frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 05 April 2024
DOI 10.3389/fcomp.2024.1379925

OPEN ACCESS

EDITED BY
Mitja Lustrek,
Institut Jožef Stefan (IJS), Slovenia

REVIEWED BY
Gašper Slapničar,
Institut Jožef Stefan (IJS), Slovenia
Orhan Konak,
Institute, University of Potsdam, Germany

*CORRESPONDENCE
Lala Shakti Swarup Ray
✉ lalashaktiswarup.ray@dfki.de

RECEIVED 31 January 2024
ACCEPTED 22 March 2024
PUBLISHED 05 April 2024

CITATION
Ray LSS, Zhou B, Suh S and Lukowicz P (2024) A comprehensive evaluation of marker-based, markerless methods for loose garment scenarios in varying camera configurations. *Front. Comput. Sci.* 6:1379925. doi: 10.3389/fcomp.2024.1379925

A comprehensive evaluation of marker-based, markerless methods for loose garment scenarios in varying camera configurations

Lala Shakti Swarup Ray[1]*, Bo Zhou[1,2], Sungho Suh[1,2] and Paul Lukowicz[1,2]

[1]German Research Centre for Artificial Intelligence, Kaiseslautern, Germany, [2]Rhineland-Palatinate University of Technology Kaiserslautern-Landau, Kaiserslautern, Germany

In support of smart wearable researchers striving to select optimal ground truth methods for motion capture across a spectrum of loose garment types, we present an extended benchmark named DrapeMoCapBench (DMCB+). This augmented benchmark incorporates a more intricate limb-wise Motion Capture (MoCap) accuracy analysis, and enhanced drape calculation, and introduces a novel benchmarking tool that encompasses multicamera deep learning MoCap methods. DMCB+ is specifically designed to evaluate the performance of both optical marker-based and markerless MoCap techniques, taking into account the challenges posed by various loose garment types. While high-cost marker-based systems are acknowledged for their precision, they often require skin-tight markers on bony areas, which can be impractical with loose garments. On the other hand, markerless MoCap methods driven by computer vision models have evolved to be more cost-effective, utilizing smartphone cameras and exhibiting promising results. Utilizing real-world MoCap datasets, DMCB+ conducts 3D physics simulations with a comprehensive set of variables, including six drape levels, three motion intensities, and six body-gender combinations. The extended benchmark provides a nuanced analysis of advanced marker-based and markerless MoCap techniques, highlighting their strengths and weaknesses across distinct scenarios. In particular, DMCB+ reveals that when evaluating casual loose garments, both marker-based and markerless methods exhibit notable performance degradation (>10 cm). However, in scenarios involving everyday activities with basic and swift motions, markerless MoCap outperforms marker-based alternatives. This positions markerless MoCap as an advantageous and economical choice for wearable studies. The inclusion of a multicamera deep learning MoCap method in the benchmarking tool further expands the scope, allowing researchers to assess the capabilities of cutting-edge technologies in diverse motion capture scenarios.

KEYWORDS

MoCap analysis, cloth simulation, quantitative characterization, smart wearable, benchmarking tool

1 Introduction

Wearable sensing systems have attracted considerable attention in recent years, particularly in the realm of motion-tracking applications. This interest has been notably focused on a variety of technologies, including Inertial Measurement Unit (IMU) sensors (Gong et al., 2021; Jiang et al., 2022; Yi X. et al., 2022), Radio-Frequency Identification (RFID) technology (Jin et al., 2018), capacitive fabric sensors (Ray et al., 2023a; Zhou et al., 2023), computational fabrics (Liu et al., 2019), and multi-modal approaches (Liu, 2020; An et al., 2022). The continuous evolution and refinement of these motion-tracking technologies have paved the way for seamless activity recognition in diverse scenarios. This recognition serves as a critical component for a myriad of downstream tasks (Jansen et al., 2007; Behera et al., 2020), extending into the realms of deep learning applications, computer vision, and the development of large language models (Radford et al., 2021; Moon et al., 2022).

However, despite the impressive advancements witnessed in wearable sensing systems, optical marker-based motion capture (MoCap) systems persist as the gold standard. Industry standards such as Qualisys (Sweden), Vicon (USA), and OptiTrack (USA) exemplify these systems, which rely on the precise placement of optical markers on the body (Jiang et al., 2022; Yi C. et al., 2022). These markers are typically positioned in skin-tight configurations over bony areas, utilizing rigid biomechanical models to convert surface points to internal joints (Groen et al., 2012; OptiTrack, 2019). The markers themselves can be either active (Barca et al., 2006; Raskar et al., 2007), featuring built-in infrared light sources, or passive (Lee and Yoo, 2017), possessing unique visual patterns or retro-reflective properties. Optical MoCap systems deploy synchronized camera triangulation to capture marker positions on the body's surface, subsequently inferring joint motion through biomechanical models (OptiTrack, 2019). Despite their prevalence and remarkable accuracy, challenges arise when markers are placed on loose garments, leading to potential kinematic errors (McFadden et al., 2020). The demand for loose-fitting garments in wearable applications (McAdams et al., 2011; Bello et al., 2021; Zhou et al., 2023), driven by considerations such as user acceptance, comfort, and mass adoption, underscores the pressing need to overcome limitations associated with marker-based MoCap.

While optical marker-based MoCap remains dominant, video-based markerless MoCap systems have gained prominence, leveraging advanced deep learning algorithms to map semantic information to pose without the need for explicit markers (Chatzis et al., 2020; Gamra and Akhloufi, 2021; Sigal, 2021). Despite the maturity of these markerless approaches, there exists a notable gap in comprehensive comparisons between marker-based and markerless MoCap systems, particularly in the challenging context of loose garments. It is crucial to note that the precision of superficial markers in marker-based MoCap is not necessarily equivalent to the accuracy of determining joint positions inside the human body. Superficial markers may capture the external movements and postures effectively, but they might lack the depth and specificity required for precisely tracking the intricate movements of joints beneath loose garments. This limitation becomes especially pronounced when dealing with complex motions or anatomical configurations, highlighting the need for a nuanced evaluation that goes beyond the surface-level comparison of marker-based and markerless MoCap systems.

Several studies have established a comparison between marker-based and markerless MoCap by using various applications, ranging from controlling endoscopic instruments (Reilink et al., 2013) and analyzing baseball pitching biomechanics (Fleisig et al., 2022) to conducting gait analysis (Kanko et al., 2021) and assessing clinical usability (Ancans, 2021) as the base metric to compare the accuracy of both methods. While marker-based MoCap generally exhibits slightly higher accuracy, markerless systems emerge as viable alternatives, particularly in clinical settings where patient comfort and ease of use are prioritized (Nakano et al., 2020). Because of the unattainability nature of the task, existing studies often focus on factors such as complexity, ease of use, and overall performance, lacking in-depth quantitative precision comparisons, particularly in the nuanced realm of loose garments. The reason behind the lack of quantitative precision comparison is due to the absence of anatomic motion reference, hence no evaluation is done considering loose garments to the level of casual apparel. In practical terms, it is unfeasible to perform an accurate quantitative comparison of MoCap methods for loose garments due to the inability to non-invasively capture true anatomical motion beneath the clothing and replicate precise motion sequences across diverse body shapes and attires. Because

1. The anatomical true motion underneath the garment and skin is required to quantitatively compare different MoCap methods, which is unknown in the real world because even marker-based MoCap uses biomechanical approximation from surface markers.
2. The exact motion sequences need to be reproduced precisely in multiple scenarios with persons of different body shapes wearing different garments.

Largely due to these challenges, existing quantitative reviews of markerless methods use marker-based MoCap as reference (Wang et al., 2021), which itself has substantial error from the anatomic joints due to the biomechanical approximation.

To solve this problem, we introduce an extended version of DMCB (Ray et al., 2023d) called DMCB+, featuring a different pose estimation method, cloth drape calculation, and a MoCap limb-wise accuracy analysis. This upgraded benchmark builds upon the foundation of DMCB that leverages 3D physics-based simulation to benchmark and compare marker-based and markerless MoCap systems, incorporating advanced capabilities to further refine the evaluation of motion capture systems. The drape calculation helps us classify different garments based on looseness while real-world MoCap datasets are utilized to generate inputs for both methods, enabling a quantitative comparison against common anatomical true motion. The benchmarking process encompasses diverse motion types and garment drape levels, providing practitioners with valuable insights into choosing optimal MoCap solutions for specific applications. Through this meticulously designed approach, we aim to provide a comprehensive evaluation framework that effectively bridges the existing gap between marker-based and markerless MoCap systems, particularly in scenarios involving loose garments.

As compared to DMCB, This work incorporates several advancements including (a) a multi-camera deep learning-based pose estimation technique, (b) improved drape calculation robustness, and (c) a thorough analysis of limb-wise accuracy. DMCB+ builds upon the groundwork laid by DMCB, which utilizes 3D physics-based simulation to assess and compare marker-based and markerless motion capture (MoCap) systems. By integrating sophisticated features, DMCB+ enhances the evaluation of motion capture systems, offering a more refined understanding of their performance across various scenarios. The refined drape calculation and limb-wise accuracy analysis provided by DMCB+ offer nuanced insights into the capabilities of both marker-based and markerless MoCap techniques. This work also introduces enhancements such as a comprehensive analysis of MoCap performance both overall and at the limb level, leveraging the newly improved drape calculations.

In particular, we make the following contributions:

1. We introduce a benchmark for simulating garment and soft body physics visualized in Figure 1, designed to assess the performance of marker-based and markerless MoCap systems across various camera configurations. This evaluation is conducted with individuals of different body types executing identical movements while wearing diverse garments with varying degrees of drape. By utilizing real-world motion datasets for input generation in both MoCap methods, we quantitatively compare their outcomes to the established anatomical true motion.
2. Our benchmark encompasses a wide range of motion types and garment drape levels. Through a comprehensive comparison, this benchmark can aid practitioners in selecting the most suitable MoCap system for generating ground truth in wearable experiments tailored to their specific applications. This decision-making process can consider factors such as garment designs, types of motion, cost, time overhead, and precision.

2 Related work

The landscape of MoCap research is rich with diverse methodologies, ranging from traditional marker-based systems to cutting-edge markerless approaches. This section explores the evolution of MoCap technologies, focusing on the distinct realms of Marker-Based Systems and Markerless deep learning (DL) Systems, further sub-categorized into Single-camera systems and Multi-camera systems.

Optical marker-based MoCap systems, exemplified by industry leaders such as Qualisys, Vicon, and OptiTrack, have long been revered for their unparalleled precision and accuracy. These systems rely on strategically placed optical markers on the human body, enabling the capture and triangulation of motion through synchronized cameras. Extensive studies have validated the effectiveness of marker-based MoCap in diverse applications, including biomechanical analysis, sports science, and clinical assessments. Furthermore, the widespread adoption of optical marker-based MoCap systems is evident in the realm of data-driven research, where numerous datasets (Joo et al., 2015; Plappert et al., 2016; Trumble et al., 2017) rely on these systems for capturing intricate motion details. Their ability to provide high-precision and accurate motion data has made them indispensable tools in various fields, contributing to the robustness and reliability of datasets utilized in areas such as artificial intelligence, machine learning, and computer graphics.

In recent years, a paradigm shift toward video-based markerless MoCap systems (Xu et al., 2020; Gong et al., 2023; Zhao et al., 2023) has emerged. These systems leverage advanced deep learning algorithms to infer pose without relying on explicit markers. While markerless approaches offer advantages in user comfort and ease of use, a comprehensive comparison with marker-based systems, especially in scenarios involving loose garments, remains conspicuously absent from the existing literature. Some markerless MoCap systems leverage a single camera for capturing and interpreting motion, employing sophisticated computer vision techniques and deep learning algorithms. Studies and applications utilizing single-camera markerless systems have shown promise in diverse scenarios, including endoscopic instrument control, biomechanics analysis of baseball pitching, gait analysis, and clinical usability assessments (Dubey and Dixit, 2023). On the other hand, multi-camera markerless MoCap systems (Tu et al., 2020; Dong et al., 2021; Liu et al., 2023) utilize synchronized camera arrays to capture motion from different perspectives. The data from multiple cameras are then processed to reconstruct three-dimensional (3D) motion information. While offering increased coverage and potential accuracy, challenges related to calibration and synchronization are paramount in multi-camera systems. Numerous investigations have contrasted the efficacy of marker-based and marker-less Motion Capture (MoCap) across various domains, including the control of endoscopic instruments (Reilink et al., 2013), biomechanical analysis of baseball pitching (Fleisig et al., 2022), gait assessment (Kanko et al., 2021), and clinical applicability (Ancans, 2021). While marker-based MoCap typically demonstrates marginally superior accuracy, markerless systems are increasingly recognized as a feasible alternative, particularly in clinical environments where patient comfort and usability are paramount considerations (Nakano et al., 2020). Conducting a meticulous quantitative comparison of motion capture methodologies for loose garments poses significant challenges. The inability to non-invasively capture underlying anatomical motion beneath clothing and replicate precise motion sequences across diverse body shapes and clothing types renders such comparisons impractical. Compounding these challenges, extant quantitative reviews of marker-less methods typically rely on marker-based MoCap as the reference standard (Wang et al., 2021), despite the inherent errors in the approximation of biomechanical joint movements. Also, literature on quantitative precision, especially in scenarios involving loose garments, is scarce. A significant challenge arises when dealing with loose garments, as the conventional use of skin-tight marker configurations may introduce kinematic errors. Despite their undeniable accuracy, the limitations associated with marker-based systems in accommodating loose-fitting attire underscore the need for alternative solutions.

Addressing the noticeable gap in existing research, our proposed work introduces a groundbreaking approach. We employ

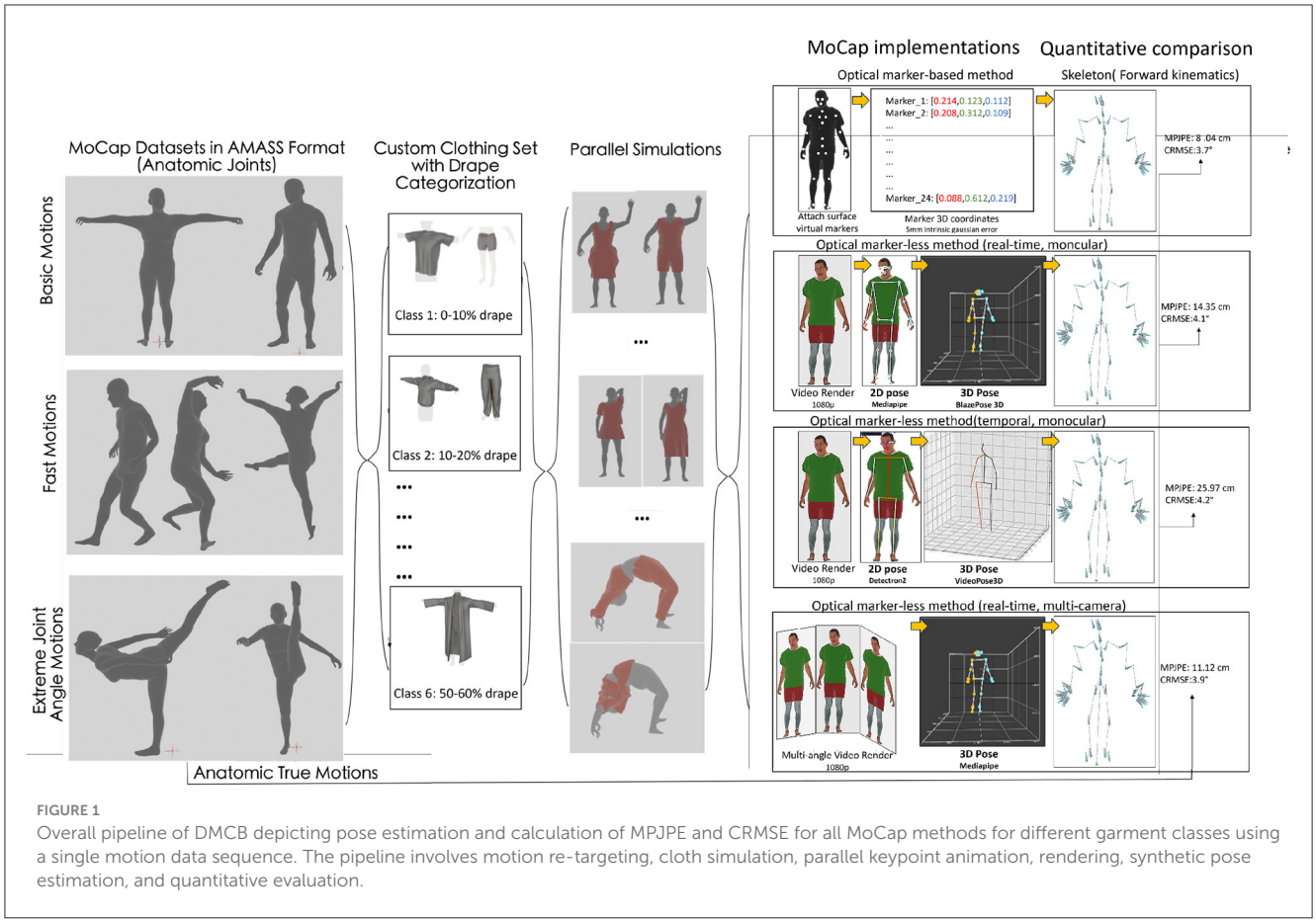

FIGURE 1
Overall pipeline of DMCB depicting pose estimation and calculation of MPJPE and CRMSE for all MoCap methods for different garment classes using a single motion data sequence. The pipeline involves motion re-targeting, cloth simulation, parallel keypoint animation, rendering, synthetic pose estimation, and quantitative evaluation.

a 3D physics-based simulation explicitly for benchmarking and comparing both marker-based and markerless MoCap systems. This pioneering method utilizes real-world MoCap datasets to generate inputs for both approaches, enabling a quantitative evaluation against common anatomical true motion.

Our benchmark encompasses diverse motion types and garment drape levels, aiming to provide a comprehensive framework for practitioners. This framework enables informed decisions based on holistic considerations such as garment design, motion types, cost, time overhead, and precision. By undertaking this endeavor, our goal is not only to address but also to significantly advance the current state of the literature. We aim to offer nuanced insights into the suitability of marker-based and markerless MoCap systems, particularly in scenarios involving loose garments.

3 Proposed method

The proposed benchmark methodology introduces a holistic approach to the evaluation of MoCap methods, addressing the challenges associated with replicating precise human motion in real-world scenarios. By leveraging 3D physics simulation, we solve the reality challenge that the exact motion cannot be perfectly reproduced to establish quantitative comparisons of different scenarios.

3.1 Simulation pipeline

The simulation pipeline, a crucial component of the methodology, ensures fidelity to real-world conditions by incorporating true-to-specification inputs for all MoCap methods. For marker-based kinematic methods, the inclusion of 3D surface marker locations is vital, while markerless vision models receive high-resolution 1,080 p image sequences. This meticulous adherence to accurate inputs sets the stage for a reliable and realistic evaluation of MoCap methodologies.

Within the 3D physics simulation phase, implemented using Blender3D (Blender Foundation, 2023) and the SMPL-X Blender addon (Pavlakos et al., 2019), motion sequences from the MoCap dataset are transformed into volumetric human bodies of varying builds. The bodies undergo realistic dressing using the Simplycloth plugin (Simplycloth, 2022). High-resolution simulated garments exhibit near-realistic properties closely mirroring those of actual cloth, and have gained widespread adoption for generating data in the realm of realistic virtual try-ons (Cho et al., 2023). This process not only accounts for the physical properties of garments but also considers the interaction between the garments and the dynamic human body. The incorporation of soft tissue dynamics using Mosh++ (Loper et al., 2014) further enhances the fidelity of the simulation, ensuring that realistic garment deformation occurs during dynamic activities with minimal artifacts.

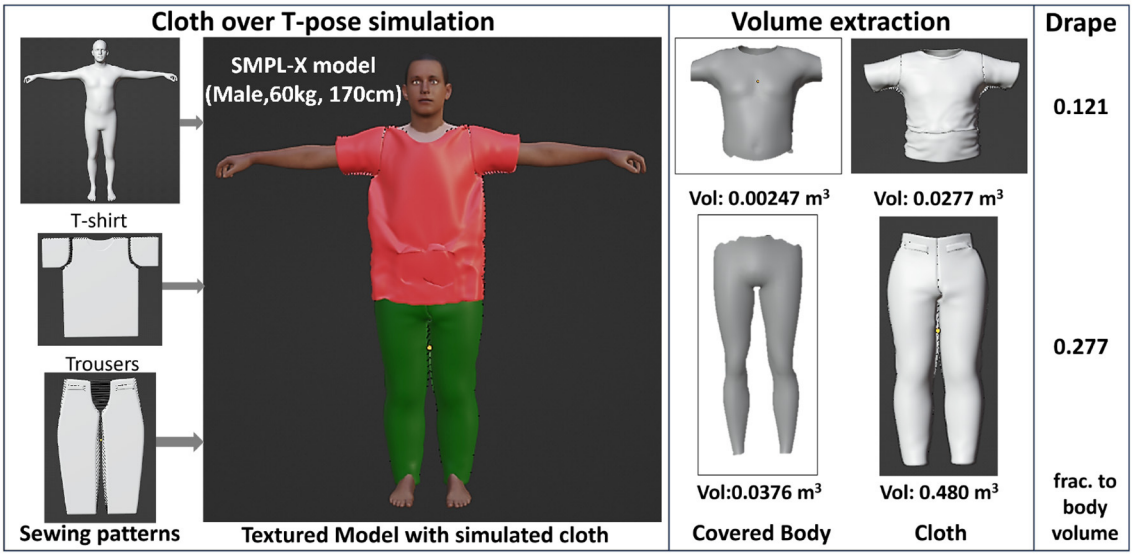

FIGURE 2
Quantifying drape for t-shirt and trousers over T-Pose by first simulating the cloth over the SMPL body for a particular pose then calculating the difference between the volume of the cloth with that of the underlying body of the mesh.

3.2 Motion source dataset

The Motion Source Dataset section emphasizes the utilization of the AMASS framework (Mahmood et al., 2019) along with the SMPL body model (Loper et al., 2015) to curate a diverse dataset encompassing different motion categories. This curated dataset includes the following sequences:

Basic motions

- 30 samples of a total of 36,210 frames of around 20 minutes.
- Walking sequences sourced from TotalCapture (Trumble et al., 2017).
- Gesture sequences sourced from HumanEva (Sigal et al., 2010).

Fast motions

- 33 samples of a total of 70,985 frames of around 40 minutes.
- Rom and Freestyle sequences sourced from TotalCapture.
- Hasaposerviko and Pentozali dancing sequences sourced from DanceDB (University of Cyprus, 2023).

Extreme joint angle motions

- 36 samples having a total of 66,560 frames of around 37 minutes.
- Extreme joint bending motions sourced from PosePrior (Akhter and Black, 2015).
- Yoga articulation sequences sourced from PresSim (Ray et al., 2023c).

By considering these varied motion scenarios, the benchmark methodology aims to offer a comprehensive evaluation of MoCap methods that goes beyond basic locomotion and accounts for the challenges posed by fast, dynamic movements and extreme joint articulation.

3.3 Quantifying drape of loose garments

In our study, we employed 3D assets encompassing a diverse range of apparel from commonly available categories for both genders, utilizing the Simplycloth plugin (Simplycloth, 2022) for garment simulation. The garments varied in style, ranging from skin-tight with minimal drape to very loose with maximal drape. The term "looseness" is highly subjective, and its interpretation depends on the relative volume of the garment as compared to the wearer. For instance, a shirt that is normally sized may be perceived as loose if worn by a slender individual. To account for this variability, we employed a quantitative measurement of drape and organized our findings into drape classes. Initially, we calculated the drape amount by considering only T-Pose of the body using the formula:

$$Drape = \frac{Volume_{\text{garment}} - Volume_{\text{CoveredBody}}}{Volume_{\text{CoveredBody}}}$$

This calculation was based on the extra volume occupied by the garment compared to the underlying body of the person wearing the garment as visualized in Figure 2. The formula is designed to address the subjective nature of assessing looseness in garments. By comparing the volume of the garment to the volume of the wearer's body it covers, the formula captures the relative excess fabric, which can vary significantly depending on the wearer's size. Normalizing the drape measurement by dividing it by the volume of the covered body ensures that the assessment is not solely dependent on the garment size but also accounts for the wearer's proportions. This

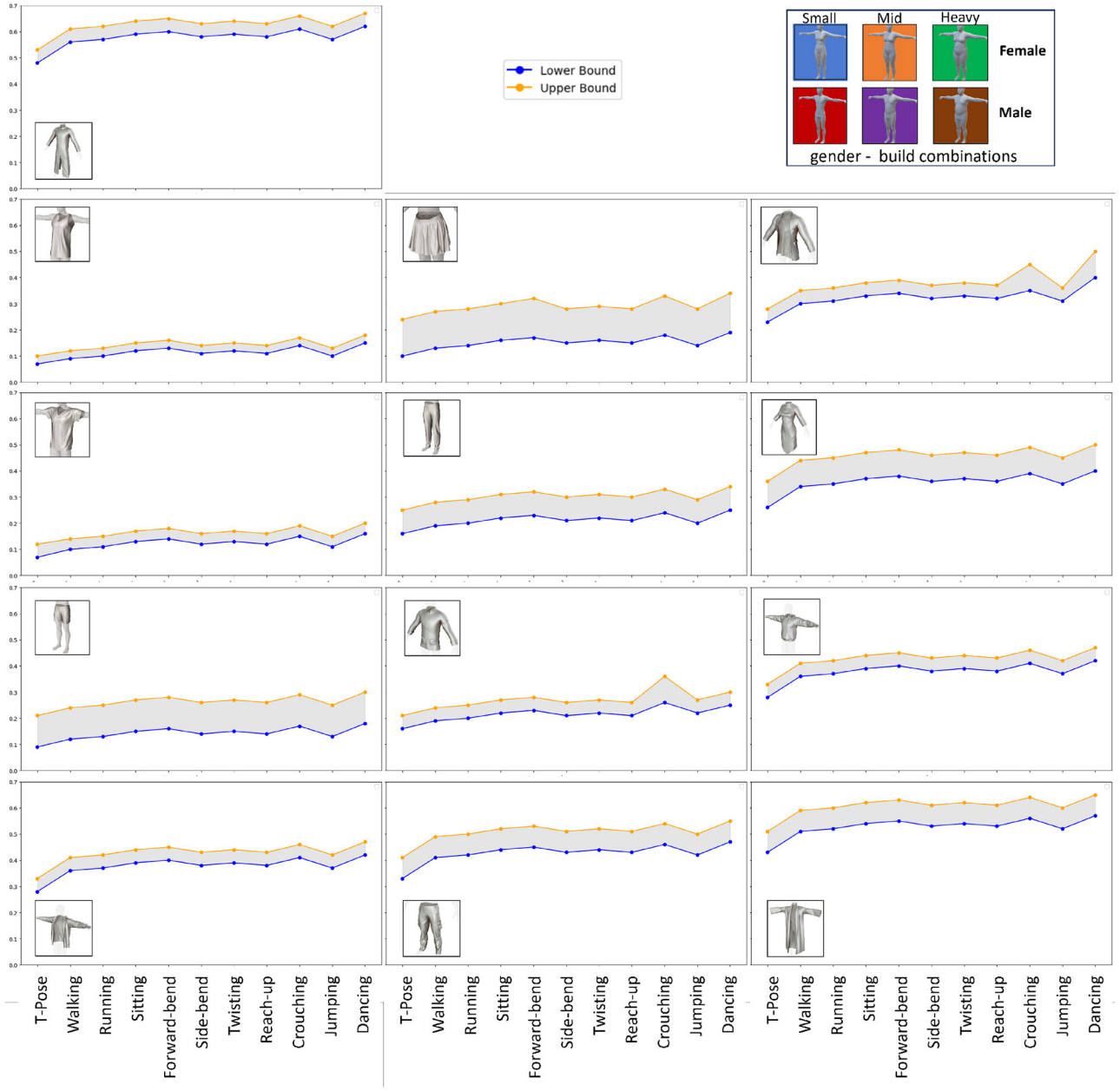

FIGURE 3
Drape range for eleven different garments over varying genders and builds from minimal to maximal possible drape for that cloth type in different poses calculated for the improved looseness quantification.

approach provides a standardized and quantitative measure of how the garment drapes on different individuals, offering a more objective evaluation of looseness. Additionally, the accompanying visualization aids in understanding how the calculation accounts for the relationship between the garment volume and the covered body volume.

However, acknowledging the limitations of the approach previously introduced in DMCB (Ray et al., 2023d), particularly for garments like skirts or dresses that display significant movement beyond the static standing posture, we refined our methodology in DMCB+. We introduced a more robust drape measurement that considers various body postures, including T-pose, walking, Running, Sitting, Forward-bend, Side-bend, Twisting, Reach-up, Crouching, Jumping, and Dancing. For each garment, we captured different postures of the body and calculated the mean drape value. This enhanced approach in DMCB+ provides a more comprehensive and accurate representation of the overall drape of the cloth under varied conditions as given in Figure 3 which changed the initial score by drape percentage from a range of 0.015 to 0.075 as given in Table 1. Each garment is reassigned based on the new drape value to fit the appropriate drape class.

In practice, for all garments except uni-cloths, we selected separate pieces for the upper and lower body, matching a particular

build and sharing a combined drape class ranging from 1 to 6, to dress the SMPL body mesh.

3.4 Marker-based method

The marker-based method (MB) involves a set of 48 markers strategically placed on the body, with 24 pairs located on both the front and back. These markers are associated with specific joints within the SMPL skeleton, including the pelvis, left leg root, right leg root, lower back, left knee, right knee, upper back, left ankle, right ankle, thorax, left toes, right toes, lower neck, left clavicle, right clavicle, upper neck, left arm root, right arm root, left elbow, right elbow, left wrist, right wrist, left hand, and right hand. Each marker's position is carefully considered over garments or skin, taking into account optimal real-world marker placement.

To replicate real-world conditions, a 5 mm error is introduced as Gaussian noise. In addition to this, biomechanical constraints are applied to these specified joints, considering more realistic angular ranges to better align with human joint capabilities during everyday movements:

- **pelvis:** angular range: -30° to 30°
- **left leg root, right leg root:** angular range: -45° to 45°
- **lower back:** angular range: -30° to 30°
- **left knee, right knee:** angular range: 0° to 120°
- **upper back:** angular range: -30° to 30°
- **left ankle, right ankle:** angular range: -45° to 45°
- **thorax:** angular range: -30° to 30°
- **left toes, right toes:** angular range: 0° to 45°
- **lower neck:** angular range: -30° to 30°
- **left clavicle, right clavicle:** angular range: -30° to 30°
- **upper neck:** angular range: -30° to 30°
- **left arm root, right arm root:** angular range: -45° to 45°
- **left elbow, right elbow:** angular range: 0° to 180°
- **left wrist, right wrist:** angular range: -45° to 45°
- **left hand, right hand:** angular range: 0° to 45°.

These realistic angular ranges account for natural anatomical limitations, and the biomechanical constraints expressed using quaternions to represent angular orientations, play a critical role in refining captured data. They ensure that estimated joint positions not only adhere to realistic human movement patterns but also account for the inherent variability in joint flexibility.

3.5 Monocular 3D pose estimation models

We considered two markerless models:

- a temporal semi-supervised 3d pose estimation model VideoPose3D (TML) (Pavllo et al., 2019)
- a lightweight real-time 3d pose estimation model BlazePose3D (IML) (Bazarevsky et al., 2020).

We applied SMPL textures to the bodies derived from SMPLitex (Casas and Comino-Trinidad, 2023). Videos (1,920 × 1,080) were rendered from the simulation scene in a simple white background, then fed into Detectron2 (Wu et al., 2019) + VideoPose3D or BlazePose3D to extract multi-joint poses relative to the video frame. The 17-joint Human3.6M skeleton and the 31-joint Mediapipe skeleton are given as input to joint2smpl (Zuo et al., 2021) to give the 24-joint SMPL skeleton (MML). They are then rescaled to the original size of the body (170 cm height) and converted to BVH files using Motion matching (Dittadi et al., 2021). Since these monocular pose estimation models calculate camera relative joint positions to make them absolute the starting pose of each pose sequence is re-positioned and reoriented to an origin identical to that of the ground truth.

TABLE 1 Average drape for each cloth type and difference from T-pose.

Cloth type	Mean drape using different poses	Difference from when using only T-pose
Sleeveless	0.085	+0.015
T-shirt	0.095	+0.015
Shorts	0.15	+0.06
Skirt	0.175	+0.075
Shirt	0.205	+0.04
Dress	0.185	+0.02
Trousers	0.255	+0.015
Jacket	0.31	+0.035
Hoodie	0.305	+0.025
Cardigan	0.305	+0.025
Cargo	0.37	+0.035
Robe	0.47	+0.025
Trench Coat	0.505	+0.025

3.6 Multi-camera 3D pose estimation models

Unlike DMCB (Ray et al., 2023d), we introduced support for multicamera markerless MoCap methods in the pipeline along with evaluation. Our simulation leverages SynthCal (Ray et al., 2023b), a system that utilizes simulated data and employs a multicamera calibration pipeline to generate input for our multi-camera markerless model. To incorporate processed videos rendered from the simulation scene, without the need for explicit markers, we created a synthetic charuco board and put it inside the scene along with the SMPL body mesh that helps in synchronization of the three cameras strategically placed in front and on both sides of the person to find the absolute position of the SMPL mesh in the scene, enhancing the accuracy of our motion capture system with the integration of BodyPose3D pipeline (Batpurev, 2021) which employs camera triangulation to find the absolute 3D position of the skeleton from multiple perspectives. A similar approach like before is employed to convert the 31-joint mediapipe pose into 24 joint SMPL pose using joint2smpl.

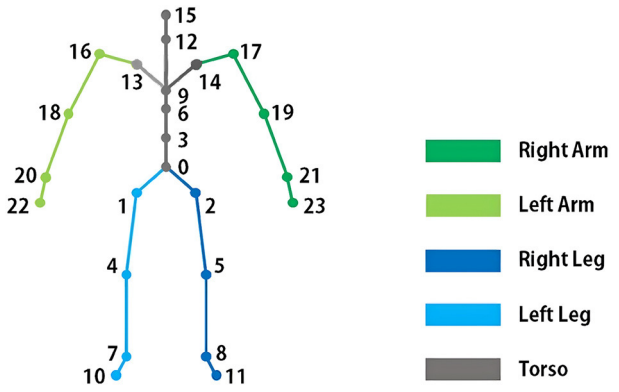

FIGURE 4
Limb-based classification of SMPL joints for more nuanced comparison of the accuracy of different pose estimation methods in different drape classes.

4 Evaluation

4.1 Evaluation metrics

The Evaluation Metrics section underscores the significance of employing two well-established metrics in the MoCap field: Mean Per Joint Position Error (MPJPE) and Circular Root Mean Squared Error (CRMSE) (Equations 1, 2). These metrics offer a quantitative assessment of MoCap accuracy, considering different aspects of 24 joint positions and pose angles. The comprehensive measurements obtained from the simulation allow for the straightforward calculation of these metrics, providing a robust foundation for the evaluation of MoCap methods.

$$MPJPE = \frac{1}{n}\sum_{i=1}^{n} ||\mathbf{P}_i - \hat{\mathbf{P}}_i|| \tag{1}$$

where n is the number of joints, $\mathbf{P}_i$ is the ground truth position of the i-th joint, $\hat{\mathbf{P}}_i$ is the estimated position of the i-th joint, and $||\cdot||$ denotes the Euclidean distance.

$$CRMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}\left(1 - \cos(\theta_i - \hat{\theta}_i)\right)} \tag{2}$$

Here, N represents the total number of joint angles, θ_i represents the ground truth angle for the i-th joint, and $\hat{\theta}_i$ represents the corresponding predicted angle. Due to the availability of comprehensive measurements about human models and garments obtained from the simulation, it is straightforward to calculate MPJPE using the 3D estimated joints in Euler space. On the other hand, the CRMSE involves estimating joint angles by applying forward kinematics and then computing the error.

4.2 Limb-based classification

In our comprehensive analysis of human body parts, we have systematically classified them into three distinct sections: torso, limbs, and legs. The arms section encompasses joints such as arm roots, elbows, wrists, and hands, a total of eight joints while the legs section includes leg roots, knees, ankles, and toes total of eight joints. The torso section comprises joints like the pelvis, lower back, upper back, thorax, lower neck, upper neck, and clavicles total of 8 joints as visualized in Figure 4. This meticulous categorization facilitates a nuanced assessment of model accuracy, as we can focus on specific limb parts to gauge the effectiveness of each model. By examining the performance of models in isolating and predicting joints within distinct body regions, we gain valuable insights into their capabilities and limitations, ultimately enhancing our understanding of human pose estimation.

4.3 Results

In the results section of our study, we conducted a comprehensive evaluation of various MoCap methods, specifically focusing on marker-based (MB) and monocular markerless (TML, IML) techniques. We also included Multicamera-realtime markerless (MML) methods in our analysis, employing both quantitative metrics and a holistic comparison.

In the initial comparison using unclothed bodies on the TotalCapture and PosePrior datasets, we observed that the Mean Per Joint Position Error (MPJPE) for marker-based methods was 4.7 cm, whereas for monocular markerless methods, it was 8.2 cm. These results align with existing literature that compares markerless models with marker-based MoCap as a reference, as evidenced by studies such as Kanazawa et al. (2018), Ostrek et al. (2019), Qiu et al. (2019), and Wang et al. (2021). This validation supports the realism and accuracy of our model in pose estimation over real data for the same datasets.

In the context of comprehensive full-body joint analysis given in Figure 5, our investigation revealed that the minimum joint-position error occurred with drape class 1 garments, as assessed through marker-based methods, surpassing 10 cm. Previous attempts at such comparisons were limited to our simulation pipeline, as obtaining anatomic joint coordinates in real-world scenarios with non-invasive methods like surface markers or video analysis proved challenging. Everyday loose garments, falling into drape class 2 or 3, exhibited MPJPE ranging from 15 to 35 cm and CRMSE ranging from 6○ to 11○ for both marker-based and markerless methods.

Becuase of complex drape nature of cloths of being loose at some places and tight in other places and to have a better understanding of their effect on different MoCap methods our study delved into various portions of the body, and our examination of the torso section given in Figure 6, always covered by garments ranging from drape class 1 to 6, revealed intriguing findings. Notably, marker-based methods exhibited a slight advantage in accuracy when focusing solely on the torso compared to the whole-body analysis. This suggests that the biomechanical constraints and reference points provided by markers contribute to enhanced precision, particularly in the context of torso movements. However, even with this marginal advantage, multicamera markerless DL methods demonstrated superior performance, surpassing marker-based techniques, particularly in the 2–3

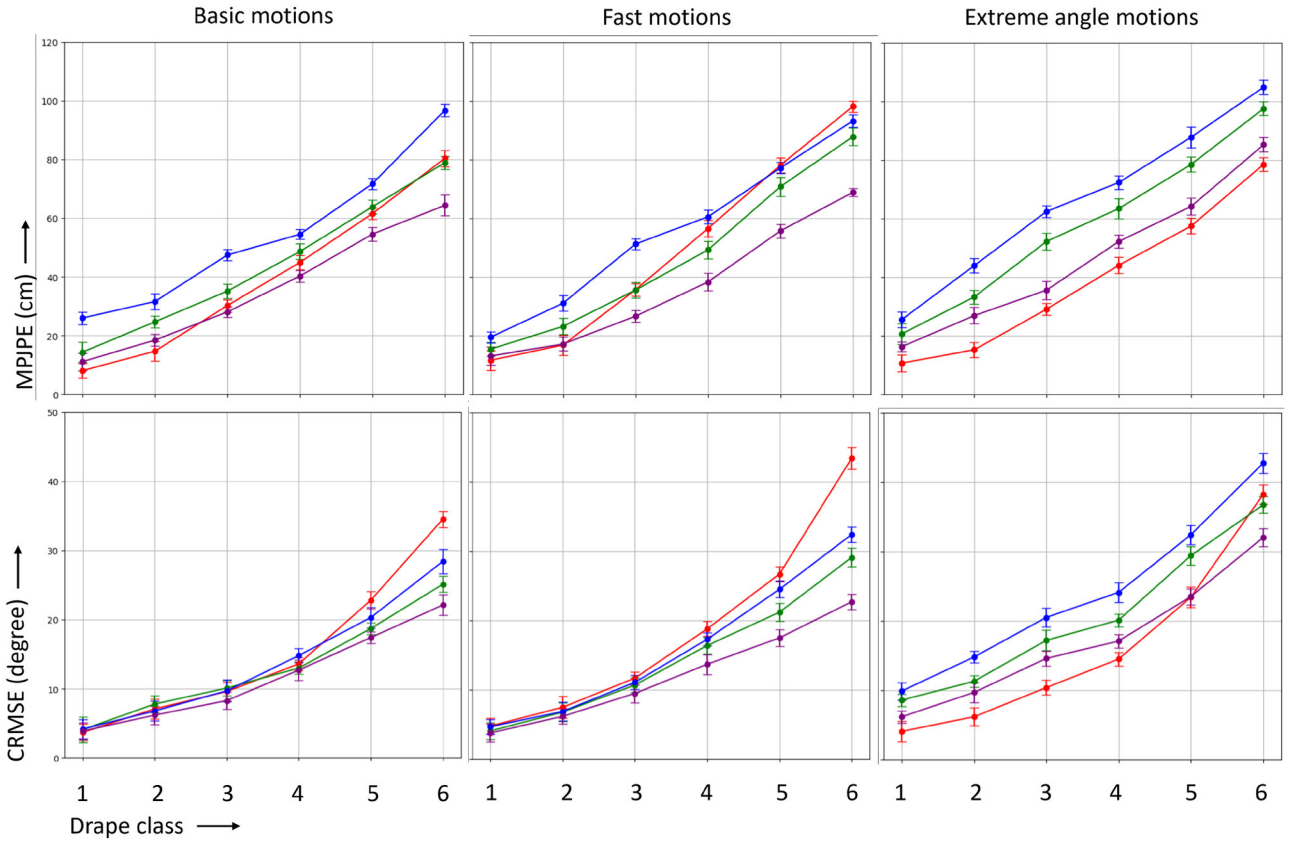

FIGURE 5
MPJPE (in centimeter) and CRMSE (in degree) for the four different MoCap methods marker-based (MB), temporal markerless (TML), instantaneous markerless (IML), and multi-camera markerless (MML) considering the **whole body**.

drape class range. This outcome suggests that for experiments concentrating exclusively on torso movements, multicamera markerless DL methods may be considered ideal, offering a more advanced and effective solution for capturing nuanced motions, even in scenarios involving moderate to substantial garment draping.

The disparity between marker-based and markerless methods becomes more pronounced when considering the simultaneous capture of both arms given in Figure 7 and legs given in Figure 8 which more or less follows a similar trend. As we transition to garments that entirely cover the arms and legs, typically around drape class 3, a minor spike in error is observed across all methods in comparison to drape class 1 or 2. This increase in error is attributed to the inherent challenges posed by garments that introduce complexity in limb movements and occlusions.

4.4 Analysis

In optimal conditions with zero drape, marker-based methods generally yields the highest accuracy, followed by multi-camera markerless, then temporal monocular markerless, and finally instantaneous monocular markerless methods. The accuracy of deep learning models correlates with their complexity: MML, which considers multiple viewpoints, typically outperforms TML, which analyzes sequences for 2D poses, and IML, which relies on single-frame 3D pose generation (Zheng et al., 2023). As drape increases, the degradation rate of accuracy varies among modalities. MB experiences a higher degradation rate per drape increment compared to DL-based methods. Additionally, the impact of drape percentage differs across motion types. In fast motions, the cloth topology takes longer to register the underlying body's topology in each frame, resulting in greater disparity between cloth and body topology compared to other motion types. Since marker-based pose estimation relies solely on surface markers, the performance suffers more in fast motions under similar levels of drape compared to other motion types (Puthenveetil et al., 2013).

Absolute MPJPE was influenced by errors in joint hierarchy alignment, including shifting and rotation, while CRMSE was impacted by the relative angle of bones. Despite the overall marker set being constrained by biomechanical considerations in marker-based methods, instances such as robes and trench coats demonstrated limited adherence to the wearer's motion, resulting in increased errors. Markerless methods showed gradually increasing errors with more draped garments, although the degradation rate was slightly lower for basic and fast motions compared to marker-based methods. It is noteworthy that monocular markerless methods, especially VideoPose3D, exhibited greater stability as the drape increased. This enhanced stability could be attributed to

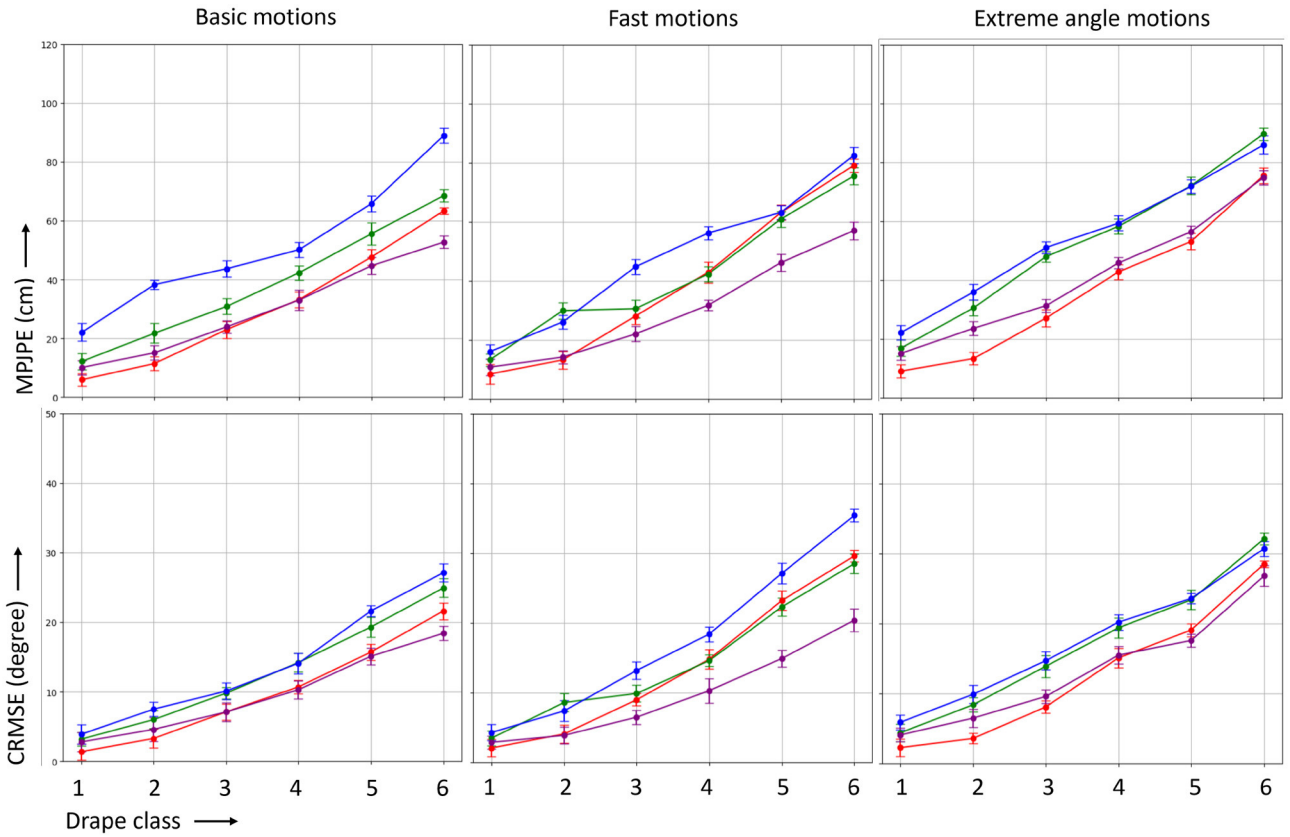

FIGURE 6
MPJPE (in centimeter) and CRMSE (in degree) for the four different MoCap methods marker-based (MB), temporal markerless (TML), instantaneous markerless (IML), and multi-camera markerless (MML) considering the **torso section**.

their ability to detect semantic segments of body parts, showcasing less variation in MoCap accuracy across different types of motion. Furthermore, DL-based markerless models performed better on basic and fast motions, given their primary training on datasets consisting of such motions. In contrast, marker-based methods did not exhibit such limitations, benefiting from forward kinematics and biomechanical constraints derived from markers, irrespective of the complexity of the motion or extreme joint angles. In our investigation, the MML method emerged as particularly noteworthy, demonstrating superior accuracy compared to other methods, especially as the drape class increased to around 2. Notably, MML outperformed both marker-based and monocular markerless methods in capturing motion details under these conditions. As the level of garment drape advanced to levels 5–6, we observed a significant decline in the accuracy of marker-based methods, marking them as less ideal for very loose garments. This decrease in accuracy suggests that the rigid constraints imposed by marker-based techniques become more limiting when dealing with extensive garment draping, highlighting a clear advantage for multicamera markerless methods in scenarios involving highly draped or flowing attire. The enhanced precision of multicamera markerless approaches underlines their potential applicability in contexts where capturing nuanced and complex motions, even in the presence of substantial garment drape, is crucial. Our comprehensive analysis identified a notable exception to the overall trend. In scenarios involving extreme joint angles and complex poses, marker-based MoCap methods demonstrated superior performance across almost all drape levels when compared to markerless alternatives. The intricate and varied nature of extreme joint angle motions posed challenges for markerless models, as the training data may not encompass the full spectrum of such complex poses. The inherent limitations in capturing the nuances of extreme joint angles and intricate movements through markerless approaches highlight a unique strength of marker-based methods in these specific scenarios. Despite the clear advantages of markerless methods in capturing motion details under draped conditions, the complex and extreme joint angle motions present a domain where the robustness and biomechanical constraints of marker-based MoCap prove invaluable, showcasing their effectiveness in situations where markerless models may face challenges due to the absence of specific training data for such intricate poses. In difficult scenarios where there is a significant presence of drapes (drape class 4–5), both marker-based and markerless methods experience a notable decline in accuracy. The noise introduced by these drapes severely impacts the performance of both methods, rendering them ineffective for normal pose estimation use, as their accuracy becomes highly inaccurate. However, the degradation of markerless deep learning methods is lower compared to marker-based methods. This outcome is expected, as marker-based models excel in scenarios with minimal

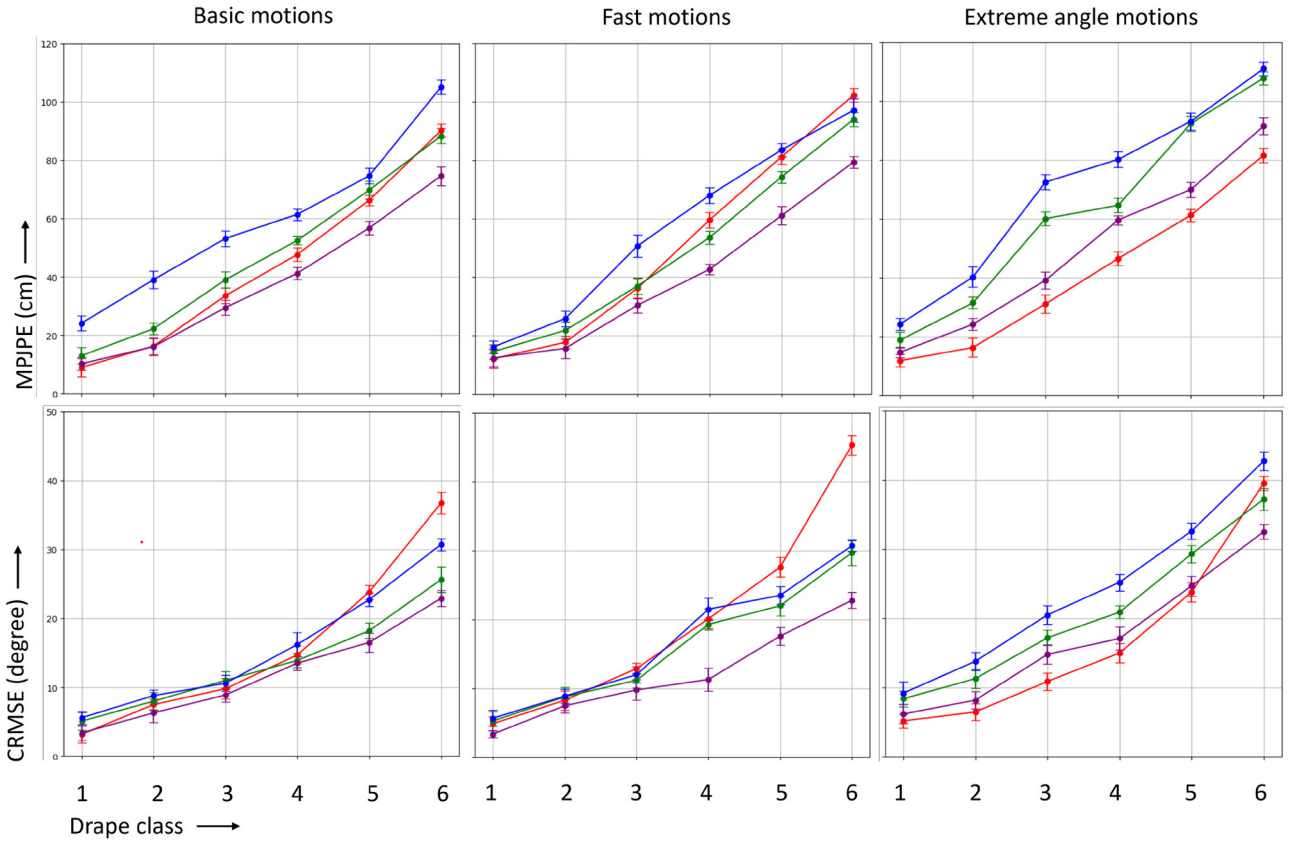

FIGURE 7
MPJPE (in centimeter) and CRMSE (in degree) for the four different MoCap methods marker-based (MB), temporal markerless (TML), instantaneous markerless (IML), and multi-camera markerless (MML) considering the **both arms**.

to zero drape, but become inferior to deep learning methods as drape presence increases. This difference can be attributed to the nature of the input for each method; marker-based methods rely solely on surface markers, whereas markerless methods consider images with pixels which is supported by the fact that other research works conducted with loose wearable sensors prefer to use markerless method as compared to marker-based one to generate their ground truth (Bello et al., 2021; Zhou et al., 2023). These finding underscores the robustness of our comparative analysis, indicating that even in scenarios where garments cover both arms and legs, the markerless methods maintain their competitive edge over marker-based counterparts.

In addition to conducting quantitative comparisons among various MoCap methods, our study delved into a comprehensive and holistic assessment of these techniques, given in Table 2. Recognizing the multifaceted nature of motion capture, we sought to go beyond mere numerical metrics and encompass a qualitative evaluation of the overall performance. By embracing a more inclusive perspective, we aimed to provide a nuanced understanding of the strengths and limitations of each MoCap method, enabling a more well-rounded appreciation of the practical implications and applications of various MoCap technologies in diverse contexts.

Cost-benefit analysis plays a pivotal role in selecting motion capture (MoCap) methods, as different applications have varying requirements and constraints. For instance, in the context of marker-based MoCap systems, considering a scenario where a research team is conducting a biomechanical study to analyze the gait patterns of athletes. In this case, the high precision offered by marker-based systems is indispensable, as even slight inaccuracies can affect the validity of the findings. Despite the substantial upfront investment required for specialized cameras and synchronization systems, the accuracy provided by marker-based MoCap outweighs the cost, making it the preferred choice for such research endeavors where precision is paramount.

Conversely, the benefits of monocular markerless MoCap methods in a different context. Imagine a game development studio working on a virtual reality (VR) game that simulates outdoor sports activities like hiking or rock climbing. In this scenario, the studio aims to capture realistic motion data of users engaging in these activities within various natural environments. Here, the cost-effectiveness and flexibility of monocular markerless methods shine through. By utilizing readily available consumer-grade cameras, the studio can significantly reduce expenses without compromising the quality of motion capture. Moreover, the ability to capture motion in diverse outdoor settings aligns perfectly with the studio's requirements, making monocular markerless MoCap the ideal choice for their VR game development project.

Furthermore, considering the case of multi-camera markerless MoCap systems in the field of wearable technology research.

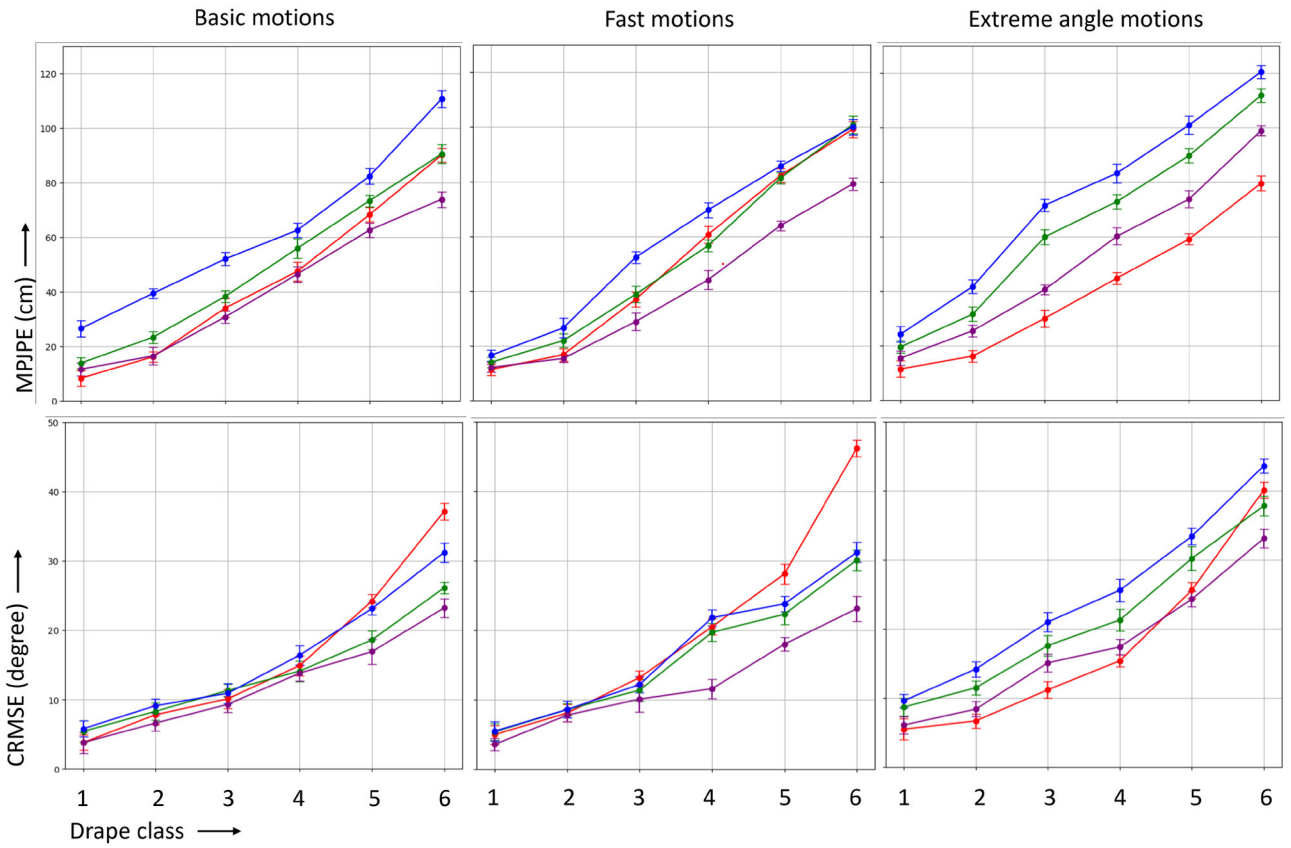

FIGURE 8
MPJPE (in centimeter) and CRMSE (in degree) for the four different MoCap methods marker-based (MB), temporal markerless (TML), instantaneous markerless (IML), and multi-camera markerless (MML) considering the **both legs**.

Suppose a team of researchers is developing a smart garment that monitors posture and movement to prevent musculoskeletal injuries in office workers. In this scenario, the researchers need a MoCap solution that offers a balance between accuracy and affordability as well as invariant to garment looseness. Multi-camera markerless systems provide the necessary precision for analyzing subtle body movements associated with poor posture while being more cost-effective than marker-based alternatives. Additionally, the ability to use these systems in real-world office environments, without the need for specialized setups, facilitates the integration of MoCap technology into everyday workplace wellness initiatives.

4.5 Limitations

A significant limitation in cloth modeling and motion capture (MoCap) is the inherent variability in garment shape, often resulting in tightness in specific regions despite overall volumetric accuracy. This variation presents a challenge in accurately capturing cloth behavior, particularly in areas where the fit is snug or restrictive. To address this issue, we propose calculating the drape while considering different sections of the cloth as well as the body individually. By doing so, we aim to enhance the precision and comprehensiveness of cloth modeling and motion capture techniques, ultimately mitigating the limitations posed by the inherent variability in garment shape as well as classifying the garments to more accurate drape classes.

Simulation is extensively utilized in cloth-related research, such as virtual try-on (Cho et al., 2023) and dataset generation (Bertiche et al., 2020), where it serves as a surrogate for real cloth physics. However, currently, there are no notable metrics or benchmarks that can effectively gauge the accuracy of simulations against real-world cloth counterparts. In forthcoming endeavors, we intend to address this gap by employing 3D scanning in tandem with parallel cloth simulation. This approach will enable us to systematically evaluate the fidelity of simulations by comparing them directly with real-world data. By conducting such evaluations, we aim to establish robust metrics and benchmarks that can accurately quantify the correctness of cloth simulations, thereby advancing the reliability and applicability of virtual cloth-related research.

4.6 Future works

Future work in the realm of state-of-the-art markerless pose estimation models entails enhancing their accuracy through the integration of synthetic drape accurately clothed body models,

TABLE 2 Holistic comparisons for MoCap methods.

	Marker-based (MB)	Monocular markerless (TML & IML)	Multi-camera markerless (MML)
Requirement	Dedicated volume space, multiple (typical 6–12) special cameras and synchronization systems, high throughput computer, active camera or active marker, multiple markers (typical 39–57) of tight placement for full body pose.	Single common digital camera (e.g., smartphone camera), AI-capable computing hardware for model inference, Sufficient subject/background contrast and lighting conditions	Multiple digital cameras (typical 2–6), Camera calibration pattern, AI capable computing hardware for interface, sufficient subject-background contrast and lightning
Setup time	High (marker placement, calibration)	Low	Medium (calibration)
Accuracy	Highly precise for MoCap with skin-tight clothing, bad for loosely fit daily use clothing	Mediocre for both skin-tight clothing, and loosely fit daily use clothing	Comparatively better for both skin-tight clothing, and loosely fit daily use clothing
Cost	High (24–72 disposable markers, 8–12 specialized high-speed motion cameras, one dedicated system)	Low (one smartphone level camera, a deep learning supported system/platform)	Low (Multiple smartphone level cameras, a manual synchronization and calibration system, a deep learning supported system/platform)
Flexibility	Restricted to dedicated space	Can be used in the wild	Can be used in the wild with passive calibration (presence of calibration pattern in the videos)
Remarks	Preferred if the marker placement requirements specified by the producer's manual can be met (e.g., skin-tight clothing), for example, medical or sports evaluations.	Good for in-the-wild captures	Sufficient in most daily activities with loose casual apparel for wearable technology research. The performance on extreme angle motions may be improved in time with ongoing computer vision research

coupled with precise drape amount and 3D pose from camera view. This approach aims to generate datasets with meticulously controlled drape variations, thereby facilitating the training of a drape-invariant pose estimation model by passing the amount of drape as a parameter to the model along with input images through a data-driven approach. By leveraging this methodology, the objective is to bolster the robustness of pose estimation models, particularly in scenarios involving loose garments and intricate movements. Through the amalgamation of synthetic clothed body models and meticulous dataset generation, strides can be made toward achieving more accurate and adaptable pose estimation systems, capable of accommodating the complexities inherent in diverse human movements and attire.

5 Conclusion

In conclusion, our benchmark methodology offers a comprehensive framework for evaluating MoCap methods realistically. The use of 3D physics simulation with true-to-specification inputs ensures accurate representation of human motion and garment dynamics, overcoming challenges associated with real-world scenarios. Implemented through Blender3D and the SMPL-X Blender addon, the simulation pipeline faithfully represents human movements, including soft tissue dynamics and realistic garment deformations. We evaluate marker-based and markerless approaches, including monocular and multi-camera markerless setups. The Evaluation Metrics employ MPJPE and CRMSE for quantitative assessment, ensuring a robust evaluation. In essence, our benchmark methodology establishes a foundation for understanding MoCap methods in realistic scenarios, considering garment dynamics and diverse evaluation metrics. It provides valuable insights into the strengths and limitations of various MoCap techniques, guiding advancements in the field and promoting the development of more accurate MoCap technologies.

Looking ahead, we plan to leverage the rich dataset that can be generated by DMCB+ to train a more robust deep learning model. This model aims to predict more accurate poses, particularly in scenarios involving loose garments. By utilizing the improved benchmark data, we anticipate achieving better performance and advancing state-of-the-art pose prediction within dynamic and challenging environments.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

LR: Conceptualization, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. BZ: Conceptualization, Formal analysis, Investigation, Methodology, Supervision, Writing – original draft, Writing – review & editing. SS: Formal analysis, Investigation, Methodology, Supervision, Writing – review

& editing. PL: Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The research reported in this paper was supported by the BMBF (German Federal Ministry of Education and Research) in the project VidGenSense (01IW21003).

Acknowledgments

We acknowledge the use of ChatGPT, for aiding in text editing and rephrasing during the writing of the paper while following the guidelines provided by Frontiers.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

References

Akhter, I., and Black, M. J. (2015). "Pose-conditioned joint angle limits for 3d human pose reconstruction," in *Proceedings of the IEEE conference on computer vision and pattern recognition* (New York, NY: IEEE), 1446–1455. doi: 10.1109/CVPR.2015.7298751

An, S., Li, Y., and Ogras, U. (2022). mRI: multi-modal 3d human pose estimation dataset using mmwave, rgb-d, and inertial sensors. *Adv. Neural Inf. Process. Syst.* 35, 27414–27426. doi: 10.48550/arXiv.2210.08394

Ancans, A. (2021). Wearable sensor clothing for body movement measurement during physical activities in healthcare. *Sensors*. 21, 2068. doi: 10.3390/s21062068

Barca, J. C., Rumantir, G., and Li, R. K. (2006). "A new illuminated contour-based marker system for optical motion capture," in *2006 Innovations in Information Technology* (New York, NY: IEEE), 1–5.

Batpurev, T. (2021). *bodypose3d*. Available online at: https://github.com/TemugeB/bodypose3d (accessed January 1, 2024).

Bazarevsky, V., Grishchenko, I., Raveendran, K., Zhu, T., Zhang, F., and Grundmann, M. (2020). Blazepose: on-device real-time body pose tracking. *arXiv* [preprint]. doi: 10.48550/arXiv.2006.10204

Behera, A., Wharton, Z., Keidel, A., and Debnath, B. (2020). Deep cnn, body pose, and body-object interaction features for drivers' activity monitoring. *IEEE Transact. Intell. Transport. Syst.* 23, 2874–2881. doi: 10.1109/TITS.2020.3027240

Bello, H., Zhou, B., Suh, S., and Lukowicz, P. (2021). "Mocapaci: posture and gesture detection in loose garments using textile cables as capacitive antennas," in *Proceedings of the 2021 ACM International Symposium on Wearable Computers* (New York, NY: ACM), 78–83.

Bertiche, H., Madadi, M., and Escalera, S. (2020). "Cloth3d: clothed 3d humans," in *European Conference on Computer Vision* (New York, NY: Springer), 344–359.

Blender Foundation (2023). *Blender*. Computer Software. Amsterdam.

Casas, D., and Comino-Trinidad, M. (2023). SMPLitex: a generative model and dataset for 3D human texture estimation from single image," in *British Machine Vision Conference (BMVC)* (Durham: BMVA).

Chatzis, T., Stergioulas, A., Konstantinidis, D., Dimitropoulos, K., and Daras, P. (2020). A comprehensive study on deep learning-based 3d hand pose estimation methods. *Appl. Sci.* 10:6850. doi: 10.3390/app10196850

Cho, Y., Ray, L. S. S., Thota, K. S. P., Suh, S., and Lukowicz, P. (2023). "Clothfit: Cloth-human-attribute guided virtual try-on network using 3d simulated dataset," in *2023 IEEE International Conference on Image Processing (ICIP)* (New York, NY: IEEE), 3484–3488.

Dittadi, A., Dziadzio, S., Cosker, D., Lundell, B., Cashman, T. J., and Shotton, J. (2021). "Full-body motion from a single head-mounted device: generating smpl poses from partial observations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (New York, NY: IEEE), 11687–11697.

Dong, J., Fang, Q., Jiang, W., Yang, Y., Huang, Q., Bao, H., et al. (2021). Fast and robust multi-person 3d pose estimation and tracking from multiple views. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 6981–6992. doi: 10.1109/TPAMI.2021.3098052

Dubey, S., and Dixit, M. (2023). A comprehensive survey on human pose estimation approaches. *Multim. Syst.* 29, 167–195. doi: 10.1007/s00530-022-00980-0

Fleisig, G. S., Slowik, J. S., Wassom, D., Yanagita, Y., Bishop, J., and Diffendaffer, A. (2022). Comparison of marker-less and marker-based motion capture for baseball pitching kinematics. *Sports Biomech.* 1–10. doi: 10.1080/14763141.2022.2076608

Gamra, M. B., and Akhloufi, M. A. (2021). A review of deep learning techniques for 2d and 3d human pose estimation. *Image Vis. Comput.* 114:104282. doi: 10.1016/j.imavis.2021.104282

Gong, J., Foo, L. G., Fan, Z., Ke, Q., Rahmani, H., and Liu, J. (2023). "Diffpose: toward more reliable 3d pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (New York, NY: IEEE), 13041–13051.

Gong, J., Zhang, X., Huang, Y., Ren, J., and Zhang, Y. (2021). Robust inertial motion tracking through deep sensor fusion across smart earbuds and smartphone. *Proc. ACM Interact. Mobile Wear. Ubiq. Technol.* 5, 1–26. doi: 10.1145/3463517

Groen, B. E., Geurts, M., Nienhuis, B., and Duysens, J. (2012). Sensitivity of the olga and vcm models to erroneous marker placement: effects on 3d-gait kinematics. *Gait Post.* 35, 517–521. doi: 10.1016/j.gaitpost.2011.11.019

Jansen, B., Temmermans, F., and Deklerck, R. (2007). "3d human pose recognition for home monitoring of elderly," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (New York, NY: IEEE), 4049–4051.

Jiang, Y., Ye, Y., Gopinath, D., Won, J., Winkler, A. W., and Liu, C. K. (2022). "Transformer inertial poser: Real-time human motion reconstruction from sparse imus with simultaneous terrain generation," in *SIGGRAPH Asia 2022 Conference Papers* (New York, NY: ACM), 1–9.

Jin, H., Yang, Z., Kumar, S., and Hong, J. I. (2018). Towards wearable everyday body-frame tracking using passive rfids. *Proc. ACM Interact. Mobile Wear. Ubiq. Technol.* 1, 1–23. doi: 10.1145/3161199

Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., et al. (2015). "Panoptic studio: a massively multiview system for social motion capture," in *Proceedings of the IEEE International Conference on Computer Vision* (New York, NY: IEEE), 3334–3342.

Kanazawa, A., Black, M. J., Jacobs, D. W., and Malik, J. (2018). "End-to-end recovery of human shape and pose," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (New York, NY: IEEE), 7122–7131.

Kanko, R. M., Laende, E. K., Davis, E. M., Selbie, W. S., and Deluzio, K. J. (2021). Concurrent assessment of gait kinematics using marker-based and markerless motion capture. *J. Biomech.* 127:110665. doi: 10.1016/j.jbiomech.2021.110665

Lee, Y., and Yoo, H. (2017). Low-cost 3d motion capture system using passive optical markers and monocular vision. *Optik* 130, 1397–1407. doi: 10.1016/j.ijleo.2016.11.174

Liu, R., Shao, Q., Wang, S., Ru, C., Balkcom, D., and Zhou, X. (2019). Reconstructing human joint motion with computational fabrics. *Proc. ACM Interact. Mobile Wear. Ubiq. Technol.* 3, 1–26. doi: 10.1145/3314406

Liu, S. (2020). A wearable motion capture device able to detect dynamic motion of human limbs. *Nat. Commun.* 11, 5615. doi: 10.1038/s41467-020-19424-2

Liu, Y., Yan, J., Jia, F., Li, S., Gao, A., Wang, T., et al. (2023). "Petrv2: a unified framework for 3d perception from multi-camera images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (New York, NY: IEEE), 3262–3272.

Loper, M., Mahmood, N., and Black, M. J. (2014). Mosh: motion and shape capture from sparse markers. *ACM Trans. Graph.* 33, 220–221. doi: 10.1145/2661229.2661273

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., and Black, M. J. (2015). SMPL: A skinned multi-person linear model. *ACM Trans. Graphics* 248:16. doi: 10.1145/2816795.2818013

Mahmood, N., Ghorbani, N., Troje, N. F., Pons-Moll, G., and Black, M. J. (2019). "Amass: archive of motion capture as surface shapes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (New York, NY: IEEE), 5442–5451.

McAdams, E., Krupaviciute, A., Géhin, C., Grenier, E., Massot, B., Dittmar, A., et al. (2011). "Wearable sensor systems: The challenges," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (New York, NY: IEEE), 3648–3651.

McFadden, C., Daniels, K., and Strike, S. (2020). The sensitivity of joint kinematics and kinetics to marker placement during a change of direction task. *J. Biomech.* 101:109635. doi: 10.1016/j.jbiomech.2020.109635

Moon, S., Madotto, A., Lin, Z., Dirafzoon, A., Saraf, A., Bearman, A., et al. (2022). Imu2clip: multimodal contrastive learning for imu motion sensors from egocentric videos and text. *arXiv* [preprint]. doi: 10.48550/arXiv.2210.14395

Nakano, N., Sakura, T., Ueda, K., Omura, L., Kimura, A., Iino, Y., et al. (2020). Evaluation of 3d markerless motion capture accuracy using openpose with multiple video cameras. *Front. Sports Active Living* 2:50. doi: 10.3389/fspor.2020.00050

OptiTrack (2019). *Skeleton Tracking*. Oregon: OptiTrack.

Ostrek, M., Rhodin, H., Fua, P., Ḿuller, E., and Sṕorri, J. (2019). Are existing monocular computer vision-based 3d motion capture approaches ready for deployment? A methodological study on the example of alpine skiing. *Sensors* 19:4323. doi: 10.3390/s19194323

Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A. A., Tzionas, D., et al. (2019). "Expressive body capture: 3d hands, face, and body from a single image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (New York, NY: IEEE), 10975–10985.

Pavllo, D., Feichtenhofer, C., Grangier, D., and Auli, M. (2019). "3d human pose estimation in video with temporal convolutions and semi-supervised training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (New York, NY: IEEE), 7753–7762.

Plappert, M., Mandery, C., and Asfour, T. (2016). The kit motion-language dataset. *Big Data* 4, 236–252. doi: 10.1089/big.2016.0028

Puthenveetil, S. C., Daphalapurkar, C. P., Zhu, W., Leu, M. C., Liu, X. F., Chang, A. M., et al. (2013). "Comparison of marker-based and marker-less systems for low-cost human motion capture," in *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Vol. 55867* (New York, NY: American Society of Mechanical Engineers), V02BT02A036.

Qiu, H., Wang, C., Wang, J., Wang, N., and Zeng, W. (2019). "Cross view fusion for 3d human pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (New York, NY: IEEE), 4342–4351.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021). "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning* (New York, NY: PMLR), 8748–8763.

Raskar, R., Nii, H., Dedecker, B., Hashimoto, Y., Summet, J., Moore, D., et al. (2007). Prakash: lighting aware motion capture using photosensing markers and multiplexed illuminators. *ACM Transact. Graph.* 26, 36-es. doi: 10.1145/1276377.1276422

Ray, L. S. S., Geißler, D., Zhou, B., Lukowicz, P., and Greinke, B. (2023a). "Capafoldable: Self-tracking foldable smart textiles with capacitive sensing," in *UbiComp/ISWC '23 Adjunct* (New York, NY: Association for Computing Machinery), 197.

Ray, L. S. S., Zhou, B., Krupp, L., Suh, S., and Lukowicz, P. (2023b). Synthcal: a synthetic benchmarking pipeline to compare camera calibration algorithms. *arXiv* [preprint]. doi: 10.48550/arXiv.2307.01013

Ray, L. S. S., Zhou, B., Suh, S., and Lukowicz, P. (2023c). "Pressim: an end-to-end framework for dynamic ground pressure profile generation from monocular videos using physics-based 3d simulation," in *2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)* (New York, NY: ACM), 484–489.

Ray, L. S. S., Zhou, B., Suh, S., and Lukowicz, P. (2023d). "Selecting the motion ground truth for loose-fitting wearables: benchmarking optical mocap methods," in *Proceedings of the 2023 ACM International Symposium on Wearable Computers* (New York, NY: ACM), 27–32.

Reilink, R., Stramigioli, S., and Misra, S. (2013). 3d position estimation of flexible instruments: marker-less and marker-based methods. *Int. J. Comput. Assist. Radiol. Surg.* 8, 407–417. doi: 10.1007/s11548-012-0795-1

Sigal, L. (2021). "Human pose estimation," in *Computer Vision: A Reference Guide* (New York, NY: Springer), 573–592.

Sigal, L., Balan, A. O., and Black, M. J. (2010). Humaneva: synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *Int. J. Comput. Vis.* 87:4. doi: 10.1007/s11263-009-0273-6

Simplycloth (2022). *Simplycloth*. Available from: https://blendermarket.com/products/simply-cloth (accessed May 23, 2023).

Trumble, M., Gilbert, A., Malleson, C., Hilton, A., and Collomosse, J. (2017). "Total capture: 3d human pose estimation fusing video and inertial sensors," in *2017 British Machine Vision Conference (BMVC)* (Durham: BMVA).

Tu, H., Wang, C., and Zeng, W. (2020). "Voxelpose: towards multi-camera 3d human pose estimation in wild environment," in *Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I 16* (New York, NY: Springer), 197–212.

University of Cyprus (2023). *Dance Motion Capture Database*. University of Cyprus. Available online at: http://dancedb.cs.ucy.ac.cy (accessed January 1, 2024).

Wang, J., Tan, S., Zhen, X., Xu, S., Zheng, F., He, Z., et al. (2021). Deep 3d human pose estimation: a review. *Comp. Vis. Image Understand.* 210:103225. doi: 10.1016/j.cviu.2021.103225

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., and Girshick, R. (2019). *Detectron2*. Available online at: https://github.com/facebookresearch/detectron2 (accessed January 1, 2024).

Xu, J., Yu, Z., Ni, B., Yang, J., Yang, X., and Zhang, W. (2020). "Deep kinematics analysis for monocular 3d human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (New York, NY: IEEE), 899–908.

Yi, C., Wei, B., Ding, Z., Yang, C., Chen, Z., and Jiang, F. (2022a). A self-aligned method of imu-based 3-dof lower-limb joint angle estimation. *IEEE Trans. Instrum. Meas.* 71, 1–10. doi: 10.1109/TIM.2022.3194935

Yi, X., Zhou, Y., Habermann, M., Shimada, S., Golyanik, V., Theobalt, C., et al. (2022b). "Physical inertial poser (pip): physics-aware real-time human motion tracking from sparse inertial sensors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (New York, NY: IEEE), 13167–13178.

Zhao, Q., Zheng, C., Liu, M., Wang, P., and Chen, C. (2023). "Poseformerv2: exploring frequency domain for efficient and robust 3d human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (New York, NY: IEEE), 8877–8886.

Zheng, C., Wu, W., Chen, C., Yang, T., Zhu, S., Shen, J., et al. (2023). Deep learning-based human pose estimation: a survey. *ACM Comp. Surv.* 56, 1–37. doi: 10.1145/3603618

Zhou, B., Geissler, D., Faulhaber, M., Gleiss, C. E., Zahn, E. F., Ray, L. S. S., et al. (2023). Mocapose: motion capturing with textile-integrated capacitive sensors in loose-fitting smart garments. *Proc. ACM Interact. Mobile Wear. Ubiq. Technol.* 7, 1–40. doi: 10.1145/3580883

Zuo, X., Wang, S., Zheng, J., Yu, W., Gong, M., Yang, R., et al. (2021). Sparsefusion: dynamic human avatar modeling from sparse rgbd images. *IEEE Transact. Multim.* 23, 1617–1629. doi: 10.1109/TMM.2020.3001506