Towards Unstructured Unlabeled Optical Mocap: *A Video Helps!*

Nicholas Milef
nicholas.milef@tamu.edu
Texas A&M University
College Station, Texas, USA

John Keyser
keyser@cse.tamu.edu
Texas A&M University
College Station, Texas, USA

Shu Kong[*]
skong@um.edu.mo,shu@tamu.edu
University of Macau
Macau, China
Texas A&M University
College Station, Texas, USA

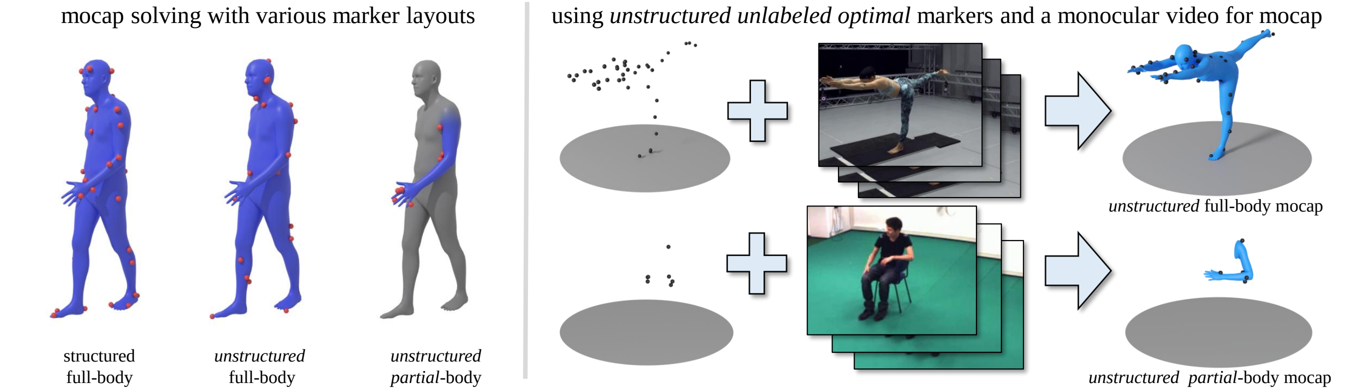
mocap solving with various marker layouts
structured full-body
unstructured full-body
unstructured partial-body
using *unstructured unlabeled optimal* markers and a monocular video for mocap
unstructured full-body mocap
unstructured partial-body mocap

Figure 1: We solve the problem of unstructured unlabeled optical (UUO) motion caption (mocap), in which retroreflective optical markers are placed in an unstructured way on the body. Importantly, markers are not manually labeled. UUO mocap reduces human effort to set up mocap environments but is more challenging than typical mocap settings that either manually label markers or place markers w.r.t some predefined structured layouts. To solve UUO mocap, we leverage a monocular video captured alongside markers and use it to extract an initial body model as a prior for subsequent optimization for body pose, shape, global translation, and rotation.

ABSTRACT

Optical motion capture (mocap) requires accurately reconstructing the human body from retroreflective markers, including pose and shape. In a typical mocap setting, marker labeling is an important but tedious and error-prone step. Previous work has shown that marker labeling can be automated by using a *structured* template defining specific marker placements, but this places additional recording constraints. We propose to relax these constraints and solve for *Unstructured Unlabeled Optical* (UUO) mocap. Compared to the typical mocap setting that either labels markers or places them w.r.t a structured layout, markers in UUO mocap can be placed anywhere on the body and even on one specific limb (e.g., right leg for biomechanics research), hence it is of more practical significance. It is also more challenging. To solve UUO mocap, we exploit a monocular video captured by a single RGB camera, which does not require camera calibration. On this video, we run an off-the-shelf method to reconstruct and track a human individual, giving strong visual priors of human body pose and shape. With both the video and UUO markers, we propose an optimization pipeline towards marker identification, marker labeling, human pose estimation, and human body reconstruction. Our technical novelties include multiple hypothesis testing to optimize global orientation, and marker localization and marker-part matching to better optimize for body surface. We conduct extensive experiments to quantitatively compare our method against state-of-the-art approaches, including marker-only mocap and video-only human body/shape reconstruction. Experiments demonstrate that our method resoundingly outperforms existing methods on three established benchmark datasets for both full-body and partial-body reconstruction.

CCS CONCEPTS

• **Computing methodologies → Motion capture**; *Reconstruction*; *Animation*.

KEYWORDS

motion capture, human body reconstruction, partial-body reconstruction.

ACM Reference Format:
Nicholas Milef, John Keyser, and Shu Kong. 2024. Towards Unstructured Unlabeled Optical Mocap: *A Video Helps!*. In *Special Interest Group on Computer*

[*] Corresponding author.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

SIGGRAPH Conference Papers '24, July 27-August 1, 2024, Denver, CO, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0525-0/24/07.
https://doi.org/10.1145/3641519.3657522

Graphics and Interactive Techniques Conference Conference Papers '24 (SIGGRAPH Conference Papers '24), July 27-August 1, 2024, Denver, CO, USA. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3641519.3657522

1 INTRODUCTION

Human reconstruction is a crucial component for creating realistic humans in movies and games [Bregler 2007; Holden 2018; West III 2019], biomechanics analysis [Averta et al. 2021; Camargo et al. 2021; Moeslund et al. 2006; Roetenberg et al. 2009; van der Zee et al. 2022], and computer vision applications [Kocabas et al. 2020; Rempe et al. 2021; Wang et al. 2023]. This is a challenging problem as individuals have different body shapes and can express various poses. Optical motion capture (mocap) systems have been the de facto system to capture pose and body shape due to their high accuracy in determining 3D marker positions [Merriaux et al. 2017]. These systems use multi-view infrared cameras to recover the positions of retroreflective markers placed on the body. Subsequently, one can fit a 3D body model (e.g., SMPL [Loper et al. 2015]) to the marker positions if the placement and corresponding body parts of markers are known [Loper et al. 2014; Pavlakos et al. 2019].

In optical mocap, accurate body reconstruction typically requires manually labeling markers and consistent placement [Loper et al. 2014; Mahmood et al. 2019]. When provided labeled markers, approaches such as HuMoR [Rempe et al. 2021] and VPoser [Pavlakos et al. 2019] can fit a 3D human body to marker locations by solving for pose and body shape through optimization. However, manually labeling markers is prone to errors and is time consuming; without labels for markers, approaches like HuMoR and VPoser can fail to reconstruct pose as optimization easily gets stuck on bad local minima. Therefore, some methods propose to automate marker labeling [Ghorbani and Black 2021; Ghorbani et al. 2019]. These methods are trained on some predetermined marker layouts and struggle to label markers placed w.r.t unseen layouts. Such layouts could come from a more user-friendly setup that allows markers to be placed anywhere on the body, i.e. *unstructured* mocap. Finally, current approaches are not able to handle partial marker layouts such as markers placed on only the left leg or right shoulder. *Partial-body* reconstruction is critical for biomechanics research [Averta et al. 2021; Camargo et al. 2021; Moeslund et al. 2006; Roetenberg et al. 2009; van der Zee et al. 2022] which often seeks to minimize the number of unnecessary markers. However, it is difficult to precisely understand how markers cover the body part without marker labels or a full body reference.

We summarize the present dilemma of mocap: *accurate labeling is crucial to accurate pose and body reconstruction, yet accurate labeling relies on accurate pose and body shape, which is challenging, if not impossible, with unstructured markers.*

This dilemma motivates our work of solving *Unstructured Unlabeled Optical* (UUO) mocap, aiming for simultaneous human body reconstruction and pose estimation. Inspired by recent advances in human reconstruction from monocular videos [Goel et al. 2023], we leverage a video captured by a commodity camera (such as a cellphone) along with UUO markers for mocap. It is worth noting that the UUO mocap setup only requires the video to be temporally synchronized w.r.t optical markers. It does not require (1) marker identification from video frames, and (2) camera calibration between the camera and multi-view infrared cameras in the mocap studio. In other words, we exploit the monocular video to obtain a human body prior to aid mocap. Hence, methods developed in this setup can generalize across a wide range of optical mocap systems.

We leverage the following insights to assist in combining UUO markers and the corresponding monocular video for solving mocap. First, modern pose estimation techniques from monocular video, though they struggle to predict global translation and absolute size, tend to produce relatively accurate poses and correctly estimate proportional body shape. We therefore use such estimations to serve as pose priors for 3D model fitting. Second, part-based segmentation of markers, which assigns a body part label to each marker, is relatively easy to solve. We leverage a statistical human model together with the insight that body parts are relatively rigid to help find optimal marker fits. By finding motion correspondence between markers and video-estimated human motion, we can jointly label the markers and solve for human pose and shape. Importantly, our method does not expect a structured marker layout; the markers can be placed anywhere or just on part of the body. Partial-body mocap is especially useful for biomechanics mocap [Averta et al. 2021; Camargo et al. 2021; van der Zee et al. 2022] and animal mocap (where marker templates may not be available) [Abson and Palmer 2015; Zhang et al. 2018].

We make three major contributions (cf. §3).

- **Problem statement.** We introduce the problem of Unstructured Unlabeled Optical (UUO) mocap, aiming for simultaneous body reconstruction and pose estimation using UUO markers. It relaxes the constraints of marker placement and requires no manual work of labeling markers.
- **New strategy.** To solve UUO mocap, we leverage a monocular video captured alongside markers to extract a body prior using an off-the-shelf method of monocular human body reconstruction. We use this body prior for optimizing body/part pose, size, and marker locations.
- **Novel techniques.** We propose a UUO mocap pipeline consisting of multiple novel techniques such as (1) a multi-stage fitting process for temporally-stable motion reconstruction, (2) identifying and localizing markers by finding their correspondence to individual body parts, and (3) multiple hypothesis testing for rotational alignment of body mesh.

Code and data are at https://github.com/NicholasMilef/UUO-Mocap.

2 RELATED WORK

2.1 Statistical Human Models

Mocap markers are placed near the surface of the skin, so one can use the marker 3D locations to model the body mesh. Various statistical methods propose to model the human body [Loper et al. 2015; Pavlakos et al. 2019; Xu et al. 2020] by introducing different vertex offsets from a based mesh template (e.g., blend shapes) that can be controlled through a learned parameter space. SMPL [Loper et al. 2015] is a well-established statistical human model that has gained popularity in various applications, The SMPL model is parameterized by body shape $\beta \in \mathbb{R}^{10}$, pose $\Theta \in \mathbb{R}^{23}$, global translation $\Gamma \in \mathbb{R}^3$, and global orientation $\Phi \in \mathbb{R}^3$. The SMPL model is differentiable w.r.t vertex positions V and joint positions J defined as function $\mathcal{M}$: $[J, V] = \mathcal{M}(\Phi, \Theta, \beta) + \Gamma$. To solve for human pose and shape reconstruction from markers, various methods fit an SMPL model to the labeled markers. For example, VPoser [Pavlakos et al. 2019], a conditional variational autoencoder, learns a pose prior from the AMASS [Mahmood et al. 2019] dataset and fits SMPL to labeled keypoints. HuMoR [Rempe et al. 2021] extends this idea by using a motion prior to assist in keypoint fitting for both image and motion capture applications. However, both of these approaches struggle to find strong initialization priors using just *unlabeled* marker point clouds. Our work also uses SMPL [Loper et al. 2015] to represent a 3D body but optimizes it for solving mocap and human body reconstruction in the *UUO mocap setting*.

2.2 Motion Capture Solving

Motion capture solving typically uses *labeled* markers to determine body pose and/or shape via optimization [Loper et al. 2014; Mahmood et al. 2019], which fits a body model to markers by minimizing distances between labeled markers and vertices of the body model. Recent approaches perform mocap solving and marker denoising via deep learning [Chen et al. 2021; Han et al. 2018; Pan et al. 2023]. These approaches assume the marker inputs to be labeled already. Some prior works attempt to mitigate the need of marker labeling. Han et al. [2018] use a deep neural network and bipartite matching to label mocap markers placed on a hand and assume a structured marker layout. Holden [2018] uses a residual network to jointly denoise marker positions and solve for pose for each frame. MocapSolver [Chen et al. 2021] estimates marker offsets from each skeletal joint, bone lengths, and the pose using a window of frames consisting of mocap marker positions. Other follow-up works adopt deep learning and optimization to improve mocap using unlabeled markers [Pan et al. 2023; Tang et al. 2023]. However, these methods all require a known marker layout. Our work distinguishes from existing ones in that we, for the first time, solve mocap using *unstructure unlabeled optical (UUO) markers.*

2.3 Automatic Marker Labeling

Traditional mocap workflows require technicians to manually label markers, which is time-consuming and error-prone. Hence, some works study marker auto-labeling. Among plenty of marker auto-labeling and mocap solvers, Meyer et al. [2014] propose an online labeling solution, but they require the actor to perform a T-pose for initialization, making it unsuitable for archival data. Schubert et al. [2015] propose an automatic mocap solver and marker location finder with a reasonably dense marker layout. However, their method needs a database of human pose/shape templates, limiting the method to poses and shapes present in the database. Alexanderson et al. [2017] propose an algorithm that solves unlabeled markers on the hands and head but does not temporally lock the marker labels and requires knowledge of the orientation of these body parts before running. Recent works propose to train neural networks on a database of defined human pose/shape layouts and use the trained networks to automatically label markers and solve mocap [Clouthier et al. 2021; Ghorbani and Black 2021; Ghorbani et al. 2019]. For example, SOMA [Ghorbani and Black 2021] labels unlabeled mocap markers through a per-frame self-attention network. After labeling, the markers can then be used to optimize for body shape and pose using MoSh++ [Mahmood et al. 2019]. While SOMA trains a "SuperSet" that can work for a variety of known marker layouts, it does not generalize to unseen or partial-body layouts. In sum, existing approaches require a database of defined layouts to train networks for marker labeling and mocap, limiting their generalization to new marker layouts. Our problem of UUO mocap does not provide marker labels and requests the study of mocap with unstructured markers, hence solutions to this problem is of practical significance in mocap systems.

2.4 Monocular Video Mocap

Markerless mocap has become a popular alternative to traditional optical marker-based mocap systems. Mocap from monocular RGB video is the most accessible form of markerless mocap due to the ubiquity of RGB cameras. For monocular video mocap, recent methods adopt model-based optimization [Bogo et al. 2016; Pavlakos et al. 2019; Rempe et al. 2021] and deep learning [Goel et al. 2023; Kanazawa et al. 2018; Zhang et al. 2023, 2021], resulting in reasonably accurate poses and proportional body shape (e.g., absolute measures such as height may be inaccurate). However, recovering global position and rotation in the world is still difficult during monocular reconstruction [Ye et al. 2023; Yuan et al. 2022]. Some approaches have sought to augment different forms of tracking data. One popular form of mocap, due to cheap cost and less setup, is Inertial Measurement Unit (IMU) based motion capture. Combining video and IMU data has been shown to be more effective than just using video or IMU data alone [Pearl et al. 2023; Tan et al. 2022]. The combination even approaches optical mocap accuracy [Shin et al. 2023b]. However, IMU sensors have a tendency to drift that video cannot fully fix, causing lower accuracy [Van der Kruk and Reijne 2018]. Another approach has been to use depth maps with optical mocap marker positions. Some approaches [Chatzitofis et al. 2022, 2021] use a few low-cost multi-view depth cameras to jointly label markers and solve for pose but theydo not match the performance of professional optical mocap systems. In our work, we show that monocular video can *assist* in optical marker-based mocap. Experiments demonstrate that our mocap solver, which exploits both UUO markers and the corresponding monocular video, outperforms methods based on either, approaching the performance of a labelled-marker based solver.

3 PROBLEM DEFINITION AND METHODOLOGY

We first present the formal problem definition of Unstructured Unlabeled Optical (UUO) mocap, then explain our method for solving UUO mocap, and finally present important implementation details.

3.1 Problem Definition

The problem of *unlabeled and unstructured optical (UUO) mocap* aims to reconstruct a full/partial body pose and shape from a sequence of *unlabeled* markers, which are placed without a predefined structure on an individual's body or body part. The problem relaxes some unfriendly constraints in existing mocap: (1) it does not require manual labeling for the markers (so alleviating manual intervention), and (2) it does not require placing markers on a predefined layout (so reducing human effort in mocap setup and allowing reconstructing a partial body). While solutions to UUO

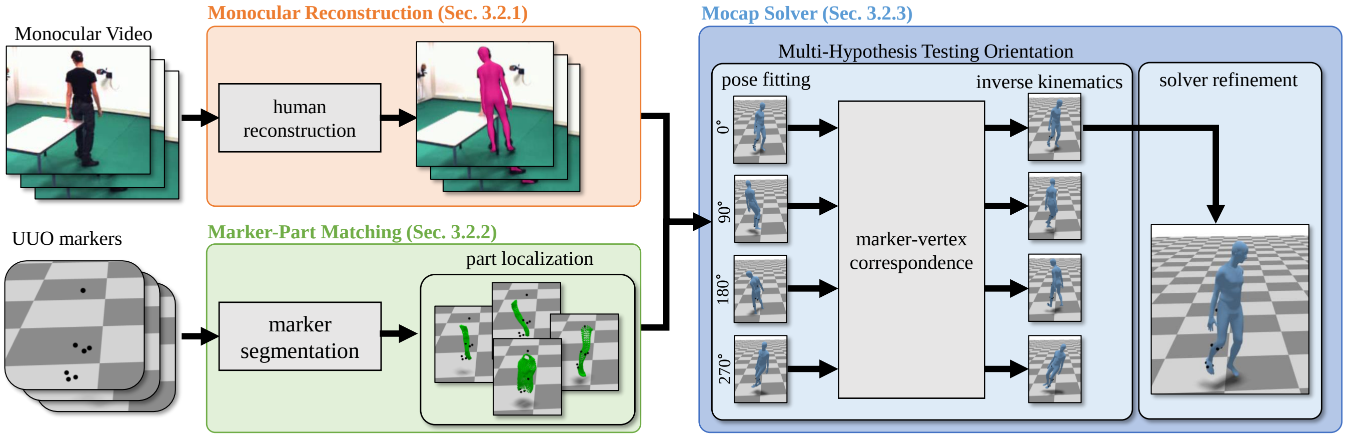
Monocular Video
Monocular Reconstruction (Sec. 3.2.1)
human reconstruction
Mocap Solver (Sec. 3.2.3)
Multi-Hypothesis Testing Orientation
pose fitting
inverse kinematics
solver refinement
0°
90°
180°
270°
marker-vertex correspondence
UUO markers
Marker-Part Matching (Sec. 3.2.2)
part localization
marker segmentation

Figure 2: The proposed pipeline of our UUO mocap solver consists of three modules (cf. details in Section 3). Our method takes as input monocular video and UUO markers to jointly predict marker labels, pose, and body shape. First, we use an off-the-shelf method (HMR2.0 [Goel et al. 2023]) to generate a human prior from the video. Then, we segment the 3D mocap markers to estimate the number of bones that need to be reconstructed. Then, we search for the best-fitting body part. Finally, we solve for the pose and body shape through a novel optimization process.

mocap are of practical significance in mocap, solving mocap under the UUO setting is more challenging than existing settings due to the lack of marker labels and any predefined structured marker placement. Next, we elaborate our method for UUO mocap.

3.2 The Proposed UUO Mocap Method

The core insight of our method is using a monocular video to reconstruct human body as a prior for subsequent mocap solving. Our method takes as input the sequence of UUO markers and the corresponding video, and maps them into SMPL [Loper et al. 2015] parameters. We denote M markers from T frames as a set of 3D points by $m \in \mathbb{R}^{T\times|M|\times 3}$. Fig. 2 sketches the pipeline of our method, consisting of three modules. First, *monocular reconstruction* produces an initial SMPL model from the video that extracts pose, shape, and relative rotation across time (§3.3). Second, *marker-part matching* finds the best correspondences between UUO markers and the initial SMPL model (§3.4). Third, *mocap solver* takes the output of the marker-part correspondence and optimizes for the final result (§3.5). We elaborate them in the following three subsections.

3.3 Monocular Reconstruction from Video

Monocular video contains important visual cues that are not present in raw marker data and can serve as a strong prior for marker fitting. Our strategy of exploiting this video is to estimate an initial body from it. Human reconstruction from monocular video is a well-studied area and various methods have been proposed in the literature. In this work, we use the state-of-the-art method HMR2.0 [Goel et al. 2023] as an off-the-shelf method, which returns parameters of the well-established SMPL model [Loper et al. 2015], e.g., pose Θ and shape β. It is worth noting that, while SMPL is widely used in mocap from (labeled) markers, we make the first attempt of using it to represent a body prior extracted from a monocular video for UUO mocap. As a monocular video has ambiguities in body size, orientation, and world translation, the SMPL model estimated from it unlikely aligns with real marker positions. Therefore, we adopt the next modules to incorporate it for mocap solving.

3.4 Marker-Part Matching

Mocap essentially requires finding the correspondence of markers and body (or vertices of body surface). With the initial SMPL model estimated from the monocular video in the previous module, intuitively one can optimize this SMPL model by fitting it to the UUO markers. However, directly solving this optimization problem can easily get stuck to bad local minima. Therefore, we aim to provide a better initialization for mocap solving with a marker-part matching module, which finds correspondences of markers and body parts. This module is not only important for full-body mocap but also particularly crucial for partial-body mocap, because without marker-body correspondence, it is difficult, if not impossible, to find a good matched body part for markers without signals from the full body (cf. Table 2).

In Marker-Part Matching, we first use the median marker position to align the SMPL mesh. This provides a good initialization of global translation (for both full-body and partial-body mocap). Next, we adopt a two-step process to find marker-part correspondence.

3.4.1 Step 1: marker segmentation. Recall that in SMPL, each vertex v on the body surface has an associated linear blend skinning (LBS) weight that is the summation of the bones. The maximal LBS weight for the vertex can be used to indicate which bones the vertex belongs to; vertices belonging to the same bones are approximately related by a rigid transformation. Therefore, finding marker-bone correspondence largely simplifies mocap solving for body part reconstruction. But searching for the best correspondence requires testing all combinations of markers and SMPL bones and hence computationally expensive. Therefore, to reduce the search space, we propose to group markers and presume each group is corresponding to a specific bone. For example, for partial-body mocap of a leg (i.e., thigh, calf, foot), we would only be interested in searching for kinematic chains consisting of 3 bones. Inspired by marker clustering [De Aguiar et al. 2006], we use the standard deviation of the Euclidean distance between every pair of markers across all frames in the sequence. Note that one can easily obtain marker-marker correspondence across time by tracking them. Then we construct an affinity matrix that consists of these standard deviations. Lastly we use agglomerative clustering with average linkage [Pedregosa et al. 2011] and distance threshold of 5mm, resulting into K segmented groups of markers for each timestamp.

3.4.2 Step 2: multiple hypothesis testing for part localization. We adopt a search-based method to determine where the markers are located on the body. Note that the K groups of segmented markers can be interpreted as a kinematic chain $\mathcal{S}$ that contains a group of K bones. Therefore, we aim to find the best match between the K groups of markers and a kinematic chain from a pool of K-bone candidate chains generated from the initial SMPL. For the pool of candidate chains, we extract all possible chains with length K from the hierarchy of SMPL bones. Then, for each of the candidate chain $\mathcal{S}$, whose vertices are denoted as $V_{\mathcal{S}} \subset V$, we fit the vertices of $V_{\mathcal{S}}$ to all the markers M. Concretely, for a marker m at time-t (i.e., $m^{(t)} \in M^{(t)}$), whose corresponding part of the candidate chain $\mathcal{S}$ at time-t is denoted as $V^{(t)}$, we find the closest vertex $v^{(t)} \in V^{(t)}$. Lastly, we select the candidate chain that produces the minimum fitting error, which is defined as $E_{\mathcal{S}} = \lambda_{\overrightarrow{3D}} E_{\overrightarrow{3D}} + \lambda_\beta E_\beta$ that consists of two terms explained below. $E_{\overrightarrow{3D}}$ is a single-directional Chamfer distance loss between markers and vertices of $V_{\mathcal{S}}$:

$$E_{\overrightarrow{3D}} = \frac{1}{|T| \cdot |M|} \sum_{t=1}^{T} \sum_{m^{(t)} \in M^{(t)}} \min_{v^{(t)} \in V^{(t)}} \|v^{(t)} - m^{(t)}\|^2 \tag{1}$$

E_β is a mean-squared error that regularizes the body shape against the initial SMPL body shape β^{img}:

$$E_\beta = (\hat{\beta} - \beta^{\text{img}})^2 / |\beta^{\text{img}}| \tag{2}$$

The number of chain candidates increases with the number of bones K. Searching over a large number of chains, e.g., for full-body reconstruction, can be computationally expensive. Fortunately, for full-body reconstruction, the exact kinematic chain selection is less relevant as our system can generally recover from poor initialization in later stages. Moreover, we reduce searching computation by removing redundant candidate chains which contain ≥90% of the same bones to other candidates. Running this module produces correspondences of UUO markers to vertices of the initial SMPL model, allowing for the subsequent mocap solving towards body pose and shape reconstruction.

3.5 Mocap Solving

Given UUO markers positions and the SMPL model (containing parameters of body shape β, pose Θ, global translation Γ, and global orientation Φ from the previous two modules), we solve mocap by fitting the SMPL model to the mocap marker positions. Intuitively, one can optimize them altogether, but the solution is easily stuck in bad local minima due to the difficulty of optimizing rotation Φ. Therefore, we break up the mocap solving process into a sequence of optimization stages and propose a multiple hypothesis testing for solving rotation.

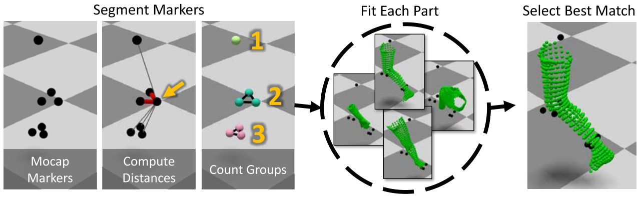

Figure 3: Our Marker-Part Matching first computes the standard deviation of distances between every other marker across all frames, then uses them to construct an affinity matrix to clustering markers into groups, and conducts hypothesis testing to select the best match that produces the minimum fitting error w.r.t the initial body model obtained from the monocular video.

3.5.1 Stage 1: multiple hypothesis testing for root rotation. We first predefine a grid of initialized rotational offsets. We optimize each of them alongside the rest of the SMPL parameters. Then, we select the SMPL parameters from the best-fitting rotational offset. Our work assumes that the global rotations of the initial SMPL meshes and markers differ by a yaw rotation b such that $\Phi_{\text{img}} = b \times \hat{\Phi}$. To facilitate rotation optimization, we independently optimize the rotation offset b, initialized by four uniformly-distributed values $y \in \mathcal{B} = \{0°, 90°, 180°, 270°\}$ using Stages 2 to 4 (described next). We then select the best optimized rotations such that:

$$\beta, \Theta, \Gamma, \Phi = \arg\min_{y \in \mathcal{B}} E_{\overrightarrow{3D}}(\beta^y, \Theta^y, \Gamma^y, \Phi^y) \tag{3}$$

3.5.2 Stage 2: pose fitting. With an initial rotational offset $y \in \mathcal{B}$, we optimize b along with pose Θ, translation Γ, and shape β by minimizing the following:

$$E_{\text{pose}} = \lambda_{\overrightarrow{3D}} E_{\overrightarrow{3D}} + \lambda_\beta E_\beta + \lambda_\Theta E_\Theta \tag{4}$$

where $E_{\overrightarrow{3D}}$ is defined in Eq. 1, $E_{\overrightarrow{\beta}}$ is defined in Eq. 2, and E_Θ regularizes pose which is defined below:

$$E_\Theta = \frac{1}{|T| \cdot |\Theta|} \sum_{t=1}^{T} (\Theta_t^{\text{img}} - \hat{\Theta}_t)^2 \tag{5}$$

3.5.3 Stage 3: marker-vertex correspondence. To better reconstruct the human body, for each marker m, we find the vertex v^m of the optimized SMPL mesh with the closest average distance to m over the entire sequence:

$$v^m = \arg\min_{v \in V} \left(\frac{1}{|T|} \sum_{t=1}^{T} \|v^{(t)} - m^{(t)}\|_2 \right) \tag{6}$$

where $v^{(t)}$ and $m^{(t)}$ denote the vertex and marker at time-t, respectively. With this found marker-vertex correspondence, we adopt inverse kinematics below that refines body reconstruction.

3.5.4 Stage 4: inverse kinematics. With the identified marker locations on the mesh surface from Stage 3, we solve an inverse kinematics (IK) problem via optimizing pose, shape and mesh:

$$E_{\text{IK}} = \lambda_M E_M + \lambda_\beta E_\beta + \lambda_\Theta E_\Theta \tag{7}$$

where minimizing E_M will better align the SMPL mesh with markers. E_M is a squared L2 norm loss between each marker $m^{(t)}$ and its corresponding vertex position v_m:

$$E_M = \frac{1}{|T| \cdot |M|} \sum_{t=1}^{T} \sum_{m^{(t)} \in M} (\|v_m^{(t)} - m^{(t)}\|_2 - \delta)^2 \qquad (8)$$

We set $\delta = 9.5$ (in mm) which is a common mocap marker offset from the skin [Ghorbani and Black 2021].

3.5.5 Stage 5: solver refinement. Finally, we repeat the stages 3 and 4 one time to refine body reconstruction. Instead of regularizing against the initial SMPL model Θ^{img}, we regularize the pose against the output from Stage 4. This helps to reduce distance inconsistencies between markers and the corresponding vertices. We do not repeat more times as more iterations do not yield notable improvements despite more computation.

3.6 Implementation

For our mocap solver, we set the learning rate of the L-BFGS solver as 1.0 for part localization (Step 2) and inverse kinematics (Stage 4) and 0.1 for pose fitting (Stage 2). We use terminal tolerances of 1e-7 on first order optimality and 1e-9 on function value/parameter changes. We process the entire sequence at once and optimize for a maximum of 10k iterations. We tune these hyperparameters on a few random annotated examples (e.g., from UMPM). After tuning, we use the same hyperparameters throughout our experiments for all the datasets (including both full-body and partial-body mocap): $\lambda_{\overrightarrow{\text{3D}}} = 10$ and $\lambda_\beta = 0.1$ for Step 2; $\lambda_\Theta = 1$, $\lambda_{\overrightarrow{\text{3D}}} = 10$, $\lambda_\beta = 1$ at Stage 2; $\lambda_M = 1$, $\lambda_\Theta = 0.1$, $\lambda_\beta = 1$ at Stage 4.

4 EXPERIMENTS

We conduct extensive experiments to validate our method by comparing against prior art. We also show rigorous ablation studies to demonstrate the importance of each step in our proposed optimization pipeline. We start with setups of datasets and metrics, followed by comprehensive results with in-depth analyses and discussions.

4.1 Evaluation Protocols

4.1.1 Metrics. We adopt multiple well-established metrics to comprehensively evaluate methods.

- **MPJPE** measures the mean per-joint position error. It is a common metric used to evaluate human pose reconstruction. Joint position errors are computed using the Euclidean distance between predicted and reference 3D joints.
- **MPJVE** evaluates the mean estimated velocity error of each joint, computed on adjacent poses. This metric evaluates the temporal consistency of a motion.
- **V2V** [Pavlakos et al. 2019] computes the vertex-to-vertex error between the predicted and reference SMPL meshes. It measures both body shape and pose.
- **m2s** measures the marker-to-surface distance. Intuitively, m2s measures the offset of the marker from the surface (skin) of the SMPL mesh. Real mocap markers have a marker offset, so there should be a small m2s even for the reference.

MJPVE is in millimeters per second and the others are in millimeters. For all the metrics, smaller values mean better performance.

4.1.2 Datasets. We use publicly available datasets that have hardware synchronized video and mocap markers. Following the literature [Rempe et al. 2021], we downsample all datasets to 30Hz. We use MoSh++ [Mahmood et al. 2019] to generate reference data for both the CMU Kitchen Pilot dataset and the UMPM dataset. To set up the UUO mocap, we remove marker labels in these datasets. Importantly, we do not use these datasets to train any models so that this simulates the unstructured setup. Therefore, at best, methods can be trained on layouts in their training data but might not know layouts on the testing datasets.

- **CMU Kitchen** (pilot study) [De la Torre et al. 2009] is a challenging dataset due to self-occlusion and prevalence of stationary motions. Interestingly, individuals in this dataset wear backpacks and we remove the markers on the backpack as they are not near the appropriate body landmarks. It has 241 sequences and each is 15s long.
- **UMPM** [van der Aa et al. 2011] contains both markers and synchronized videos. As it uses it uses uncommon labels and marker placements, we create ground-truth label-vertex correspondences for this dataset. As we focus on single-person reconstruction, we use its single-person subset (p1) in our work. It has 12 sequences and each is 15s long.
- **MOYO** [Tripathi et al. 2023] contains many novel and difficult poses that are largely out-of-distribution compared to existing motion capture datasets. Importantly, we do not use this dataset to train models but only use its validation split for evaluation. This helps benchmark the generalization performance of different methods. It has 171 sequences and each is 3s long.

4.1.3 Compared methods. We repurpose and compare existing mocap methods for UUO mocap. First, for the video-only method, we compare HMR2.0 [Goel et al. 2023], a recent algorithm of monocular pose estimation. Because reconstruction from a monocular video contains ambiguities in global position, orientation, and scale, we evaluate its result with a rigid registration step to globally align the body w.r.t the mocap markers. We call this method HMR2.0+RR. Moreover, we repurpose well-established marker-based methods for UUO mocap using appropriate modifications. For example, SOMA [Ghorbani and Black 2021] trains on diverse marker layouts and expects to generalize to unseen layouts. We find that it has limited generalization in experiments; it trains on marker layouts that share marker labels with the CMU Kitchen and performs well on this dataset, but it performs significantly worse on other datasets (Table 1). For SOMA, we run MoSh++ on the marker labels to solve for the SMPL parameters. VPoser [Pavlakos et al. 2019] and HuMoR [Rempe et al. 2021] are not designed for UUO mocap but can be modified for it using appropriate constraints. Importantly, both VPoser and HuMoR are sensitive to having strong initialization, so we initialize them with HMR 2.0's SMPL parameters. This results in modified versions that exploit both markers and video. We denote these as VPoser+V and HuMoR+V. Finally, we use MoSh++ [Mahmood et al. 2019] as the reference achieved by using *labeled* markers.

4.2 Comparisons

Tables 1 and 2 benchmark methods for full- and partial-body reconstruction, respectively. We summarize salient conclusions here.

Table 1: Comparison of different methods for full-body reconstruction with UUO markers on three datasets. HMR2.0 is a method of reconstructing human body from a monocular video, we use rigid registration w.r.t UUO markers for its output (i.e., HMR2.0+RR) as a modified version that can serve UUO mocap. It underperforms the marker-only method SOMA. SOMA trains on marker layouts similar to CMU Kitchen so that it yields better numeric metrics on this dataset than the other two. Nevertheless, our method resoundingly outperforms all the compared approaches, approaching the performance of the reference which uses labeled markers.

Method	Modality	UMPM				MOYO				CMU Kitchen			
		m2s↓	MPJPE↓	MPJVE↓	V2V↓	m2s↓	MPJPE↓	MPJVE↓	V2V↓	m2s↓	MPJPE↓	MPJVE↓	V2V↓
VPoser	markers	204.9	713.5	2962.1	738.7	39.3	612.4	1892.0	638.2	40.4	371.4	857.9	394.5
HuMoR	markers	195.3	651.4	2464.8	689.5	42.3	607.9	1828.0	636.2	44.5	369.3	873.4	395.6
SOMA	markers	26.8	101.1	151.5	95.5	54.3	268.5	102.9	276.2	17.0	88.0	**23.7**	90.9
HMR2.0+RR	markers+video	150.1	334.1	515.1	360.7	146.6	430.5	305.8	448.3	131.9	396.3	267.8	425.4
VPoser+V	markers+video	201.5	524.4	3134.6	572.5	19.2	132.2	1299.1	150.1	33.9	321.2	436.3	404.9
HuMoR+V	markers+video	249.7	558.1	2137.6	598.2	32.5	205.9	1259.7	234.3	33.2	308.7	316.3	376.6
our method	markers+video	**11.0**	**60.8**	**81.5**	**62.6**	**15.5**	**65.2**	**37.9**	**78.1**	**10.6**	**44.2**	27.2	**47.4**
Reference	labeled markers	9.7	0.00	0.00	0.00	27.9	0.0	0.0	0.0	11.1	0.0	0.0	0.0

Table 2: Comparison of different methods for partial-body reconstruction with UUO markers on the UMPM dataset [van der Aa et al. 2011]. Different from the conclusions in full-body reconstruction (Table 1), SOMA, as well as other marker-only methods, underperforms HRM2.0+RR, showing the limitation of marker-only mocap approaches for partial-body reconstruction. Our method still performs the best. Note that our method produces lower m2s than the reference method (achieved by MoSh++ over labeled markers), but it does not indicate which one is factually better than the other because 9.5mm offsets [Mahmood et al. 2019] are expected between markers and a "ground-truth" body mesh.

Method	Modality	Left Leg			Right Arm			Left Shoulder		
		m2s↓	MPJPE↓	MPJVE↓	m2s↓	MPJPE↓	MPJVE↓	m2s↓	MPJPE↓	MPJVE↓
VPoser	markers	265.5	812.1	3137.0	165.2	652.8	2941.4	176.0	696.0	2913.2
HuMoR	markers	241.6	721.9	2618.9	157.3	641.2	2555.3	180.2	632.2	2377.0
SOMA	markers	106.6	678.1	**509.5**	53.0	451.5	625.8	53.6	716.8	610.0
HMR2.0+RR	markers+video	42.0	301.3	751.0	25.4	280.6	506.2	32.1	413.2	610.5
VPoser+V	markers+video	235.1	565.7	3413.7	210.4	620.7	3480.8	115.1	450.0	3003.5
HuMoR+V	markers+video	309.3	627.1	2310.0	250.4	655.0	2405.4	148.2	458.9	1939.8
our method	markers+video	**7.6**	**278.3**	538.6	**8.6**	**143.2**	**208.2**	**8.8**	**384.8**	**454.7**
Reference	labeled markers	11.4	0	0	10.2	0	0	8.4	0	0

First, our method outperforms the compared approaches by a large margin on all datasets for both full- and partial-body mocap. Second, SOMA, a method that involves labeling markers from a database of diverse full-body marker layouts, performs competitively against our method for full-body mocap (Table 1) but performs poorly for partial-body mocap. This demonstrates the challenge of mocap w.r.t unstructured markers, e.g. for partial-body reconstruction. Third, VPoser and HuMoR, originally designed for mocap using labeled markers, perform poorly with unlabeled markers. Assisted by monocular video, they (VPoser+V and HuMoR+V) generally achieve significant improvement for full-body reconstruction (Table 1). However, for partial-body reconstruction, all other methods perform significantly worse than ours (Table 2), further confirming the difficulty of solving mocap using unstructured markers. Our method still outperforms others for partial-body mocap.

We provide visualizations in Figs. 4, 5, and 6 to qualitatively compare different methods. We also attach video demos in the supplementary material. Visual results show that VPoser+V and HuMoR+V both struggle to find correct root alignment during optimization, even with a video prior. Both additionally struggle to generate poses that match the markers well for the UMPM dataset, often resulting in more static poses. In addition, these two methods tend to produce far more jittery motions. Recall that SOMA is trained on a few different marker layouts, it chooses marker labels a superset of positions of these layouts; as UMPM has an unusual marker layout, SOMA does not perform well on it. On a 15-second UMPM sequence data, our method takes 12min (plus 6min for video processing), while SOMA (GPU) and MoSh++(CPU) takes 20min (i5-13600k/Titan RTX).

4.3 Ablation Study

We perform two main ablations studies to demonstrate the effectiveness of different parts of our pipeline.

4.3.1 System design. We highlight some of the key design choices in Table. 3. We show that having multi-hypothesis testing (MHT) is critical for correct alignment for solving. Our method that uses $\mathcal{B} = \{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$ performs substantially better than solving for rotation with a single initial starting angle (i.e., $\mathcal{B} = \{0^\circ\}$).

Table 3: Ablation for multi-hypothesis testing (MHT) for root orientation (Sec. 3.5, Stage 1) on the UMPM dataset [van der Aa et al. 2011]. "- MHT" means that we remove MHT and only optimize the initial orientation. Using multiple initial rotations avoids local minima during mesh alignment and significantly improves mocap performance.

Method	m2s↓	MPJPE↓	MPJVE↓	V2V↓
our method (that uses $\lvert\mathcal{B}\rvert$=4)	**11.0**	**60.8**	**81.5**	**62.6**
- MHT (i.e., $\lvert\mathcal{B}\rvert$=1)	26.6	297.0	421.5	395.0

Table 4: Stage ablations on the UMPM [van der Aa et al. 2011] dataset for the full-body reconstruction task (Sec. 3.5). We progressively evaluate the result after each stage. Note that Stage 3 finds marker-vertex correspondence and does not involve optimization. Results show that the stage-wise optimization clearly leads to better mocap performance.

Stage	m2s↓	MPJPE↓	MPJVE↓	V2V↓
Marker-part matching	69.8	240.4	651.5	283.4
Stage 2: pose fitting	15.7	88.1	620.2	88.3
Stage 4: inverse kinematics	11.6	62.1	89.3	63.6
Stage 5: solver refinement	**11.0**	**60.8**	**81.5**	**62.6**

4.3.2 Stage ablations. Table 4 measures the errors produced after each stage in the third module Mocap Solver (Sec. 3.5). We use the best angle as determined after MHT for root orientation to evaluate stages 2 and 4. Stage 2, the pose fitting stage, is dominated by a Chamfer distance loss and finds best fit on a per-frame basis. Unfortunately, this stage adds considerable jitter, which is reflected by the high MPJVE error. The Stage 4 inverse kinematics effectively removes this jitter because the marker-surface correspondence is locked during the optimization process. Finally, Stage 5 solver refinement helps to mitigate the effect of incorrect marker locations.

5 LIMITATIONS

Our approach sometimes struggles with aligning the correct part to markers, especially when they lack certain identifiable characteristics (Fig. 8). For example, if an actor only has markers on the left leg and jumps with both legs, then the wrong leg could be aligned to the markers. In practice, this may be less of an issue because the use of unilateral partial-body reconstruction is often for isolation movements in biomechanical analysis. Another issue is that very sparse layouts (e.g., UMPM) can be labeled incorrectly due poor coverage. Incorporating physical scene constraints could help (e.g. the floor) improve reconstruction.

6 CONCLUSION

We motivate the problem of mocap with unstructured unlabeled optical (UUO) markers. We propose to exploit a monocular video, captured alongside markers, to estimate a human body prior. Concretely, we extract an initial SMPL model from the video, and use it to optimize human body shape, pose, global translation, and rotation by fitting the SMPL model to the UUO markers. We introduce a pipeline of optimization techniques and show its superior performance over prior art on three benchmark UUO mocap datasets for both full-body and partial-body mocap.

ACKNOWLEDGMENTS

Authors thank Dr. Lingjie Liu for insightful discussions. Shu Kong is partially supported by the University of Macau (SRG2023-00044-FST). The CMU Kitchen data used in this paper was obtained from kitchen.cs.cmu.edu and the data collection was funded in part by the National Science Foundation under Grant No. EEEC-0540865.

REFERENCES

Karl Abson and Ian Palmer. 2015. Motion capture: capturing interaction between human and animal. *The Visual Computer* 31 (2015), 341–353.

Simon Alexanderson, Carol O'Sullivan, and Jonas Beskow. 2017. Real-time labeling of non-rigid motion capture marker sets. *Computers & graphics* 69 (2017), 59–67.

Giuseppe Averta, Federica Barontini, Vincenzo Catrambone, Sami Haddadin, Giacomo Handjaras, Jeremia PO Held, Tingli Hu, Eike Jakubowitz, Christoph M Kanzler, Johannes Kühn, et al. 2021. U-Limb: A multi-modal, multi-center database on arm motion control in healthy and post-stroke conditions. *GigaScience* 10, 6 (2021), giab043.

Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. 2016. Keep It SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 561–578.

Chris Bregler. 2007. Motion capture technology for entertainment [in the spotlight]. *IEEE Signal Processing Magazine* 24, 6 (2007), 160–158.

Romain Brégier. 2021. Deep Regression on Manifolds: A 3D Rotation Case Study. In *2021 International Conference on 3D Vision (3DV)*. 166–174. https://doi.org/10.1109/3DV53792.2021.00027

Jonathan Camargo, Aditya Ramanathan, Will Flanagan, and Aaron Young. 2021. A comprehensive, open-source dataset of lower limb biomechanics in multiple conditions of stairs, ramps, and level-ground ambulation and transitions. *Journal of Biomechanics* 119 (2021), 110320.

Anargyros Chatzitofis, Georgios Albanis, Nikolaos Zioulis, and Spyridon Thermos. 2022. A Low-Cost & Real-Time Motion Capture System. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 21453–21458.

Anargyros Chatzitofis, Dimitrios Zarpalas, Petros Daras, and Stefanos Kollias. 2021. DeMoCap: Low-cost marker-based motion capture. *International Journal of Computer Vision* 129, 12 (2021), 3338–3366.

Kang Chen, Yupan Wang, Song-Hai Zhang, Sen-Zhe Xu, Weidong Zhang, and Shi-Min Hu. 2021. MoCap-Solver: A neural solver for optical motion capture data. *ACM Trans. Graph. (TOG)* 40, 4 (2021), 1–11.

Allison L Clouthier, Gwyneth B Ross, Matthew P Mavor, Isabel Coll, Alistair Boyle, and Ryan B Graham. 2021. Development and validation of a deep learning algorithm and open-source platform for the automatic labelling of motion capture markers. *IEEE Access* 9 (2021), 36444–36454.

Edilson De Aguiar, Christian Theobalt, and Hans-Peter Seidel. 2006. Automatic learning of articulated skeletons from 3d marker trajectories. In *Advances in Visual Computing: Second International Symposium*.

Fernando De la Torre, Jessica Hodgins, Adam Bargteil, Xavier Martin, Justin Macey, Alex Collado, and Pep Beltran. 2009. Guide to the Carnegie Mellon University Multimodal Activity (CMU-MMAC) Database. (2009).

Nima Ghorbani and Michael J. Black. 2021. SOMA: Solving Optical Marker-Based MoCap Automatically. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 11117–11126.

Saeed Ghorbani, Ali Etemad, and Nikolaus F. Troje. 2019. Auto-labelling of Markers in Optical Motion Capture by Permutation Learning. In *Advances in Computer Graphics*, Marina Gavrilova, Jian Chang, Nadia Magnenat Thalmann, Eckhard Hitzer, and Hiroshi Ishikawa (Eds.). Springer International Publishing, Cham, 167–178.

Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. 2023. Humans in 4D: Reconstructing and Tracking Humans with Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 14783–14794.

Shangchen Han, Beibei Liu, Robert Wang, Yuting Ye, Christopher D Twigg, and Kenrick Kin. 2018. Online optical marker-based hand tracking with deep labels. *ACM Trans. Graph. (TOG)* 37, 4 (2018), 1–10.

Daniel Holden. 2018. Robust solving of optical motion capture data by denoising. *ACM Trans. Graph. (TOG)* 37, 4 (2018), 1–12.

Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. 2018. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7122–7131.

Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. 2020. VIBE: Video Inference for Human Body Pose and Shape Estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5253–5263.

Matthew Loper, Naureen Mahmood, and Michael J Black. 2014. MoSh: motion and shape capture from sparse markers. *ACM Trans. Graph.* 33, 6 (2014), 220–1.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34, 6 (Oct. 2015), 248:1–248:16.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. 2019. AMASS: Archive of Motion Capture as Surface Shapes. In *International Conference on Computer Vision*. 5442–5451.

Pierre Merriaux, Yohan Dupuis, Rémi Boutteau, Pascal Vasseur, and Xavier Savatier. 2017. A study of vicon system positioning performance. *Sensors* 17, 7 (2017), 1591.

Johannes Meyer, Markus Kuderer, Jörg Müller, and Wolfram Burgard. 2014. Online marker labeling for fully automatic skeleton tracking in optical motion capture. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*. 5652–5657. https://doi.org/10.1109/ICRA.2014.6907690

Thomas B. Moeslund, Adrian Hilton, and Volker Krüger. 2006. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding* 104, 2 (2006), 90–126.

Xiaoyu Pan, Bowen Zheng, Xinwei Jiang, Guanglong Xu, Xianli Gu, Jingxiang Li, Qilong Kou, He Wang, Tianjia Shao, Kun Zhou, and Xiaogang Jin. 2023. A Locality-based Neural Solver for Optical Motion Capture. , Article 117 (2023), 11 pages. https://doi.org/10.1145/3610548.3618148

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 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 (CVPR)*. 10975–10985.

Owen Pearl, Soyong Shin, Ashwin Godura, Sarah Bergbreiter, and Eni Halilaj. 2023. Fusion of video and inertial sensing data via dynamic optimization of a biomechanical model. *Journal of Biomechanics* 155 (2023), 111617.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. 2022. Tracking people by predicting 3D appearance, location and pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2740–2749.

Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. 2021. HuMoR: 3D Human Motion Model for Robust Pose Estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 11488–11499.

Daniel Roetenberg, Henk Luinge, Per Slycke, et al. 2009. Xsens MVN: Full 6DOF human motion tracking using miniature inertial sensors. *Xsens Motion Technologies BV, Tech. Rep* 1 (2009), 1–7.

Tobias Schubert, Alexis Gkogkidis, Tonio Ball, and Wolfram Burgard. 2015. Automatic initialization for skeleton tracking in optical motion capture. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*. 734–739. https://doi.org/10.1109/ICRA.2015.7139260

Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. 2023a. WHAM: Reconstructing World-grounded Humans with Accurate 3D Motion. *arXiv preprint arXiv:2312.07531* (2023).

Soyong Shin, Zhixiong Li, and Eni Halilaj. 2023b. Markerless Motion Tracking With Noisy Video and IMU Data. *IEEE Transactions on Biomedical Engineering* 70, 11 (2023), 3082–3092. https://doi.org/10.1109/TBME.2023.3275775

Tian Tan, Dianxin Wang, Peter B Shull, and Eni Halilaj. 2022. IMU and smartphone camera fusion for knee adduction and knee flexion moment estimation during walking. *IEEE Transactions on Industrial Informatics* 19, 2 (2022), 1445–1455.

Jilin Tang, Lincheng Li, Jie Hou, Haoran Xin, and Xin Yu. 2023. A Divide-and-conquer Solution to 3D Human Motion Estimation from Raw MoCap Data. In *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. 767–768. https://doi.org/10.1109/VRW58643.2023.00226

Shashank Tripathi, Lea Müller, Chun-Hao P. Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. 2023. 3D Human Pose Estimation via Intuitive Physics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4713–4725.

N.P. van der Aa, X. Luo, G.J. Giezeman, R.T. Tan, and R.C. Veltkamp. 2011. UMPM benchmark: A multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. 1264–1269. https://doi.org/10.1109/ICCVW.2011.6130396

Eline Van der Kruk and Marco M Reijne. 2018. Accuracy of human motion capture systems for sport applications; state-of-the-art review. *European journal of sport science* 18, 6 (2018), 806–819.

Tim J van der Zee, Emily M Mundinger, and Arthur D Kuo. 2022. A biomechanics dataset of healthy human walking at various speeds, step lengths and step widths. *Scientific data* 9, 1 (2022), 704.

Jingbo Wang, Ye Yuan, Zhengyi Luo, Kevin Xie, Dahua Lin, Umar Iqbal, Sanja Fidler, and Sameh Khamis. 2023. Learning Human Dynamics in Autonomous Driving Scenarios. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 20796–20806.

Thomas J West III. 2019. Going Ape: Animacy and affect in Rise of the Planet of the Apes (2011). *New Review of Film and Television Studies* 17, 2 (2019), 236–253.

Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. 2020. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 6184–6193.

Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. 2023. Decoupling human and camera motion from videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21222–21232.

Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. 2022. GLAMR: Global Occlusion-Aware Human Mesh Recovery With Dynamic Cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 11038–11049.

He Zhang, Sebastian Starke, Taku Komura, and Jun Saito. 2018. Mode-adaptive neural networks for quadruped motion control. *ACM Trans. Graph.* 37, 4 (2018), 1–11.

H. Zhang, Y. Tian, Y. Zhang, M. Li, L. An, Z. Sun, and Y. Liu. 2023. PyMAF-X: Towards Well-Aligned Full-Body Model Regression From Monocular Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 10 (oct 2023), 12287–12303. https://doi.org/10.1109/TPAMI.2023.3271691

Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. 2021. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11446–11456.

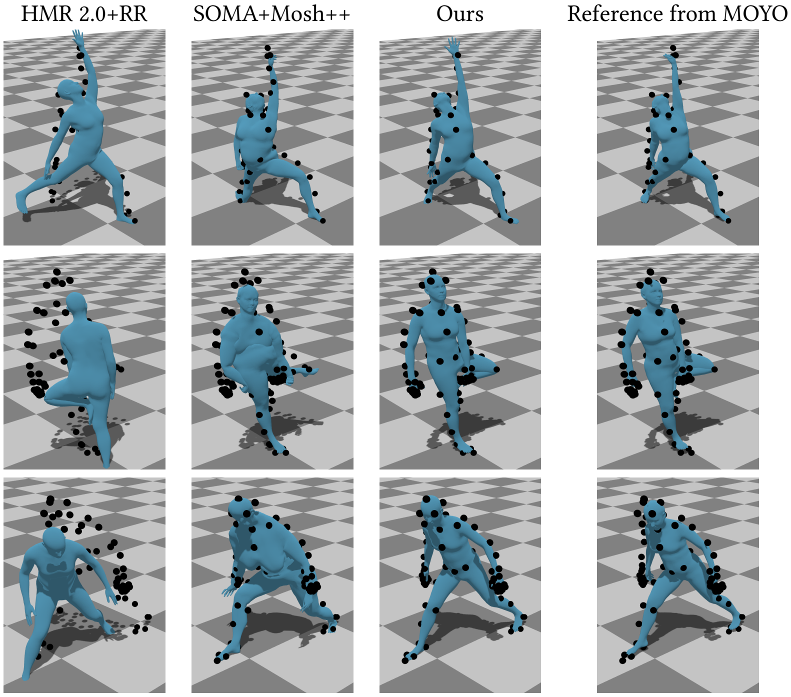

Figure 4: Qualitative results for the validation split of the MOYO dataset [Tripathi et al. 2023]. This dataset is challenging that has unique and difficult poses. Furthermore, markers are densely packed, which can present ambiguity for labeling. SOMA struggles to accurately label the markers, resulting in poor quality reconstruction. Our method produces better visual results.

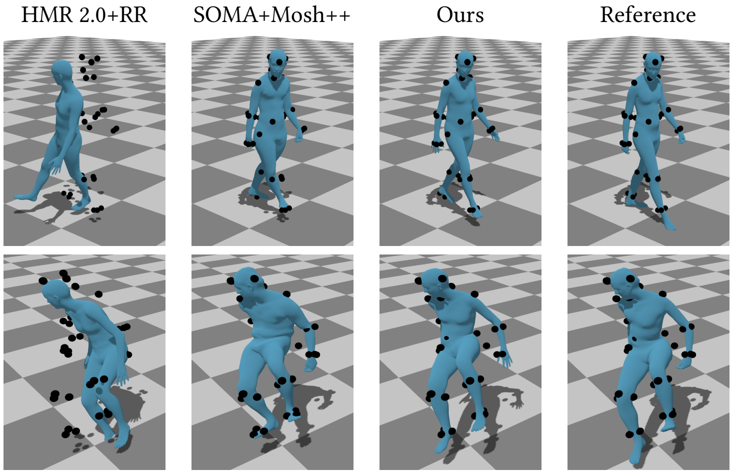

Figure 5: Qualitative results for the validation split of the UMPM dataset [van der Aa et al. 2011]. HMR2.0+RR contains alignment issues, and SOMA produces an incorrect joint position at the right knee. In contrast, our method produces better visual results.

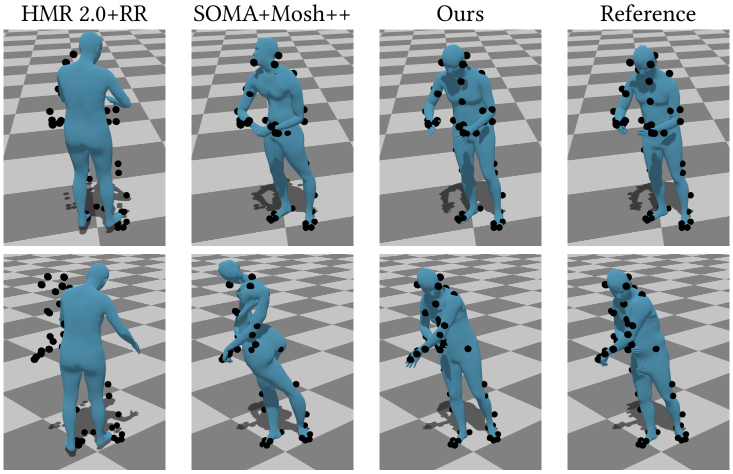

Figure 6: Qualitative results for the validation split of the CMU Kitchen dataset [De la Torre et al. 2009]. Our approach does aligns better to the markers compared to HMR 2.0+RR and produces a closer body shape and poser to the reference compared to SOMA.

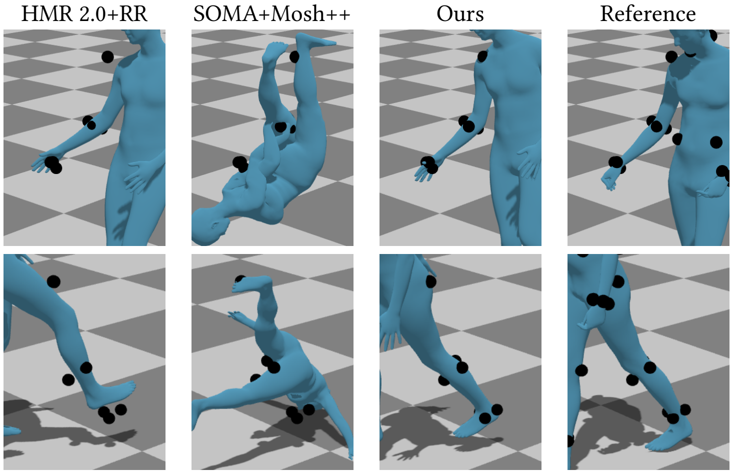

Figure 7: Partial-body reconstruction for the UMPM dataset [van der Aa et al. 2011] for right arm (top row) and left leg (bottom row). SOMA is unable to handle partial body reconstruction; HMR 2.0+RR aligns the correct part due to using our part localization but has noticeable gaps between the markers and the surface and incorrect alignment. Our method produces better visual results.

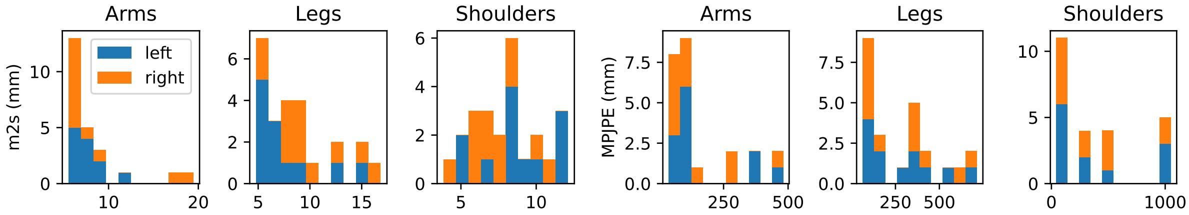

Figure 8: Frequency of errors (left: m2s, right: MPJPE) for individual sequences for partial-body reconstruction. One limitation of our algorithm is that matching the wrong part can lead to higher errors.

A PARTS

For part evaluation, we use the part definitions in Table 5. Each part consists of multiple bones (i.e., SMPL "joints").

B COMPARISONS

In this section, we discuss experimental setups of comparison techniques.

B.1 Reference

B.1.1 UMPM and CMU Kitchen. . Our reference data is computed by using labeled markers with MoSh++ [Mahmood et al. 2019]. However, one problem is that there is variation with the marker placements. While MoSh++ can handle some variation in placement, too much variation can cause error in reconstruction. In our dataset, we observe some errors in fitting markers to joint positions.

For the UMPM [van der Aa et al. 2011] dataset, we manually create the marker-label correspondence between UMPM marker labels and SMPL-X [Pavlakos et al. 2019] vertices. During manual labeling, we cross referenced various videos of the actors and the manual [van der Aa et al. 2011] to determine marker placement. Markers are often placed along a band around limbs, but the orientation of this band can vary and cause some errors in reconstruction. Thus, we report m2s as well as other traditional pose metrics.

The CMU Kitchen Pilot [De la Torre et al. 2009] dataset uses common marker labels, so we use the labels provided in the MoSh++ source code. One problem with this dataset is that the actors wear a backpack with 7 markers (LBWT, NEWLBAC, NEWRBAC, RBAC, RBWT, T10, T8), on the exterior of the backpack. These marker labels traditionally correspond to markers placed on the back, but the backpack adds offsets that distort the body shape considerably. Thus, we test both with and without these markers. While this dataset used hardware synchronization, we found that the source files are not synchronized. Each clip has a synchronization event in which the actor turns on and off a light bulb at the start and end of each trial. We manually synchronize the video and mocap data using these events. Furthermore, we found that the video data is closer to 29.97Hz while the source mocap data is at 120Hz (which we downsample to 30Hz). The discrepancy between frequencies becomes an issue for longer video sequences. To account for this, we insert a duplicate video frame to change the frequency to 30Hz.

For all of the datasets, we use 12 evenly-spaced frames (starting with the first frame and ending with the last frame) to perform the first stage in MoSh++ [Mahmood et al. 2019].

B.1.2 MOYO. . For the MOYO [Tripathi et al. 2023] dataset, we use the SMPL-X models provided by the authors.

B.1.3 Conversion. . For all three datasets, we need to convert from the SMPL-X model to the neutral SMPL body model for evaluation. More specifically, we convert SMPL-X models to neutral SMPL models via the official conversions tools (https://github.com/vchoutas/smplx).

B.2 HuMoR

For motion capture solving [Rempe et al. 2021], HuMoR requires marker labels in the form of vertex correspondences. However, they also test on RGB-D datasets. In this case, they use a Chamfer distance loss. We apply a single-directional Chamfer distance loss without robust weighting as we found this to help with reconstruction. Additionally, as HuMoR is more efficient for optimizing smaller time window, we adopt their sequence splitting and merging method wherein we split the sequence into overlapping sequences of 2 seconds. We found it necessary to tune some of the parameters used in their approach:

```
 --data-fps 30

--prox-batch-size 8  # 2 for MOYO, 8 for UMPM and CMU Kitchen

--prox-seq-len 60
--prox-overlap-len 1
--point3d-weight 100000.0 100000.0 100000.0
--pose-prior-weight 0.01 0.01 0.0
--shape-prior-weight 0.1 0.1 0.1

--joint3d-smooth-weight 0.1 0.1 0.0

--motion-prior-weight 0.0 0.0 1e-5
--motion-optim-shape

--init-motion-prior-weight 0.0 0.0 0.0

--joint-consistency-weight 0.0 0.0 100.0
--bone-length-weight 0.0 0.0 2000.0

--contact-vel-weight 0.0 0.0 100.0
--contact-height-weight 0.0 0.0 10.0

--floor-reg-weight 0.0 0.0 1.0

--lr 1.0
--num-iters 30 70 70

--stage3-tune-init-num-frames 15
--stage3-tune-init-freeze-start 30
--stage3-tune-init-freeze-end 55
```

B.3 SOMA

SOMA [Ghorbani and Black 2021] labels markers from generally structured marker layouts. We compare their SuperSet model, which is trained to classify 89 common marker labels. The idea behind the SuperSet is that there are many common marker layouts that share selected keypoints on the body for humans. Note that the SuperSet model is generally less accurate than fine-tuned models per layout, but it is necessary to use this model when the marker layout is unknown. SOMA provides these discrete marker labels, and then MoSh++ uses these labels to compute the SMPL-X parameters for the human. MoSh++ uses two stages of optimization. The goal of the first stage is to estimate the body shape and the marker locations on the surface. The goal of the second stage is to estimate the pose. We repeat both stages for every sequence. Furthermore, MoSh++ requires 12 representative frames for the first stage. For this we simply select 12 frames uniformly spaced across the entire sequence.

Table 5: Part definitions with corresponding SMPL bone names

Part name	Joints
Left arm	left_shoulder, left_elbow, left_wrist
Left leg	left_hip, left_knee, left_ankle, left_foot
Left shoulder	spine3, left_collar, left_shoulder, left_shoulder, left_elbow
Right arm	right_shoulder, right_elbow, right_wrist
Right leg	right_hip, right_knee, right_ankle, right_foot
Right shoulder	spine3, right_collar, right_shoulder, right_shoulder, right_elbow

C ALIGNMENT OF MONOCULAR VIDEO AND MOCAP MARKERS

HMR 2.0 [Goel et al. 2023] provides SMPL and camera parameters. However, it does not provide accurate root translation, as root translation in world-space is a difficult problem to solve for monocular video [Shin et al. 2023a; Ye et al. 2023; Yuan et al. 2022]. The root orientation from HMR 2.0 does not necessarily align with the root orientation of the mocap markers. The difference is mostly due to a single yaw-rotational offset between the root orientations of HMR and mocap markers. Optimizing for root orientation is prone to local minima, particularly with respect to front and back of the SMPL mesh.

D DATA PREPROCESSING

The motion capture data generally is high quality in all three datasets. However, we needed to handle some edge cases in preprocessing for a small number of sequences. During some frames, markers could reset to the origin, which may have been caused by tracking errors. These markers are masked for the problematic frames during the optimization process. HMR 2.0 [Goel et al. 2023] with PHALP [Rajasegaran et al. 2022] sometimes drops tracking for some frames. If these frames are at the beginning or end of the sequences, we use the closest known SMPL parameters. If the frames are in the middle of the sequences, we linearly interpolate β and Γ and perform spherical linear interpolation [Brégier 2021] Φ and Θ. However, we mask out these frames with our method when finding the marker-vertices correspondences.

D.1 Video Processing

All three datasets have multiple cameras with different labels. Because we only evaluate with monocular vision, we select one camera for each dataset. Furthermore, we down-sample each dataset to reduce video processing time. The video properties and resolutions are shown in Table 6.

Table 6: We list the configurations for the community to reproduce results and fairly benchmark results.

	UMPM	CMU Kitchen	MOYO
Camera name	l	7151062	YOGI_Cam_06
Resolution	644 × 486	1024 × 768	1028 × 752

E ADDITIONAL RESULTS

E.1 Synthetic Marker Placement

To stress-test our algorithm, we randomly place markers to simulate different marker layouts. To get marker placement, we uniformly sample the surface based on surface area and add an offset of 9.5mm to the surface of the ground-truth SMPL mesh. We acquire layouts with 20, 30, 40, and 50 markers (see Fig. 10 for reconstructions). We only generate the 50-marker layout and then progressively remove 10 markers to get the other layouts. We do this with 10 different random seeds, effectively producing 10 unique layouts for each number of markers.

As seen in Fig. 9, we test different numbers of markers to show that our technique generally has lower errors with more markers. Importantly, because the markers are randomly placed, they may not be placed in optimal positions.

E.2 Video Reconstruction Error Robustness

While HMR 2.0 [Goel et al. 2023] provides accurate results in general, it does fail in certain cases. For example, we observed problems in reconstruction when self-occlusions are present. Additionally, sometimes tracking can temporarily fail (see Fig. 12). Our approach generally recovers well from these issues because we mostly rely on the video reconstruction results for initialization.

E.3 Marker Tracking Loss

Our method is robust against markers that get lost during tracking (e.g., from occlusions). In our implementation, markers with lost tracking are masked out during optimization, so they only contribute to frames in which they are visible (i.e., the marker position $m^{(t)} \neq (0, 0, 0)$).

To test robustness for marker loss, we simulate marker loss by randomly dropping markers. The results of these experiments are shown in Fig. 13. Each frame, we hide markers with probabilities of (0.0%, 0.2%, 1.0%) and keep them hidden for 10 frames. Even with multiple dropped markers, our approach can still perform accurate mocap solving.

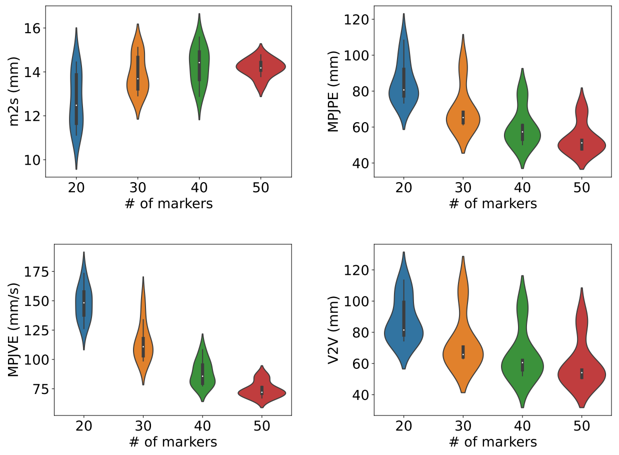

Figure 9: Synthetic benchmark. We generate and evaluate synthetic examples for 10 different synthetic layouts with 20, 30, 40, and 50 markers. Our method produces lower MPJPE, MPJVE, and V2V with a higher number of markers. As the layouts are randomly generated, the reconstruction error can vary depending on the marker placement.

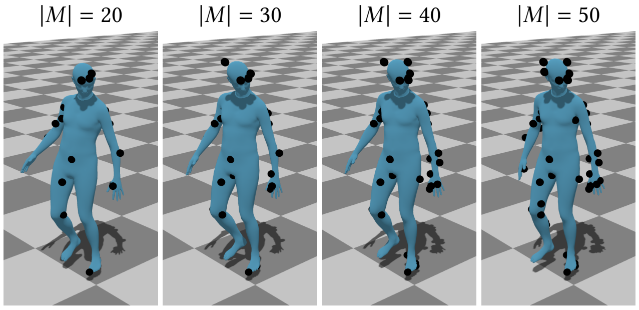

Figure 10: Our reconstruction for synthetic mocap data for layouts with 20, 30, 40, and 50 markers. While our method works better with more markers, it can still solve even with lower numbers of markers.

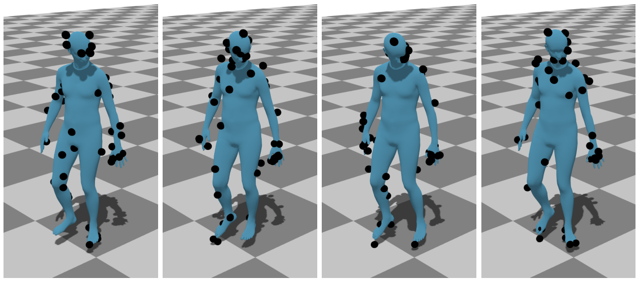

Figure 11: Our reconstruction with different seeds for marker placement with 50 markers. Our method is robust against different marker layouts, even if marker placement is sub optimal.

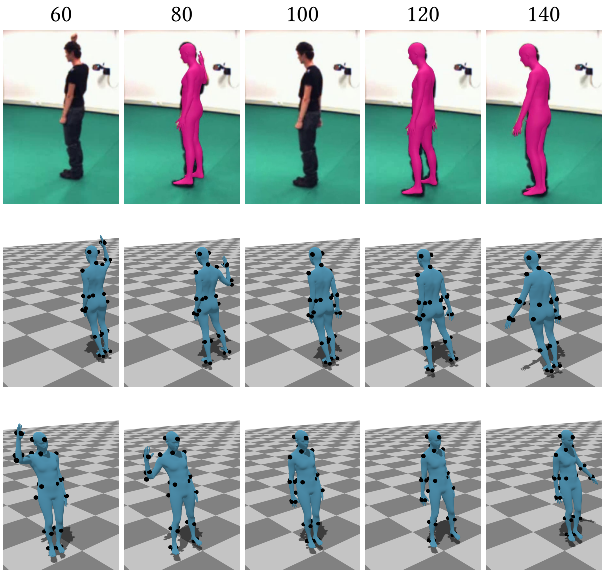

Figure 12: Qualitative results for a sequence from UMPM [van der Aa et al. 2011] at frames {60, 80, 100, 120, 140}. First row: the pink reprojected SMPL overlay shows the tracked person by HMR 2.0 [Goel et al. 2023]. Images without the pink overlay show tracking failure. Second and third rows: successful reconstruction with our method from two different camera angles.

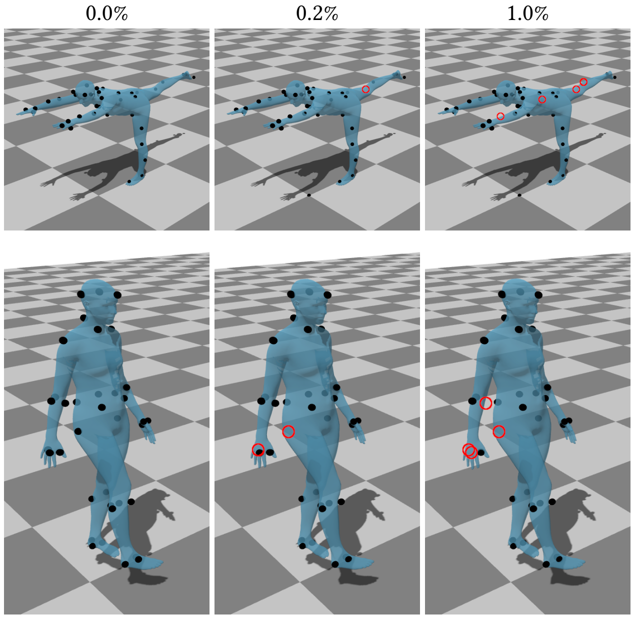

Figure 13: Reconstruction results for simulated marker loss. The red circles show places where the marker should be located but tracking failed. Even with a few lost markers, our approach can accurately reconstruct results.