fit into the above categories. Direct mesh regression aims to infer vertex coordinates of a standard topology mesh from images (163). Peeled-depth layering estimates multiple depth layers per pixel (65), and point-based representations can capture loose garments without explicit templates (103). Sparse-coding approaches rely on dictionary learning for compressed shape representations (170), and part-based graphical models distribute parameters across articulated segments (203).

5. Parametric Body Shape Modeling

Parametric human body models enable the representation of body shape, pose, and surface geometry in a convenient mathematical form. These models rely on data obtained from 3D body scans or motion capture. They organize the scan data into a consistent reference (template mesh), then extract shape and pose through data-driven statistical or geometric methods. This section reviews a range of such models, compares their formulations, and highlights their common elements. It first covers key ideas in data-driven shape modeling, then explains how early approaches factor the body surface into shape and pose, and later reviews methods that refine these factors for large datasets, dynamic deformations, or further generalization.

5.1. Data Representation and Registration

Three-dimensional body scans are used as the input for data-driven modeling, where each scan is a triangulated surface of a subject in a particular pose. Therefore, a common mesh template $T = \{v_i \in \mathbb{R}^3 : i = 1, \ldots, n\}$ is used. For each subject's scan, a non-rigid deformation aligns T to the scan surface to produce T'. This step is known as *registration* and establishes a consistent vertex correspondence across subjects and poses. Once all scans share the same mesh topology, each scan can be vectorized as $\mathbf{x} \in \mathbb{R}^{3n}$, enabling direct application of dimensionality reduction or factorization methods.

Allen *et al.* (7) introduced such non-linear optimization to register a high-resolution template mesh to each scan via per-vertex affine transformations $\mathbf{A}_i \in \mathbb{R}^{4\times4}$. The global objective function,

$$E \;=\; \alpha\, E_{\text{data}} + \beta\, E_{\text{smooth}} + \gamma\, E_{\text{marker}},$$

balances data fidelity (E_{data}), measured by distances from deformed template vertices to the observed scan, with smoothness constraints (E_{smooth}) enforcing local regularity, and marker constraints (E_{marker}) matching anthropometric landmarks. First, the partial derivatives of E with respect to each $\mathbf{A}_i$ are determined, and an iterative nonlinear solver updates $\{\mathbf{A}_i\}$ until convergence. Additionally, multiple passes with different weight schedules (α, β, γ) are used to refine the final alignment. Once registration is complete, each scan is turned into a deformation $T' \in \mathbb{R}^{3n}$ of the same template.

5.2. Low-Dimensional Shape Spaces

Empirical observations indicate that human body shape variation lies in a low-dimensional space. Principal Component Analysis (PCA) is therefore often used as a simple and sufficient method to approximate such distribution of $\{\mathbf{x}_j\} \subset \mathbb{R}^{3n}$, where $\mathbf{x}_j$ is the vectorized registration of subject j. PCA can approximate the body shape distribution as:

$$\mathbf{x}_j \;\approx\; \overline{\mathbf{x}} \;+\; \sum_{k=1}^{K} p_{j,k}\, \mathbf{c}_k,$$

where $\overline{\mathbf{x}}$ is the mean shape, $\{\mathbf{c}_k\}_{k=1}^{K}$ are the first K principal directions, and $p_{j,k}$ are coefficients. Let $\sigma_1^2 \geq \sigma_2^2 \geq \ldots$ be the corresponding eigenvalues.

Many studies choose K such that 95–99% of the variance is captured, though K can also be tuned to a fixed dimensionality in practice (see Figure 2) . Formally, one interprets

$$\mathbf{x}_j \;\sim\; \mathcal{N}\Big(\overline{\mathbf{x}},\; \Sigma\Big), \qquad \Sigma = \mathbf{C}\,\Lambda\,\mathbf{C}^{\top},$$

where $\mathbf{C} = [\mathbf{c}_1 | \cdots | \mathbf{c}_K]$ and $\Lambda = \text{diag}(\sigma_1^2, \ldots, \sigma_K^2)$.

In practical applications, separate models for males and females can further improve the fidelity (7). Additionally, this approach allows also semantic editing, which involves manipulating specific principal directions that correspond to interpretable traits such as height or weight (180).

5.3. Separating Shape and Pose Contributions

In order to augment PCA-based static shape representations with pose variation, many body models introduce parametric functions that map joint angles to local

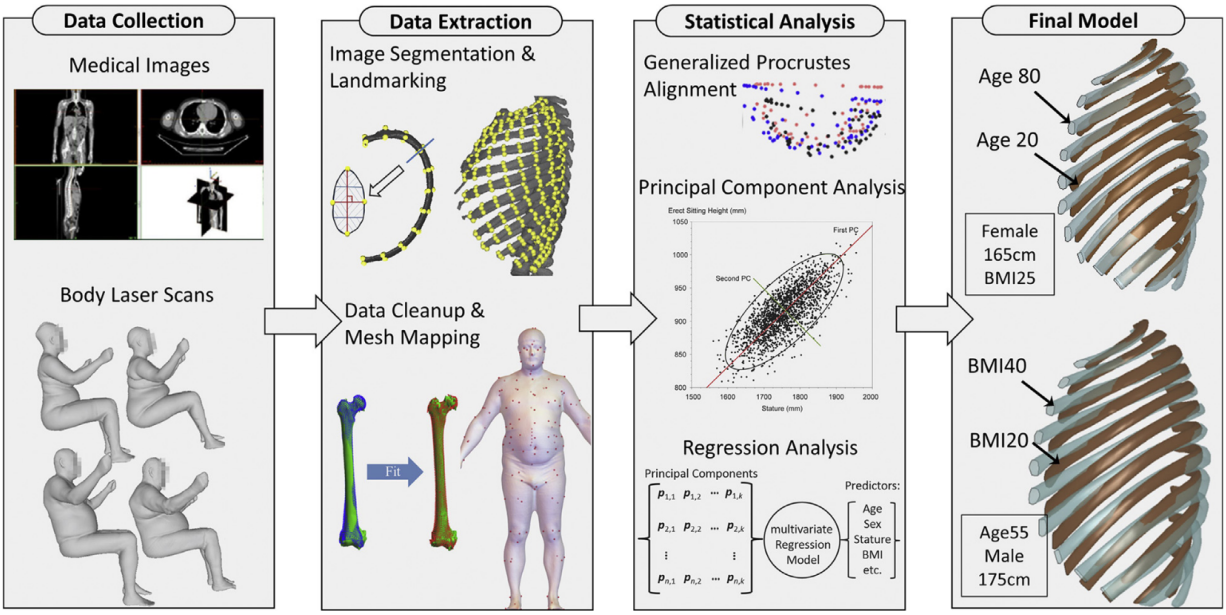

Figure 1. The figure (59) provides an Overview of the parametric human body modeling pipeline. The process starts with a 3D body scan, which is registered to a template mesh. The shape and pose parameters are then extracted using a statistical model, such as PCA.

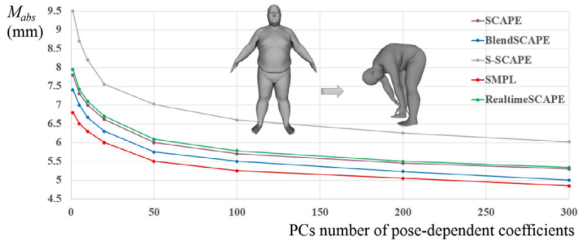

Figure 2. The figure from (29) demonstrates how the mean absolute vertex (Mabs) error varies across five typical models, depending on the number of principal components (PCs) representing pose-dependent coefficients. The test evaluates how effectively a shape in an A-pose (left) adapts to a new pose (right).

surface deformations. Rigid transformations alone (e.g. rotating limbs) cannot realistically mirror soft-tissue changes at the joints or the subtle shift in body contour when a limb is flexed. Additional pose-dependent offsets are introduced to encode these non-rigid effects.

SCAPE. SCAPE (9) factors inter-person shape variation and pose deformation. The shape component is a low-dimensional PCA subspace (as in Allen *et al.* (7)), whereas pose corrections are learned from scans of one individual in multiple poses. Each triangular face in a registered template has an affine transformation for shape and another for pose, followed by a rigid rotation from the skeleton. Since SCAPE relies on scans from a single person to learn pose, it does not generalize pose corrections across different identities.

The core SCAPE formulation for a triangle edge $\hat{v}_{k,j}$ is

$$y_{k,j} = R_{\ell(k)}\, S_k^{(\text{body})}\, Q_k^{(\text{pose})}\, \hat{v}_{k,j},$$

where $Q_k^{(\text{pose})} \in \mathbb{R}^{3\times 3}$ is a linear function of twist coordinates (joint angle differences) and $S_k^{(\text{body})} \in \mathbb{R}^{3\times 3}$ encodes per-face shape adjustments. This approach is piecewise affine, therefore a *per-face* optimization is further used to avoid discontinuities. SCAPE does not place global priors on improbable poses or handle multi-subject pose variability.

Unified Shape-and-Pose Subspaces. Hasler *et al.* (53) aggregate scans from many subjects in multiple poses and build a single PCA space for shape and pose. The method encodes local deformations using rotation-invariant features: each face has a rotation $\mathbf{R}_k$ and a stretch/shear matrix $\mathbf{S}_k$. The transformations are normalized against rigid motion, and the resulting codes are then factored by PCA. This allows the learned subspace to mix both subject-specific body shape and pose-induced geometry changes, though success depends on sufficient coverage in shape–pose space.

Further improvements. Realistic statistical models require large amounts of training data to capture de-

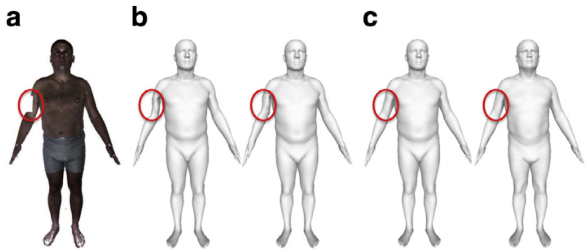

Figure 3. Comparison of SCAPE and BlendSCAPE models (29). (a) The original data exhibits gaps in the left upper arm highlighted by the red circle. (b) Rigid part rotations can occasionally introduce significant artifacts early in the fitting process, which co-registration struggles to eliminate. (c) BlendSCAPE's blend rotations result in smoother, less pronounced artifacts that co-registration can quickly correct.

mographic and anthropometric variation. Pishchulin *et al.* (125) address this via an iterative bootstrapping procedure, in which a pre-trained SCAPE model provides initialization for non-rigid alignment. Additionally, posture normalization (109; 171) reposes each scan to a common stance, preventing minor arm or leg pose differences from contaminating the PCA shape space with trivial pose variation.

Neophytou and Hilton (109) developed the Shape and Pose Space Deformation (SPSD) model, extending SCAPE to multiple people and multiple poses. SPSD uses a blending approach in the joint shape–pose space: for a new subject shape and pose, weights are computed by comparing local shape and pose distances to the training examples. This local interpolation scheme is able to capture subject-dependent pose corrections more robustly (109). The registration process, was further improved by Hirshberg *et al.* (56), who unified the registration of multiple scans with the learning of a SCAPE-like model into a single objective. The co-registration approach couples data alignment terms with a shape-and-pose prior, avoiding inconsistencies that can arise if scans are registered independently. The resulting BlendSCAPE model reduces artifacts at joints and simultaneously estimates per-subject shape offsets and per-triangle pose corrections (see Figure 3).

Lastly, SCAPE-based methods are not able to capture temporal changes in the body surface. To address the gap, Dyna (126) introduces time-varying offsets that depend on velocities, accelerations, and an autoregressive model of soft tissue. The dynamic offsets $\mathbf{D}_n(\boldsymbol{\delta}_k)$ are learned from high-speed 3D sequences. An important extension is the use of the subject's BMI to weight subject-specific autoregressive models, reflecting heavier or leaner subjects' different tissue jiggle behavior.

5.4. SMPL and beyond

The Skinned Multi-Person Linear model (SMPL) (102) replaces SCAPE's per-face transformations with a simpler vertex-based approach and standard linear blend skinning (LBS). SMPL decomposes shape into PCA blend-shapes and pose into small corrective displacements. Each vertex belongs to one or more skeleton joints with certain blend weights $\mathbf{W} \in \mathbb{R}^{N \times K}$. Let $\boldsymbol{\beta}$ be the shape coefficients, and $\boldsymbol{\theta}$ be the axis-angle pose parameters:

$$\mathbf{M}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \mathcal{W}\Big(\overline{\mathbf{T}} + B_S(\boldsymbol{\beta}) + B_P(\boldsymbol{\theta}),\ \mathbf{J}(\boldsymbol{\beta}),\ \boldsymbol{\theta},\ \mathbf{W}\Big),$$

where $\overline{\mathbf{T}}$ is a shared rest-pose template with N vertices, B_S (shape blend-shapes) is a PCA mapping, and B_P (pose blend-shapes) corrects bending artifacts. SMPL's *pose blend-shapes* are linear in the rotation matrix elements (minus identity) to ensure deformations depend consistently on joint orientation. Training SMPL involves two main phases:

1. learning pose-corrective terms from multi-pose data (keeping shape fixed), and
2. learning the shape PCA space from multi-subject data.

5.5. Improvements and Extensions of SMPL

SCAPE and SMPL differ primarily in the internal representation of pose corrections. SMPL is more friendly to standard graphics and animation software due to compatibility with linear blend skinning pipelines and is therefore widely adopted in the community. Numerous extensions improve upon the initial design to address the its limitations such as sparse corrections (STAR (3)), dynamic offsets (DMPL (102)), or multilinear shape–pose factorizations (Chen *et al.* (28)). Further developments also include anatomical detail. SKEL (76) replaces SMPL's approximate joints with biomechanically plausible bones and DoFs. STMPL (1) adds a learned volumetric layer that deforms under contact. Importantly, these improvements remain compatible with SMPL's blend-skinning foundation, therefore preserving its ease of use and compatibility with existing tools.

Hands. Romero *et al.* (64) propose MANO, a PCA-based hand model with per-joint pose offsets. They merge MANO into SMPL, obtaining SMPL+H, which accurately captures finger articulation and body pose in a unified representation, addressing the problem of simplified hand geometry in standard SMPL. Addtionally, HTML (129) provides a full parametric texture model of the hand (not just shape and pose) that is compatible

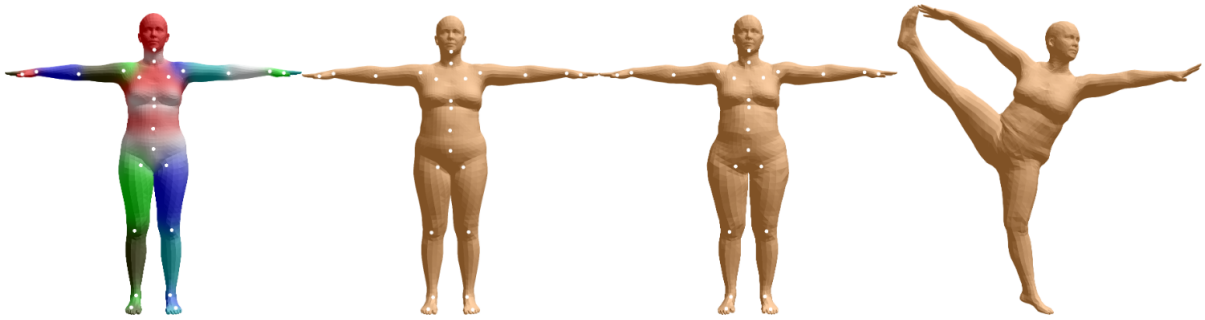

Figure 4. Figure (102) illustrates the Skinned Multi-Person Linear model (SMPL).(a) The template mesh is displayed, with color-coded blend weights across the surface and skeletal joints indicated in white. (b) Here, only the identity blendshapes are active, so both vertices and joints move linearly with changes in the shape vector. (c) Pose blendshapes are then introduced for the impending split pose, causing the hips to widen. (d) Finally, dual quaternion skinning repositions the vertices into the split pose.

with MANO's mesh. As a result, HTML is able to recover both personalized hand shape and realistic appearance from a single RGB image. Later approach (127), provides high-resolution hand geometry based on much greater amount of scans also pushing the scores on benchmarks on datasets like Freihand (199).

Face. FLAME (156) builds on SMPL's approach for the head region, adding a jaw, neck, and eyes, in addition to a linear expression space. The model captures both identity differences (e.g. head shape) and facial expression changes. It is trained on thousands of head scans and 4D expression captures. FLAME does not provide a full mouth interior or tongue, but it can model a broad range of facial configurations. In contrast, a more recent and orders of magnitude (20x) smaller model JNR (164) trained only on 94 scans is able to match the performance of FLAME on the BU-3DFE scans (194).

Full-Body Integration. Frank and Adam (51) as well as SMPL-X (117) unify body, face, and hands in a single vertex-based mesh. SMPL-X adds face and hand articulation to SMPL, capturing correlations among body shape, facial expressions, and finger motion. GHUM & GHUML (58) in addition offer high-resolution data also for body, face, and hands in an end-to-end trainable pipeline.

Garments. Alldieck *et al.* (5) add a displacement field $\mathbf{D} \in \mathbb{R}^{3N}$ to the SMPL body, enabling it to represent clothing or hair geometry outside the body. This approach, called SMPL+D, is therefore able to fit both the underlying body and the clothing displacements from monocular videos. The final body mesh includes both large garment deformations and the standard SMPL shape. SO-SMPL (166), which introduced a two-layer offset framework on SMPL-X (117), allowing a clean separation between the unclothed body and garments, i.e. the body offset $\mathbf{O}_h$ is added first, and then a garment offset $\mathbf{O}_c$ is applied only to the subset of vertices belonging to clothing. This design simplifies garment replacement, removing geometric entanglement between body and clothing layers. Hewitt *et al.* (55) follow a similar layered approach to reconstruct entire humans with garments in multi-camera settings or single images. Their pipeline combines learned 2D regressors for landmarks with a layered parametric model that can be iteratively fit to real or synthetic images. Each garment layer is topologically separated, avoiding overlap with the underlying body or other garments.

5.6. Conclusion

Parametric human body models evolved from early PCA-based static-shape representations to methods that incorporate pose dependence, clothing, dynamic soft tissue, and detailed face or hand geometry. They share a common foundation:

- Register scans to a shared template,
- reduce dimensionality with PCA or similar factorizations,
- represent pose transformations via skeleton-based methods or local affine transformations.

Many modern models combine all these elements into a single pipeline, sometimes trained end-to-end on large databases. These advances enable realistic synthesis,

motion capture, and editing of full-body human geometry, benefitting applications in vision, graphics, biomechanics, and beyond.

6. SMPL in Human Pose Estimation

Early work on single-image 3D human pose and shape estimation using parametric body models often combined traditional optimization steps with learned functions, such as SMPLify (19), which relies on a pretrained 2D keypoint detector and then iteratively fits SMPL parameters by minimizing a reprojection error between detected joints and the model's projected joints. This purely optimization-based strategy enforces pose and shape priors during fitting, achieving $\approx 79.9\,\mathrm{mm}$ of average 3D joint error on HumanEva (148) and $82.3\,\mathrm{mm}$ on Human3.6M (61). By contrast, HMR (71) takes a regression-centric approach, predicting SMPL parameters in a single feedforward pass while making use of a learned adversarial prior to encourage plausible pose estimates. This approach bypasses direct iterative fitting and attains a reconstruction error of approximately $56.8\,\mathrm{mm}$ on Human3.6M.

Later, SPIN (80) sought to combine both paradigms in a closed-loop system. It uses a feedforward regressor similar to HMR but refines the regressed parameters via an optimization procedure that aligns the predicted 3D joints with 2D keypoints. The refined estimates then supervise the regressor, allowing iterative feedback between the model-fitting steps and the learning component. This hybrid approach obtains around $41.1\,\mathrm{mm}$ mean per-joint position error on Human3.6M and $59.2\,\mathrm{mm}$ on 3DPW (165). Taken together, these works highlight three major themes in early parametric recovery of pose and shape: fully iterative optimization (SMPLify (19)), purely feedforward regression (HMR), and a hybrid scheme that integrates both (SPIN). Research have been expanded on these ideas by exploring additional constraints, leveraging multiple views, and incorporating temporal information in subsequent methods.

6.1. Pose Estimation Taxonomy

Several complementary design choices emerge. These include the balance between purely optimization-based versus direct regression models, the use of single-frame versus temporal inputs, methods focused on single-person scenarios versus multi-person scenes, deterministic versus probabilistic frameworks, and the variety of domain-specific constraints and in-the-wild training strategies.

Optimization vs. Regression vs. Hybrid. Classical optimization approaches, e.g. SMPLify (19), rely on an iterative energy minimization with a parametric model, thereby offering flexibility but often requiring multiple iterations and reliable 2D detections. Regression-based pipelines (e.g., HMR (71)) can be faster at inference and integrate well with large-scale datasets. However, pure regression can struggle under out-of-distribution poses, occlusions, or limited data. Hybrids, such as SPIN (80), Neural Descent (187), and HybrIK (86), aim to combine the best of both worlds. They learn a regressor, then refine or guide it with either an analytical or iterative update, improving robustness to unusual poses and partial occlusions.

Single-Frame vs. Video. Single-frame solutions often produce pose jitter when run frame-by-frame on a video. Temporal methods like VIBE (78), TCMR (30), or HuMoR (134) significantly improve motion continuity by modeling temporal dependencies or learning generative motion priors. Nevertheless, such approaches can require more complex architectures, additional training data (e.g., motion capture sequences), and reliable temporal bounding-box tracking. They also must handle variable sequence lengths and potential viewpoint changes over time.

Single-Person vs. Multi-Person. Some estimators specialize in reconstructing a single subject in isolation, while multi-person methods handle occlusions, inter-person overlaps, and the need to identify who is who in the scene. Multi-person pipelines often feature extra modules, such as detection-based segmentation, multi-person refinement blocks, or set-prediction transformers (e.g., PSVT (131)), and must address collisions or interactions (for instance, verifying that no two meshes penetrate each other). This additional complexity can improve reconstructions in group scenarios but also makes training and inference more computationally demanding.

Multi-View vs. Single-View. Multi-view methods leverage synchronized cameras and geometric cues (e.g., epipolar consistency) to reduce depth ambiguity and occlusion. For example, Chun et al.(32) predict partial vertices from each view, then fuse them into a unified 3D mesh with explicit multi-view constraints. Though single-view pipelines are simpler, they remain sensitive to occlusion and lack geometric grounding compared with multi-view setups.

Deterministic vs. Probabilistic. Most pipelines deliver a single best pose or shape per image or video. However, some more recent methods, such as

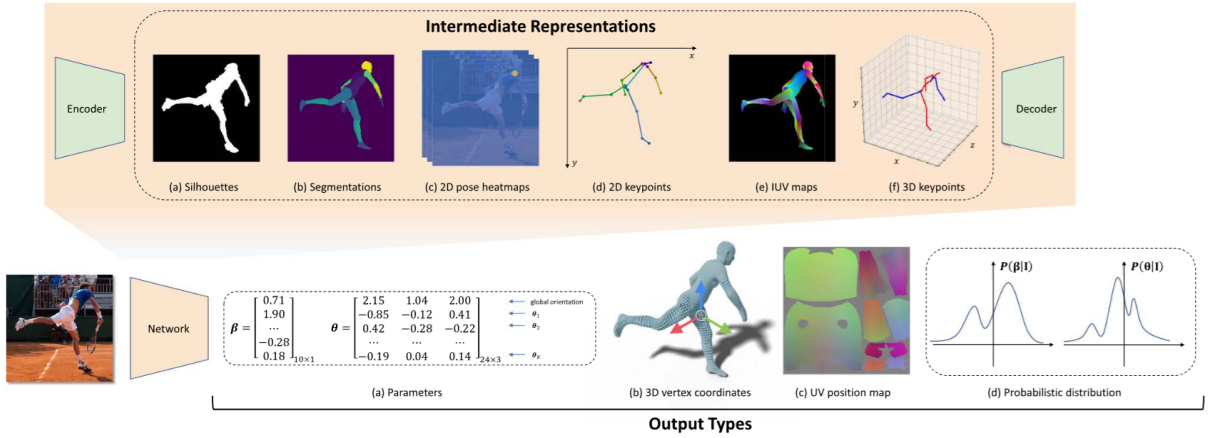

Figure 5. This figure (155) illustrates different output types and intermediate representations utilized in regression networks. We explore four distinct output formats: (a) parametric representation; (b) 3D coordinates of mesh vertices; (c) UV position maps; and (d) probability distributions over pose and/or shape parameters. The multi-stage frameworks incorporate various intermediate representations, including (a) silhouettes; (b) segmentation maps; (c) 2D pose heatmaps; (d) 2D keypoint coordinates; (e) IUV maps; and (f) 3D keypoint coordinates, which function either as simplified input representations or as guiding signals.

ProHMR (81) or HuManiFlow (144), treat pose estimation as a distribution, sampling multiple plausible 3D configurations consistent with the same 2D evidence. Such probabilistic schemes can capture ambiguities and self-occlusions, offering uncertainty estimates or multiple hypotheses. While more flexible, they also require additional modeling and can be more challenging to deploy at scale.

In-the-Wild Data vs. Domain-Specific Constraints. To handle diverse real-world settings, many methods exploit data augmentation or specialized data sources. Techniques such as EFT (80) refine a baseline model via fitting to large-scale 2D datasets, generating pseudo-3D labels. LASOR (178) integrates synthetic occlusion data and silhouette rendering to deal with human-human overlaps. Sensors like event cameras (200) or polarization (202) supplement regular RGB, enabling robust estimation under challenging lighting or motion. Others embed physical constraints, e.g. biomechanical assumptions in BioPose (79) or domain-specific occupant priors inside vehicles (77), improving realism and plausibility. (**?**) also claims higher accuracy on fine-tuned models for specific domains. Such strategies expand pose estimation beyond generic image-to-mesh regression, focusing on specialized scenarios or constraints that further improve accuracy and practicality.

6.2. Single-Image Regression

Single-image methods extend or refine the foundational pipeline of estimating SMPL parameters by direct regression, often building on or comparing against pure optimization (SMPLify (19)) and baseline regression (71; 88). The main focus is how each approach addresses the limitations of single-image inference, such as global orientation ambiguity, reliance on sparse keypoints, and lack of direct 3D ground truth. Several methods introduce specialized constraints (e.g., clothing segmentation, biomechanical priors), while others learn uncertainty models or multiple-hypothesis distributions. SMPLify (19) remains a baseline for optimization-driven fitting of SMPL to 2D joints. Many newer methods still reference it as a component in training loops or as a benchmark for single-view performance.

Location and Global Orientation. A recurring issue in top-down pipelines is the loss of global positioning when images are cropped. CLIFF (88) addresses this by preserving the subject's original bounding-box information and enforcing a full-image reprojection loss. This improves global rotation estimation compared to earlier methods.

Data Augmentation and Fine-Tuning. EFT (68) augments large 2D datasets with 3D pseudo-labels by individually refining a pretrained regressor on each image. This allows building large-scale 3D sets (e.g., from COCO (97)) and benefits methods that require 3D supervision. HybrIK (86) integrates analytical inverse kinematics with a network-predicted twist angle. This approach is similar to SMPLify but streamlines the fit by only learning the twist, which is more difficult to infer from joints alone.

Pose and Shape Constraints. Several works encode stronger constraints into the pose or shape estimation pipeline. These strategies improve over classic 2D keypoint reprojection alone by introducing auxiliary supervision, such as silhouettes, clothing masks, or anatomical priors. HEMlets PoSh (196) introduces local depth-ordering cues via part-centric triplets, bridging 2D and 3D more effectively than standard keypoint supervision. BioPose (79) places a neural IK module in the loop, enforcing biomechanical consistency in joint angles and body scale. DSR (39) (Differentiable Semantic Rendering) uses clothing segmentation and a differentiable renderer to refine the alignment of SMPL meshes with high-level body-part information. LASOR (178) synthesizes occlusion-aware silhouettes and employs a neural mesh renderer to supervise shape estimation under various occlusions.

Uncertainty and Multi-Hypothesis Modeling. Another direction focuses on capturing the inherent ambiguity in single-view 3D recovery. CUPS (190) and POCO (41) embed uncertainty estimation within a standard regression framework, providing calibrated error bounds for each joint. This helps with tasks like filtering low-confidence predictions and guiding decision-making. ProHMR (81) and HuManiFlow (144) instead learn probability distributions over the SMPL pose parameters, relying on normalizing flows. They allow sampling multiple plausible configurations and extracting the most likely or a set of likely solutions. 3D Multi-Bodies (17) applies a multi-hypothesis neural network approach, ensuring at least one solution matches the observed 2D joints. (142) extends this idea by parameterizing pose through a tree of matrix-Fisher distributions, leading to a structured representation of pose uncertainty. These probabilistic methods address self-occlusions or depth ambiguities by distributing likelihood across multiple plausible poses, in contrast to a single deterministic outcome.

Continuous Field Approaches. Neural Localizer Fields (NLF) (139) is designed to handle any type of body point annotation—2D keypoints, DensePose UVs (50), partial 3D scans—using a single, continuous localizer function. Instead of predicting only a fixed set of joint coordinates or SMPL parameters, NLF learns a mapping for any queried body point (e.g., a vertex on a mesh or a pixel labeled in DensePose) for which a learned hypernetwork configures a convolutional layer to pinpoint its 3D coordinates. Because this "localizer" can be trained on heterogeneous supervision signals (2D part labels, partial scans, etc.), it can fuse multiple annotation types in one framework. Once the 3D locations are recovered for the relevant points, a simple fitting step to SMPL or SMPL-X produces a parametric surface. This novel design stands in contrast to direct SMPL regressors and makes NLF more practical and readily adaptable to new or partial labels.

Overall, these single-image methods demonstrate the range of strategies for refinement beyond a basic CNN regressor. Some directly incorporate geometric or semantic constraints (39; 79; 178; 196), others refine bounding-box or orientation cues (88), and several integrate explicit or learned uncertainty (41; 190) or multi-hypothesis sampling (17; 81; 144).

6.2.1 More Constraints, Specialized Domains, and Refined Supervision

A number of approaches extend beyond standard single-image pipelines or multi-view setups by leveraging uncommon sensors, specialized domain knowledge, or additional semantic feedback. These signals may be physical (polarization, EM sensors), synthetic (cars with known seating), or high-level (segmentation-based feedback). These approaches show that extra data or priors help resolve depth ambiguities and refine shape details, especially in challenging environments and result in higher accuracy in niche contexts or richer 3D reconstructions in common ones (39).

Alternative Sensors and Physical Observations. EventHPE (200) makes use of sparse, high-temporal-resolution event stream, merging learned optical flow (derived from event frames) with a shape-estimation network in an RNN-based inference. By modeling flow coherence across frames, it outperforms conventional frame-only systems in scenarios with fast motion or dim lighting. In a similar direction, (202) incorporates polarization images to infer per-pixel surface normals, guiding a refinement step that adds detailed clothing contours. Meanwhile, EMDB (73) captures global 3D motion outdoors via electromagnetic sensors on the subject, producing high-precision pose data despite real-world occlusions or camera motion.

Domain-Specific or Constrained Settings. Certain works consider niche domains, embedding specialized priors about body configurations. (77) inserts a VAE-based "in-vehicle" pose prior tailored to seated occupants, refining pose parameters consistent with typical car interiors. The design accounts for partial visibility (e.g., legs under dashboards) or unusual vantage points. By contrast, the pipeline proposed by Pavlakos et al. (119) originally used silhouettes and 2D joints in more general scenes, demonstrating that combining shape or silhouette constraints with 2D keypoints can make early single-image solutions more robust.

Differentiable Rendering and Dense Constraints. Some methods build additional refinement loops by rendering the estimated mesh back to 2D. DenseRaC (176) employs IUV maps to link image pixels with surface coordinates, then compares rendered results to the original correspondences. This dense alignment goes beyond sparse joints, improving shape fidelity. Another example is Neural Descent / HUND (187), a "learning-to-optimize" scheme where a recurrent network simulates iterative fitting steps in (6; 117); the pipeline uses differentiable rendering to align silhouettes, part segments, and 2D keypoints. These strategies allow stronger 2D constraints (beyond just joints) and can backpropagate errors at the pixel or semantic-segmentation level.

6.3. Multi-View Approaches

Multi-view methods aim to deal with depth and scale ambiguities by relying on geometric consistency across multiple cameras. In contrast to single-view frameworks, these methods incorporate cross-view constraints or triangulation to improve pose accuracy, shape plausibility, and overall robustness to occlusions.

UPose3D (37) leverages uncertainty-aware 2D keypoint estimates from each camera, refines these via a pose compiler that enforces cross-view and temporal consistency. Unlike many multi-view approaches that rely on direct 3D ground truth, UPose3D only requires synthetic multi-view 2D data for training. By modeling per-joint uncertainties with normalizing flows, it fuses noisy 2D points across multiple viewpoints, returning robust 3D estimates without explicit in-the-wild 3D labels. Another approach (32) synthesizes multi-view features to predict a reduced set of mesh vertices, then fits SMPL to these partial vertex predictions. By focusing on a sub-vertex mesh rather than the full SMPL surface, LMT reduces memory overhead while retaining enough spatial constraints to solve for the final pose and shape. A visibility module ensures that each camera only contributes features for visible regions of the mesh, increasing robustness to occlusions and conflicting viewpoints. (90) proposes a hybrid training loop that fuses a regression-based CNN with an extended multi-view SMPLify (19). The CNN regresses 3D body parameters from each view, which then initialize a multi-view optimization enforcing joint alignment across all cameras. The refined parameters subsequently supervise the CNN, improving its predictions. By using multi-view data only during training, this system can later infer from a single image at test time while still benefiting from the learned multi-view constraints. A different design is *SkelFormer* (38), which first obtains 2D keypoints from each view and then triangulates them to produce coarse 3D joints. A skeletal transformer then maps these noisy 3D keypoints to parametric SMPL pose and shape. Another advantage of this approach is, that it distinctly decouples 2D detection from 3D regression: the network learns an inverse-kinematics-like solution, handling imperfect triangulation and reducing failure cases that arise from end-to-end multi-view training.

6.4. Multi-Person and Interaction Methods

When multiple individuals appear in a scene or when the goal is to capture fine-grained interactions between body parts and external objects, the 3D pose and shape estimation pipeline becomes more complex. These methods must deal with overlapping bounding boxes, occlusions, scene constraints, and, in some cases, object-hand contact modeling.

Common Multi-Person Pipelines. One approach is to decompose multi-person estimation into a sequence of tasks and then refine globally. Cha et al. (26) propose a three-stage procedure: first, a network estimates 3D skeletons from a single image under heavy occlusions (skeleton-first). Next, the skeletons are mapped to body meshes via an inverse kinematics solver. Finally, a Transformer-based module refines pose and shape by leveraging interactions (intra-person consistency, inter-person collisions). This coarse-to-fine design alleviates ambiguity from nearby individuals, iteratively adjusting poses where traditional single-person fits might fail. RoboSMPLX (115) makes use of bounding-box localization and full-body (body + hands + face) reconstruction. It uses a localization module that explicitly estimates part segmentation, a contrastive feature extraction to remain invariant under bounding-box shifts, and a differentiable renderer to align final predictions pixel-wise. As a result, it can handle crowd scenes or partial occlusions, refining pose and shape predictions even if bounding-box detections are imperfect. Zanfir et al. (188) integrates explicit scene constraints when multiple people share the camera view. In addition to typical 2D keypoint reprojection terms, they impose a shared ground-plane constraint and penalize volume intersections among different subjects. This makes use of higher-level information—such as a common floor plane and non-overlapping body volumes—to handle extreme occlusions in crowd scenes. Such constraints can significantly reduce false positives for individuals partially overlapping in an image.

Multi-person Video Transformers. Qiu et al. (131) propose a single-stage multi-person video approach (PSVT) built on Transformer architectures. They treat each frame's individuals as query tokens, refining both pose and shape over time. The key is a progressive decoding mechanism that updates pose estimates in a

temporal manner, preventing the accumulation of errors over video sequences. This avoids separate detection/tracking modules while relying on spatio-temporal attention to manage person-to-person overlap and motion consistency. AiOS (151) further consolidates multi-person detection and SMPL-X fitting into a DETR-style transformer pipeline. It uses dedicated tokens for body parts, hands, and face, capturing each region with specialized queries in a single forward pass. This approach achieves multi-person, expressive reconstructions without external detectors or multi-stage cropping. Transformer-based self-attention accounts for interactions between people, preserving global context and local detail in a single integrated framework.

Object and Hand Interaction. Methods dealing with fine-grained hand-object contact face further complexity. GRIP (152) models realistic hand-object interactions by focusing on contact fidelity: a two-stage pipeline refines arm trajectories first, then predicts hand poses using distance-based sensors for the object's local geometry. A consistency step enforces physically correct contact while minimizing hand-object penetrations. This extends beyond multi-person settings to multi-entity interaction, ensuring the final hand configuration remains plausible and stable over time.

6.5. Video-Based and Temporal Consistency

Methods designed for video input take advantage of motion cues and temporal structure, in contrast to single-frame pipelines that treat every image in isolation. These approaches result in smoother pose trajectories, fewer temporal artifacts, and more plausible motion dynamics. Although they vary in how they integrate time: some rely on recurrent layers or Transformers, others on adversarial motion priors, and still others on explicit generative or optimization-based refinements. They typically improve over single-image estimates by integrating frame-to-frame dependencies or by applying learned or model-based motion priors.

Predicting Pose Using Past and Future Frames. A central idea is to exploit neighboring frames explicitly, sometimes excluding the current frame's appearance features to force reliance on temporal continuity. TCMR (30) does this by training two extra GRUs that exclusively forecast from past or future frames—excluding the current frame's features entirely—thus learning to infer poses by observing context. This design reduces jitter and promotes coherent pose sequences, avoiding abrupt frame-to-frame changes that can occur when single-image models are independently applied.

Recurrent Architectures and Adversarial Motion Priors. Another perspective is to refine per-frame predictions with a recurrent module and a discriminator that enforces realistic motion. VIBE (78) encodes each frame through a CNN, aggregates these representations in a bidirectional GRU, and adds a motion discriminator. By distinguishing "real" 3D pose trajectories (from large-scale mocap) from those generated by the model, the discriminator guides the recurrent generator to produce anatomically coherent, temporally smooth sequences. Such adversarial supervision results in a tangible improvement in realism, especially for in-the-wild videos lacking direct 3D labels.

Self-Similarity Attention and Hierarchical Refinements. To ensure stable motion across longer clips, some approaches employ attention-based modules. MPS-Net (169) introduces self-similarity matrices and hierarchical attentive feature integration, aggregating key features across multiple frames so that poses remain consistent over time. GLoT (147) similarly divides temporal modeling into two stages: a global transformer that learns long-range dependencies (masking certain frames) and a local refinement step that corrects fine pose details in short subsequences. Such designs deal with both — large temporal windows (for broad motion context) and local corrections (for subtle adjustments at the frame level). Other strategies incorporate uncertainty modeling to handle noisy or corrupted frames. UNSPAT (84) fuses a temporal transformer with a module that identifies high-uncertainty image regions, subsequently downweighting their influence on final pose estimates. This focus on uncertainty helps maintain consistent predictions in videos with frequent occlusions or low visibility of limbs, emphasizing frames or spatial regions that are more reliable.

Generative Motion Models at Test Time. Going beyond purely feed-forward solutions, HuMoR (134) integrates a generative prior (CVAE-based) on human motion to optimize pose sequences at inference. Rather than only smoothing frames within a neural network, HuMoR runs a short test-time optimization procedure guided by the learned motion distribution, enforcing physically consistent states (e.g., correct ground contact, plausible transitions) even if keypoint detections are noisy or incomplete. This approach reconstructs longer trajectories with fewer artifacts and higher fidelity to realistic human motion.

6.6. Full-Body Models and Scaling

A recent line of work extended SMPL-type body models to include the head, hands, while also scaling the

amount of training data. The following methods seek to cover more expressive articulations (e.g., SMPL-X) and to leverage large-scale image collections.

Expressive SMPL-X Models. Techniques such as HYRE (87) and AiOS (151) address the full SMPL-X parameter space. HYRE merges a parametric branch (SMPL) with a non-parametric mesh-regression branch, including shared attention cues for upper- and lower-body refinements. AiOS frames multi-person SMPL-X estimation as a DETR-style transformer, where the network predicts both coarse body bounding boxes and specialized tokens for hand and facial landmarks. RoboSMPLX (115) also integrates expressive SMPL-X estimation (body, hands, face), with special emphasis on bounding-box stability and contrastive feature alignment to handle partial views or misalignment.

Large-Scale Learning. Scaling data resources can boost performance beyond architectural details. SMPLest-X (183) reportedly trains on tens of millions of images across dozens of datasets, using a simplified single-stage architecture but getting the benefits from the massive training set. Meanwhile, HuManiFlow (144) focuses on multi-hypothesis distributions for pose (and hand/face articulations), sampling from learned conditional flows on the SO(3) manifold. This approach is another option to introduce "scale": capturing plausible body part configurations under uncertainty, rather than single deterministic estimates. Overall, expressive full-body modeling poses new challenges in obtaining reliable hand and face annotations, spanning higher-dimensional parametric spaces. Yet the demonstrated advantage of large, diverse data—whether from curated multi-dataset aggregates or synthetic expansions—suggests that scaling up may win over narrower body model improvements. Methods that are able to make use of broad training sets thus represent a promising path forward.

6.7. Evaluation

Quantitative human pose estimation evaluation requires emplys several standardized evaluation procedures. This section reviews commonly used datasets, quantitative metrics, and practical challenges associated with evaluating pose estimation methods.

6.7.1 Datasets

Publicly available datasets lie at the core of evaluating 3D human pose estimation models. They provide diverse scenes, ground-truth annotations, and standardized tasks that enable exhaustive comparison across methods.

In 3D human pose estimation, two main categories of datasets are commonly used:

1. those with *direct 3D measurements* (e.g., motion capture) and
2. those relying on *synthetic or pseudo-labeled* ground truth.

Below are summarized the most widely adopted datasets, highlighting how each obtains its reference pose or shape. For a more comprehensive overview, see Table 3.

Human3.6M. (61) is a large-scale, indoor benchmark where subjects perform diverse motions (e.g., walking, sitting, discussion). Ground truth is derived via multi-camera *marker-based* motion capture. Reflective markers placed on the subject's clothing are tracked to produce accurate 3D joint coordinates (120 Hz). Although controlled and richly annotated, it lacks real-world backgrounds and heavy occlusions.

MPI-INF-3DHP. (105) combines indoor and green-screen outdoor captures. Ground truth 3D poses come from markerless multi-view capture with a calibrated camera rig. This setup includes varied clothing and more natural actions than strictly lab-based tasks, but backgrounds are still somewhat constrained.

3DPW. (165) is an in-the-wild dataset featuring everyday outdoor scenes (streets, university campus). Its ground truth is acquired using inertial measurement units (IMUs) strapped on each limb, fused with synchronized video in a bundle-adjustment approach. It features more realistic backgrounds and interactions than purely indoor motion capture. De facto considered a standard test for generalization and occlusion-handling.

AGORA. (128) is a large-scale synthetic dataset featuring multi-person scenarios, randomized backgrounds, and parametric ground truth from SMPL/SMPL-X. Because the 3D meshes are generated via rendering pipeline, the pose and shape annotations are considered "perfect" in theory. AGORA tests algorithms' ability to handle multi-person overlap, complex lighting, and widely varying body shapes.

SURREAL. (162) is another synthetic dataset created by rendering the SMPL model in Blender with randomized clothing, body shapes, backgrounds, and lighting. It provides pixel-wise ground truth for depth, body-part segmentation, and 2D/3D keypoints. Although purely synthetic, it introduces large-scale, diverse training data and supports dense supervision.

LSP / COCO / MPII. (66; 97) are traditional 2D pose benchmarks labeled by human annotators. They lack direct 3D ground truth but are routinely pseudo-labeled with 3D parameters (via SMPLify (20) or EFT (69)) to expand the variety of real-world images. Despite potential label noise, these 2D sets help in-the-wild generalization.

Acquisition of Ground Truth. The gold standard for 3D pose has historically been marker-based motion capture, resulting in sub-centimeter accuracy at high frame rates. However, it requires specialized studio setups, reflective markers, and controlled backgrounds (61). To capture more natural environments, markerless multi-camera rigs use structure-from-motion or silhouette constraints to triangulate each joint (105). In some cases, inertial sensors (IMUs) worn on the subject's limbs provide drift-corrected pose parameters via sensor fusion (165). Alternatively, fully synthetic pipelines (AGORA, SURREAL) provide "perfect" 3D but may not fully capture the complexity of real clothing or background clutter. Where only 2D keypoints exist, parametric model-fitting can approximate 3D pose and shape labels, albeit with some noise.

Usage and Splits. It is common practice to train on a combination of indoor 3D datasets (Human3.6M, MPI-INF-3DHP) plus pseudo-labeled 2D data (COCO, MPII) to enrich viewpoint and background diversity (80; 89). The 3DPW dataset is often held out for final testing due to its realistic scenes. Synthetic sets like AGORA can further improve shape accuracy, especially for underrepresented body types or multi-person poses. Each dataset typically has also recommended training/validation/testing splits to standardize evaluation. When multiple datasets are combined, cross-dataset evaluation helps measure generalization.

Dataset Considerations and Pitfalls.

- **Pose Diversity vs. Overfitting.** Some sets (Human3.6M) have many frames but fewer subjects and repeated indoor actions. Models can overfit to these poses if not combined with more varied data.
- **Quality of Annotations.** Marker-based or multi-view setups have generally high-fidelity 3D coordinates but can be expensive and often times domain-limited. Synthetic or pseudo-labeled data (e.g., SURREAL, EFT-labeled COCO) scale more easily yet introduce label noise or realism gaps.
- **Clothing and Shape Variation.** Datasets with minimal clothing (lab attire) do not fully test shape estimation. Real-world sets with varied outfits, e.g. 3DPW, or synthetic with randomized clothing, e.g. SURREAL/AGORA, better stress-test shape recovery.
- **Occlusion and Multi-person Scenes.** Datasets with heavier occlusion or multi-person interactions (AGORA) are more representative of real deployment scenarios. Performance often drops in these challenging settings.
- **Pose Parametrization Differences.** Some datasets provide only skeleton-based 3D joints, others supply SMPL/SMPL-X parameters. Conversions between different parametric representations can introduce small misalignments.

Robust single-view 3D pose estimators commonly employ a blend of real indoor data, in-the-wild 2D images (with pseudo-3D labels), and synthetic data for coverage. Future datasets may integrate additional modalities (e.g., depth, event cameras, or inertial sensors) to capture an even wider range of poses, shapes, and real-world conditions.

6.7.2 Evaluation Metrics

Human pose estimation methods rely on quantitative metrics to measure accuracy, robustness, and temporal consistency. Although the precise choice of metric can vary by dataset and application setting (e.g. single-person vs. multi-person, 2D vs. 3D, or static images vs. video), there are a handful of metrics widely recognized in the literature. We briefly outline the most prevalent ones and explain why they are used.

2D Metrics.

- **PCK (Percentage of Correct Keypoints).** A predicted 2D keypoint is considered correct if it lies within a certain distance threshold from its ground-truth counterpart. Common thresholds are a fraction of the head size or the torso size, or a fraction of the bounding-box dimension. **Variants.**
  - **PCKh:** Uses the human head size as a reference, popular in MPII.
  - **PCK@α:** Distance threshold is $\alpha \times (\text{reference\_dimension})$, e.g. 0.2 of torso length.
- **mAP (Mean Average Precision).** Derived from COCO Keypoint Challenge; computes Average Precision across a range of OKS (Object Keypoint Similarity) thresholds. Rewards not just localization but also the confidence ranking of detections. High mAP indicates robust detection of all keypoints across different difficulty levels.

- **PCP (Percentage of Correct Parts).** Treats each limb (e.g. upper arm, lower leg) as a line segment (part). A limb is correct if both endpoints are within a specified tolerance of ground truth. It is an older metric, but still referenced in some single-person benchmarks such as LSP.
- **AUC (Area Under the Curve).** Integrates PCK over multiple distance thresholds, capturing overall performance rather than at a single cutoff. It provides a single number to compare methods that might have different sensitivity at small vs. large distances.

3D Metrics.

- **MPJPE (Mean Per Joint Position Error).** Measures the average Euclidean distance (in millimeters) between predicted 3D joints and ground-truth joints, aggregated over all frames and joints. A direct measure of 3D skeleton accuracy. Typically evaluated after translating the subject so root joints coincide but without rotating or scaling. Good absolute measure if camera intrinsics are known, but heavily penalizes global orientation or scale errors if not.
- **PA-MPJPE (Procrustes-Aligned MPJPE).** MPJPE after rigid alignment (rotation, translation, possibly scaling) between predicted and ground-truth skeletons. Focuses on the correctness of the pose rather than penalizing differences in global position/orientation. Often called "reconstruction error."
- **N-MPJPE (Normalized MPJPE).** An MPJPE variant that includes alignment of global scale, in addition to rotation and translation. Crucial when comparing methods that might estimate humans of different sizes or uncertain camera focal length. Helps measure pure pose accuracy without scale bias.
- **PVE (Per-Vertex Error).** The mean distance between each vertex of the predicted 3D mesh (e.g. SMPL or SMPL-X) and the corresponding vertex of a ground-truth mesh, in millimeters. More fine-grained than a skeleton-based measure; captures shape accuracy. Particularly relevant for methods regressing full surface meshes.
- **3DPCK / PCK3D.** A 3D extension of PCK. A 3D joint is considered correct if it is within a specified distance (e.g. 150 mm) of the ground truth. "PCK3D@150mm" or "relative to torso length." AUC can similarly be measured by integrating PCK over multiple distance thresholds.
- **NMVE / NMJE (Normalized Mean Vertex/Joint Error).** Proposed in AGORA to handle multi-person and scale ambiguities. Normalizes each subject's shape/pose so that body-size differences are accounted for. In large-scale or multi-person benchmarks, ensures that short and tall subjects contribute equally in measuring shape/pose accuracy.

Temporal or Video-based Metrics. Several approaches extend single-frame 3D pose estimation to videos, adding consistency across frames. As a result, the following additional metrics can be reported:

- **Acceleration / Velocity Error.** Compares the acceleration (or velocity) of predicted joints against ground truth, measuring jitter or temporal smoothness. High 3D accuracy but choppy frame-to-frame transitions can degrade realism, so motion coherence is also assessed.
- **MOTA / MOTP (Multiple Object Tracking Accuracy / Precision).** For multi-person or group scenarios, these track how consistently individuals are identified and localized over time. It is also sometimes adapted to 3D skeleton tracking in crowd or multi-person videos.

Special-Purpose or Derived Metrics.

- **Silhouette IoU / Part Segmentation Accuracy.** For methods that also predict 2D masks or 3D occupancy, overlap (IoU) with ground-truth silhouettes or part segmentations (e.g. LSP) can be reported.
- **Penetration / Contact Metrics.** Some advanced methods ensure physically plausible results by measuring how much body parts overlap or if the feet realistically contact the ground. Minimal *penetration volume* or correct foot–floor contact indicates better realism.

Typical Usage in Literature.

- **2D pose methods** commonly report PCK (often PCKh) or mAP (COCO style).
- **3D skeleton-based approaches** almost always include MPJPE or PA-MPJPE, with an increasing trend toward 3DPCK/AUC for broader comparison.
- **Full mesh regressors** (e.g. SMPL-based) frequently show PVE or PA-MPJPE to cover both shape and pose errors.

- **AGORA or multi-person** tasks rely on NMVE / NMJE, since subject-specific scale is vital and large crowd scenes add complexity.
- **Video-based** methods (e.g. TCMR, VIBE) might add temporal smoothness or acceleration error to measure stability across frames.

In summary, when evaluating single-view 3D pose estimation, one should select the metric(s) most relevant to the target setting—e.g. absolute global accuracy vs. shape accuracy vs. purely skeleton configuration—and consider whether a rigid alignment step (Procrustes) or scale normalization is desired.

6.7.3 Comparative Analysis

Evaluations typically compare methods along both accuracy and computational cost. Comparisons often appear in tables listing MPJPE, PA-MPJPE, or PVE on Human3.6M, MPI-INF-3DHP, and 3DPW. We provide a comparison of selected single-view methods on Human3.6M in Table 1 and on 3DPW in Table 2.

7. SMPL in Action & Activity Recognition

The use of parametric body models in action and activity recognition has gained traction in recent years and its use cases go well-beyond simple visualization purposes. In this section we explore how SMPL and SMPL-X models have been integrated into action recognition pipelines, used to generate synthetic data for training, and improve the robustness of action recognition models.

7.1. Input Representations.

The adoption of mesh or skeleton inputs from parametric models addresses some central problems in activity recognition.

First, pose estimation errors from 2D keypoints can be mitigated, because 3D parametric fitting imposes body priors that increase consistency across frames. The main advantage of body-model-based representations lies in their robustness to viewpoint changes and occlusions. By aligning raw input to mesh template, methods such as 3DMesh-GAR (138) first reconstruct the 3D body from an RGB frame, then extract pose and shape parameters as low-dimensional features for group activity recognition. Similarly, (25) obtains high-fidelity SMPL-X meshes from video frames, then samples preselected set of joints from these meshes for action classification, outperforming methods that rely on sensor-captured skeletons. STMT (198) preserves the full 3D mesh structure over time, allowing a transformer architecture to learn spatio-temporal motion cues directly from vertex sequences without discarding local surface details.

Second, shape cues become readily available for other tasks that depend on anthropometric attributes.

Third, parametric models unify representation across datasets, supporting reuse of pre-trained networks on new subjects or environments. Studies leveraging these models have indicated improvements in recognition accuracy for single-person, multi-person, and group activities, suggesting that parametric body representations can serve as a solid foundation for classification, localization, and forecasting tasks.

FICTION (11) uses SMPL-based scene representations to anticipate upcoming object interactions and body configurations, fusing video observations with 3D context for long-horizon prediction. Moving beyond single-actor predictions, ReGenNet (173) models multi-person interactions by encoding two interacting SMPL bodies in a diffusion-based network capturing spatiotemporal relations. Several approaches adopt this parametric representation to stabilize the input space and to achieve invariance to viewpoint or subject appearance.

Body-model–derived inputs extend naturally to multi-person scenarios (193). In cases where multiple individuals are present, fitting parametric bodies to each subject and combining them at the feature level can simplify inference (92), by aggregating the resulting features into a subsequent classifier, consequently enabling the recognition of group activities. This strategy allows the model to sidestep bounding-box heuristics and focus on body shape and motion details (138). Similarly, researchers have used parametric mesh fitting as a precursor to more specialized pipelines for gaze analysis in dynamic scenes in real time (111).

Temporal action analysis has also incorporated parametric models to address tasks beyond basic classification. For instance, a transformer-based framework uses SMPL annotations to learn both action boundaries and labels in 3D motion data, improving localization and recognition of complex sequences (150).

Such a representation supports accurate modeling of social interactions because inter-person distances and pose relations are recorded in the same parametric space (174). These approaches tend to reduce ambiguity from overlapping bounding boxes and are advantageous for subsequent interaction or group-level classification steps.

7.2. Generative and Synthetic Data Approaches

Recent developments in action recognition extend parametric modeling also to generative pipelines for motion creation, editing, and augmentation. Parametric body models (e.g., SMPL, SMPL-X) serve as low-

dimensional bases that can be combined with deep generative architectures to produce or modify human motions with precise pose and shape control. T his includes methods that embed text prompts or action labels to guide synthesis (124; 135; 154), multi-action editing through iterative refinements (168), and multi-person interaction modeling with distance-based latent constraints (173). These frameworks leverage robust priors, ensuring anatomically coherent outputs and promoting data diversity for training or testing recognition algorithms.

Parallel research explores synthetic data generation, using parametric avatars retargeted with motion-capture data or adapted from user-supplied videos (27; 57; 136). Such approaches create large, labeled corpora under varied poses, environments, and camera settings, reducing domain gaps and mitigating annotation costs. Related strategies incorporate body-specific pressure maps for sensor-rich tasks (133), or expand 3D motion libraries with action-conditioned sampling (124; 154). By systematically controlling factors such as subject shape, viewpoint, and lighting, researchers can address class imbalances, validate model robustness, and refine real-time pipelines. These generative and synthetic data methods demonstrate the versatility of parametric models for broader tasks, including creative motion editing, data augmentation, and advanced simulation in fields where diverse or scarce datasets hamper progress.

7.3. Specialized Use Cases and Domains.

Human body models such as SMPL or SMPL-X have been adapted for a range of tasks where action and activity recognition demand specialized solutions across distinct environments, user populations, and interactive contexts. They facilitate richer annotations of complex movements and enable new forms of data fusion in domains that include healthcare, human-robot interaction, sports, multi-person group analysis, and creative applications.

Medicine. In clinical environments, monitoring patients in non-contact settings requires handling frequent occlusions and privacy constraints. Video-based frameworks that incorporate body models can assist in tasks such as in-bed movement assessment, where models can mitigate the impact of blankets and low lighting by imposing human pose priors (72). In seizure analysis and related patient monitoring, pose estimation pipelines that use body or skeleton tracking help discriminate subtle motor events, although mesh-based reconstructions remain less common (4). Beyond direct video processing, synthetic pressure maps derived from parametric body models can serve as augmented data to improve activity classification under challenging occlusions or limited real data (133).

Robotics. Parametric body models also support action observation frameworks, where a robot "watches" human motion to learn tasks or gestures. This includes gaze-aware methods that fit body models in real time and map attention to human body parts (111). When transferring human motion to robotic manipulators, parametric hand and body reconstructions (such as MANO or SMPL-X) can provide consistent retargeting from video to robot joint space, as shown in approaches that learn dexterous manipulation skills by analyzing large video corpora (146). More broadly, surveying large-scale instructional data for robot policy learning benefits from human pose pipelines to align, classify, and segment human demonstrations (42). Similar modeling strategies also improve gesture recognition for human-robot collaborations by capturing 3D meshes, which aid in mitigating viewpoint variations typical of real-world HRI settings (197).

Sports and Fitness. Several application domains involve non-routine or off-ground motions. Climbing analysis, for example, requires specialized handling of vertical poses and frequent body-surface contact. Datasets like CIMI4D register SMPL parameters to multi-sensor data while enforcing scene constraints to capture physically plausible climbing motions (177). Similarly, synthetic data platforms such as EgoSim can simulate diverse body shapes and poses for multi-view activity recognition and pose estimation in dynamic and athletic contexts (57). Annotating fitness-oriented motions also benefits from mesh-based ground truth, as seen in datasets focusing on pull-ups, push-ups, and sit-ups (189). These representations let researchers examine contact points, non-canonical orientations, and a broader range of kinematic constraints than in standard walking or running datasets.

Social Interaction and Group Activities. Action recognition in multi-person or crowd contexts can also leverage mesh-based representations. (138) regresses 3D meshes of multiple individuals from single images, then aggregates or processes them for high-level group activity classification. In environments where individuals interact over extended periods, such as kitchen scenes with multiple participants, registering everyone to a shared parametric representation allows modeling of social interaction cues and spatiotemporal predictions (153). Across these use cases, deploying meshes instead of skeletons can offer finer granularity of intra-person motions and closer alignment to 3D scene con-

straints.

Creative Applications. Motion generation, animation, and creative editing profit from parametric human models that unify motion data with learned priors. Generative pipelines address tasks from text-to-motion (135; 154) to multi-action synthesis (168) and fall-motion generation (120), each requiring plausible body kinematics that align with user-defined constraints or semantic descriptions. Other approaches integrate body models with high-level language embeddings to bridge motion and textual semantics (99), enabling richer editing, compositional generation, or style transfer.

These frameworks illustrate how parametric body models converge with neural architectures to produce new forms of animated content, interactive VR experiences, and large-scale data augmentation for action recognition or creative applications.

8. Discussion

This review has surveyed human body modeling with a focus on parametric approaches such as SMPL (101) and its variants, discussing their role in both 3D human pose estimation and action/activity recognition. Substantial progress has been made in monocular and multi-view pose recovery, where methods either optimize over model parameters (e.g., SMPLify (19)), directly regress them (e.g., HMR (71)), or combine both paradigms (e.g., SPIN (80)). Similarly, in action recognition, several pipelines leverage parametric meshes to obtain robust, viewpoint-invariant features. These methods have demonstrated promising accuracy improvements and broadened applications in multi-person scenes, clinical settings, and complex human-object interactions.

The findings align with emerging efforts in human modeling that emphasize holistic scene understanding rather than pose estimation alone (27; 198). In particular, recent research integrates human meshes with environment geometry, motion priors, and multi-modal data (e.g., inertial signals, gaze tracking) to address complexities in occlusion, domain shifts, and real-time deployment (111; 133). These directions expand upon classical pose-estimation frameworks, where the focus was predominantly on joint localization. The shared emphasis across pose estimation and action recognition on larger datasets (e.g., SMPLest-X (**?**), M3Act (27), HuMMan (21)) and domain generalization (14) reflects broader trends in computer vision toward scale and diversity. At the same time, researchers increasingly embed biomechanical or physical constraints (79; 152) to enhance plausibility and support specialized applications such as sports analysis or patient monitoring.

8.1. Strengths and Limitations of Parametric Models in Action Recognition

Strengths. Parametric modeling brings interpretability and explicit shape–pose separation, providing anatomically plausible meshes that extend traditional skeletal joints. This advantage improves action recognition by capturing subtle aspects of body shape and posture (138). Furthermore, the low-dimensional parameter space of models like SMPL enables efficient manipulation, making them suitable for data augmentation, multi-person tracking, and generative motion modeling (124; 154).

Limitations. Despite these strengths, parametric fitting methods can be computationally expensive, creating challenges for real-time systems (72; 111). Large transformer-based pose estimation models also demand significant memory, restricting on-device deployment. Certain categories of motion, including extreme activities or interactions with elaborate objects, remain difficult to reconstruct using only standard SMPL-based pipelines. Additionally, privacy considerations, such as re-identification risks from anthropometric data should also be considered (72; 132).

Dataset Constraints. While robust to some occlusions and variations, current approaches rely heavily on data coverage. Imbalances in motion categories or lack of domain-specific annotations reduce generalization (14; 27), and many existing benchmarks do not capture complex group behavior or multi-modal sensor streams (21; 161).

8.2. Applications

The synthesis of parametric body modeling with pose estimation affects among others safety, healthcare, robotics, and content creation. In clinical monitoring, for instance, mesh-based representations help identify patient movements despite frequent occlusions from blankets or medical apparatus (72). In real-time scenarios such as driver monitoring or industrial settings, multi-camera systems can leverage robust 3D pose estimates to track posture and reduce accident risks (77). Action recognition with mesh parameters equally accelerates research on human–robot collaboration: robots can interpret human motions accurately and adapt to them through shared parametric spaces (42). In animation, generative motion models that build on SMPL support workflows for film, gaming, and VR, allowing controllable creation of realistic human actions (168).

8.3. Open Questions and Outlook

Handling Complex Occlusions and Interactions. Current pipelines struggle in scenes with extreme self-occlusions or crowded multi-person scenarios (144; 152). Future research might integrate multi-hypothesis modeling and advanced collision constraints to better capture full-body contact with objects or other individuals.

Physical Plausibility and Multi-Level Semantics. While preliminary steps have integrated physics-based or biomechanical priors (79), bridging high-level semantic understanding with low-level mechanics remains largely open. Research is needed on physically informed data augmentation and parametric models that automatically encode contact forces, object manipulation, and environmental affordances.

Dataset Expansion and Standardization. Efforts on large-scale 3D datasets (e.g., EgoSim (57), HuMMan (21), M3Act (27)) must continue, but consistent annotation standards, especially for complex multi-person or fine-grained activities, are crucial (14). Benchmark suites that unify parametric annotations with new sensor data (e.g., IMUs, event cameras) can further accelerate progress.

Lightweight and Real-Time Solutions. Achieving real-time performance remains a major technical obstacle (77; 86). Compressed or partially learned parametric models, hardware-specific optimization, and distillation strategies can broaden the applicability of SMPL-like pipelines to embedded domains.

Privacy and Ethical Considerations. Parametric meshes offer some protection by removing facial texture, yet body shape and anthropometric cues may inadvertently reveal personal information (132). Investigating encryption, anonymization, and user-consent mechanisms for mesh data is an ongoing ethical challenge that must be addressed in future real-world deployments given the sensitivity of visual data.

9. Conclusion

In this review, we surveyed current advancements in parametric human body modeling for both 3D pose estimation and action/activity recognition. We traced how foundational body models, such as SMPL (101), evolved to capture rich shape and pose information, thereby enabling enhanced performance in action classification (138), group activity tracking, and motion forecasting. A core takeaway is that parametric representations unify geometry, appearance, and temporal cues, facilitating relatively reliable reconstructions even under partial occlusions and offering a framework for tasks ranging from in-bed patient monitoring (72) to high-level motion generation (168).

These findings underscore the growing emphasis on comprehensive scene understanding, multi-view and sensor fusion strategies (77; 202), and physically informed priors (79; 152). Nevertheless, several limitations persist, including computational complexity in real-time deployments, data imbalances that impede generalization, and ethical concerns regarding personal-identifiable shape information (132). Moving forward, greater standardization across datasets, integration of object and environment interactions, advances in hardware, and efficient model architectures will likely propel the field toward more robust, biomechanically consistent, and real-time solutions. Ultimately, the versatility of parametric body models and their increasing alignment with advanced sensing techniques hold promise for analysis of human behaviors, interactions, and the broader scope of physical and social contexts.

References

[1] Anton Agafonov and Lihi Zelnik-Manor. Stmpl: Human Soft-Tissue Simulation. *arXiv.org*, 2024. 7

[2] J. K. Aggarwal and Q. Cai. Human motion analysis: A review. *Computer Vision and Image Understanding*, 73(3):428–440, March 1999. 2

[3] A. A. Osman Ahmed, Bolkart Timo, and J. Black Michael. Star: Sparse Trained Articulated Human Body Regressor. *European Conference on Computer Vision*, 2020. 7

[4] David Ahmedt-Aristizabal, Mohammad Ali Armin, Zeeshan Hayder, Norberto Garcia-Cairasco, Lars Petersson, Clinton Fookes, Simon Denman, and Aileen McGonigal. Deep learning approaches for seizure video analysis: A review. *Epilepsy Behavior*, 154:109735, 2024. 18

[5] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1175–1186, 2019. 8

[6] Thiemo Alldieck, Hongyi Xu, and Cristian Sminchisescu. imghum: Implicit generative models of 3d human shape and articulated pose. In *CVPR*. IEEE, 2021. 4, 12

[7] Brett Allen, Brian Curless, and Zoran Popović. The space of human body shapes: reconstruction and parameterization from range scans. *ACM TOG*, 22(3):587–594, 2003. 5, 6

[8] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings*

of the IEEE Conference on computer Vision and Pattern Recognition, pages 3686–3693, 2014. 34

[9] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: Shape completion and animation of people. *ACM Transactions on Graphics (TOG)*, 24(3):408–416, 2005. 1, 2, 4, 6

[10] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *arXiv preprint*, 2019. No explicit conference or DOI given in the text. 29

[11] Kumar Ashutosh, Georgios Pavlakos, and Kristen Grauman. Fiction: 4d future interaction prediction from video. *arXiv preprint arXiv:2412.00932*, 2024. 17

[12] First Author, Second Author, and Others. Combining 3d morphable models: A large scale face-and-head model. In *Proceedings of the XYZ Conference*, pages 1–10, 202X. 4

[13] A. AuthorOne, B. AuthorTwo, and C. AuthorThree. A semantic parametric model for 3d human body reshaping. In *Proceedings of the XYZ Conference on Computer Vision*, 2023. 4

[14] Kristijan Bartol, David Bojanić, Tomislav Petković, Nicola D'Apuzzo, and Tomislav Pribanić. A review of 3d human pose estimation from 2d images. In *3DBODY.TECH 2020*, 2020. 1, 19, 20

[15] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *European Conference on Computer Vision (ECCV), LNCS 12347*, pages 311–329. Springer, 2020. 5

[16] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15935–15946, 2022. 30

[17] Benjamin Biggs, Sébastien Ehrhardt, Hanbyul Joo, Benjamin Graham, Andrea Vedaldi, and David Novotny. 3D Multi-bodies: Fitting Sets of Plausible 3D Human Models to Ambiguous Image Data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 11

[18] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8726–8737, 2023. 35

[19] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, 2016. 1, 2, 9, 10, 12, 19

[20] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3d human pose and shape from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 561–578, 2016. 15, 29

[21] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, et al. Humman: Multi-modal 4d human dataset for versatile sensing and modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2023. https://caizhongang.github.io/projects/HuMMan/. 19, 20

[22] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Yanjun Wang, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. Smpler-x: Scaling up expressive human pose and shape estimation. In *NeurIPS 2023 Datasets and Benchmarks Track*, 2023. 29

[23] Zhongang Cai, Mingyuan Zhang, Jiawei Ren, Chen Wei, Daxuan Ren, Zhengyu Lin, Haiyu Zhao, Lei Yang, Chen Change Loy, and Ziwei Liu. Playing for 3d human recovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 37

[24] Enrico Calabrese, Gemma Taverni, Christopher Awai Easthope, Sophie Skriabine, Federico Corradi, Luca Longinotti, Kynan Eng, and Tobi Delbruck. Dhp19: Dynamic vision sensor 3d human pose dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 32

[25] Junuk Cha, Muhammad Saqlain, GeonU Kim, Mingyu Shin, and Seungryul Baek. Learning multi-person 3d pose and shape with inverse kinematics and refinement. *arXiv preprint arXiv:2210.13529*, 2022. 17

[26] Junuk Cha, Muhammad Saqlain, GeonU Kim, Mingyu Shin, and Seungryul Baek. Multi-person 3d pose and shape estimation via inverse kinematics and refinement. *arXiv preprint arXiv:2210.13529*, 2022. 12

[27] Che-Jui Chang, Danrui Li, Deep Patel, Parth Goel, Honglu Zhou, Seonghyeon Moon, Samuel S Sohn, Sejong Yoon, Vladimir Pavlovic, and Mubbasir Kapadia. Learning from synthetic human group activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 18, 19, 20

[28] Yinpeng Chen, Zicheng Liu, and Zhengyou Zhang. Tensor-based human body modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 105–112, 2013. 7

[29] Zhi-Quan Cheng, Yin Chen, Ralph R Martin, Tong Wu, and Zhan Song. Parametric modeling of 3d human body shape—a survey. *Computers & Graphics*, 71:88–100, 2018. 1, 2, 6, 7

[30] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 9, 13

[31] Chih-Hsing Chu, Ya-Tien Tsai, Charlie C. L. Wang, and Tsz-Ho Kwok. Exemplar-based statistical model for semantic parametric design of human body. *Computers in Industry*, 61(6):541–549, 2010. 4

[32] Sungho Chun, Sungbum Park, and Ju Yong Chang. Learnable human mesh triangulation for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF*

Winter Conference on Applications of Computer Vision (WACV), 2023. 9, 12

[33] Deepseek Contributors. Llm deepseek: A semantic search engine for large language models. https://github.com/deepseek/LLM-Deepseek, 2023. Accessed: 2023-03-23.

[34] Enric Corona, Guillem Alenyà, Gerard Pons-Moll, and Francesc Moreno-Noguer. Layernet: High-resolution semantic 3d reconstruction of clothed people. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):1257–1272, 2024. 5

[35] Steve Dias Da Cruz, Oliver Wasenmüller, Hans-Peter Beise, Thomas Stifter, and Didier Stricker. An overview of the sviro dataset and benchmark. In *Proceedings of Computer Science in Cars Symposium*, 2019. 30

[36] Yudi Dai, YiTai Lin, XiPing Lin, Chenglu Wen, Lan Xu, Hongwei Yi, Siqi Shen, Yuexin Ma, and Cheng Wang. Sloper4d: A scene-aware dataset for global 4d human pose estimation in urban environments. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 682–692, 2023. 33

[37] Vandad Davoodnia, Saeed Ghorbani, Marc-Andre Carbonneau, Alexandre Messier, and Ali Etemad. UPose3D: Uncertainty-Aware 3D Human Pose Estimation with Cross-View and Temporal Cues, 2024. 12

[38] Vandad Davoodnia, Saeed Ghorbani, Alexandre Messier, and Ali Etemad. Skelformer: Markerless 3d pose and shape estimation using skeletal transformers. *arXiv preprint arXiv:2404.12625*, 2024. 12

[39] Sai Kumar Dwivedi, Nikos Athanasiou, Muhammed Kocabas, and Michael J. Black. Learning to regress bodies from images using differentiable semantic rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 11

[40] Sai Kumar Dwivedi, Nikos Athanasiou, Muhammed Kocabas, and Michael J Black. Learning to regress bodies from images using differentiable semantic rendering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11250–11259, 2021. 29

[41] Sai Kumar Dwivedi, Cordelia Schmid, Hongwei Yi, Michael J. Black, and Dimitrios Tzionas. Poco: 3d pose and shape estimation with confidence. In *2024 International Conference on 3D Vision (3DV)*, 2024. 11, 29

[42] Chrisantus Eze and Christopher Crick. Learning by watching: A review of video-based learning approaches for robot manipulation. arXiv preprint arXiv:2402.07127, 2024. 1, 18, 19

[43] Matteo Fabbri, Fabio Lanzi, Simone Calderara, Andrea Palazzi, Roberto Vezzani, and Rita Cucchiara. Learning to detect and track visible and occluded body joints in a virtual world. In *Proceedings of the European conference on computer vision (ECCV)*, pages 430–446, 2018. 33

[44] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12943–12954, 2023. 30

[45] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7214–7223, 2020. 31

[46] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Learning complex 3d human self-contact. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1343–1351, 2021. 32

[47] Mihai Fieraru, Mihai Zanfir, Silviu Cristian Pirlea, Vlad Olaru, and Cristian Sminchisescu. AIFit: Automatic 3D Human-Interpretable Feedback Models for Fitness Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9919–9928, 2021. 31

[48] Nahuel E. Garcia-D'Urso, Pablo Ramon Guevara, Jorge Azorin-Lopez, and Andres Fuster-Guillo. 3d human body models: Parametric and generative methods review. In *Lecture Notes in Computer Science*. 2023. 2

[49] Saeed Ghorbani, Kimia Mahdaviani, Anne Thaler, Konrad Kording, Douglas James Cook, Gunnar Blohm, and Nikolaus F Troje. Movi: A large multi-purpose human motion and video dataset. *Plos one*, 16(6):e0253157, 2021. 30

[50] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7297–7306, 2018. 11

[51] Joo H., Simon T., and Sheikh Yaser. Total Capture: A 3d Deformation Model for Tracking Faces, Hands, and Bodies. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 8

[52] Nils Hasler, Carsten Stoll, Bodo Rosenhahn, Thorsten Thormaehlen, and Hans-Peter Seidel. Estimating body shape of dressed humans. *Computers & Graphics*, 33(3):211–216, 2009. 4

[53] Nils Hasler, Carsten Stoll, Martin Sunkel, Bodo Rosenhahn, and H-P Seidel. A statistical model of human pose and body shape. In *Computer graphics forum*, volume 28, pages 337–346. Wiley Online Library, 2009. 6

[54] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2282–2292, 2019. 35

[55] Charlie Hewitt, Fatemeh Saleh, Sadegh Aliakbarian, Lohit Petikam, Shideh Rezaeifar, Louis Florentin, Zafirah Hosenie, Thomas J. Cashman, Julien Valentin, Darren Cosker, and Tadas Baltrušaitis. Look ma, no markers: holistic performance capture without the hassle. *ACM TOG*, 43(6), 2024. 8

[56] David A Hirshberg, Matthew Loper, Eric Rachlin, and Michael J Black. Coregistration: Simultaneous alignment and modeling of articulated 3d shape. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October*

7-13, 2012, Proceedings, Part VI 12, pages 242–255. Springer, 2012. 7

[57] Dominik Hollidt, Paul Streli, Jiaxi Jiang, Yasaman Haghighi, Changlin Qian, Xintong Liu, and Christian Holz. EgoSim: An egocentric multi-view simulator and real dataset for body-worn cameras during motion and activity. *Neural Information Processing Systems (NeurIPS)*, 2024. 18, 20

[58] Xu Hongyi, Gabriel Bazavan Eduard, Zanfir Andrei, Freeman W., Sukthankar R., and Sminchisescu C. Ghum GHUML: Generative 3d Human Shape and Articulated Pose Models. *Computer Vision and Pattern Recognition*, 2020. 8

[59] Jingwen Hu. Parametric human modeling. In *Basic Finite Element Method as Applied to Injury Biomechanics*, chapter 10, pages 417–445. Elsevier Inc., 2018. 1, 2, 6

[60] Chun-Hao P Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J Black. Capturing and inferring dense full-body human-scene contact. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13274–13285, 2022. 33

[61] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 9, 14, 15, 29

[62] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 36(7):1325–1339, 2014. 36

[63] Umar Iqbal, Anton Milan, and Juergen Gall. Posetrack: Joint multi-person pose estimation and tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2011–2020, 2017. 34

[64] Romero Javier and Tzionas Dimitrios. Embodied Hands : Modeling and Capturing Hands and Bodies Together * * Supplementary Material * *. 2017. 7

[65] Sai Sagar Jinka, Rohan Chacko, Avinash Sharma, and P. J. Narayanan. Peeledhuman: Robust shape representation for textured 3d human body reconstruction. In *3DV*. IEEE, 2020. 5

[66] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *bmvc*, volume 2, page 5. Aberystwyth, UK, 2010. 15

[67] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE international conference on computer vision*, pages 3334–3342, 2015. 32

[68] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting: Towards in-the-wild 3d human pose estimation. In *2021 International Conference on 3D Vision (3DV)*, pages 42–52. IEEE, 2021. 10

[69] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting: Towards in-the-wild 3d human pose estimation. arXiv preprint arXiv:2004.03686, 2021. 15, 29

[70] Shen Ka, Guo Chen, Kaufmann Manuel, Jose Zarate Juan, Valentin Julien, Song Jie, and Hilliges Otmar. X-Avatar: Expressive Human Avatars. *Computer Vision and Pattern Recognition*, 2023. 30

[71] Angjoo Kanazawa, Michael J. Black, David Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, pages 7122–7131, 2018. 1, 2, 9, 10, 19, 29

[72] Tamas Karacsony, Laszlo A. Jeni, Fernando De La Torre Frade, and Jo
ao Paulo Silva Cunha. Deep learning methods for single camera based clinical in-bed movement action recognition. TechRxiv Preprint, 2023. 18, 19, 20

[73] Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. EMDB: The Electromagnetic Database of Global 3D Human Pose and Shape in the Wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 11

[74] Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. Emdb: The electromagnetic database of global 3d human pose and shape in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14632–14643, 2023. 33

[75] Marilyn Keller, Keenon Werling, Soyong Shin, Scott Delp, Sergi Pujades, C Karen Liu, and Michael J. Black. From skin to skeleton: Towards biomechanically accurate 3d digital humans. *ACM Transactions on Graphics*, 42(6), 2023. 4

[76] Marilyn Keller, Keenon Werling, Soyong Shin, Scott Delp, Sergi Pujades, C Karen Liu, and Michael J Black. From skin to skeleton: Towards biomechanically accurate 3d digital humans. *ACM Transactions on Graphics (TOG)*, 42(6):1–12, 2023. 7

[77] Kwang-Lim Ko, Jun-Sang Yoo, Chang-Woo Han, Jungyeop Kim, and Seung-Won Jung. Pose and shape estimation of humans in vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 25(1):402–416, 2024. 10, 11, 19, 20

[78] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5249–5259, 2020. 9, 13

[79] Farnoosh Koleini, Muhammad Usama Saleem, Pu Wang, Hongfei Xue, Ahmed Helmy, and Abbey Fenwick. Biopose: Biomechanically-accurate 3d pose estimation from monocular videos. arXiv preprint arXiv:2501.07800, 2025. 10, 11, 19, 20

[80] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 9, 10, 15, 19, 29

[81] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayara-

man, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 10, 11

[82] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 29

[83] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6050–6059, 2017. 36

[84] Minsoo Lee, Hyunmin Lee, Bumsoo Kim, and Seunghwan Kim. Unspat: Uncertainty-guided spatiotemporal transformer for 3d human pose and shape estimation on videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024. 13

[85] Jiefeng Li, Siyuan Bian, Chao Xu, Zhicun Chen, Lixin Yang, and Cewu Lu. HybrIK-X: Hybrid Analytical-Neural Inverse Kinematics for Whole-body Mesh Recovery. *arXiv preprint arXiv:2304.05690*, 2023. 29

[86] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 5, 9, 10, 20, 29

[87] Wenhao Li, Mengyuan Liu, Hong Liu, Bin Ren, Xia Li, Yingxuan You, and Nicu Sebe. HYRE: Hybrid Regressor for 3D Human Pose and Shape Estimation. *IEEE Transactions on Image Processing*, 34:235–246, 2025. 14, 29

[88] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. *arXiv preprint arXiv:2208.00571*, 2022. 10, 11

[89] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. *arXiv preprint arXiv:2208.00571v2*, 2022. 15, 29

[90] Zhongguo Li, Magnus Oskarsson, and Anders Heyden. 3d human pose and shape estimation through collaborative learning and multi-view model-fitting. In *WACV*, 2021. 12

[91] Zhongguo Li, Magnus Oskarsson, and Anders Heyden. 3D Human Pose and Shape Estimation Through Collaborative Learning and Multi-view Model-fitting. Preprint, 2023. Code available at https://github.com/leezhongguo/MVSPIN_NEW. 29

[92] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Interhuman: A 3d multi-person dataset with realistic interactions for multi-human motion generation. *arXiv preprint arXiv:2304.05684*, 2023. 17

[93] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36:25268–25280, 2023. 32

[94] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 29

[95] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21159–21168, 2023. 35

[96] Siyou Lin, Hongwen Zhang, Zerong Zheng, Ruizhi Shao, and Yebin Liu. Learning Implicit Templates for Point-Based Clothed Human Modeling. *European Conference on Computer Vision*, 2022. 5

[97] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. 10, 15, 34

[98] Sheng Liu, Liangchen Song, Yi Xu, and Junsong Yuan. Nech: Neural clothed human model. In *2021 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 1–5. IEEE, 2021. 4

[99] Xinpeng Liu, Yong-Lu Li, Ailing Zeng, Zizheng Zhou, Yang You, and Cewu Lu. Bridging the gap between human motion and action semantics via kinematic phrases. *arXiv preprint arXiv:2310.04189*, 2023. 19

[100] Zhao Liu, Jianke Zhu, Jiajun Bu, and Chun Chen. A survey of human pose estimation: The body parts parsing based methods. *Journal of Visual Communication and Image Representation*, 32:10–19, 2015. 2

[101] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):248, 2015. 1, 2, 4, 19, 20

[102] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):1–16, 2015. 7, 8

[103] Qianli Ma, Jinlong Yang, Michael J. Black, and Siyu Tang. Neural point-based shape modeling of humans in challenging clothing. In *International Conference on 3D Vision (3DV)*, pages 679–689. IEEE, 2022. 5

[104] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 35

[105] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, pages 506–516.

IEEE, 2017. 14, 15, 36, 37

[106] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 international conference on 3D vision (3DV)*, pages 120–130. IEEE, 2018. 34

[107] Marko Mihajlovic, Yan Zhang, Michael J. Black, and Siyu Tang. Leap: Learning articulated occupancy of people. In https://neuralbodies.github.io/LEAP, 2021. 4

[108] Lea Muller, Ahmed AA Osman, Siyu Tang, Chun-Hao P Huang, and Michael J Black. On self-contact and human pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9990–9999, 2021. 33

[109] Alexandros Neophytou and Adrian Hilton. Shape and pose space deformation for subject specific animation. In *2013 International Conference on 3D Vision-3DV 2013*, pages 334–341. IEEE, 2013. 7

[110] Alexandros Neophytou and Adrian Hilton. A layered model of human body and garment deformation. In *2014 Second International Conference on 3D Vision*, pages 171–178. IEEE, 2014. 4

[111] Shuji Oishi, Kenji Koide, Masashi Yokozuka, and Atsuhiko Banno. 4d attention: Comprehensive framework for spatio-temporal gaze mapping. *IEEE Robotics and Automation Letters*, 6(4):7240–7247, 2021. 17, 18, 19

[112] Ahmed A. A. Osman, Timo Bolkart, and Michael J. Black. Star: Sparse trained articulated human body regressor. In *European Conference on Computer Vision (ECCV)*, pages 598–613. Springer, 2020. 4

[113] Ahmed A. A. Osman, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Supr: A Sparse Unified Part-Based Human Representation. *European Conference on Computer Vision*, 2022. 4

[114] Pablo Palafox, Aljaz Bozic, Justus Thies, Matthias Niessner, and Angela Dai. Npms: Neural parametric models for 3d deformable shapes. In *European Conference on Computer Vision*. Springer, 2021. 5

[115] Hui En Pang, Zhongang Cai, Lei Yang, Qingyi Tao, Zhonghua Wu, Tianwei Zhang, and Ziwei Liu. Towards robust and expressive whole-body human pose and shape estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 12, 14

[116] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*. IEEE, 2019. 4, 33

[117] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, pages 10975–10985, 2019. 8, 12

[118] Georgios Pavlakos, Nikos Kolotouros, and Kostas Daniilidis. TexturePose: Supervising Human Mesh Estimation with Texture Consistency. University of Pennsylvania, Technical Report, 2020. Project page: https://seas.upenn.edu/~pavlakos/projects/texturepose. 29

[119] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 459–468, 2018. 11

[120] Siyuan Peng, Kate Ladenheim, Snehesh Shrestha, and Cornelia Fermüller. Generation of novel fall animation with configurable attributes. In *Proceedings of the 9th International Conference on Movement and Computing (MOCO '24)*. ACM, 2024. 19

[121] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021. 4

[122] Xavier Perez-Sala, Sergio Escalera, Cecilio Angulo, and Jordi González. A survey on model based approaches for 2d and 3d visual human pose recovery. *Sensors*, 14(3):4189–4210, 2014. 2

[123] Radostina Petkova, Ivaylo Bozhilov, Desislava Nikolova, Ivaylo Vladimirov, and Agata Manolova. Taxonomy and survey of current 3d photorealistic human body modelling and reconstruction techniques for holographic-type communication. *Electronics*, 12(22):4705, 2023. 1, 2

[124] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-Conditioned 3D Human Motion Synthesis with Transformer VAE. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10985–10995, 2021. 18, 19

[125] Leonid Pishchulin, Stefanie Wuhrer, Thomas Helten, Christian Theobalt, and Bernt Schiele. Building statistical shape spaces for 3d human modeling. *Pattern Recognition*, 67:276–286, 2017. 7

[126] Gerard Pons-Moll, Javier Romero, Naureen Mahmood, and Michael J Black. Dyna: A model of dynamic human shape in motion. *ACM Transactions on Graphics (TOG)*, 34(4):1–14, 2015. 7

[127] Rolandos Alexandros Potamias, Stylianos Ploumpis, Stylianos Moschoglou, Vasileios Triantafyllou, and Stefanos Zafeiriou. Handy: Towards a high fidelity 3d hand shape and appearance model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4670–4680, 2023. 8

[128] Patel Priyanka, Huang C., Tesch J., T. Hoffmann David, Tripathi Shashank, and J. Black Michael. Agora: Avatars in Geography Optimized for Regression Analysis. *Computer Vision and Pattern Recognition*, 2021. 14, 35

[129] Neng Qian, Jiayi Wang, Franziska Mueller, Florian Bernard, Vladislav Golyanik, and Christian Theobalt. Html: A parametric hand texture model for 3d hand reconstruction and personalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 54–71. Springer, 2020. 7

[130] Shenhan Qian, Jiale Xu, Ziwei Liu, Liqian Ma, and Shenghua Gao. Unif: United Neural Implicit Functions

for Clothed Human Reconstruction and Animation. *European Conference on Computer Vision*, 2022. 4

[131] Zhongwei Qiu, Qiansheng Yang, Jian Wang, Haocheng Feng, Junyu Han, Errui Ding, Chang Xu, Dongmei Fu, and Jingdong Wang. PSVT: End-to-end multi-person 3d pose and shape estimation with progressive video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 9, 12

[132] Siddharth Ravi, Pau Climent-Pérez, and Francisco Florez-Revuelta. A review on visual privacy preservation techniques for active and assisted living. *Multimedia Tools and Applications*, 83:14715–14755, 2024. 19, 20

[133] Lala Shakti Swarup Ray, Vitor Fortes Rey, Bo Zhou, Sungho Suh, and Paul Lukowicz. Pressuretransfernet: Human attribute guided dynamic ground pressure profile transfer using 3d simulated pressure maps. arXiv preprint arXiv:2308.00538, 2023. 18, 19

[134] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 9, 13

[135] Daniel Ribeiro, Alice Gomes, and Others. MotionGPT: Large language models for motion generation. arXiv preprint arXiv:2401.01234, 2024. 18, 19

[136] Ana Romero, Pedro Carvalho, Luís Côrte-Real, and Américo Pereira. Synthesizing human activity for data generation. *J. Imaging*, 9(10):204, 2023. 18

[137] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2020. 4

[138] Muhammad Saqlain, Donguk Kim, Junuk Cha, Changhwa Lee, Seongyeong Lee, and Seungryul Baek. 3DMesh-GAR: 3d human body mesh-based method for group activity recognition. *Sensors*, 22(4):1464, 2022. 1, 17, 18, 19, 20

[139] István Sárándi and Gerard Pons-Moll. Neural localizer fields for continuous 3d human pose and shape estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 11

[140] István Sárándi and Gerard Pons-Moll. Neural Localizer Fields for Continuous 3D Human Pose and Shape Estimation. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS)*, 2024. 29

[141] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Synthetic training for accurate 3d human pose and shape estimation in the wild. In *BMVC*, 2020. 35

[142] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Hierarchical kinematic probability distributions for 3d human shape and pose estimation from images in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 11

[143] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Hierarchical kinematic probability distributions for 3d human shape and pose estimation from images in the wild. arXiv preprint / unpublished, 2023. 29

[144] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Humaniflow: Ancestor-conditioned normalising flows on so(3) manifolds for human pose and shape distribution estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 10, 11, 14, 20

[145] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Humaniflow: Ancestor-conditioned normalising flows on so(3) manifolds for human pose and shape distribution estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 29

[146] Kenneth Shaw, Shikhar Bahl, and Deepak Pathak. Videodex: Learning dexterity from internet videos. In *6th Conference on Robot Learning (CoRL)*, 2022. 18

[147] Xiaolong Shen, Zongxin Yang, Xiaohan Wang, Jianxin Ma, Chang Zhou, and Yi Yang. Global-to-local modeling for video-based 3d human pose and shape estimation. In *CVPR*, 2023. 13

[148] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1):4–27, 2010. 9, 34

[149] Dae-Young Song, HeeKyung Lee, Jeongil Seo, and Donghyeon Cho. Difu: Depth-guided implicit function for clothed human reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2023. 4

[150] Jiankai Sun, Bolei Zhou, Michael J. Black, and Arjun Chandrasekaran. Locate: End-to-end localization of actions in 3d with transformers. *arXiv preprint*, 2022. 17

[151] Qingping Sun, Yanjun Wang, Ailing Zeng, Wanqi Yin, Chen Wei, Wenjia Wang, Haiyi Mei, Chi-Sing Leung, Ziwei Liu, Lei Yang, and Zhongang Cai. Aios: All-in-one-stage expressive human pose and shape estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 13, 14

[152] Omid Taheri, Yi Zhou, Dimitrios Tzionas, Yang Zhou, Duygu Ceylan, Soren Pirk, and Michael J. Black. GRIP: Generating interaction poses using spatial cues and latent consistency. In *Proceedings of the IEEE/CVF International Conference on 3D Vision (3DV)*, 2024. 13, 19, 20, 34

[153] Julian Tanke, Oh-Hun Kwon, Felix B. Mueller, Andreas Doering, and Juergen Gall. Humans in kitchens: A dataset for multi-person human motion forecasting with scene context. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023. 18

[154] Guy Tevet. MotionCLIP: Exposing human motion generation to CLIP space. *European Conference on Computer Vision (ECCV)*, 2022. 18, 19

[155] Yating Tian, Hongwen Zhang, Yebin Liu, and Limin Wang. Recovering 3D Human Mesh From Monocular Images: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):15406–15425, 2023. 2, 10

[156] Li Tianye, Bolkart Timo, J. Black Michael, Li Hao, and

Romero J. Learning a model of facial shape and expression from 4d scans. *ACM Transactions on Graphics*, 2017. 8

[157] Andrea C Tricco, Jesmin Antony, Wasifa Zarin, Lisa Strifler, Marco Ghassemi, John Ivory, Laure Perrier, Brian Hutton, David Moher, and Sharon E Straus. A scoping review of rapid review methods. *BMC medicine*, 13:1–15, 2015. 2

[158] Shashank Tripathi, Agniv Chatterjee, Jean-Claude Passy, Hongwei Yi, Dimitrios Tzionas, and Michael J Black. Deco: Dense estimation of 3d human-scene contact in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8001–8013, 2023. 31

[159] Shashank Tripathi, Lea Müller, Chun-Hao P. Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. 3d human pose estimation via intuitive physics. In *CVPR*, 2023. 29

[160] Shashank Tripathi, Lea Müller, Chun-Hao P Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. 3d human pose estimation via intuitive physics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4713–4725, 2023. 32

[161] Neel Trivedi, Anirudh Thatipelli, and Ravi Kiran Sarvadevabhatla. Ntu-x: An enhanced large-scale dataset for improving pose-based recognition of subtle human actions. In *Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*. ACM, 2021. 19

[162] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 109–117, 2017. 14

[163] Abbhinav Venkat, Chaitanya Patel, Yudhik Agrawal, and Avinash Sharma. Humanmeshnet: Polygonal mesh recovery of humans. *arXiv preprint arXiv:1907.12555*, 2019. 5

[164] Noranart Vesdapunt, Mitch Rundle, HsiangTao Wu, and Baoyuan Wang. Jnr: Joint-based neural rig representation for compact 3d face modeling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 389–405. Springer, 2020. 8

[165] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, pages 601–617, 2018. 9, 14, 15, 32

[166] Jionghao Wang, Yuan Liu, Zhiyang Dou, Zhengming Yu, Yongqing Liang, Cheng Lin, Rong Xie, Li Song, Xin Li, and Wenping Wang. Disentangled clothed avatar generation from text descriptions. In *European Conference on Computer Vision*, pages 381–401. Springer, 2025. 8

[167] Jinbao Wang, Shujie Tan, Xiantong Zhen, Shuo Xu, Feng Zheng, Zhenyu He, and Ling Shao. Deep 3d human pose estimation: A review. *Computer Vision and Image Understanding*, 210:103225, 2021. 2

[168] Weiqiang Wang, Xuefei Zhe, Qiuhong Ke, Di Kang, Tingguang Li, Ruizhi Chen, and Linchao Bao. NEURAL MARIONETTE: A Transformer-based Multi-action Human Motion Synthesis System. *arXiv preprint arXiv:2209.13204*, 2023. 18, 19, 20

[169] Wen-Li Wei, Jen-Chun Lin, Tyng-Luh Liu, and Hong-Yuan Mark Liao. Capturing humans in motion: Temporal-attentive 3d human pose and shape estimation from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 13

[170] Yiu-Bun Wu, Bin Liu, Xiuping Liu, Xiuping Li, and Charlie C. L. Wang. Data-driven human modeling by sparse representation. *Computer-Aided Design*, 2020. 5

[171] Stefanie Wuhrer, Chang Shu, and Pengcheng Xi. Posture-invariant statistical shape analysis using laplace operator. *Computers & Graphics*, 36(5):410–416, 2012. 7

[172] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. Econ: Explicit clothed humans optimized via normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2023. 5

[173] Lingwei Xu, Huajun Cao, Miao Li, and Bo Chen. ReGenNet: Generating human motions via dual regression of pose and shape. *arXiv preprint arXiv:2401.01234*, 2024. 17, 18

[174] Liang Xu, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. InterGen: Diffusion-based multi-human motion generation under complex interactions. *arXiv preprint arXiv:2304.05684*, 2023. 17

[175] Yinghao Xu, Yifan Wang, Alexander W. Bergman, Menglei Chai, Bolei Zhou, and Gordon Wetzstein. Efficient 3d articulated human generation with layered surface volumes. *arXiv preprint arXiv:2307.05462*, 2023. 5

[176] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 12

[177] Ming Yan, Xin Wang, Yudi Dai, Siqi Shen, Chenglu Wen, Lan Xu, Yuexin Ma, and Cheng Wang. CIMI4D: A Large Multimodal Climbing Motion Dataset Under Human-Scene Interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 18

[178] Kaibing Yang, Renshu Gu, Maoyu Wang, Masahiro Toyoura, and Gang Xu. LASOR: Learning accurate 3d human pose and shape via synthetic occlusion-aware data and neural mesh rendering. *IEEE Transactions on Image Processing*, 31:1938–1948, 2022. 10, 11

[179] Kaibing Yang, Renshu Gu, Maoyu Wang, Masahiro Toyoura, and Gang Xu. Lasor: Learning accurate 3d human pose and shape via synthetic occlusion-aware data and neural mesh rendering. *IEEE Transactions on Image Processing*, 31:1938–1948, 2022. 29

[180] Yipin Yang, Yao Yu, Yu Zhou, Sidan Du, James Davis, and Ruigang Yang. Semantic parametric reshaping of

human body models. In *2014 2nd International Conference on 3D Vision*, volume 2, pages 41–48. IEEE, 2014. 5

[181] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating holistic 3d human motion from speech. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 469–480, 2023. 36

[182] Wanqi Yin, Zhongang Cai, Ruisi Wang, Fanzhou Wang, Chen Wei, Haiyi Mei, Weiye Xiao, Zhitao Yang, Qingping Sun, Atsushi Yamashita, et al. Whac: World-grounded humans and cameras. In *European Conference on Computer Vision*, pages 20–37. Springer, 2024. 37

[183] Wanqi Yin, Zhongang Cai, Ruisi Wang, Ailing Zeng, Chen Wei, Qingping Sun, Haiyi Mei, Yanjun Wang, Hui En Pang, Mingyuan Zhang, et al. Smplest-x: Ultimate scaling for expressive human pose and shape estimation. *arXiv preprint arXiv:2501.09782*, 2025. 14, 29

[184] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17016–17027, 2023. 31

[185] Zhengdi Yu, Shaoli Huang, Yongkang Cheng, and Tolga Birdal. Signavatars: A large-scale 3d sign language holistic motion dataset and benchmark. In *European Conference on Computer Vision*, pages 1–19. Springer, 2024. 36

[186] Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Neural descent for visual 3d human pose and shape. In *Unspecified (not stated in the provided text)*, 2020. No DOI/arXiv info provided. 29

[187] Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Neural descent for visual 3d human pose and shape. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 9, 12

[188] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes–the importance of multiple scene constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2148–2157, 2018. 12

[189] Chuanlei Zhang, Lixin Liu, Minda Yao, Wei Chen, Dufeng Chen, and Yuliang Wu. HSiPu2: A New Human Physical Fitness Action Dataset for Recognition and 3D Reconstruction Evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021. 18

[190] Harry Zhang and Luca Carlone. CUPS: Improving Human Pose-Shape Estimators with Conformalized Deep Uncertainty. *arXiv preprint arXiv:24xx.xxxxx*, 2024. 11

[191] Hongwen Zhang, Siyou Lin, Ruizhi Shao, Yuxiang Zhang, Zerong Zheng, Han Huang, Yandong Guo, and Yebin Liu. Closet: Modeling clothed humans on continuous surface with explicit template decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2023. 5

[192] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. PyMAF: 3D Human Pose and Shape Regression with Pyramidal Mesh Alignment Feedback Loop. In *Unpublished/ArXiv*, 2021. 29

[193] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J. Black, and Siyu Tang. PLACE: Proximity learning of articulation and contact in 3d environments. In *International Conference on 3D Vision (3DV)*, pages 642–651, 2020. 17

[194] Zheng Zhang, Jeff M Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, et al. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3438–3446, 2016. 8

[195] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3170–3184, 2022. 5

[196] Kun Zhou, Xiaoguang Han, Nianjuan Jiang, Kui Jia, and Jiangbo Lu. HEMlets PoSh: Learning Part-Centric Heatmap Triplets for 3D Human Pose and Shape Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2999–3014, 2022. 11

[197] Xiaoyu Zhu, Celso M. de Melo, and Alexander Hauptmann. Leveraging body pose estimation for gesture recognition in human-robot interaction using synthetic data. In *Synthetic Data for Artificial Intelligence and Machine Learning: Tools, Techniques, and Applications*, volume 12529, page 125290Z. SPIE, 2023. 18

[198] Xiaoyu Zhu, John Smith, and Jane Doe. Stmt: Spatio-temporal motion transformer for 3d pose refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 17, 19

[199] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *ICCV*, pages 813–822, 2019. 8

[200] Shihao Zou, Chuan Guo, Xinxin Zuo, Sen Wang, Pengyu Wang, Xiaoqin Hu, Shoushun Chen, Minglun Gong, and Li Cheng. Eventhpe: Event-based 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 10, 11

[201] Shihao Zou, Xinxin Zuo, Sen Wang, Yiming Qian, Chuan Guo, and Li Cheng. Human pose and shape estimation from single polarization images. *IEEE Transactions on Multimedia*, 25:3560–3572, 2022. 36

[202] Shihao Zou, Xinxin Zuo, Sen Wang, Yiming Qian, Chuan Guo, and Li Cheng. Human Pose and Shape Estimation From Single Polarization Images. *IEEE Transactions on Multimedia*, 25:3560–3572, 2023. 10, 11, 20

[203] Silvia Zuffi and Michael J. Black. The stitched puppet: A graphical model of 3d human shape and pose. In *CVPR*. IEEE, 2015. 5

A. Comparisons of 3D Pose Estimation Methods

Table 1. Quantitative comparison of selected single-view, single-person RGB 3D HPE on **Human3.6M** (61). MPJPE = mean per-joint position error (in mm), PA-MPJPE = Procrustes-aligned MPJPE.

Method	Year	MPJPE	PA-MPJPE	Project Link
SMPLify (20)	2016	82.3	82.3	–
HMR (71)	2018	87.9	56.8	–
Arnab et al. (10)	2019	77.8	54.3	–
SPIN (80)	2019	62.5	41.1	link
HUND (186)	2020	69.5	52.6	–
TexturePose (118)	2020	51.3	49.7	link
DSR (40)	2021	60.9	40.3	link
ProHMR (82)	2021	–	41.2	link
EFT (69)	2021	–	43.7	link
PyMAF (192)	2021	57.7	40.5	link
HybrIK (86)	2021	54.4	34.5	link
CLIFF (89)	2022	47.1	32.7	–
IPMAN-R (159)	2023	60.7	41.1	link
MVSPIN (91)	2023	64.8	43.8	link
HybrIK-X (R34) (85)	2023	55.3	33.7	link
HybrIK-X (W48) (85)	2023	47.0	29.8	link
Neural Localizer Fields (140)	2024	∼40.0	∼29.0	link
HYRE (87)	2025	64.8	42.1	–

Table 2. Quantitative comparison of selected single-view, single-person RGB 3D HPE on **3DPW**. MPJPE = mean per-joint position error (in mm), PA-MPJPE = Procrustes-aligned MPJPE. Some methods reported also PVE (per-vertex error).

Method	Year	MPJPE	PA-MPJPE	PVE	Project Link
HMR (71)	2018	130.0	81.3	–	–
Arnab et al. (10)	2019	–	72.2	–	–
SPIN (80)	2019	96.9	59.2	116.4	link
HUND (186)	2020	81.4	57.5	–	–
DSR (40)	2021	85.7	51.7	99.5	link
EFT (69)	2021	–	54.2	–	link
HybrIK (86)	2021	80.0	48.8	94.5	link
ProHMR (82)	2021	–	59.8	–	link
PyMAF (192)	2021	92.8	58.9	110.1	link
CLIFF (89)	2022	69.0	43.0	81.2	–
LASOR (179)	2022	–	52.4	108	link
MVSPIN (91)	2023	93.4	58.6	–	link
SMPLer-X-L32 (22)	2023	75.2	50.5	–	link
Sengupta (Hier. Kinematic) (143)	2023	84.7	59.2	–	–
HuManiFlow (145)	2023	83.9	53.4	–	–
HybrIK-X (85)	2023	71.6	41.8	82.3	link
OSX (94)	2023	74.7	45.1	–	link
Neural Localizer Fields (140)	2024	**54.1**	**33.7**	**63.7**	link
POCO (41)	2024	70.9	43.3	84.6	link
HYRE (87)	2024	59.8	42.1	80.5	–
SMPLest-X-H40 (183)	2025	76.0	46.5	–	link

B. Pose Estimation Datasets

Table 3. Comparison of popular datasets for (primarily) 3D human pose estimation.

Dataset	Type	Acquisition Method	Modalities	Size	Scene	Annotations	Applications	Accuracy	Splits	Contribution	Link
SURREAL (70)	Syn.	SMPL-based avatars posed by CMU MoCap data (fitted via MoSh)	RGB (synthetic), Depth, Segmentations, Optical Flow, Normals	>6M frames; single-person; highly varied poses, shapes, clothing, lighting, indoor backgrounds (LSUN)	Single person over random indoor image background; minor self-occlusions	2D/3D joints, SMPL parameters, segmentation, depth, normals, flow	Single-person tasks: pose estimation, segmentation, depth	Perfect synthetic GT from SMPL fits (very accurate)	Recommended train/test split: 115 training subjects, 30 test subjects	Large-scale synthetic dataset with pixel-wise ground truth; effective for pre-training	http://www.di.ens.fr/willow/research/surreal/
SVIRO (35)	Syn.	Synthetic rendering of vehicle interiors (photo-realistic 3D models); fixed textures, lighting	RGB, simulated IR (grayscale), Depth, Instance segmentation, 2D pose	10 vehicle types; 16k training sceneries, 4k test; occupant classes (adult, child seat, empty), small objects	In-vehicle passenger compartments; partial occlusions from seats and child seats	2D bounding boxes, 2D keypoints, instance masks, depth maps, IR images	Rear-seat occupant detection/classification; occupant pose	Automatically generated from 3D pipeline; consistent ground truth	Split: 16k train, 4k test	Realistic synthetic car interior domain; multiple sensor modalities	https://sviro.kl.dfki.de
MoVi (49)	Real	Marker-based optical motion capture (Qualisys) + multi-view cameras + IMU	Multi-view RGB (4 cameras: 2 stationary Grasshopper, 2 handheld iPhones), 3D marker data, IMU	90 participants; ∼9h MoCap, 17h video, 6.6h IMU; 21 actions	Indoor mocap lab; single-person; partial occlusions from body/clothing	3D skeletons (Visual3D), SMPL/DMPL shape parameters (MoSh++), 2D frames, IMU measurements	Full-body pose/shape estimation, action recognition, motion modeling, gait analysis	∼0.8 cm reprojection error for markers	No formal train/test split; 5 capture sequences (F, S1, S2, I1, I2)	Large multi-modal dataset with synchronized video, IMU, and body shape for the same 90 actors	Research only; https://www.biomotionlab.ca/movi/
ARCTIC (44)	Real	Marker-based MoCap (54 Vicon cams) + 8 static RGB + 1 egocentric RGB	Multi-view RGB (2800×2000@30fps, synchronized	2.1M frames, 10 subjects (5f/5m), 339 sequences, 11 articulated objects	Indoor lab; bimanual dexterous manipulation; heavy occlusions mitigated by multi-view	3D hand (MANO) + full-body (SMPL-X), articulated object meshes, dynamic contact	Dexterous 2-hand motion + articulated-object tasks: reconstruction, contact/interaction fields	∼mm-level from Vicon markers + model fitting (MoSh++)	Split by subjects: 8 train, 1 val (male), 1 test (female); protocols for allocentric/egocentric	First large-scale dataset of bimanual manipulation with articulated objects; detailed contact, full SMPL-X	https://arctic.is.tue.mpg.de
BEHAVE (16)	Real	Multi-view Kinect (4 cams) in natural indoor env.; markerless approach using point-cloud fits	RGB-D, 3D object scans, contact labels	∼15k frames, 8 subjects, 20 objects, 5 indoor locations	Indoor, heavy occlusions from human-object interactions	SMPL pose+shape, object mesh pose, segmentation, contact annotations	Full-body human-object 3D tracking, contact analysis, shape/pose estimation	Pseudo-GT from Kinect PC: ∼1.80,cm Chamfer (human), 2.42,cm (object)	10.7k frames (train), 4.5k (test)	First large-scale real dataset of dynamic human-object interactions with explicit 3D contacts	http://virtualhumans.mpi-inf.mpg.de/behave

Table 3 (continued)

Dataset	Type	Acquisition Method	Modalities	Size	Scene	Annotations	Applications	Accuracy	Splits	Contribution	Link
CHI3D (45)	Real	Marker-based MoCap with 10 cameras + 4 synced RGB cams. One subject wears markers (ground-truth), second subject pseudo-labeled via triangulated 2D detections	Multi-view RGB + MoCap	631 seqs, 2,525 contact events, 728,664 3D skeletons from 6 people (3 pairs)	Indoor lab; close interactions (hug, handshake, push, etc.) with partial occlusions	3D skeletons (one subject via markers, other via triangulation), region/facet-level contact annotations	3D pose/shape reconstruction focusing on human-human contact	Sub-cm error for the marked subject, some triangulation noise for the second	No formal train/test split	First large-scale two-person contact dataset with precise 3D, 8 interaction classes	http://vision.imar.ro/ci3d
FlickrCI3D (45)	Real	Single-view images from Flickr/YFCC100M, manual annotation of contact regions	Single-view RGB	11,216 images, 14,081 contact pairs, 81k facet-level correspondences, 138k contact regions	In-the-wild (indoor/outdoor), varied backgrounds, frequent occlusions	Binary contact labels (contact/no), region-based and facet-level contact signatures, 2D skeletons	Studying subtle human-human contact in unconstrained scenes, 3D reconstruction from monocular	Manual labeling (single-view) implies potential ambiguity in occluded contacts	85%/7.5%/7.5% train/val/test splits	Largest in-the-wild contact labeling dataset; detailed region-level contact ground truth	http://vision.imar.ro/ci3d
Fit3D (47)	Real	Marker-based optical MoCap (VICON with 12 cameras) + 4 synchronized RGB cameras	Multi-view RGB, 3D motion capture, 3D scans	Over 3 million images, 2.96M mocap frames, 13 subjects (varied height/weight), 37 exercises (5+ reps each)	Indoor controlled environment (VICON setup), gym equipment (dumbbells, barbell), partial self-occlusions	3D joints (VICON), GHUM body model parameters, 3D body scans, manual repetition timestamps	Full-body fitness pose/shape estimation, repetition counting	Marker-based MoCap (sub-cm precision on markers)	Train/val: 10 subjects (2.28M images), Test: 3 subjects (0.69M images)	Large-scale fitness-specific dataset; real-time feedback; covers major muscle groups	http://vision.imar.ro/fit3d
Hi4D (184)	Real	Markerless multi-view volumetric capture of close-contact interactions	Multi-view RGB, 4D textured scans	20 subject pairs, 100 sequences, 11K+ frames (¿6K with contact)	Indoor multi-person scenarios, heavy occlusions	2D/3D instance masks, SMPL fits, vertex-level contact	Multi-person 3D pose/shape estimation, contact analysis, geometry reconstruction	High-fidelity registrations (no explicit numeric error reported)	No formal train/test split	First 4D dataset with segmentations + vertex-level contact under close interaction	https://ait.ethz.ch/Hi4D
DAMON (158)	Real	Crowdsourced "painting" of 3D contact on SMPL from in-the-wild single-view images	Single-view RGB (sourced from V-COCO and HAKE via HOT)	5,522 labeled images; 84 object labels, 24 body parts	Unconstrained everyday scenes (in-door/outdoor), partial occlusions, diverse human-object contacts	Dense vertex-level 3D contact (SMPL), contact type (scene-supported vs. human-supported), object categories	3D contact detection; full-body contact reasoning in the wild	Quality-controlled vertex annotations via manual curation; SMPL from CLIFF	No formal train/test split	First large-scale in-the-wild dataset with dense full-body contact labels; beneficial for contact-based pose/shape methods	Research only; https://deco.is.tue.mpg.de

Table 3 (continued)

Dataset	Type	Acquisition Method	Modalities	Size	Scene	Annotations	Applications	Accuracy	Splits	Contribution	Link
CMU Panoptic Studio (67)	Real	Markerless dome setup with 480 synced VGA cams, sometimes Kinect data	Multi-view RGB (480 cams), partial Kinect Depth	5 vignettes (Ultimatum, Prisoner's Dilemma, Mafia, Haggling, 007-bang) with 3–8 participants, 25 Hz	Indoor dome (5.49 m radius), multi-person interactions, heavy interpersonal occlusions	15-keypoint 3D skeletons, fused from multi-view 2D detections, some Kinect references	Full-body multi-person social interaction analysis	∼3–6 cm average 3D error in evaluations	No formal train/test split	Very large multi-view coverage for natural group interactions; robust to occlusions	http://www.cs.cmu.edu/~hanbyulj/panoptic-studio
DHP19 (24)	Real	4 synchronized event cameras (DAVIS) + Vicon MoCap	Event streams (microsecond), 2D grayscale (DAVIS), 3D pose (Vicon)	17 subjects, 33 movements each repeated 10 times; ∼87k DAVIS frames/camera	Indoor therapy lab (2×2×2 m^3), partial self-occlusions	3D joint positions (13), 2D projections in event cam	Low-latency 3D pose from event data; rehab, VR, real-time	Vicon sub-mm precision; final ≈8 cm error if approximating joint centers	12 subjects train/val, 5 test	First event-based 3D pose dataset, capturing 33 motion types	https://sites.google.com/view/dhp19
3DPW (165)	Real	IMUs on limbs + single hand-held smartphone camera; jointly optimized (VIP method)	Single-view RGB (smartphone) + inertial sensors	60 sequences, ¿51k frames, 7 actors, 18 outfits	Unconstrained outdoor/indoor scenes, sometimes multi-person, moving camera, significant occlusions	3D joints, SMPL parameters (pose + shape), 3D scans	Full-body 3D pose and shape in real-world (single-/two-person) activities	Validated at 26 mm error on TotalCapture; similar accuracy implied for 3DPW	No formal train/test split	First dataset with accurate 3D pose+shape "in the wild" (IMU + single camera), diverse everyday activities, occlusions	Research only; http://virtualhumans.mpi-inf.mpg.de/3DPW
Motion-X (93)	Real (plus some animation/game footage)	Markerless capture from single-/multi-view videos; hierarchical 2D detection with ViT + SMPL-X fitting	Single-/multi-view RGB, pseudo-3D SMPL-X	81.1K clips, 15.6M frames (∼144 hrs); indoor/outdoor, diverse poses	Broad variety of real scenes (indoor/outdoor), frequent occlusions, complex actions	3D SMPL-X (body, hands, face), 2D keypoints, sequence-level & frame-level text	Text-driven motion generation, 3D whole-body mesh recovery, motion understanding	Pseudo-GT refined by multi-step optimization; e.g. ∼19.7 mm MPVPE (EHF, best config)	80% train, 5% val, 15% test; standard text-to-motion benchmarks (FID, R-Precision, etc.)	Largest expressive whole-body motion dataset (body+hands+face) with semantic & fine-grained pose text	https://motion-x-dataset.github.io
HumanSC3D (46)	Real	Marker-based motion capture + multi-view cameras + 3D body scanner	4 synchronized RGB cameras, 3D marker data, full body scan	1032 sequences, over million frames, 5058 contact events, 6 subjects	Indoor environment (standing, sitting on floor/chair); partial occlusions	3D joints, 3D shape from scans, region-level self-contact, 2D contact localization	Self-contact modeling, 3D pose/shape estimation, face-touch detection	Sub-cm for body scans, mm-level marker-based accuracy	No formal train/test split	First large-scale 3D self-contact dataset (marker-based GT + multi-view + annotated contact)	http://vision.imar.ro/sc3d
MoYo (160)	Real	Multi-view camera system (8 static RGB cams) + synchronized optical MoCap + pressure mat	Multi-view RGB (4K), MoCap, pressure sensor	1 subject, 200 complex yoga poses, ∼ 1.75 M frames total	Indoor environment, challenging self-occlusions, floor interaction	SMPL-X meshes, ground-truth CoM, pressure maps, body-floor contact	Single-person, static complex poses, stable pose analysis	∼ 53 mm CoM discrepancy vs. Vicon, Pressure IoU ∼ 0.32, CoP err. ∼ 57 mm	No formal train/test split	Unique real dataset with extreme yoga poses, synchronized pressure + CoM	https://ipman.is.tue.mpg.de

Table 3 (continued)

Dataset	Type	Acquisition Method	Modalities	Size	Scene	Annotations	Applications	Accuracy	Splits	Contribution	Link
TUCH (108)	Real	3D scans + refined AMASS poses + crowd-sourced single-view images mimicking self-contact poses (SMPLify-XMC)	Single-view RGB (in-the-wild), 3D body scans	3DCP: 1653 contact poses; MTP: 3731 images from 148 subjects; DSC: 30k images with discrete contact	Mixed indoor scans plus in-the-wild backgrounds; self-contact emphasis; moderate occlusions	SMPL-X parameters (body, hands, face), discrete contact labels, pseudo-GT fits	Full-body 3D pose/shape estimation with explicit self-contact	Pseudo GT via SMPLify-XMC; few-cm-level accuracy (shown via 3DPW metrics)	No formal train/test split	First dataset focusing on self-contact in the wild; "Mimic The Pose" crowd-sourcing and DSC for discrete contact	https://tuch.is.tue.mpg.de
RICH (60)	Real	Markerless multi-view capture (6–8 cameras at 4K) + high-resolution laser scans	Multi-view RGB at 4K, scene mesh scans	22 subjects, 5 scenes, 134 sequences, ∼85k frames, 540k images total	Indoor/outdoor, multi-person, occlusions, dynamic backgrounds	SMPL-X parameters (pose, shape), 3D scene contact (dense per-vertex)	3D pose/shape estimation, dense human-scene contact, monocular or multi-view	High-accuracy pseudo-GT from multi-view fits, laser scans for precise contact	57 training, 27 validation, 50 testing sequences; specialized subsets for unseen subjects/scenes/interactions	First real large-scale dataset with outdoor scanned scenes and dense body-scene contact labels	Research only; https://rich.is.tue.mpg.de
SLOPER4D (36)	Real	Head-mounted LiDAR + camera + 17 wearable IMUs; LiDAR-inertial SLAM for large-scale scene mapping	LiDAR point clouds, single-view RGB video, IMU data, global scene mesh	15 sequences, 12 subjects, 10 urban scenes; ∼100k LiDAR frames, 300k video frames, 500k IMU frames; area up to 30k m^2	Large-scale outdoor urban, partial occlusions, dynamic backgrounds	2D keypoints, 2D bboxes, 3D SMPL parameters (pose+shape), global translations, scene reconstructions	Global 3D human pose estimation, camera-/LiDAR-based 3D HPE, scene-aware motion capture	∼cm-level after multi-sensor optimization (mesh-to-points, scene contact)	11 sequences train, 4 test; MPJPE, PA-MPJPE, ATE, RPE for benchmarks	First large-scale urban 4D pose dataset with multi-modal capture (LiDAR + RGB + IMUs) for GHPE	http://www.lidarhumanmotion.net/sloper4d/
SMPL-X EHF (116)	Real	4D body scans of a single subject, then SMPL-X is fitted for pseudo-GT	Single-view RGB images (generated from scanning)	100 images, 1 subject, variety of poses, gestures, facial expressions	Controlled scanning environment; minimal occlusions	SMPL-X parameters (body, hands, face), 3D joints, surface mesh	Benchmark for body, hands, face from single RGB	High-fidelity 'pseudo' GT from 4D scans	No formal train/test split	Rare dataset with face and hand articulation in 3D from real scans	Research only; https://smpl-x.is.tue.mpg.de/
EMDB (74)	Real	Body-worn EM sensors + handheld iPhone (RGB-D + ARKit) in the wild	Single-view iPhone RGB-D, up to 12 EM sensors on body, global camera trajectory	81 sequences, 58 minutes, 10 participants (in-door/outdoor), 105k frames	Free-moving camera, single-person, diverse real locations	SMPL pose/shape, 3D joints, global body/camera trajectories	Full-body 3D pose/shape with global motion in unconstrained environments	≈2.3 cm pose error vs. multi-view studio reference; 5.1 cm global trajectory error	No formal train/test split	First dataset fusing EM sensors + smartphone for in-the-wild global 3D pose	https://ait.ethz.ch/emdb
JTA (43)	Syn.	Photorealistic videogame environment (GTA V) with custom mod controlling scenes	Single-view RGB (1080p@30 fps)	512 HD videos, each 30 s @30 fps (total 460k frames); up to 60 people/frame, 10 M poses	Urban settings, heavy crowd-ing/occlusions, varied illumina-tion/viewpoints	2D/3D joints (14), occlusion/self-occlusion flags, unique ID for each pedestrian	Multi-person 2D/3D pose estimation, short-term tracking, occlusion reasoning	Near-perfect engine-based 2D/3D ground truth, fully automated	256 videos for training, 256 for testing	Largest synthetic urban pose dataset with 10 M annotations; explicit occlusions	http://imagelab.ing.unimore.it/jta

Table 3 (continued)

Dataset	Type	Acquisition Method	Modalities	Size	Scene	Annotations	Applications	Accuracy	Splits	Contribution	Link
GRAB (152)	Real	Marker-based Vicon (54 cams) capturing human-object interactions	3D motion capture (IR), no standard RGB; 99 markers on body, 8+ on object	1,334 sequences, 10 people (5m/5f), 51 objects, 1.6M frames	Controlled studio, single-person object manipulation, moderate occlusions	SMPL-X, 6-DoF object poses, contact labels	Human-object full-body grasping, including face/hands contact	∼3–4 mm average error (MoSh++), robust tracking	No formal train/test split; example used 41 objects for train, 4 val, 6 test	Detailed multi-part contacts, real 3D-printed objects, per-vertex contact	Research only; https://grab.is.tue.mpg.de
HumanEva (148)	Real	Marker-based Vicon + 7 synchronized cameras (4 grayscale, 3 color) at 60 Hz	Multi-view RGB + grayscale, 3D MoCap	∼50k frames (video+MoCap sync), 4 subjects, 6 actions	Indoor lab, single-person, moderate complexity	3D marker positions, optional 2D projections, silhouettes	2D/3D pose evaluation, single-person motion tracking	Vicon-level accuracy (possible marker slip on clothing)	Standard train/val/test; test GT withheld for online eval	Early classic multi-view 3D dataset	Research only; http://humaneva.is.tue.mpg.de/
MuCo-3DHP / MuPoTS-3D (106)	Mixed	**MuCo-3DHP:** composites single-person scans from MPI-INF-3DHP into multi-person images **MuPoTS-3D:** newly recorded multi-view markerless MoCap (indoor/outdoor)	RGB images, multi-person, 3D joints from markerless MoCap	**MuCo**: 400k+ synthetic composites; 1–4 people **MuPoTS**: 20 real seqs, 8k frames, up to 3 persons	Occlusions from multiple people; MuPoTS includes real backgrounds indoor/outdoor	3D joints, 2D keypoints, part affinity, occlusion labels	Multi-person 3D pose from monocular images	Markerless capture accuracy, a few cm error	MuCo for training, MuPoTS for testing	Large multi-person 3D dataset; real test set w. heavy occlusion	https://arxiv.org/abs/1712.03453
COCO (Keypoints) (97)	Real	Photos from Flickr, manually labeled via Amazon Mechanical Turk	Single-view RGB	328k images, 2.5M labeled instances (91 categories), ∼7.7 objects/image	Everyday scenes, varied backgrounds, partial occlusions	2D keypoints for humans (17 joints), bounding boxes, instance segmentations	Object detection / segmentation, 2D keypoint pose	Manually annotated, consistent bounding boxes and masks	Train (164k), val (82k), test (82k)	Large-scale everyday object dataset, widely used, includes 2D pose	https://cocodataset.org/
PoseTrack (63)	Real	Unconstrained single-view videos (e.g. from YouTube) with manual annotation	Single-view RGB	60 videos (each 41–151 frames), 2–16 persons/video, ∼16k total poses	Indoor/outdoor, multi-person, frequent occlusions, large scale variations	14 keypoints (2D) + head bounding box, occlusion flags, unique person IDs	Multi-person 2D pose estimation and tracking	Manual labeling; evaluated with PCKh (20% threshold), MOTA, MOTP, mAP	30 videos train, 30 test; new protocol (no assumption on person count/scale)	First dataset for joint multi-person pose estimation & tracking in unconstrained videos	http://pages.iai.uni-bonn.de/iqbal_umar/PoseTrack/
MPII (Extended Pose) (8)	Real	YouTube frames covering diverse activities, no specialized hardware	Single-view RGB	>40k annotated persons in 24,920 frames from 3,913 videos, 491 activities	Everyday indoor/outdoor, varied backgrounds, partial occlusions	2D joints, partial face keypoints, head bounding box, occlusion flags, activity labels	Single-person 2D pose estimation, wide activity coverage	Manually annotated; typical 2D labeling uncertainty	∼3/4 train (28,821 persons), 1/4 test (11,701)	Adjacent frames provided for motion context, large variety of daily poses	http://human-pose.mpi-inf.mpg.de

Table 3 (continued)

Dataset	Type	Acquisition Method	Modalities	Size	Scene	Annotations	Applications	Accuracy	Splits	Contribution	Link
UBody (95)	Real	Single-view real-life videos (no specialized mocap)	Single-view RGB	>1.05M frames, 15 diverse real-life scenes (talk shows, sign language, etc.)	Heavy truncations on upper body, partial occlusions from objects/subtitles, varied lighting	2D whole-body keypoints + 3D SMPL-X (body, hands, face)	Expressive upper-body analysis (gestures, facial expressions), real-life scenarios	High-quality pseudo-GT from a specialized annotation pipeline (better than OpenPose/MediaPip	Train/test splits with intra-scene and inter-scene protocols	Large-scale real dataset emphasizing hands/face in natural daily scenes	https://osx-ubody.github.io
AMASS (104)	Real	Unifies 15 marker-based optical mocap datasets; re-fitted with MoSh++	3D body (SMPL/DMPL/H) parameters per frame; no raw images	>40 hours motion, 344 subjects, 11k+ sequences, broad motion variety	Mostly lab-based, single-person, no complex environment	SMPL pose/shape, including soft-tissue (DMPL), 3D joints	Motion modeling, synthesizing new data, training dynamic pose networks	∼7–8 mm mesh error vs. 4D scans	No formal train/test split	Largest unified MoCap-to-SMPL archive, wide variety of motions	Research only; https://amass.is.tue.mpg.de/
AGORA (128)	Syn.	3D scans of real humans registered to SMPL-X, rendered in Unreal Engine	Single-view RGB (4K), multiple people, segmentation, SMPL-X	∼19k images total, 5–15 people each, 4240 scans (incl. 257 children)	Multi-person scenes, occlusions, photorealistic clothing/hair	SMPL-X parameters (body, hands, face), 3D joints, masks	3D pose/shape estimation for multi-person, includes children	≈5 mm scan-to-SMPLX fitting error	Train/val with annotations; test GT withheld (online eval)	Unique child scans, multi-person, high-quality outfits, 4K resolution	https://agora.is.tue.mpg.de/
BEDLAM (18)	Syn.	Unreal Engine 5, motions from AMASS (104), clothes via CLO3D, hair via Character Creator	RGB (1280×720 at 30 fps), plus optional Depth/Segm.	∼380k frames, 1M bounding boxes, 271 body shapes, 111 outfits, up to 10 people/scene	95 HDRI + 8 3D scenes, multi-person, realistic garments, lighting changes	SMPL-X ground truth, 3D surfaces, depth, segmentation	3D pose/shape from single images, includes clothing dynamics, multi-person	Perfect synthetic annotation by design	Train (75%), val (20%), test (5%) with withheld GT, eval server	Large variety of shapes (incl. obese), physically simulated clothing, multi-person	Research only; https://bedlam.is.tue.mpg.de/
PROX (54)	Real	Kinect-One RGB-D + Structure Sensor (scanned indoor scenes); separate small Vicon-based subset	Single-view RGB-D (main), multi-camera marker-based MoCap subset (180 frames)	12 indoor scenes (bedrooms, living rooms, etc.), 20 subjects (4f/16m), ∼100k frames total	Real indoor rooms with furniture, partial occlusions from chairs, beds, objects	SMPL-X parameters (body, hands, face), 3D joints, pseudo-ground-truth from SMPLify-X + MoSh++ fitting	Full-body pose/shape in real indoor scenes, emphasizing contact with furniture	High fidelity from Kinect + fitted SMPL-X; a separate Vicon-based subset for GT	No formal train/test split; a 180-frame set (Vicon) used for evaluation	Unique real data with scene constraints (furniture interactions), 3D scans of environments	Research only; https://prox.is.tue.mpg.de
SSP-3D (141)	Real	Multi-frame SMPL fitting (Keypoint-RCNN + PointRend silhouettes) on Sports-1M frames	Single-view RGB	311 images, 62 subjects; wide body shape diversity (tightly-clothed sportspeople)	In-the-wild sports scenes, partial occlusions, varied lighting/backgrounds	SMPL (pose+shape), 2D/3D joints, silhouettes	Monocular 3D shape & pose estimation in diverse body shapes	Pseudo-GT from multi-frame optimization with forced shape consistency	No formal train/test split	Emphasizes body shape diversity and tight clothing for accurate shape labels	https://github.com/akashsengupta1997/STRAPS-3DHumanShapePose

Table 3 (continued)

Dataset	Type	Acquisition Method	Modalities	Size	Scene	Annotations	Applications	Accuracy	Splits	Contribution	Link
EgoBody (105)	Real	Multi-Azure Kinect + Microsoft HoloLens2 (egocentric), 2-person interactions	Multi-view RGB-D, plus egocentric RGB/D, eye gaze, hand tracking	125 sequences, 36 subjects (18m/18f), 15 indoor scenes, ∼220k frames 3rd-person, ∼200k ego frames	Indoor, frequent body truncations in first-person, object interactions	SMPL-X, 3D scene scans, 2D keypoints, gaze	Social interaction, 3D full-body from wearable device, AR/VR	Multi-step optimization + LEMO refinement; a few cm error	Train/val/test with subject separation	First large-scale egocentric + multi-view dataset with 3D ground truth	https://sanweiliti.github.io/egobody/egobody.html
Human3.6M (62)	Real	10 Vicon cameras @120 Hz + 4 DV cameras + ToF depth in a lab	Multi-view RGB, 3D MoCap, Depth, body scans	3.6M frames, 11 subjects, 15 everyday actions, up to 4 cameras	Indoor lab, single-person, moderate complexity, some objects (chair/table)	3D joint positions, 2D bounding boxes, silhouettes, full body scans	3D pose estimation, single-person baseline, widely used	Vicon ∼ mm-level error, strong GT	7 subjects for train/val, 4 for test; standard MPJPE, etc.; online eval	Very large, multi-modal, well-known benchmark in controlled lab	Research only; http://vision.imar.ro/human3.6m
PHSPD (201)	Real	One polarization camera + 3 or 5 Kinect v2 cameras; multi-view depth + iterative SMPL fitting	RGB-D, Polarization images, up to 4–6 cams	∼527k frames total, 21 subjects, 31 actions, 9.5 hours	Indoor, single-person, some objects, partial self-occlusions	SMPL (pose+shape), 2D/3D joints, segmentation, possibly normals	3D body shape/pose from polarization + depth, clothing detail	Multi-view depth alignment $\Longrightarrow$ cm-level shape accuracy	Typical subject-split (e.g. 12 train, 5 test), or v1/v2 with 117k test samples	First large-scale 3D dataset with polarization data for shape estimation	https://github.com/JimmyZou/PolarHumanPoseSha
TalkSHOW (181)	Real	Single-view in-the-wild videos, refined SMPLify-X ("SHOW") pipeline	Single-view RGB + audio	26.9h from 4 speakers, 30FPS, split into $\leq$10s clips	In-the-wild, diverse backgrounds, minimal occlusion	SMPL-X (face, body, hands), pseudo-GT	3D speech-driven motion (face/body/hands), style modeling	Improved SMPLify-X p-GT, no explicit numeric error given	80%-10%-10% train-val-test	First large-scale "holistic" dataset (face/body/hands) for speech-to-3D motion	https://talkshow.is.tue.mpg.de/
SignAvatars (185)	Real	Public SL videos (single-/multi-view), markerless fitting (SMPL-X) with biomechanical constraints	RGB; final 3D meshes (SMPL-X / MANO), 2D/3D keypoints	70k videos (8.34M frames), 153 signers, 117 h; isolated & continuous signing	Diverse online/studio clips, occlusions from co-articulation	SMPL-X (body/face/hands), MANO subset, 2D/3D joints	3D SL recognition/production, multi-lingual sign tasks	∼12.9 mm PA-MPVPE on EHF	Follows original splits + new 3D SLP benchmark	First large-scale multi-prompt 3D SL dataset (text, word, HamNoSys); robust auto-annotation	Research only; https://signavatars.github.io/
UP-3D (83)	Real	Single-view real images from LSP, LSP-extended, MPII, and FashionPose; SMPLify fitting + human validation of silhouettes	Single-view RGB	After curation: $\sim 7k$ training + $\sim 1.4k$ test images; merged from $\sim 27.6k$ initial fits	Single-person, varied indoor/outdoor sports/street scenes, partial occlusions	SMPL pose + shape, 91 dense 2D landmarks, classic 14-joint skeleton, 31-part segmentation	Full-body 3D pose/shape estimation, 2D landmark detection, semantic body-part segmentation	Human-verified SMPL fits; shown $\sim$ 1.6 mm improved MPJPE on Human3.6M vs. baseline	Retains original dataset splits (LSP, MPII, etc.)	Unifies multiple 2D sets into a holistic 3D dataset, enabling iterative self-improvement in 3D fitting	http://up.is.tuebingen.mpg.de/

Table 3 (continued)

Dataset	Type	Acquisition Method	Modalities	Size	Scene	Annotations	Applications	Accuracy	Splits	Contribution	Link
WHAC-A-Mole (182)	Syn.	Fully synthetic rendering with SMPL-XL parametric humans plus algorithmic cinematic camera motions (arc, push/pull, tracking, panning)	Single-view RGB; ground-truth SMPL-X (body, hands, face) and camera poses	2434 sequences (∼1.46M frames/crops), multi-person from AMASS, DLP-MoCap, DD100	Diverse virtual 3D environments, interactive motions, multi-person, varied occlusions	SMPL-X parameters, 3D joints, camera translation/rotation, track IDs, contact labels	World-grounded EHPS with full body and camera trajectories; multi-person interactive scenarios	Perfect synthetic GT from parametric modeling and known geometry	80% train, 20% test; standard MPJPE, PVE, trajectory metrics	Large-scale synthetic dataset bridging body and camera in world coordinates; realistic interactions	https://wqyin.github.io/projects/WHAC/
GTA-Human (23)	Syn.	Game engine (GTA-V) with custom mod controlling scenes; SMPLify on 3D keypoints	Single-view RGB videos (30 fps), optional depth/semantic	1.4M frames, 20k video sequences, 600+ subjects, 20k actions	Wide variety of indoor/outdoor in-game locations, weather/time variation, environment interaction	2D/3D joints, SMPL pose+shape, occlusion flags, depth/semantic maps	Single-person 3D pose/shape estimation, video-based methods	Derived from near-perfect engine 3D keypoints plus temporal SMPLify	No formal train/test split; typically used as large-scale training data	Massive synthetic dataset with strong SMPL supervision, highly diverse outdoor scenes	https://caizhongang.github.io/projects/GTA-Human/
MMHPSD (105)	Real	A multi-camera rig with an event camera, one polarization camera, five RGB-D cameras; about 15 FPS for grayscale	Event stream, polarization, RGB-D	240k frames, 15 subjects, 21 actions, single-person; 12 short clips/subject	Indoor environment, moderate complexity, no multi-person	SMPL (body pose/shape), 2D/3D keypoints, segmentations	3D pose/shape from multi-modal data (event, polar, depth)	Fitted with multi-view approach, uncertain exact error	No formal train/test split	Largest event-based 3D dataset, combining multiple modalities	https://github.com/JimmyZou/EventHPE
MPI-INF-3DHP (105)	Real	Marker-less multi-camera system (The Captury) in green-screen studio; up to 14 cameras	Multi-view RGB, possible background compositing	>1.3M frames, 8 subjects (4m/4f), 8 activity sets, plus test sets (incl. outdoor)	Mostly studio + green screen, also separate test set with indoor/outdoor	3D joints (universal skeleton), 2D projections, silhouettes, clothing masks	Full-body 3D pose estimation (single-person), generalizable to in-the-wild	State-of-the-art marker-less system, uncertain exact error	Training data (indoor), test data with green screen / no green screen / outdoor	Chroma-key masks, realistic augmentation, single-person but broad coverage	http://gvv.mpi-inf.mpg.de/3dhp-dataset/