NeMo: 3D *Ne*ural *Mo*tion Fields from Multiple Video Instances of the Same Action

Kuan-Chieh Wang[†], Zhenzhen Weng, Maria Xenochristou, João Pedro Araújo, Jeffrey Gu,
C. Karen Liu[‡], Serena Yeung[ᵀ]
Stanford University
[†]wangkua1@stanford.edu, [‡]karenliu@cs.stanford.edu, [ᵀ]syyeung@stanford.edu

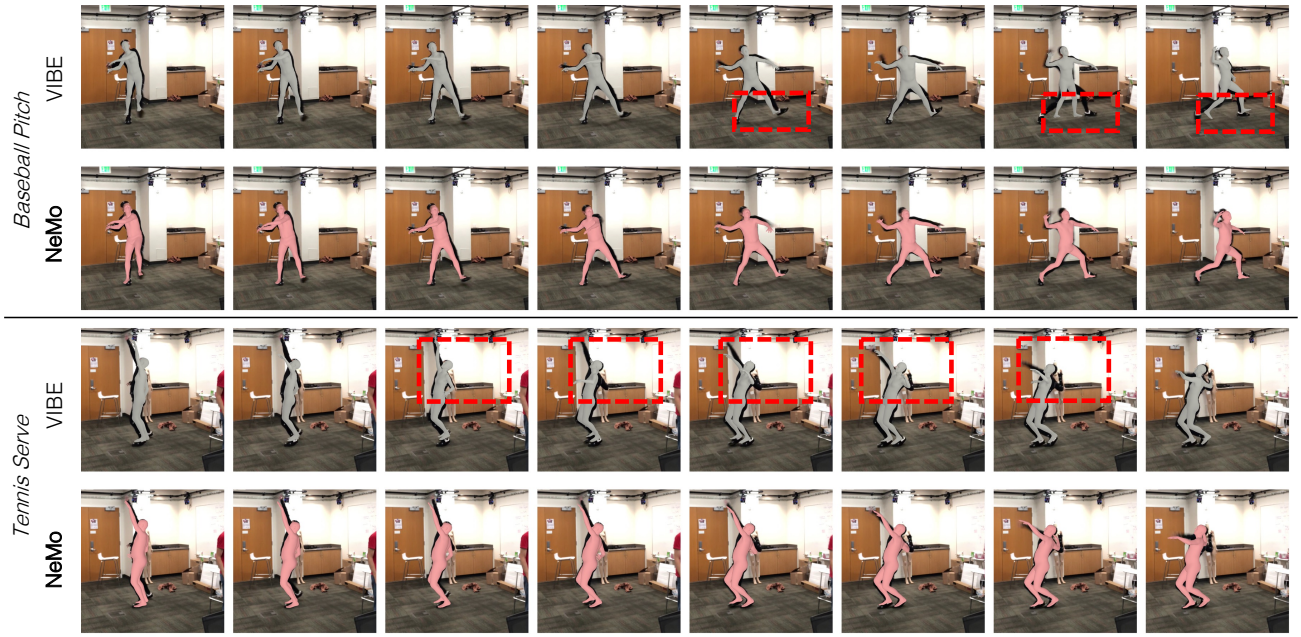

Figure 1. We propose the ***Ne***ural ***Mo***tion (**NeMo**) field that learns the global 3D motion from multiple video instances of the same action. NeMo can accurately recover the *dynamic* ranges of athletic motion. To illustrate, VIBE, a baseline video-based HMR method, fails to capture the large step taken by the subject in the "Baseball Pitch" example, and swaps the arms in "Tennis Serve". Visit our project page for video visualization, our code, and the NeMo-MoCap dataset: https://sites.google.com/view/nemo-neural-motion-field.

Abstract

The task of reconstructing 3D human motion has wide-ranging applications. The gold standard Motion capture (MoCap) systems are accurate but inaccessible to the general public due to their cost, hardware, and space constraints. In contrast, monocular human mesh recovery (HMR) methods are much more accessible than MoCap as they take single-view videos as inputs. Replacing the multi-view MoCap systems with a monocular HMR method would break the current barriers to collecting accurate 3D motion thus making exciting applications like motion analysis and motion-driven animation accessible to the general public. However, the performance of existing HMR methods degrades when the video contains challenging and dynamic motion that is not in existing MoCap datasets used for training. This reduces its appeal as dynamic motion is frequently the target in 3D motion recovery in the aforementioned applications. Our study aims to bridge the gap between monocular HMR and multi-view MoCap systems by leveraging information shared across multiple video instances of the same action. We introduce the Neural Motion (NeMo) field. It is optimized to represent the underlying 3D motions across a set of videos of the same action. Empirically, we show that NeMo can recover 3D motion in sports using videos from the Penn Action dataset, where NeMo outperforms existing HMR methods in terms of 2D keypoint detection. To further validate NeMo using 3D metrics, we collected a small MoCap dataset mimicking actions in Penn Action, and show that NeMo achieves better 3D reconstruction compared to various baselines.

1. Introduction

Reconstructing 3D human motion has wide-ranging applications from the production of animation movies like Avatar [4], human motion synthesis [17, 44, 45] and biomechanical analysis of motion [3, 10, 16, 27, 39, 40]. Traditional marker-based MoCap systems work by recording 2D infrared images of light reflected by markers placed on the human subject. Placing, calibrating, and labeling the markers are tedious, and the markers can potentially restrict the range-of-motion of the subject. Alternatively, *markerless* MoCap systems work with RGB videos and use computer vision techniques to extract the 3D human pose, which eliminates the need of placing physical markers on human subjects. Given a set of synchronized video captures from multiple views, one can run 2D keypoint detection methods like OpenPose [5] and perform triangulation to recover the 3D pose [35]. Markerless MoCap approaches, however, are still restricted by the assumption of having *synchronized multi-view* video capture of the *same instance* of the motion.

Meanwhile, there has been a rapid development of 3D *monocular* human pose estimation (HPE) and human mesh recovery (HMR) methods. These methods aim to recover the 3D human motion from a *single-view* video capture. The accessibility of monocular HMR makes it an attractive alternative to existing MoCap systems, especially in situations where setting up additional hardware is challenging like when the human subject is on a bike [13] or underwater [9]. The accuracy of monocular methods is generally lower because single-view input only provides partial information about the underlying 3D motion. The model needs to overcome complications like depth ambiguity and self-occlusion. Like other machine learning systems, HMR models overcome these difficulties by learning from paired training data. Because of the difficulty and cost of collecting MoCap data, paired datasets of videos and 3D motions are scarce. Additionally, publically available MoCap datasets are often restricted to simple everyday motions like Human3.6M [19] and AMASS [12]. As a result, existing HMR methods generalize less well in domains with less available MoCap data, such as motions in sports, a dominant application domain of 3D human motion recovery [8, 15, 33, 34]. As an example, see Figure 1 for where existing HMR methods struggled to capture the dynamic range of athletic motions.

We bridge the gap between multi-view MoCap and monocular HMR by assuming there is shared and complementary information in *multiple instances* of video captures of the same action, similar to what is in the (same instance) multi-view setup. These *multiple video instances* can be different repetitions of the same action from the same person, or even from different people executing the same action in different settings (See the left side of Figure 2 for illustration). For application domains like sports, a key feature of its motions is that they are well-defined and structured. For example, an athlete is also often instructed to practice by performing many repetitions of the same action and the variation across the repetitions is oftentimes slight. Even when we look at the executions of the same action from different athletes, these different motion "instances" often contain shared information. In this work, we aim to better reconstruct the underlying 3D motion from videos of multiple instances of the same action in sports.

We parametrize each motion as a neural network. It takes as input a scalar phase value indicating the phase/progress of the action and an instance code vector for variation and outputs human joint angles, root orientation, and translation. Since the different sequences are not synchronized, and the actions might progress at slightly different rates (e.g., a faster versus slower pitch), we use an additional learned phase network for synchronization. The neural network is shared across all the instances while other components including the instance codes and phase networks are instance-specific. All the components are learned jointly. We optimize using the 2D reprojection loss with respect to the 2D joint keypoints of the input videos and prior 3D loss w.r.t. initial predictions from HMR methods to enforce 3D prior. We call the resulting neural network a Neural Motion (**NeMo**) field. NeMo can also be seen as a new test-time optimization scheme for better domain adaptation of 3D HMR similar in spirit to SMPLify [2, 36]. A key difference is that we leverage shared 3D information at the group level of many instances to learn a canonical motion and its variations. To summarize, our contributions are:

- We propose the neural motion (NeMo) field and an optimization framework that improves 3D HMR results by jointly reasoning about different video instances of the same action.
- We optimize NeMo fields on sports actions selected from the Penn Action dataset [49]. Since the Penn Action dataset only has 2D keypoint annotations, we collected a small MoCap dataset with 3D groundtruth where the actor was instructed to mimic these motions, which we will refer to as our *NeMo-MoCap* dataset. We show improved 3D motion reconstruction compared to various baseline HMR methods using both 3D metrics, and also improved results on the Penn Action dataset using 2D metrics.
- The NeMo field also recovers *global* root translation. Compared to the recently proposed global HMR method, recovered global motion from NeMo is substantially more accurate on our NeMo-MoCap dataset.

2. Related Work

In this section, we discuss related work in 3D HMR methods, multi-view 3D modeling, and human motion datasets.

HMR Methods Our proposed methode NeMo bridges the gap between monocular HMR methods [7, 14, 21, 24–26, 38, 43, 47], and traditional multi-view MoCap systems. In a way, it can be seen as a test-time optimization (TTO) extension for finetuning predictions from existing HMR methods, much like the popular TTO algorithm SMPLify [36]. Compared to SMPLify, NeMo leverages information across multiple video instances of the same action, resulting in better 3D reconstruction. NeMo can be used in conjunction with any existing HMR methods that are video-based like VIBE [28] or framed-based like PARE [29]. Compared to most HMR methods, NeMo also recovers the global root trajectory, which is a central piece of MoCap data, while most HMR methods do not. Recently, *global* HMR is attracting attention from researchers where global root trajectory is also recovered. This is a more challenging but also a more impactful version of the HMR task. Compared to GLAMR [46], in terms of global HMR metrics, NeMo reduces the overall error in our experiments.

Multi-view 3D Human Modeling Monocular HMR is fundamentally challenging due to issues such as occlusions and depth ambiguity. Multi-view 3D models aim to overcome these issues by utilizing video captures from multiple viewpoints to gain a holistic understanding of the scene. These multi-view videos could be shot changes of the same scene in movies [37], or different viewpoints recorded by multiple synchronized cameras [18, 20, 42, 48]. In comparison, our approach uses different instances of the same action performed asynchronously by one or more humans to capture the 3D motion of a sports action. iMoCap [11] studied a problem similar to ours by curating videos from the internet and also aimed to recover the 3D motion. In contrast to our neural representation, their method fitted a fixed set of poses over time, requiring them to additionally enforce temporal smoothness and cannot naturally allow for interpolations. Furthermore, their method does not leverage recent advances from monocular HMR, which is vital for having good 3D motion prior. Lastly, they curated their videos and did not use an existing video dataset, making comparison impossible. In contrast, we apply NeMo to the Penn Action dataset and further validated it on a MoCap dataset we collected which we intend to open-source. [1]

3D Human Motion Datasets Even though datasets with 3D ground truth human motion are essential to developing reliable models for human mesh reconstruction (HMR), collecting 3D data is costly and labor-intensive. The result is that available 3D datasets are limited in the number of subjects and motions. The Human3.6M dataset [19] contains data (3D MoCap, 2D keypoints, action labels, and videos) for only 11 human subjects, while the 3DPW dataset [41] is similar in having paired video and 3D groundtruth, but was captured in an outdoor environment, using a combination of inertial measurement units (IMUs) and vision models. Both the Human3.6M and 3DPW datasets do not contain sports motion. The AMASS dataset [12] is much larger in scale, containing over 300 subjects and more than 11000 motions, but still inherits the restrictions of MoCap, and does not cover the full range of athletic motions. The lack of MoCap data for sports makes existing HMR methods suffer from the domain shift issue and perform poorly on sports sequences, especially during the dynamic segment of the motion. Many sports datasets only contain 2D joint annotations and are significantly downsampled in time [1, 6, 49].

[1] Only the raw videos were released for their project, but not the annotations, the extracted 3D motion, or the code for their method. Attempts to communicate with the authors were also unsuccessful.

3. Neural Motion (NeMo) Fields

In this section, we focus on the problem of extracting the 3D human motion for specific athletic actions, such as "baseball pitching", given a set of videos. We assume that for the same action, the underlying 3D human motion is similar across the videos. Intuitively, the 3D reconstruction task can be made easier by combining information from all the videos into a single motion with variations. This makes the 3D reconstruction problem easier than treating all the videos separately. See Figure 2 for an illustration of our method.

Problem Formulation Given multiple video instances of the same action, our goal is to recover the 3D *global motions*; namely, sequences of 3D poses (including the root orientation), $\boldsymbol{\theta}^n_{1:T}$, and root translations, $\boldsymbol{x}^n_{1:T}$ for each of the video instances. The superscript $\circ^n$ denotes the n-th video instance and the subscript $\circ_{1:T}$ denotes a sequence from time 1 to T. Our key insight is that, for many actions, the variations across multiple instances (i.e, executions) can be slight, which means we can improve our estimate of the 3D motion by solving for the motion instances jointly.

We first process the videos using off the shelf 2D and 3D pose estimators to get the initial estimates of the 2D keypoints $\tilde{\boldsymbol{j}}^n_{1:T}$ and 3D poses $\tilde{\boldsymbol{\theta}}^n_{1:T}$. We use tilde, $\tilde{\circ}$, to denote the initial estimates. We then try to optimize for the motion jointly using both the 2D and 3D initial predictions across all video instances of the same action. Our method can also be viewed as a test-time optimization algorithm for improving 3D motion like SMPLify [28, 36] that leverages the shared information at the group level. Note, most existing 3D HMR methods only output the pose/articulation (i.e., $\boldsymbol{\theta}$) and not the global root translation, $\boldsymbol{x}$. In contrast, we also aim to recover the 3D global root translation. In the following sections, we describe how we parametrize and optimize for the set of motions using a shared neural motion field.

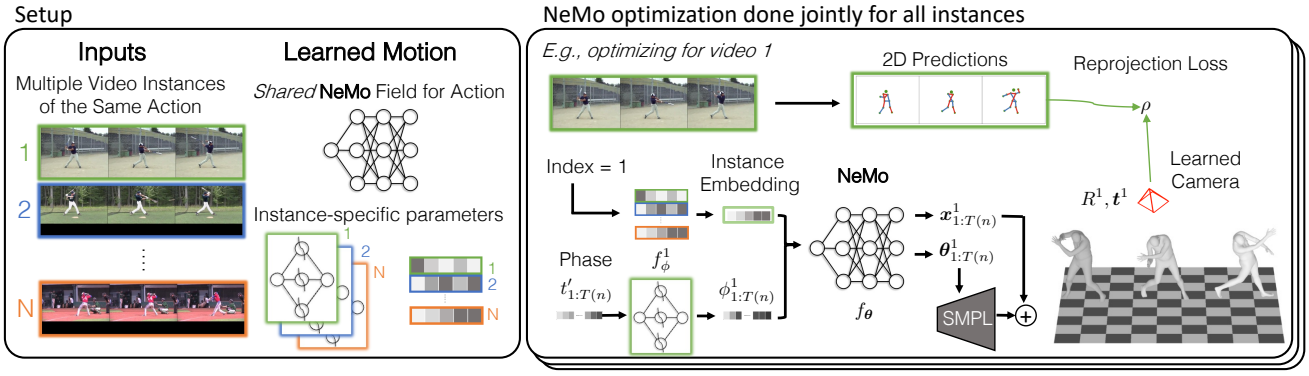

Figure 2. **System illustration of NeMo.** NeMo learns a shared canonical motion parametrized by a neural network, instance-specific phase networks, and latent vectors. Optimizing NeMo mainly relies on the 2D reprojection error. The phase networks are monotonically increasing warping functions that help synchronize the different progressions across videos. Given the warped phases along with a learnable instance embedding, the NeMo field outputs the joint angles and the root translation of the motion, which are rendered using the SMPL model [32].

Neural Motion Field We represent a 3D motion sequence using a multi-layer perceptron (MLP) and call this representation a Neural Motion (NeMo) field. The input to the network is the phase of the motion sequence, $\phi \in [0, 1]$, which can be viewed as the current progression in a time series and an instance vector $\boldsymbol{z} \in \mathbb{R}^{N_z}$ to account for the instance variation of the motion. The instance vectors are also learnable parameters that are optimized jointly with NeMo. The MLP outputs 23 joint angles, the root orientation $\boldsymbol{\theta}$ and 3D global translation $\boldsymbol{x}$, $f_{\boldsymbol{\Theta}} : \mathbb{R}^{1+N_z} \mapsto \mathbb{R}^{24\times 6+3}$. The joints use the 6D representation for rotation proposed in Zhou et al. [50] makes optimizing angles easier, and is commonly used in HMR networks [28, 30]. For convenience, we denote the sub-network that outputs global translation as $f_{\boldsymbol{x}}$ and the rest that outputs joint angles and orientation as $f_{\boldsymbol{\theta}}$.

Given the output of NeMo fields, the joint angles, and root translation, we use the Skinned Multi-Person Linear (SMPL) model [32] to represent the 3D mesh of the human body. The SMPL body model is a differentiable function $f_m : \mathbb{R}^{72+10} \mapsto \mathbb{R}^{6890\times 3}$ that takes a pose parameter $\boldsymbol{\theta} \in \mathbb{R}^{72}$ and shape parameter $\boldsymbol{\beta} \in \mathbb{R}^{10}$, and returns the body mesh $\boldsymbol{m}$ with 6890 vertices. In this work, we assume a neutral shape, which is fixed to the constant vector of zeros, i.e. $\boldsymbol{\beta} = \mathbf{0}$, and drop it in what follows for simplicity. A linear regressor W can be fitted to get the major body joints in 3D, $\boldsymbol{p} \in \mathbb{R}^{J\times 3}$ and $\boldsymbol{p} = W\boldsymbol{m}$, where each joint is a linear combination of the mesh vertices. To get the 3D body joints given an input phase ϕ, the combination of the NeMo field and SMPL is used as follows:

$$\boldsymbol{p} = W\Big(f_{\boldsymbol{m}}\big(f_{\boldsymbol{\theta}}([\phi; \boldsymbol{z}])\big) + f_{\boldsymbol{x}}([\phi; \boldsymbol{z}])\Big), \tag{1}$$

where $[\cdot;\cdot]$ denotes concatenation.

Phase Networks Since the videos are not synchronized, and the different motion instances can progress at different rates, we allow the phases for the different sequences to vary. We introduce a self-normalized monotonic neural network, $f_{\boldsymbol{\phi}} : \mathbb{R} \mapsto \mathbb{R}$, which takes as input the linearly normalized time index, $t' = \frac{t}{T}$, where T is the total length of a given motion sequence and outputs the phase, ϕ. A monotonic neural network can be composed by summing K shifted and scaled sigmoid function. The full phase network is written as:

$$\phi = f_{\boldsymbol{\phi}}(t') = \frac{g(t') - g(0)}{g(1) - g(0)}, \tag{2}$$

where

$$g(t') = \frac{1}{K}\sum_{k=1}^{K} \sigma(f_{\text{ReLU}}(a_k)(t' - f_{\text{ReLU}}(b_k))). \tag{3}$$

We define $\sigma(\cdot)$ as the logistic function, $f_{\text{ReLU}}(\cdot)$ as the ReLU activation function, and $\{a_k, b_k\}_{k=1}^{K}$ as the learnable shift and scale parameters. To ensure the phase starts at 0 and ends at 1, self-normalization is added (Equation 2). The ReLU function ensures the sigmoid functions are increasing.

NeMo Optimization NeMo optimization goes through two main stages. In both stages, optimization is done jointly across all videos. In the first stage, the pose component of the NeMo field (i.e., $f_{\boldsymbol{\theta}}$ is optimized w.r.t the initial 3D estimate $\tilde{\boldsymbol{\theta}}$ to mimic the prediction for the 3D pose estimator). In the second stage, the warmed-up NeMo field, along with all the other parameters, are jointly optimized using 2D reprojection loss, which we describe below. In addition to a NeMo field, instance vectors, and phase networks, we also fit the cameras. Each camera has its own extrinsic parameters including a rotation matrix, R, a translation vector, $\boldsymbol{t}$, and intrinsic parameters. We fix the intrinsic parameters and learn the extrinsic parameters; namely, how the cameras are

placed in the 3D world. The optimization of NeMo can be written as:

$$\min_{f_{\boldsymbol{\theta}}, f_{\boldsymbol{x}}, \{R^n, \boldsymbol{t}^n, f_{\boldsymbol{\phi}^n}, \boldsymbol{z}^n\}_{n=1}^N} \sum_{n=1}^{N} \Big(\frac{1}{T(n)} \sum_{t=1}^{T(n)} \rho(\boldsymbol{j}_t^n, \tilde{\boldsymbol{j}}_t^n) \Big), \tag{4}$$

where,

$$\boldsymbol{j}_t^n = P\Big(R^n f_{3d}(\boldsymbol{p}_t^n, \boldsymbol{z}^n) - \boldsymbol{t}^n\Big), \tag{5}$$

$$\boldsymbol{p}_t^n = W\Big(f_{\boldsymbol{m}}\big(f_{\boldsymbol{\theta}}([\phi_t^n; \boldsymbol{z}^n])\big) + f_{\boldsymbol{x}}([\phi_t^n; \boldsymbol{z}^n])\Big), \tag{6}$$

$$\phi_t^n = f_{\boldsymbol{\phi}}^n(\frac{t}{T(n)}). \tag{7}$$

We use P to denote the perspective projection and $\rho(\cdot)$ the error function for 2D points. We use the Geman-McClure error function, which is more robust to outliers than the mean squared errors. $T(n)$ indicates the length of the n-th video.

4. Experiments

In this section, we validate our proposed method NeMo on two datasets: the NeMo-MoCap dataset we collected, and the Penn Action dataset [49]. We report standard 3D evaluation metrics for HMR and metrics for global HMR on our NeMo-MoCap dataset (Section 4.1). On the Penn Action dataset where only 2D groundtruth is available, we report 2D metrics and show qualitative results (Section 4.2). Since our work focuses on dynamic and athletic motion, results are best viewed in videos. Please visit our project page for rendered results: https://sites.google.com/view/nemo-neural-motion-field.

Methods We used VIBE [28] for our initial 3D estimate and OpenPose [5] for our 2D psuedo-groundtruth. NeMo used an architecture of 3 hidden-layer MLP. We ran 300 steps using the 3D loss as warmup and 2000 steps using the 2D loss in the second stage. The same hyperparameters were used for all motions across all datasets. For more details please refer to Appendix A. For baselines, we compared NeMo to the following HMR methods:

- **VIBE** [28] – a video-based HMR method that we also used as our initial 3D estimate.
- **VIBE+SMPLify** [28, 36] – this method combines VIBE with SMPLify which finetunes the results using 2D reprojection loss.
- **PARE** [29] – a framed-based HMR method that trained on multiple datasets, including using 3D pseudo-groundtruth extracted by EFT [22].
- **GLAMR** [46] – a state-of-the-art global HMR method that infers global root trajectory based on initial estimates from HybrIK [31].

Penn Action Dataset The Penn Action Dataset [49] contains thousands of video sequences of different athletic actions with both action and 2D joint annotations for each sequence. We use this dataset as an example where 1. 3D groundtruth was not collected, and 2. using traditional MoCap to collect 3D groundtruth would be too expensive or infeasible because of constraints from the environment, motion, and availability of human experts.

Specifically, we focus on five actions: "Baseball Swing", "Baseball Pitch", "Tennis Serve", "Tennis Forehand", and "Golf Swing". The reason is that these actions are representative of our targeted problem: actions that are well-defined and repeatable. Other actions like "Playing Guitar" in the Penn Action dataset are not as well defined.

Our NeMo-MoCap Dataset Since the Penn Action dataset only contains 2D keypoint groundtruth and not 3D groundtruth, it is not enough for us to validate our reconstructed 3D motion. To validate our proposed method, we collected a MoCap dataset with their corresponding videos. We collected 8 repetitions/motion instances for each of the 5 actions above from different camera views. The human actor is instructed to mimic the motion shown in the Penn Action dataset. Visit our website for visualization and access to our NeMo-MoCap dataset.

Metrics The following metrics were used for evaluation.

- **MPJPE / MPVPE** – mean per joint/vertex position error is commonly used for evaluating 3D HMR methods. MPJPE computes the distance from a predicted joint to the groundtruth joint in 3D and MPVPE computes distances for all vertices. Results are reported in millimeters (mm).
- **Global-MPJPE / MPVPE** – the global version of MPJPE/MPVPE measures the error taking into account the predicted global root translation and orientation. This is in contrast to the non-global version where the prediction is root-centered.
- **2D Recon. Err.** – 2D reconstruction error measures the 2D error predicted and groundtruth joints in 2D. Results are reported in terms of the number of pixels.
- **PCK** – the percentage of correct keypoints is a measure of accuracy. It thresholds 2D reconstruction error by 10% of the bounding box size of the target human.

More experimental details can be found in Appendix A.

4.1. Results on our NeMo-MoCap dataset

In this section, we quantitatively evaluate the ability of NeMo to recover global 3D motion on our NeMo-MoCap dataset. We compare NeMo to the video-based HMR

Method	Baseball Pitch	Baseball Swing	Tennis Serve	Tennis Swing	Golf Swing	Mean
			MPJPE (mm, ↓)			
VIBE [28]	101.2 / 141.3	84.5 / 120.7	94.4 / 129.9	69.5 / 96.5	87.5 / 114.1	87.4 / 120.5
VIBE+SMPLify [28]	111.3 / 153.8	89.9 / 128.7	99.5 / 137.7	79.4 / 108.7	103.5 / 132.7	96.7 / 132.3
GLAMR [46]	99.7 / 131.8	100.4 / 139.3	116.0 / 155.1	80.2 / 106.3	114.0 / 150.2	102.1 / 136.5
PARE [29]	97.7 / 134.8	84.5 / 129.9	89.5 / 118.0	73.2 / 97.6	96.7 / 132.4	88.3 / 122.5
NeMo (Ours)	**85.8 / 108.7**	**65.3 / 91.3**	**80.6 / 97.9**	**65.4 / 85.4**	**78.9 / 94.2**	**75.2 / 95.5**
			MPVPE (mm, ↓)			
VIBE [28]	126.7 / 178.8	101.1 / 149.2	117.6 / 164.4	86.5 / 122.4	108.9 / 141.7	108.2 / 151.3
VIBE+SMPLify [28]	139.5 / 196.2	108.8 / 160.1	124.8 / 174.5	101.5 / 141.7	123.9 / 157.6	119.7 / 166.0
GLAMR [46]	129.0 / 168.6	126.9 / 182.1	149.7 / 201.6	107.5 / 142.8	156.0 / 201.6	133.8 / 179.3
PARE [29]	122.8 / 170.4	102.5 / 163.9	113.5 / 150.8	93.6 / 127.8	121.2 / 163.6	110.7 / 155.3
NeMo (Ours)	**112.5 / 147.3**	**77.9 / 118.9**	**95.5 / 121.1**	**83.3 / 116.8**	**96.4 / 118.1**	**93.1 / 124.4**

Table 1. **3D evaluation on our NeMo-MoCap dataset.** Both errors over the entire sequence (*left* in a cell) and over the dynamic range of a sequence (*right* in a cell) are reported. The improvement is more pronounced during the dynamic range of the motions where performance of existing HMR methods degrade.

MoCap	Method	Baseball Pitch	Baseball Swing	Tennis Serve	Tennis Swing	Golf Swing	Mean
Global-MPJPE (mm, ↓)	GLAMR [46]	**144.01**	114.43	198.71	**121.87**	128.26	141.46
	NeMo (Ours)	151.45	**91.92**	**146.06**	148.49	**95.46**	**126.68**
Global-MPVPE (mm, ↓)	GLAMR [46]	**159.78**	130.42	218.59	**134.6**	158.13	160.3
	NeMo (Ours)	163.56	**96.22**	**149.36**	147.82	**100.79**	**131.55**

Table 2. **Global 3D evaluation on our NeMo-MoCap dataset.**

method with and without test-time optimization, state-of-the-art frame-based HMR method, and a recent global HMR method.

Evaluation with Standard HMR Metrics Table 1 shows that NeMo outperforms baseline methods in terms of 3D metrics (MPJPE/MPVPE) across all actions. The improvement is even more pronounced during the dynamic ranges of the motion [2]. This validates our original hypothesis that existing HMR methods are less robust for videos containing dynamic and athletic motion, which is an important application domain for 3D human motion recovery. In the dynamic ranges of the motion, NeMo improves MPJPE from the best-performing baseline VIBE from 120.5 mm to 95.5 mm, a 20.8% improvement. Also, worth noting is the comparison with VIBE+SMPLify which also performs test-time optimization using 2D reprojection loss. In a way, it can be seen as an ablation of NeMo that does not learn from multiple instances jointly. Interestingly, VIBE+SMPLify does not always improve the results from VIBE since the 2D keypoint predicted from OpenPose from a single video might not add more information. This stresses the importance of using the joint optimization proposed for NeMo.

In Appendix B, we include results using 2D evaluation metrics for the same experiment. Note, while NeMo still performed the best overall, the performance between NeMo and baselines was much closer than they were in 3D evaluation because 2D projection is a lossy process and many erroneous 3D poses can be projected to the same 2D pose. This speaks to the importance of using 3D evaluation for 3D motion recovery and our collected MoCap dataset.

Evaluation with *Global* HMR Metrics An important additional benefit of using NeMo is that the recovered motion contains global root information. This is essential for applications like animation, viewpoint-free synthesis, and motion analysis. Global HMR is a recent task, and most of the baselines do not perform global HMR. Compared to GLAMR, a recent method for global HMR [46], NeMo improves the global MPJPE from 141 mm to 127 mm (see Table 2). In Figure 3 and 1B, we show example comparisons between recovered global root trajectory from NeMo and GLAMR. The results from NeMo are much less jittery and better represent the motion. For example, in Figure 1B, the trajectory

[2] See Appendix A for the definition of dynamic range.

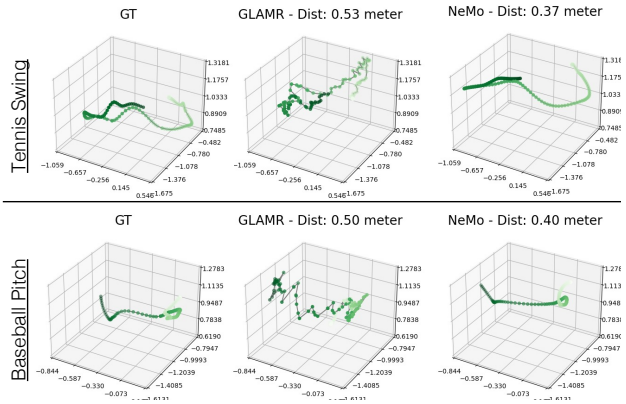

Figure 3. **Visualization of 3D global root trajectory on our NeMo-MoCap dataset.** The brightness of the color denotes temporal progression (from dark to bright).

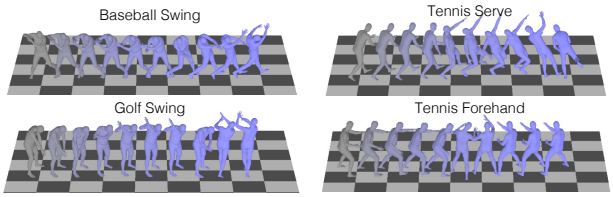

Figure 4. **Qualitative motion rollout from optimized NeMo fields on Penn Action.** Actions progress from left to right.

for the tennis serve captures the jump in the serving motion (i.e., the large increase and decrease in the z-axis). In Figure 3, we can see smooth and large steps taken during both of the actions from the NeMo motion, but not from the recovered motion using GLAMR. Being able to capture the global movement is critical in the eventual goal of replacing MoCap systems with HMR methods. Additional qualitative comparisons using rendered videos between NeMo and baselines are included on the website.

4.2. Results on Penn Action

2D Evaluation The Penn Action dataset only has 2D keypoint annotations but not 3D groundtruth MoCap. This is commonly the case as 2D annotations can be done post-hoc for most videos, but 3D MoCap can only be captured in a laboratory. We use the Penn Action dataset as a demonstration that NeMo can be applied to existing real-world video captures. While we cannot validate the results using 3D metrics, Table 3 shows that in terms of 2D metrics, the NeMo outperformed existing 2D and 3D pose estimators overall. To examine the realism of the recovered motion in 3D, we show qualitative results in the next paragraph.

Qualitative Results Figure 4 contains four instances of optimized NeMo fields on four distinct sports actions. Qualitatively, the optimized NeMo fields conform with our expectations of their respective actions and capture many details in their execution. Combined with the better performance in terms of 2D metrics in the previous paragraph, these results demonstrate the NeMo can be applied to real-world videos like those in the Penn Action dataset and recover realistic 3D motion. In Figure 5, we visualize the learned NeMo fields from Penn Action and our NeMo-MoCap dataset beside each other. One can qualitatively observe the difference in recovered motion. Penn Action videos often capture advanced athletes whose motions are more dynamic and exaggerated compared to an amateur. Such as in the Baseball Swing motion, many players in Penn Action let go of their hands at the end of the swing whereas in our NeMo-MoCap dataset, the subject human did not. In the Baseball Pitch motion, the player from Penn Action keeps their throwing arm back while they step forward, which gives their pitch more power. These results highlight the importance of being able to achieve motion recovery from in-the-wild videos. Additional results that show more variations of the learned motion in 3D can be found in Appendix B.

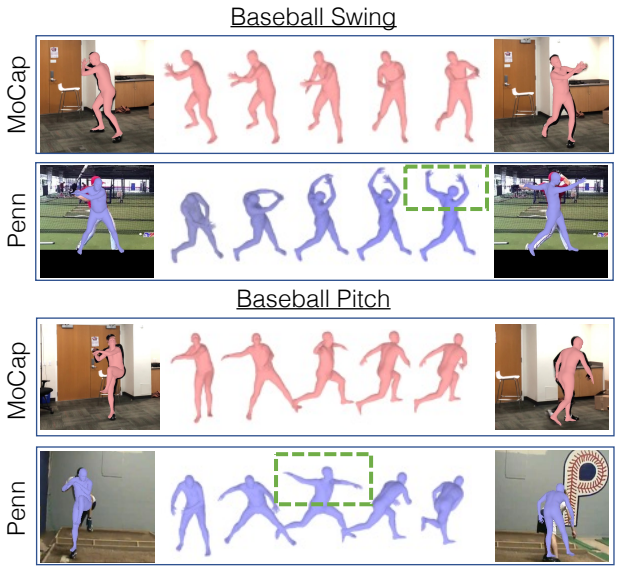

Figure 5. **Comparison of learned NeMo fields from our NeMo-MoCap dataset and Penn Action.** Action in Penn Action is often executed by an advanced athlete whose motion is dynamic and exaggerated. Differences are highlighted by the green box.

Ablations Table 4 shows that removing 3D loss using the initial predictions from the existing 3D HMR method hurts 3D reconstruction but the 2D metric can still appear good. This shows the importance of using an initial 3D estimate to enforce a good 3D prior. Removing the instance-specific parameters degrades the accuracy of the 3D reconstruction

Penn Action	Method	Baseball Pitch	Baseball Swing	Tennis Serve	Tennis Swing	Golf Swing	Mean
Recon. Err. (↓)	OpenPose [5]	**12.71**	12.93	11.64	10.19	8.98	11.29
	VIBE [28]	14.2	**7.35**	20.74	8.65	**5.04**	11.2
	VIBE+SMPLify [28]	16.15	20.56	9.74	20.15	7.23	14.77
	NeMo (Ours)	13.31	9.85	**4.99**	**4.36**	7.97	**8.1**
PCK (↑)	OpenPose [5]	**88.25**	88.91	92.1	93.93	93.81	91.4
	VIBE [28]	83.42	89.63	80.92	89.4	94.69	87.61
	VIBE+SMPLify [28]	79.67	82.23	88.58	88.86	93.13	86.5
	NeMo (Ours)	80.16	**90.93**	**95.99**	**97.46**	**95.37**	**91.98**

Table 3. **2D evaluation on the Penn Action dataset.**

	MPJPE (↓)	MPJPE (Dynamic, ↓)	2D PCK (↑)
NeMo (full)	77.96	95.70	99.08
- 3D HMR	125.22	161.82	98.40
- Phase-networks	76.53	93.62	99.00
- Instance-specific	134.85	194.61	89.57

Table 4. **Ablations.** The second row removes 3D supervision from the HMR initial estimates. The third row removes the phase networks. The last row removes the instance-specific learnable parameters (i.e. phase networks and latent codes).

which is reflected in both the 3D and 2D metrics. When we remove just the phase networks, it performed as well as our full model, which suggests that using the instance latent codes alone was enough to account for all variations. Conceptually, having both the phase networks and latent codes could allow for disentangled representation of 'progression' and 'motion variability'. In practice, having the latent code alone was able to achieve good accuracy.

5. Limitations & Future Directions

One limitation of the NeMo model is the assumption of a fixed camera. While there are many sport videos that have an almost still camera where NeMo is applicable, as shown in Section 4, many sports videos are captured with a moving camera. This is especially true for motion that covers a lot of grounds, like a volleyball spike, or a basketball layup. Often the full action can only be captured by moving a camera to track the athlete. Extending NeMo to account for a moving camera will further improve its applicability. Another worthy future direction is in using the learned NeMo fields as a data augmentation tools for improving regression-based HMR methods, similar to what was proposed in EFT [23]. Currently, NeMo works with multiple video instances of the same action and is a test-time optimization algorithm. While it produces more accurate 3D results than existing HMR methods, it is slow and limited to repeatable actions. Using it as a data collection tool to then finetune HMR methods can potentially lead to a more accurate HMR methods that is also efficient.

6. Conclusion

We proposed NeMo, a neural motion representation and an optimization framework for extracting 3D motions given a set of different video instances of the same sports action. Compared to existing HMR methods whose performance degrades in sports videos due to domain shift, NeMo can better recover the 3D motion of athletic motion by leveraging shared information across different video instances. To validate NeMo, we collected a MoCap dataset mimicking the Penn Action dataset and show that NeMo outperformed a range of HMR baselines – frame-based, video-based, test-time optimization algorithm, and global HMR method. We also evaluated NeMo on the Penn Action dataset using 2D metrics, and show qualitative results. Furthermore, NeMo can recover a much more faithful 3D root trajectory when compared to a recently proposed global HMR method.

This project falls under the umbrella of works that aim to improve 3D reconstruction in the wild using computer vision. Collectively, our society has already built a massive database of videos that capture the human experience in the form of movies, sports event broadcasts, news media, and more. Technology that can transform this existing data into their 3D reconstruction will take us closer to a realistic virtual experience. By using existing videos, we might even reconstruct events in the past, like Michael Jordan winning his first NBA title in 1991, and see it happen from anywhere on the basketball court.

Acknowledgement

We acknowledge the support of this work by the Wu Tsai Human Performance Alliance: https://humanperformancealliance.org/.

References

[1] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5167–5176, 2018. 3

[2] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European conference on computer vision*, pages 561–578. Springer, 2016. 2

[3] Melissa Ann Boswell, Lukasz Kidzinski, Jennifer L Hicks, Scott David Uhlrich, Antoine Falisse, and Scott L Delp. Smartphone videos of the sit-to-stand test predict osteoarthritis and health outcomes in a nationwide study. *medRxiv*, 2022. 2

[4] James Cameron, Sam Worthington, Zoe Saldana, and Sigourney Weaver. *Avatar*. Twentieth Century Fox Home Entertainment, 2009. 2

[5] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019. 2, 5, 8, 11

[6] Xin Chen, Anqi Pang, Wei Yang, Yuexin Ma, Lan Xu, and Jingyi Yu. Sportscap: Monocular 3d human motion capture and fine-grained understanding in challenging sports videos. *International Journal of Computer Vision*, 129(10):2846–2864, 2021. 3

[7] Junhyeong Cho, Kim Youwang, and Tae-Hyun Oh. Cross-attention of disentangled modalities for 3d human mesh recovery with transformers. In *European Conference on Computer Vision*, pages 342–359. Springer, 2022. 3

[8] Steffi L Colyer, Murray Evans, Darren P Cosker, and Aki IT Salo. A review of the evolution of vision-based motion analysis and the integration of advanced computer vision methods towards developing a markerless system. *Sports medicine-open*, 4(1):1–15, 2018. 2

[9] Neil J Cronin, Timo Rantalainen, Juha P Ahtiainen, Esa Hynynen, and Ben Waller. Markerless 2d kinematic analysis of underwater running: A deep learning approach. *Journal of biomechanics*, 87:75–82, 2019. 2

[10] Samarjit Das, Laura Trutoiu, Akihiko Murai, Dunbar Alcindor, Michael Oh, Fernando De la Torre, and Jessica Hodgins. Quantitative measurement of motor symptoms in parkinson's disease: a study with full-body motion capture data. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 6789–6792. IEEE, 2011. 2

[11] Junting Dong, Qing Shuai, Yuanqing Zhang, Xian Liu, Xiaowei Zhou, and Hujun Bao. Motion capture from internet videos. In *European Conference on Computer Vision*, pages 210–227. Springer, 2020. 3

[12] Mahmood et al. Amass: Archive of motion capture as surface shapes. In *ICCV*, 2019. 2, 3

[13] Anthony A Gatti, Peter J Keir, Michael D Noseworthy, Marla K Beauchamp, and Monica R Maly. Hip and ankle kinematics are the most important predictors of knee joint loading during bicycling. *Journal of Science and Medicine in Sport*, 24(1):98–104, 2021. 2

[14] Shanyan Guan, Jingwei Xu, Yunbo Wang, Bingbing Ni, and Xiaokang Yang. Bilevel online adaptation for out-of-domain human mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10472–10481, 2021. 3

[15] Joseph Hamill, Kathleen M Knutzen, and Timothy R Derrick. Biomechanics: 40 years on. *Kinesiology Review*, 10(3):228–237, 2021. 2

[16] Nicos Haralabidis, David John Saxby, Claudio Pizzolato, Laurie Needham, Dario Cazzola, and Clare Minahan. Fusing accelerometry with videography to monitor the effect of fatigue on punching performance in elite boxers. *Sensors*, 20(20):5749, 2020. 2

[17] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)*, 35(4):1–11, 2016. 2

[18] Buzhen Huang, Yuan Shu, Tianshu Zhang, and Yangang Wang. Dynamic multi-person mesh recovery from uncalibrated multi-view cameras. In *2021 International Conference on 3D Vision (3DV)*, pages 710–720. IEEE, 2021. 3

[19] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 2, 3

[20] Umar Iqbal, Pavlo Molchanov, and Jan Kautz. Weakly-supervised 3d human pose learning via multi-view images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5243–5252, 2020. 3

[21] Umar Iqbal, Kevin Xie, Yunrong Guo, Jan Kautz, and Pavlo Molchanov. Kama: 3d keypoint aware body mesh articulation. In *2021 International Conference on 3D Vision (3DV)*, pages 689–699. IEEE, 2021. 3

[22] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. *arXiv preprint arXiv:2004.03686*, 2020. 5

[23] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *2021 International Conference on 3D Vision (3DV)*, pages 42–52. IEEE, 2021. 8

[24] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 3

[25] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 371–386, 2018.

[26] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5614–5623, 2019. 3

[27] Łukasz Kidziński, Bryan Yang, Jennifer L Hicks, Apoorva

Rajagopal, Scott L Delp, and Michael H Schwartz. Deep neural networks enable quantitative movement analysis using single-camera videos. *Nature communications*, 11(1):4054, 2020. 2

[28] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5253–5263, 2020. 3, 4, 5, 6, 8, 11

[29] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *Proc. International Conference on Computer Vision (ICCV)*, pages 11127–11137, Oct. 2021. 3, 5, 6, 11

[30] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2252–2261, 2019. 4

[31] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3383–3393, 2021. 5

[32] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 4

[33] Lars Mündermann, Stefano Corazza, and Thomas P Andriacchi. The evolution of methods for the capture of human movement leading to markerless motion capture for biomechanical applications. *Journal of neuroengineering and rehabilitation*, 3(1):1–11, 2006. 2

[34] Gergely Nagymáté and Rita M Kiss. Application of optitrack motion capture systems in human movement analysis: A systematic literature review. *Recent Innovations in Mechatronics*, 5(1.):1–9, 2018. 2

[35] Nobuyasu Nakano, Tetsuro Sakura, Kazuhiro Ueda, Leon Omura, Arata Kimura, Yoichi Iino, Senshi Fukashiro, and Shinsuke Yoshioka. Evaluation of 3d markerless motion capture accuracy using openpose with multiple video cameras. *Frontiers in sports and active living*, 2:50, 2020. 2

[36] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3, 5, 11

[37] Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Human mesh recovery from multiple shots. *arXiv preprint arXiv:2012.09843*, 2020. 3

[38] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Hierarchical kinematic probability distributions for 3d human shape and pose estimation from images in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11219–11229, 2021. 3

[39] Scott D Uhlrich, Antoine Falisse, Łukasz Kidziński, Julie Muccini, Michael Ko, Akshay S Chaudhari, Jennifer L Hicks, and Scott L Delp. Opencap: 3d human movement dynamics from smartphone videos. *bioRxiv*, 2022. 2

[40] Scott D Uhlrich, Antoine Falisse, Łukasz Kidziński, Julie Muccini, Michael Ko, Akshay S Chaudhari, Jennifer L Hicks, and Scott L Delp. Opencap: 3d human movement dynamics from smartphone videos. *bioRxiv*, pages 2022–07, 2022. 2

[41] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 601–617, 2018. 3

[42] Yangang Wang, Yebin Liu, Xin Tong, Qionghai Dai, and Ping Tan. Outdoor markerless motion capture with sparse handheld video cameras. *IEEE Transactions on Visualization and Computer Graphics*, 24(5):1856–1866, 2018. 3

[43] Zhenzhen Weng, Kuan-Chieh Wang, Angjoo Kanazawa, and Serena Yeung. Domain adaptive 3d pose augmentation for in-the-wild human mesh recovery. *arXiv preprint arXiv:2206.10457*, 2022. 3

[44] Shihong Xia, Lin Gao, Yu-Kun Lai, Ming-Ze Yuan, and Jinxiang Chai. A survey on human performance capture and animation. *Journal of Computer Science and Technology*, 32(3):536–554, 2017. 2

[45] Kevin Xie, Tingwu Wang, Umar Iqbal, Yunrong Guo, Sanja Fidler, and Florian Shkurti. Physics-based human motion estimation and synthesis from videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11532–11541, 2021. 2

[46] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 5, 6

[47] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *arXiv preprint arXiv:2207.06400*, 2022. 3

[48] Jianfeng Zhang, Yujun Cai, Shuicheng Yan, Jiashi Feng, et al. Direct multi-view multi-person 3d pose estimation. *Advances in Neural Information Processing Systems*, 34, 2021. 3

[49] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2248–2255, 2013. 2, 3, 5

[50] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019. 4