

Marker Data Enhancement For Markerless Motion Capture

Antoine Falisse, Scott D. Uhlich, Akshay S. Chaudhari, Jennifer L. Hicks, and Scott L. Delp

Abstract—Objective: Human pose estimation models can measure movement from videos at a large scale and low cost; however, open-source pose estimation models typically detect only sparse keypoints, which leads to inaccurate joint kinematics. OpenCap, a freely available service for researchers to measure movement from videos, addresses this issue using a deep learning model—the marker enhancer—that transforms sparse keypoints into dense anatomical markers. However, OpenCap performs poorly on movements not included in the training data. Here, we create a much larger and more diverse training dataset and develop a more accurate and generalizable marker enhancer. **Methods:** We compiled marker-based motion capture data from 1176 subjects and synthesized 1433 hours of keypoints and anatomical markers to train the marker enhancer. We evaluated its accuracy in computing kinematics using both benchmark movement videos and synthetic data representing unseen, diverse movements. **Results:** The marker enhancer improved kinematic accuracy on benchmark movements (mean error: 4.1° , max: 8.7°) compared to using video keypoints (mean: 9.6° , max: 43.1°) and OpenCap’s original enhancer (mean: 5.3° , max: 11.5°). It also better generalized to unseen, diverse movements (mean: 4.1° , max: 6.7°) than OpenCap’s original enhancer (mean: 40.4° , max: 252.0°). **Conclusion:** Our marker enhancer demonstrates both accuracy and generalizability across diverse movements. **Significance:** We integrated the marker enhancer into OpenCap, thereby offering its thousands of users more accurate measurements across a broader range of movements.

Index Terms—Deep learning, markerless motion capture, musculoskeletal modeling and simulation, pose estimation, trajectory optimization.

I. INTRODUCTION

MARKERLESS motion capture has become popular for biomechanical analysis of human movement because it reduces the cost and time associated with marker-based motion capture and facilitates large-scale, out-of-lab studies. Multi-camera video systems have achieved kinematic accuracy within approximately five degrees of marker-based motion capture for movements including walking, running, cycling, squatting, sit-to-stand, and drop jump [1]–[3]. These systems typically identify, for each video, the two-dimensional (2D)

position of keypoints on the body using pose estimation models [4], reconstruct their three-dimensional (3D) position using triangulation algorithms like Direct Linear Transformation [5], and compute joint kinematics using, for example, multi-body models and inverse kinematics [6].

Several open-source pose estimation models can estimate 2D keypoints from video (e.g., OpenPose [7], HRNet [8]–[11], VITPose [12], and others [13]). These models, trained on datasets like COCO [14] and MPII [15] which have a limited number of body points labeled by non-experts in human anatomy, usually only detect a sparse set of joint centers [16]. This sparsity combined with the inherent noise of pose estimation models [17]–[19] make it difficult to estimate joint kinematics accurately. In previous work that compared a dual-camera system to marker-based motion capture [1], we found joint kinematic errors of up to 43° when using keypoints estimated from HRNet with a multi-body model and inverse kinematics. The large errors were primarily in the lumbar extension and hip flexion degrees of freedom, because of the few keypoints identified on the torso (shoulder joints), pelvis (hip joints), and femurs (knee joints). These keypoints fail to sufficiently constrain the kinematic redundancy of the multi-body model, where various poses can produce identical joint center positions.

OpenCap is an open-source multi-camera system to compute the kinematics and kinetics of human movement from video [1]. To mitigate the problems that arise from keypoint sparsity, OpenCap incorporates a Long Short-Term Memory (LSTM) model—named the marker enhancer—that extrapolates the 3D position of 43 denser anatomical markers from the 3D position of 20 sparser keypoints (Fig. 1). We found that using the extrapolated anatomical markers reduced kinematic errors by 4.2° on average and by up to 35.9° at certain degrees of freedom compared to using video keypoints directly [1]. OpenCap, however, has lower performance on movements not included in the dataset used to train the marker enhancer (see examples in Fig. 2, column Uhlich et al. (2023)). This is particularly noticeable in movements where the individual is lying prone, supine, or not executing movements aligned with the forward direction of the world frame, since the training set only includes movements performed upright and forward-facing. Ruescas-Nicolau et al. developed a comparable marker enhancer model using a different dataset and also observed varying performance across movements, with worse results on running compared to walking, squatting, and jumping [20].

The objective of this study was to improve OpenCap’s ability to capture the kinematics of a broad range of movements (i.e., generalizability) while preserving its previously

Manuscript submitted on July 12, 2024. This study was funded by the U.S. National Institutes of Health (NIH) under grant IP41EB027060 and the Joe and Clara Tsai Foundation through the Wu Tsai Human Performance Alliance.

Antoine Falisse, Scott D. Uhlich, and Jennifer L. Hicks are with the Department of Bioengineering, Stanford University, Stanford, CA, 94305, USA (email: afalisse@stanford.edu; suhlich@stanford.edu; jenhicks@stanford.edu).

Akshay S. Chaudhari is with the Department of Radiology and Biomedical Data Science, Stanford University, Stanford, CA, 94305, USA (email: akshaysc@stanford.edu).

Scott L. Delp is with the Department of Bioengineering, Mechanical Engineering, and Orthopaedic Surgery, Stanford University, Stanford, CA, 94305, USA (email: delp@stanford.edu).

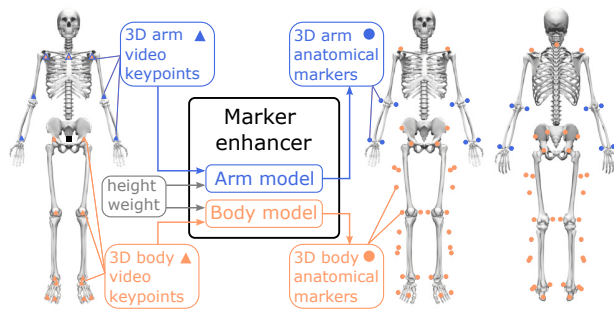


Fig. 1. The marker enhancer model predicts the 3D position of 43 anatomical markers (colored dots on right skeletons) from 20 video keypoint (colored triangles on left skeleton) positions. It consists of two models: the *arm model* predicts the 3D position of eight arm-located anatomical markers from seven arm and shoulder keypoint positions (blue) and the *body model* predicts the 3D position of 35 anatomical markers located on the shoulder, torso, and lower-body from 15 shoulder and lower-body keypoint positions (orange). Both models include the subject's height and weight as input, and all marker positions are expressed with respect to a root marker (the midpoint of the hip keypoints; black square on left skeleton).

reported accuracy on a set of benchmark movements. We first created an enhanced dataset with over twice the data, or ten times more when including augmented data, from a much broader range of movements. We then trained three marker enhancers with different deep learning architectures: linear, LSTM, and transformer. We evaluated the performance of the three marker enhancers when computing joint kinematics from videos on a set of four benchmark movements. We then evaluated their ability to reconstruct joint kinematics from an unseen dataset containing a wide array of movements, thereby assessing their ability to generalize. Finally, we used muscle-driven tracking simulations to estimate dynamics from videos and evaluated the effect of more accurate joint kinematics on dynamic quantities. We integrated the best performing marker enhancer as part of the web-deployed version of OpenCap (<https://www.opencap.ai/>), and shared code and data to facilitate broader usage and reproduction of our work. A preliminary version of this work has been reported at [21].

II. METHODS

A. Dataset

We compiled a large dataset of expert-processed marker-based motion capture data, and synthesized corresponding 3D video keypoints ($n=20$) and anatomical markers ($n=43$) to train the marker enhancers. We first processed the marker data from 16 movement datasets [22]–[39] with OpenSim [40] and AddBiomechanics [41] to obtain scaled OpenSim models and coordinate files (i.e., kinematic data). We then added virtual markers to the scaled models corresponding to the video keypoints and anatomical markers (Fig. 1). We finally extracted the 3D trajectory of each virtual marker for each coordinate file to create the synthetic dataset. To augment the experimental data, we simulated shorter and taller subjects by uniformly scaling each OpenSim model in the dataset by 90, 95, 105, and 110%, thereby quintupling the dataset size.

Prior to model training, we sampled the data at 60 Hz, split them into overlapping (50%) time-sequences of 1 s, and

TABLE I
DATASET DISTRIBUTION

Datasets	Tasks	# subjects	# hours (%)
[28]–[30], [38]	Treadmill walking	110	60.5 (27%)
[34], [36]	Overground walking	98	7.4 (3%)
[31], [32]	Running	31	3.2 (1%)
[33], [35]	Cycling	29	3.4 (2%)
[37]	Karate	175	15.6 (7%)
[39]	Tennis	15	2.9 (1%)
	Other		
[22]–[27]	(e.g., dance, cuts, stairs, squats, ...)	718	128.3 (58%)
Total		1176	221.3

balanced the contribution of the individual datasets to obtain a wide movement distribution (Table I). In total, we included data from 1176 subjects and used over 221 hours of data (excluding data rotation, see next section). The dataset consisted of diverse movements, with the *Other* category, which includes a broad range of movements (e.g., dancing, cutting, stair climbing, squatting, jumping, rebounding in basketball, kicking and heading a soccer ball, performing hamstring exercises, balancing, sitting, boxing, and practicing yoga), accounting for 58% of the dataset (Table I). In comparison, OpenCap's original marker enhancer had been trained on 108 hours of data from 336 subjects, from which over 80% was treadmill gait data and 45% was from a single dataset.

B. Model architecture and training

We trained two models: one to predict anatomical markers located on the arms (*arm model*) and one to predict anatomical markers located on the shoulders, torso, and lower-body (*body model*) (Fig. 1). Since not all individual datasets included arms, we used a subset of the data (about 152 hours, i.e., 69% of total data) to train the *arm model*. The *body model* takes as input the 3D position of 15 shoulder and lower-body keypoints and predicts the 3D position of 35 anatomical markers, whereas the *arm model* uses the position of seven arm and shoulder keypoints to predict the position of eight anatomical markers. Both models include the subject's height and weight as input.

During training, we expressed the 3D position of each marker with respect to a root marker (the midpoint of the hip keypoints), normalized the root-centered 3D positions by the subject's height, added Gaussian noise (standard deviation: 18 mm) to each time step of the normalized positions [17]–[19], and standardized the data to have zero mean and unit standard deviation. Finally, to represent movements performed in different directions, we rotated each sample eight times; we applied six rotations about the vertical axis, and two random Euler rotation sequences to represent movements performed in any orientation. In total, when including rotations (i.e., data augmentation), we used 1433 hours of data for training the model, compared to the 85 hours used in the training set of the original OpenCap model.

We compared three deep learning architectures for our marker enhancer: linear, LSTM, and transformer models. For the LSTM models, we performed a random search to find the learning rate ($6e-5$), the number of LSTM layers (4), and the number of units (128) of the *body model* (498,409 trainable


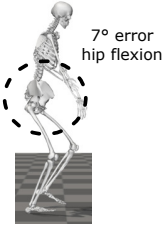
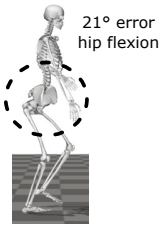

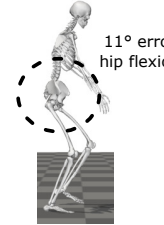

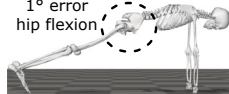
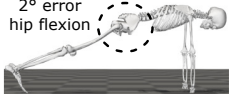

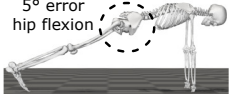

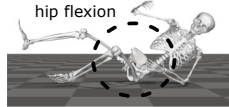
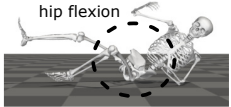

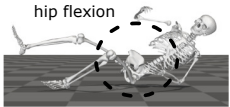

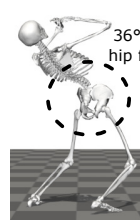

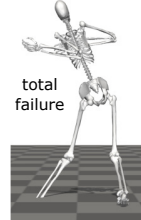
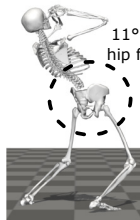
	Reference	Keypoints (no noise)	Keypoints (noise)	Uhlrich et al. (2023)	LSTM (ours)
Free walking					
Push-ups					
Rolling					
Acting					

Fig. 2. Kinematics of diverse movements computed from different synthetic marker sets: reference anatomical markers, video keypoints without and with noise, and anatomical markers predicted from noisy video keypoints using the marker enhancer from Uhlrich et al. (2023) [1] and our marker enhancer (LSTM) trained on the enhanced dataset. The instances shown are selected from the *generalizability* task dataset.

parameters). We used the same learning rate and number of units for the *arm model*, and performed a grid search to find the number of LSTM layers (5; 607,256 parameters). We used the same learning rate when training the linear models (5,040 and 576 parameters for the *body* and *arm models*, respectively). For the transformer models, we used a positional embedding layer followed by an encoder stack consisting of identical layers with a multi-head self-attention sub-layer and a feed-forward sub-layer [42]. We set the embedding dimension (256), internal dimension of the feed-forward network (1024), and attention and value key size (64) based on previous work with movement data [43], and performed a grid search to find the number of layers (2) in the encoder stack and the number of self-attention heads (4). We used a custom learning rate scheduler according to [42] and the same hyper-parameters for both *arm* (1,591,832 parameters) and *body* (1,618,793 parameters) *models*. For all three model architectures, we used a batch size of 64, the Adam optimizer [44], and a weighted mean squared error as loss function. We observed improved kinematic accuracy by doubling the weight on the error associated with the three markers located on each foot, while keeping the weight for all other markers at one. This can be explained by the sensitivity of the ankle and subtalar angles to small changes in foot marker positions. All results presented here are based on this loss function, with doubled weights for the foot markers.

For each dataset, we divided the data into three subsets: training (80%), validation (10%), and test (10%) sets. This partitioning was conducted on a per-subject basis to ensure that data from a subject was exclusively assigned to one set. We used an early stopping criterion (delta of 0 and patience of 3) based on the validation loss and fine-tuned hyperparameters to minimize the root mean squared error (RMSE) on the validation set. Each model was trained twice, and we selected the one with the lowest RMSE on the validation set. Finally, we assessed the performance of the chosen models on the test set using RMSE of non-normalized data. We trained the models in Python 3.11, using Tensorflow 2.12, and one NVIDIA GeForce RTX 3090 GPU.

C. Performance evaluation

We evaluated the performance of the three marker enhancers through two tasks. First, we assessed their accuracy when computing joint kinematics from videos on a set of benchmark movements. Second, we assessed their ability to generalize across a broad range of movements. We refer to these tasks as *accuracy* and *generalizability* tasks, respectively.

1) *Accuracy task*: We incorporated the marker enhancers as part of OpenCap to estimate joint kinematics from videos. The OpenCap pipeline consists of four main steps: 1) pose estimation to identify 2D keypoints from videos, 2) triangulation to reconstruct the 3D position of the keypoints,

3) marker enhancement to predict the 3D position of the anatomical markers, and 4) scaling and inverse kinematics in OpenSim to compute joint kinematics from anatomical marker positions. We compared video- (i.e., OpenCap-) based joint kinematics against reference values from a traditional marker-based pipeline in OpenSim using the OpenCap dataset [1], which was not used to train the marker enhancers. The dataset contains synchronized videos, marker-based motion capture marker trajectories, and force plate data from 10 subjects performing four types of activities: walking, squat, sit-to-stand, and drop jump. Each activity is performed in a self-selected manner and in a modified way to simulate the effect of musculoskeletal conditions (e.g., walking naturally and with a trunk sway gait modification). In total, the dataset includes 160 trials (six walking trials, two times five squats and five sit-to-stands, and six drop jumps per subject). We computed joint kinematics from videos using the three marker enhancers (linear, LSTM, and transformer), along with the marker enhancer from [1]. Additionally, we evaluated the direct use of keypoints as input to scaling and inverse kinematics in OpenSim, bypassing marker enhancement, to assess its significance in the pipeline. Both video- and marker-based pipelines employed the same musculoskeletal model [1], [29], [45], [46], comprising 33 degrees of freedom. Our analysis focuses on 21 lumbar and lower-body degrees of freedom (lumbar [3], pelvis in the ground frame [6], hips [2x3], knees [2x1], and ankles [2x2]). For the video-based pipeline, we utilized HRNet (person model: `faster_rcnn_r50_fpn_coco`, pose model: `hrnet_w48_coco_wholebody_384x288_dark_plus`) from mmpose [47] as the pose estimation model and videos captured from two iPhones (12 Pro; Apple Inc., Cupertino, CA, USA) positioned at approximately $\pm 45^\circ$ from the subject's forward-facing direction. We evaluated video-against marker-based pipelines using RMSE of joint kinematics.

2) *Generalizability task*: We used marker-based motion capture data from the Total Capture dataset [48], which was not used to train the marker enhancers, to create synthetic video keypoints and evaluate the generalizability of the marker enhancers across a range of movements. We incorporated data from five subjects engaged in movements including performing various ranges of motion, walking freely within a capture volume (i.e., not only walking straight along the world frame's forward direction), acting, and engaging in freestyle movements. The latter especially covered a broad range of movements (e.g., rolling and doing push ups), on which OpenCap was known to perform poorly. In total, we included data from 38 trials, corresponding to about 0.6 hours of movement. We used a similar approach as for the marker enhancer dataset to generate corresponding synthetic video keypoints and anatomical markers. First, we processed the marker data from motion capture with AddBiomechanics to obtain scaled OpenSim models and coordinate files. We then attached virtual markers corresponding to the video keypoints and (reference) anatomical markers (Fig. 1) to the scaled models and extracted the 3D trajectory of each virtual marker for each coordinate file. Next, we added Gaussian noise

(standard deviation: 18 mm) to each time step of the keypoint positions [17]–[19] and passed these positions as input to the marker enhancers to predict anatomical marker positions. Finally, we computed joint kinematics from anatomical marker positions using scaling and inverse kinematics in OpenSim. We used RMSE to compare joint kinematics computed from the reference anatomical markers and from the predicted anatomical markers. We also evaluated the direct use of video keypoints, with and without noise, as input to scaling and inverse kinematics in OpenSim to compute joint kinematics.

D. Muscle-driven tracking simulations

We evaluated the effect of improved kinematics on the accuracy of dynamic measures (i.e., ground reaction forces and joint moments). To compute dynamics, we generated tracking simulations of joint kinematics using the musculoskeletal model [1], [49]–[51]. The model is driven by 80 muscles actuating the lower-limb joints and 13 ideal torque motors actuating the lumbar, shoulder, and elbow joints. External forces are modeled through six foot-ground contact spheres attached to the foot segments. We formulated the simulations as optimal control problems that aim to identify muscle excitations that minimize a cost function subject to constraints describing muscle and skeleton dynamics. The cost function J includes effort terms (squared muscle activations a and excitations of the ideal torque motors e_T) and kinematic tracking terms (squared difference between simulated and experimental data), namely tracking of experimental joint positions \tilde{q} , velocities $\dot{\tilde{q}}$, and accelerations $\ddot{\tilde{q}}$:

$$J = \int_{t_0}^{t_f} \left(\underbrace{w_1 a^2 + w_2 e_T^2}_{\text{Effort terms}} + \underbrace{w_3 \|\tilde{q} - q\|_2^2 + w_4 \|\tilde{\dot{q}} - \dot{q}\|_2^2 + w_5 \|\tilde{\ddot{q}} - \ddot{q}\|_2^2}_{\text{Tracking terms}} \right) dt, \quad (1)$$

where t_0 and t_f are initial and final times, w_i with $i = 1, \dots, 5$ are weights, and t is time. More details about the problem formulation can be found in [1]. We formulated the problems in Python (v3.9) with CasADi [52] (v3.5.5), using direct collocation and implicit dynamics [49]. We used algorithmic differentiation to compute derivatives [50] and IPOPT to solve the resulting nonlinear programming problems [53].

We generated tracking simulations of joint kinematics using data from the *accuracy* task (i.e., four movements from the OpenCap dataset). We filtered the kinematic data (walking: 6 Hz, squat: 4 Hz, sit-to-stand: 4 Hz, and drop jump: 30 Hz) and manually tuned the cost term weights following a heuristic process. We further tailored each problem formulation to the movement of interest. The walking simulations were from right heel strike to left toe-off, incorporating time buffers at both the beginning and end of the simulations. These buffers were disregarded during data analysis, but they enhanced the simulation results within the intended time period by providing contextual boundaries. Without buffers, we observed

instability at the beginning and end of the simulations. For the simulations of squat, sit-to-stand, and drop jump, which involve large hip and knee flexion, we excluded the contribution of passive muscle forces and added reserve actuators to the hip rotation degree of freedom. We found this to be necessary to prevent non-physiological muscle force contribution in deep flexion. For the sit-to-stand simulations, we added a cost term penalizing the model from lifting its heels, thereby incorporating knowledge of the task-specific human objective into the problem formulation. We also added time buffers, following the same rationale as for walking. For the squat simulations, we segmented the repetitions based on the vertical pelvis position, and imposed periodic pelvis position and speed (i.e., same position and speed at the beginning and end of the repetition). Periodic constraints provide contextual boundaries, and we therefore did not add time buffers. Finally, the drop jump simulations were conducted from 0.3 seconds prior to landing to 0.3 seconds after take-off, encompassing a time buffer around the contact phase, which was the period of focus.

To assess the impact of kinematic accuracy on simulated dynamic measures, we conducted simulations tracking joint kinematics computed from marker data derived through three different methods: 1) predicted from videos using the marker enhancer from [1], 2) predicted from videos using the best performing marker enhancer (LSTM) trained on the enhanced dataset, and 3) measured with marker-based motion capture in the laboratory. Comparing the first two cases allows evaluating the influence of the marker enhancer model on simulated dynamic measures. Comparing simulations tracking video-versus laboratory marker-based joint kinematics helps gauge the potential gains in accuracy by refining kinematic estimates from videos (assuming laboratory marker-based kinematics are the gold standard). This analysis also aids in identifying potential limitations of muscle-driven tracking simulations for estimating dynamic measures. For all 10 subjects, we generated three simulations for each movement type (three walking trials, three squats, three sit-to-stands, and three drop jumps) for each variant (self-selected and modified). The laboratory-based marker data from walking lacked a time buffer following left toe-off, thereby not allowing us to apply the same optimal control formulation as for simulations tracking video-based kinematics. We therefore excluded walking simulations tracking laboratory marker-based kinematics from the analysis. In total, we generated 660 simulations (240 simulations for each video-based case and 180 for the laboratory marker-based case). We used RMSE of ground reaction forces (expressed in percent bodyweight, %BW) and joint moments (expressed in percent bodyweight times height, %BW*ht) as performance metrics, with force plate data and joint moments from laboratory-based inverse dynamics as reference. Note that inverse dynamics results include non-physical pelvis residual forces and moments, whereas muscle-driven simulations are dynamically consistent. Differences between joint moments from inverse dynamics and dynamic simulations are therefore not entirely attributable to errors in the simulation pipeline.

III. RESULTS

A. Marker enhancer model training

The transformer model achieved the lowest RMSEs on the test set (*body model*: 8.6 mm; *arm model*: 15.3 mm), marginally outperforming the LSTM model (*body model*: 10.0 mm; *arm model*: 16.3 mm). The linear model performed worst, with RMSEs about twice as large as those of the transformer model (*body model*: 16.5 mm; *arm model*: 31.3 mm).

B. Performance evaluation

1) *Accuracy task*: The LSTM and transformer models performed best for rotational degrees of freedom (mean RMSEs: $4.1 \pm 0.3^\circ$ and $4.4 \pm 0.4^\circ$, respectively) (Table II). The linear and LSTM models performed best for pelvis translations (mean RMSEs: 12.4 ± 1.1 mm and 12.8 ± 1.8 mm, respectively). All three models (linear, LSTM, transformer) trained on the enhanced dataset outperformed mean results obtained with the marker enhancer from [1]. Estimating kinematics from keypoints directly resulted in higher mean RMSEs ($9.6 \pm 1.5^\circ$ and 24.6 ± 1.8 mm), with RMSEs for some degrees of freedom as large as 43.1° (lumbar extension) and 45.0 mm (pelvis anterior-posterior translation).

2) *Generalizability task*: The LSTM model achieved the lowest mean RMSEs on the *generalizability* task (Table II) for both rotational degrees of freedom ($4.1 \pm 1.3^\circ$) and pelvis translations (6.5 ± 2.3 mm), marginally outperforming the transformer model ($4.5 \pm 1.5^\circ$ and 6.8 ± 1.4 mm). All three models (linear, LSTM, transformer) trained on the enhanced dataset outperformed results obtained with the marker enhancer from [1]. The latter achieved mean RMSEs larger than 40° (mean across rotational degrees of freedom) and 32 mm (mean across pelvis translations), and often failed to capture the movement kinematics (see examples in Fig. 2, column Uhlrich et al. (2023)). All three models also outperformed results obtained when using noisy keypoints instead of anatomical markers as input for scaling and inverse kinematics in OpenSim. When excluding noise, simulating perfect, albeit unrealistic, conditions, the RMSEs decreased but were still larger than those obtained with the LSTM and transformer models.

Overall, the LSTM and transformer models performed best on both *accuracy* and *generalizability* tasks. We incorporated the LSTM, which is easier to use across different operating systems, into OpenCap. For estimating dynamics using muscle-driven tracking simulations, we also used the kinematics produced with the LSTM as tracking data.

C. Muscle-driven tracking simulations

On average across movements, the muscle-driven simulations achieved slightly lower RMSEs when tracking the kinematics produced with the LSTM (ground reaction forces: 6.7 ± 4.3 %BW; joint moments: 1.34 ± 0.96 %BW*ht) compared to the kinematics produced with the marker enhancer from [1] (ground reaction forces: 7.2 ± 4.8 %BW; joint moments: 1.37 ± 1.04 %BW*ht) (Table III). Despite some improvements, the changes were limited and not consistent across movements

TABLE II
ROOT MEAN SQUARED ERROR (RMSE) BETWEEN REFERENCE AND ESTIMATED JOINT KINEMATICS

	<i>Accuracy task</i>		<i>Generalizability task</i>	
	Rotations (n=18, deg)	Translations (n=3, mm)	Rotations (n=18, deg)	Translations (n=3, mm)
Keypoints (no noise) ^a	/	/	4.6 ± 2.3 (1.9-9.1)	11.0 ± 2.3 (7.8-12.7)
Keypoints (noise) ^a	9.6 ± 1.5 (2.0-43.1)	24.6 ± 1.8 (6.2-45.0)	9.0 ± 3.9 (4.1-17.7)	15.5 ± 2.3 (12.3-17.7)
Uhlrich et al. 2023 [1]	5.3 ± 0.5 (2.0-11.5)	14.0 ± 0.9 (6.3-22.3)	40.4 ± 64.0 (14.4-252.0)	32.8 ± 5.5 (26.9-40.1)
Linear	5.1 ± 0.4 (1.7-12.6)	12.4 ± 1.1 (5.4-21.0)	7.4 ± 2.1 (4.8-11.5)	10.0 ± 3.6 (6.8-15.0)
LSTM	4.1 ± 0.3 (1.5-8.7)	12.8 ± 1.8 (5.8-22.6)	4.1 ± 1.3 (2.3-6.7)	6.5 ± 2.3 (4.2-9.6)
Transformer	4.4 ± 0.4 (1.6-9.9)	13.9 ± 1.9 (6.5-25.0)	4.5 ± 1.5 (2.4-7.8)	6.8 ± 1.4 (5.1-8.6)

In the *accuracy* task, errors for each activity were averaged over trials and degrees of freedom, and the reported RMSE is an average over activities. In the *generalizability* task, errors were averaged over trials and degrees of freedom. Kinematic errors are presented as the mean ± one standard deviation and range over the degrees of freedom. Bold numbers indicate lowest errors across conditions.

^a In the *accuracy* task, keypoints are identified from videos, inherently incorporating noise (keypoints without noise do not exist). In the *generalizability* task, keypoints are synthesized from motion data. Keypoints without noise represent perfect, albeit unrealistic, conditions, while keypoints with noise simulate real-world data.

and dynamic quantities. RMSEs consistently showed lower values for drop jumps and higher values for walking, albeit with relatively small changes. Mixed results were observed for squat and sit-to-stand.

Tracking laboratory-based kinematics did not dramatically and consistently improve the accuracy of dynamic estimates. When tracking kinematics computed from laboratory- instead of video-based marker data, the muscle-driven simulations achieved lower vertical and anterior-posterior ground reaction force RMSEs, but higher medio-lateral RMSEs for squat, sit-to-stand, and drop jumps. This resulted in higher joint moment RMSEs for squat, and lower RMSEs for sit-to-stand and drop jumps.

Out of 660 simulations, seven did not converge and were excluded from the analysis. All other simulations converged to an optimal solution, and we used the same settings for all simulations of a same movement type (e.g., squat).

IV. DISCUSSION

We created a diverse dataset that enabled the creation of a marker enhancer capable of accurately capturing joint kinematics for a broad range of movements. We demonstrated this generalizability using an unseen dataset comprising a wide array of movements. Furthermore, we showed that the marker enhancer increases accuracy on a benchmark task of estimating joint kinematics from videos. We integrated the marker enhancer in OpenCap, an open-source software platform used by thousands of researchers to measure movement from smartphone videos. Our marker enhancer improves OpenCap's performance, particularly in tasks where it previously exhibited low accuracy, like when a subject is prone or supine, or changes direction.

We compared different model architectures and found that both LSTM and transformer models achieved similar performance levels on our evaluation tasks, surpassing the performance of a linear model. We hypothesize that the task—predicting a dense set of 3D marker coordinates from a sparse one—is relatively straightforward, allowing both models to achieve comparable results. Both LSTM models, through their memory cells and gated architecture, and transformer models, through their attention mechanisms, leverage the temporal dynamics of time-series data. This utilization of temporal

information and the inherent ability of these models to capture non-linear relationships likely contribute to their superior performance compared to a simpler linear model.

The marker enhancer is both accurate and generalizable, and we hypothesize that further accuracy improvements will be challenging to realize without compromising generalizability. We base this assertion on several observations. First, our model performs equally well (RMSE of about four degrees) on both the *accuracy* task, which comprises a limited set of controlled movements from the training set, and the *generalizability* task, which comprises a broader range of movements not explicitly covered in training. This consistency suggests that our model is generalizable and not overfitting to specific movements. If it were, we would have expected worse performance on the *generalizability* task compared to the *accuracy* task. Second, our model surpasses the performance of our previous model [1] on the *accuracy* task, despite the fact that our previous model was trained on a smaller and less diverse dataset that comprised the movements tested in this task. This suggests that our new model has not sacrificed accuracy on specific movements for increased generalizability. Otherwise, we would have anticipated a decrease in accuracy on the *accuracy* task compared to our previous model [1]. Third, both the LSTM and transformer models demonstrated similar performance across both evaluation tasks. This outcome suggests that the specific architecture may not be a critical factor in this context, as both models achieved comparable results despite their structural differences. While further accuracy gains can be expected on a given movement (e.g., walking) by fine-tuning the model with a movement-specific dataset, we do not expect that refining the model architecture will lead to substantial accuracy gains on our evaluation tasks.

Our results emphasize the importance of using a dense set of markers to compute joint kinematics. Using sparse keypoints resulted in a larger mean kinematic error ($9.6 \pm 1.5^\circ$) and in a wider range of errors (up to 43.1°) across degrees of freedom (Table II, *Accuracy* task) compared to using dense anatomical markers. Our results on the *generalizability* task (Table II, *Generalizability* task) confirm this claim and further highlight that even under ideal conditions—where we have knowledge of the exact position of the sparse keypoints relative to the multi-body model, thereby ignoring noise inherent to pose

TABLE III
ROOT MEAN SQUARED ERROR (RMSE) BETWEEN REFERENCE AND ESTIMATED KINEMATICS AND DYNAMICS

		Marker data	Movements				
			Walking	Squat	Sit-to-stand	Drop jump	Mean
Joint kinematics	rotations n=18 (deg)	Predicted-[1]	4.9 (2.7-7.6)	4.8 (2.1-8.4)	5.4 (2.0-11.5)	6.1 (2.7-10.1)	5.3 ± 0.5
		Predicted-LSTM	4.0 (2.1-7.2)	4.0 (2.4-7.0)	3.9 (1.5-8.7)	4.6 (2.1-7.0)	4.1 ± 0.3
	translations n=3 (mm)	Predicted-[1]	13.9 (7.8-21.3)	13.8 (7.6-19.5)	15.3 (6.3-22.3)	12.9 (7.6-17.8)	14.0 ± 0.9
		Predicted-LSTM	11.3 (7.6-13.2)	12.2 (6.8-15.3)	15.8 (5.8-22.6)	11.7 (7.0-14.4)	12.8 ± 1.8
Ground reaction forces	vertical (%BW)	Predicted-[1]	10.1	6.1	5.5	29.3	12.7 ± 11.2
		Predicted-LSTM	10.4	4.7	5.4	26.0	11.6 ± 9.9
	anterior-posterior (%BW)	Laboratory ^a	/	3.4	4.6	20.0	/
		Predicted-[1]	2.8	1.3	2.2	10.5	4.2 ± 4.2
	medio-lateral (%BW)	Predicted-LSTM	3.0	1.4	1.7	9.0	3.8 ± 3.6
		Laboratory ^a	/	1.2	1.2	8.4	/
	anterior-posterior (%BW)	Predicted-[1]	1.3	5.0	3.1	8.7	4.5 ± 3.2
		Predicted-LSTM	1.6	5.1	3.3	8.6	4.7 ± 3.0
Joint moments	n=15 (%BW*ht)	Laboratory ^a	/	5.1	3.5	9.0	/
		Predicted-[1]	0.85 (0.26-1.41)	1.01 (0.12-1.73)	0.69 (0.17-1.09)	2.91 (0.91-7.14)	1.37 ± 1.04
		Predicted-LSTM	0.90 (0.27-1.46)	0.98 (0.11-1.69)	0.71 (0.15-1.17)	2.76 (0.81-6.39)	1.34 ± 0.95
		Laboratory ^a	/	1.03 (0.06-2.18)	0.56 (0.08-0.96)	2.47 (0.76-5.01)	/

Ground reaction forces and joint moments were computed using muscle-driven simulations tracking joint kinematics computed from marker data either predicted from videos using the reference marker enhancer from [1] (Predicted-[1] in Marker data column), predicted from videos using the newly trained LSTM (Predicted-LSTM), or measured in the laboratory (Laboratory). Joint kinematic errors were computed between video- and laboratory-based kinematic estimates, and dynamic errors were computed between muscle-driven simulation results and force plate data and inverse dynamic results. Errors for each activity were averaged over trials and degrees of freedom. The reported mean is an average ± one standard deviation across activities. Ground reaction forces are expressed in percent bodyweight (BW), and joint moments are expressed in percent BW times height (ht). Joint kinematic and moment errors are presented as the mean and range (minimum to maximum) over degrees of freedom. Bold numbers indicate the lowest errors observed among video-based results.

^a The laboratory-based marker data from walking did not allow applying the same optimal control formulation as for simulations tracking video-based data (see methods for details). Laboratory marker-based simulations of walking were therefore excluded from the analysis.

estimation models—using sparse keypoints does not result in higher accuracy compared to using the dense set of anatomical markers predicted by the marker enhancer. Interestingly, when using sparse keypoints with noise, thereby simulating more realistic conditions, the error increased and was similar in magnitude (nine degrees) to results on the *accuracy* task with video data. This validates our choice of noise magnitude (standard deviation: 18 mm).

Combining 2D pose estimation models with the marker enhancer is one method for obtaining a dense markerset from videos; however, other approaches are viable. Sáránci et al. [54] introduced a method that integrates diverse datasets used to train pose estimation models, accommodating variations in labeled keypoints—some datasets containing sparser or denser markersets. This approach shows potential for predicting a denser set of keypoints from a sparser one, and is worthy of comparison with our marker enhancer in future studies. Another approach is direct estimation of a dense set of markers from videos, potentially mitigating bias and error accumulation associated with our two-step process (pose estimation followed by marker enhancement). Our preliminary research in this direction [55] involved creating a synthetic video dataset based on the SMPL model [56] with known anatomical marker positions. Using this dataset, we then retrained 2D pose

estimation models to directly identify these markers from videos, yielding promising results on the OpenCap dataset used in our *accuracy* task. However, overall performance was not as high as that achieved with our marker enhancer. Future work should concentrate on refining the synthetic video dataset to enhance kinematic accuracy.

Our muscle-driven simulation results indicate that tracking more accurate kinematics does not always result in more accurate dynamics, highlighting the need for better methods to estimate forces. We found that improvements in joint kinematic accuracy obtained with the marker enhancer did not consistently lead to improved estimates of dynamic quantities across the different movements (Table III, Predicted-LSTM versus Predicted-[1]). And even when our muscle-driven simulations tracked joint kinematics computed from gold-standard, laboratory-based marker data (Table III, Laboratory) the dynamics errors, with respect to force plate data and inverse dynamics results, were comparable to those generated when tracking kinematics from video-based data.

We suggest that three main factors contribute to these outcomes: musculoskeletal modeling assumptions, optimal control problem formulation assumptions, and simulation sensitivity. First, modeling assumptions limit the ability of the musculoskeletal model to track given kinematics with physio-

logically realistic forces. This causes either the kinematics in the dynamic simulation to deviate from the reference kinematics, or the produced forces to be non realistic. For instance, we found the gluteus maximus muscles to produce excessive hip abduction and external rotation torques in deep hip flexion in activities like squat or sit-to-stand. Personalizing the models, in particular better characterizing the muscle-tendon parameters of the Hill-type muscle model (e.g., using electromyography [22], [57]) and the muscle geometries (e.g., using medical imaging [58]), might improve the muscle operating ranges and force production capabilities, resulting in more realistic force estimates for the given kinematics. Second, the formulation of the optimal control problems underlying the muscle-driven simulations might not fully capture the task constraints and the subject's motor control objective. For instance, the intention of the subject, which is modeled through the cost function, for walking likely differs from squatting. Using experimental force data with methods like inverse optimal control [59] might help optimize the cost function. It will, however, remain difficult to find a formulation that is robust against how different subjects perform a given task. Third, dynamic simulations lack robustness, which can lead to varying output quality. For example, performing simulations of walking for different trials using the same problem formulation may result in different quality outcomes. While this might relate to different underlying control strategy, we also found the simulations to produce different solutions when, for instance, slightly adjusting the time window of the problem, underlining the simulations' overall sensitivity. In this analysis, we applied a consistent scaling approach to obtain the musculoskeletal model and used the same problem formulation for all movements of a specific type (e.g., walking) across all ten subjects. Refining the model and problem formulation for each trial could potentially enhance result accuracy. However, this task is challenging without experimental force data available for comparison.

Overall, our muscle-driven tracking simulations provide ground reaction force and joint moment estimates that we previously showed were good enough for applications such as screening for disease risk and informing rehabilitation decisions [1]. Our findings suggest that to improve accuracy, the focus should shift to the methods used for obtaining dynamics rather than enhancing joint kinematic estimates from videos. For example, using data-driven models to predict dynamic quantities like ground reaction forces and inform muscle-driven simulations could be a promising approach.

V. CONCLUSION

We created a large and diverse dataset and developed a marker enhancer that is accurate and generalizable across a broad range of movements. Our results illustrated that to improve the accuracy of dynamic estimates from videos using muscle-driven tracking simulations, improving the fidelity of the musculoskeletal model and refining the optimal control problem formulation might have a bigger impact than improving the accuracy of kinematic estimates. By integrating our marker enhancer into OpenCap, we enable its thousands of users to more accurately measure a wider variety of movements from smartphone videos.

APPENDIX A CODE AND DATA AVAILABILITY

Source code, trained models, and public data will be available at <https://github.com/antoinefalisse/marker-enhancer>.

REFERENCES

- [1] S. D. Uhrlich, A. Falisse, Ł. Kidziński, J. Muccini, M. Ko, A. S. Chaudhari, J. L. Hicks, and S. L. Delp, "Opencap: Human movement dynamics from smartphone videos," *PLoS Computational Biology*, vol. 19, no. 10, p. e1011462, 2023.
- [2] R. M. Kanko, E. K. Laende, E. M. Davis, W. S. Selbie, and K. J. Deluzio, "Concurrent assessment of gait kinematics using marker-based and markerless motion capture," *Journal of Biomechanics*, vol. 127, p. 110665, 2021.
- [3] D. Pagnon, M. Domalain, and L. Reveret, "Pose2sim: An end-to-end workflow for 3d markerless sports kinematics—part 2: Accuracy," *Sensors*, vol. 22, no. 7, p. 2712, 2022.
- [4] Y. Desmarais, D. Mottet, P. Slangen, and P. Montesinos, "A review of 3d human pose estimation algorithms for markerless motion capture," *Computer Vision and Image Understanding*, vol. 212, p. 103275, 2021.
- [5] R. I. Hartley and P. Sturm, "Triangulation," *Computer Vision and Image Understanding*, vol. 68, no. 2, pp. 146–157, 1997.
- [6] S. Delp, F. Anderson, A. Arnold, P. Loan, A. Habib, C. John, E. Guendelman, and D. Thelen, "Opensim: Open-source software to create and analyze dynamic simulations of movement," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 11, pp. 1940–1950, 2007.
- [7] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, pp. 172–186, 2021.
- [8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015.
- [9] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5693–5703.
- [10] S. Jin, L. Xu, J. Xu, C. Wang, W. Liu, C. Qian, W. Ouyang, and P. Luo, "Whole-body human pose estimation in the wild," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 196–214.
- [11] F. Zhang, X. Zhu, H. Dai, M. Ye, and C. Zhu, "Distribution-aware co-ordinate representation for human pose estimation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 7091–7100.
- [12] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "ViTPose: Simple vision transformer baselines for human pose estimation," in *Advances in Neural Information Processing Systems*, 2022.
- [13] C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, N. Kehtarnavaz, and M. Shah, "Deep learning-based human pose estimation: A survey," *ACM Comput. Surv.*, vol. 56, no. 1, aug 2023.
- [14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 740–755.
- [15] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *2014 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 3686–3693.
- [16] L. Wade, L. Needham, P. McGuigan, and J. Bilzon, "Applications and limitations of current markerless motion capture methods for clinical gait biomechanics," *PeerJ*, vol. 10, p. e12995, 2022.
- [17] K. Isakov, E. Burkov, V. Lempitsky, and Y. Malkov, "Learnable triangulation of human pose," in *International Conference on Computer Vision (ICCV)*, 2019.
- [18] N. Nakano, T. Sakura, K. Ueda, L. Omura, A. Kimura, Y. Iino, S. Fukushima, and S. Yoshioka, "Evaluation of 3d markerless motion capture accuracy using openpose with multiple video cameras," *Frontiers in Sports and Active Living*, vol. 2, no. 50, 2020.
- [19] M. A. Boswell, S. D. Uhrlich, Ł. Kidziński, K. Thomas, J. A. Kolesar, G. E. Gold, G. S. Beaupre, and S. L. Delp, "A neural network to predict the knee adduction moment in patients with osteoarthritis using anatomical landmarks obtainable from 2d video analysis," *Osteoarthritis and Cartilage*, vol. 29, no. 3, pp. 346–356, 2021.

- [20] A. V. Ruescas-Nicolau, E. Medina-Ripoll, H. de Rosario, J. Sanchiz Navarro, E. Parrilla, and M. C. Juan Lizandra, "A deep learning model for markerless pose estimation based on keypoint augmentation: What factors influence errors in biomechanical applications?" *Sensors*, vol. 24, no. 6, 2024.
- [21] A. Falisse, S. D. Uhlrich, J. L. Hicks, A. S. Chaudhari, and S. L. Delp, "Marker data augmentation for robust markerless motion capture," in *Proceedings of the XIX International Symposium on Computer Simulation in Biomechanics*, Kyoto, Japan, 2023.
- [22] A. Falisse, S. V. Rossom, I. Jonkers, and F. De Groote, "EMG-driven optimal estimation of subject-specific Hill model muscle-tendon parameters of the knee joint actuators," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 9, pp. 2253–2262, 2017.
- [23] J. A. Thompson, A. A. Tran, C. T. Gatewood, R. Shultz, A. Silder, S. L. Delp, and J. L. Dragoo, "Biomechanical effects of an injury prevention program in preadolescent female soccer athletes," *American Journal of Sports Medicine*, vol. 45, no. 2, pp. 294–301, 2017.
- [24] J. A. Thompson-Kolesar, C. T. Gatewood, A. A. Tran, A. Silder, R. Shultz, S. L. Delp, and J. L. Dragoo, "Age influences biomechanical changes after participation in an anterior cruciate ligament injury prevention program," *American Journal of Sports Medicine*, vol. 46, no. 3, pp. 598–606, 2018.
- [25] B. V. Hooren, B. Vanwanseele, S. van Rossom, P. Teratsias, P. Willems, M. Drost, and K. Meijer, "Muscle forces and fascicle behavior during three hamstring exercises," *Scandinavian Journal of Medicine and Science in Sports*, vol. 32, no. 6, pp. 997–1012, 2022.
- [26] "CMU graphics lab motion capture database," <http://mocap.cs.cmu.edu/>, accessed: April 24, 2024.
- [27] G. D. Myer, "Dataset of marker-based motion capture data from anterior cruciate ligament (ACL) related activities," 2023, unpublished dataset shared with the authors.
- [28] S. D. Uhlrich, J. A. Kolesar, Ł. Kidziński, M. A. Boswell, A. Silder, G. E. Gold, S. L. Delp, and G. S. Beaupre, "Personalization improves the biomechanical efficacy of foot progression angle modifications in individuals with medial knee osteoarthritis," *Journal of Biomechanics*, vol. 144, p. 111312, 2022.
- [29] S. D. Uhlrich, R. W. Jackson, A. Seth, J. A. Kolesar, and S. L. Delp, "Muscle coordination retraining inspired by musculoskeletal simulations reduces knee contact force," *Scientific Reports*, vol. 12, p. 9842, 2022.
- [30] M. C. Rosenberg, B. S. Banjanin, S. A. Burden, and K. M. Steele, "Predicting walking response to ankle exoskeletons using data-driven models," *Journal of the Royal Society Interface*, vol. 17, no. 171, p. 20200487, 2020.
- [31] W. Swinnen, W. Hoogkamer, F. De Groote, and B. Vanwanseele, "Habitual foot strike pattern does not affect simulated triceps surae muscle metabolic energy consumption during running," *Journal of Experimental Biology*, vol. 22, no. 23, p. jeb212449, 2019.
- [32] W. Swinnen, I. Mylle, W. Hoogkamer, F. De Groote, and B. Vanwanseele, "Changing stride frequency alters average joint power and power distributions during ground contact and leg swing in running," *Medicine and Science in Sports and Exercise*, vol. 53, no. 10, pp. 2111–2118, 10 2021.
- [33] A. A. Gatti, P. J. Keir, M. D. Noseworthy, M. K. Beauchamp, and M. R. Maly, "Hip and ankle kinematics are the most important predictors of knee joint loading during bicycling," *Journal of Science and Medicine in Sport*, vol. 24, no. 1, pp. 98–104, 2021.
- [34] C. Schreiber and F. Moissenet, "A multimodal dataset of human gait at different walking speeds established on injury-free adult participants," *Scientific Data*, vol. 6, no. 111, 2019.
- [35] A. A. Gatti, P. J. Keir, M. D. Noseworthy, M. K. Beauchamp, and M. R. Maly, "Equations to prescribe bicycle saddle height based on desired joint kinematics and bicycle geometry," *European Journal of Sport Science*, vol. 22, no. 3, pp. 344–353, 2022.
- [36] T. Lencioni, I. Carpinella, M. Rabuffetti, A. Marzegan, and M. Ferrarin, "Human kinematic, kinetic and EMG data during different walking and stair ascending and descending tasks," *Scientific Data*, vol. 6, 2019.
- [37] A. Szczęśna, M. Błaszczyszyn, and M. Pawlyta, "Optical motion capture dataset of selected techniques in beginner and advanced kyokushin karate athletes," *Scientific Data*, vol. 8, 12 2021.
- [38] A. D. Koelewijn, D. Heinrich, and A. J. van den Bogert, "Metabolic cost calculations of gait using musculoskeletal energy models, a comparison study," *PLoS ONE*, vol. 14, no. 9, p. e0222037, 2019.
- [39] L. Fourel, P. Touzard, K. L. Arles, M. Fadier, K. Deghaies, S. Ozan, and C. Martin, "Relationship between force time curve variables and tennis performance in competitive tennis players," *Journal of Strength and Conditioning Research*, 2024, in press.
- [40] A. Seth, J. L. Hicks, T. K. Uchida, A. Habib, C. L. Dembia, J. J. Dunne, C. F. Ong, M. S. DeMers, A. Rajagopal, M. Millard, S. R. Hamner, E. M. Arnold, J. R. Yong, S. K. Lakshmikanth, M. A. Sherman, J. P. Ku, and S. L. Delp, "Opensim: Simulating musculoskeletal dynamics and neuromuscular control to study human and animal movement," *PLoS Computational Biology*, vol. 14, no. 7, p. e1006223, 2018.
- [41] K. Werling, N. A. Bianco, M. Raitor, J. Stingel, J. L. Hicks, S. H. Collins, S. L. Delp, and C. K. Liu, "Addbiomechanics: Automating model scaling, inverse kinematics, and inverse dynamics from human motion data through sequential optimization," *PLoS ONE*, vol. 18, no. 11, p. e0295152, 2023.
- [42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [43] Y. Jiang, Y. Ye, D. Gopinath, J. Won, A. W. Winkler, and C. K. Liu, "Transformer inertial poser: Real-time human motion reconstruction from sparse imu with simultaneous terrain generation," in *SIGGRAPH Asia 2022 Conference Papers*, 2022.
- [44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2024.
- [45] A. Rajagopal, C. Dembia, M. DeMers, D. Delp, J. Hicks, and S. Delp, "Full body musculoskeletal model for muscle-driven simulation of human gait," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 10, pp. 2068–2079, 2016.
- [46] A. K. Lai, A. S. Arnold, and J. M. Wakeling, "Why are antagonist muscles co-activated in my simulation? a musculoskeletal model for analysing human locomotor tasks," *Annals of Biomedical Engineering*, vol. 45, no. 12, pp. 2762–2774, 2017.
- [47] MMPose Contributors, "OpenMMLab pose estimation toolbox and benchmark," <https://github.com/open-mmlab/mmpose>, 2020.
- [48] M. Trumble, A. Gilbert, C. Malleson, A. Hilton, and J. Collomosse, "Total capture: 3d human pose estimation fusing video and inertial sensors," in *2017 British Machine Vision Conference (BMVC)*, 2017.
- [49] A. Falisse, G. Serranolfi, C. L. Dembia, J. Gillis, I. Jonkers, and F. De Groote, "Rapid predictive simulations with complex musculoskeletal models suggest that diverse healthy and pathological human gaits can emerge from similar control strategies," *Journal of The Royal Society Interface*, vol. 16, no. 157, p. 20190402, 2019.
- [50] A. Falisse, G. Serranolfi, C. L. Dembia, J. Gillis, and F. De Groote, "Algorithmic differentiation improves the computational efficiency of opensim-based trajectory optimization of human movement," *PLoS ONE*, vol. 14, no. 10, p. e0217730, 2019.
- [51] A. Falisse, M. Afschrift, and F. De Groote, "Modeling toes contributes to realistic stance knee mechanics in three-dimensional predictive simulations of walking," *PLoS ONE*, vol. 17, p. e0256311, 2022.
- [52] J. A. E. Andersson, J. Gillis, G. Horn, J. B. Rawlings, and M. Diehl, "Casadi : a software framework for nonlinear optimization and optimal control," *Mathematical Programming Computation*, vol. 11, pp. 1–36, 2019.
- [53] A. Wächter and L. T. Biegler, "On the implementation of primal-dual interior point filter line search algorithm for large-scale nonlinear programming," *Mathematical Programming*, vol. 106, pp. 25–57, 2006.
- [54] I. Sárándi, A. Hermans, and B. Leibe, "Learning 3D human pose estimation from dozens of datasets using a geometry-aware autoencoder to bridge between skeleton formats," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023.
- [55] Y. Gozlan, A. Falisse, S. D. Uhlrich, A. Gatti, M. J. Black, and A. S. Chaudhari, "OpenCapBench: A benchmark to bridge pose estimation and biomechanics," *arXiv preprint arXiv:2406.09788*, 2024.
- [56] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015.
- [57] D. G. Lloyd and T. F. Besier, "An EMG-driven musculoskeletal model to estimate muscle forces and knee joint moments in vivo," *Journal of Biomechanics*, vol. 36, no. 6, pp. 765–776, 2003.
- [58] L. Modenese and J. Kohout, "Automated generation of three-dimensional complex muscle geometries for use in personalised musculoskeletal models," *Annals of Biomedical Engineering*, vol. 48, pp. 1793–1804, 6 2020.
- [59] K. Mombaur, A. Truong, and J. P. Laumond, "From human to humanoid locomotion—an inverse optimal control approach," *Autonomous Robots*, vol. 28, pp. 369–383, 2010.