Measuring Physical Plausibility of 3D Human Poses Using Physics Simulation

Nathan Louis[1]
natlouis@umich.edu
Mahzad Khoshlessan[2]
mkhoshle@umich.edu
Jason J. Corso[1,2]
jjcorso@umich.edu

[1] Electrical and Computer Engineering
University of Michigan
Ann Arbor, Michigan, USA

[2] Robotics
University of Michigan
Ann Arbor, Michigan, USA

Abstract

Modeling humans in physical scenes is vital for understanding human-environment interactions for applications involving augmented reality or assessment of human actions from video (*e.g*. sports or physical rehabilitation). State-of-the-art literature begins with a 3D human pose, from monocular or multiple views, and uses this representation to ground the person within a 3D world space. While standard metrics for accuracy capture joint position errors, they do not consider physical plausibility of the 3D pose. This limitation has motivated researchers to propose other metrics evaluating jitter, floor penetration, and unbalanced postures. Yet, these approaches measure independent instances of errors and are not representative of balance or stability during motion. In this work, we propose measuring physical plausibility from within physics simulation. We introduce two metrics to capture the physical plausibility and stability of predicted 3D poses from any 3D Human Pose Estimation model. Using physics simulation, we discover correlations with existing plausibility metrics and measuring stability during motion. We evaluate and compare the performances of two state-of-the-art methods, a multi-view triangulated baseline, and ground truth 3D markers from the Human3.6m dataset.

1 Introduction

Physically grounding objects from video is vital for understanding spatial relationships [36], geometric properties [34], and contact forces [3, 18]. For humans, this allows us to understand how a person interacts within an environment and models their actions for entertainment or evaluation in sports and physical rehabilitation [4, 19, 20]. To physically ground a person from video, common learning-based approaches employ state-of-the-art 3D human pose estimation (HPE) methods [15, 22, 40, 41] and measure performance using Mean Per Joint Position Error (MPJPE), Mean Per-vertex Error (MPVE), or Probability of Correct Keypoint (PCK). But even impressive pose estimation progress on computer vision datasets [6, 11, 35] present notable flaws in pose estimates, such as floating, foot-skating, ground penetration and unnatural positions [27]. Moreover, MPJPE—the primary quantitative metric—often displays little correlation between low error and visual quality of results [14]. These

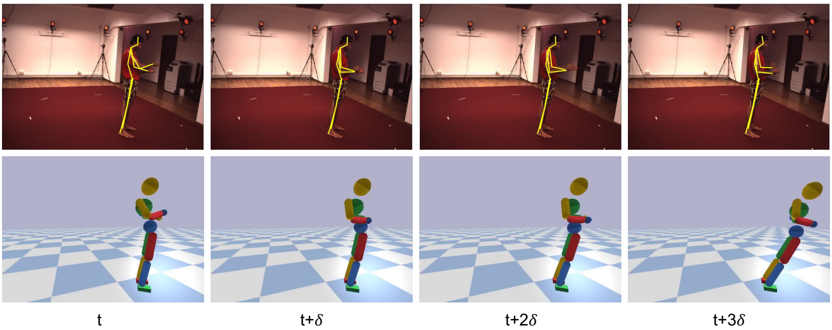


Figure 1: On the Human3.6M dataset (*S9 - Directions 1*), while the predicted pose (top row) appears plausible, in simulation we see the lean at the hip causes a loss of balance over time.

discrepancies led us to question the physical plausibility, hence physical grounding, of these predicted 3D human poses. Prior studies [10, 25] have shown that humans are sensitive to motions lacking perceptual realism which may negatively impact applications involving augmented or virtual reality and physical understanding.

Identifying these issues, researchers have explored physics-based metrics [5, 15, 27, 28, 32, 39] for 3D HPE. For measuring realistic contact with the ground, metrics such as foot-skate [27], foot slide [39], and ground penetration [28, 39] have been proposed. To highlight unnatural jittering, others have computed smoothness losses from the velocity or acceleration of joints [5, 15, 28, 39]. And for stationary stability, Shimada *et al.* [28] and Tripathi *et al.* [32] introduced terms to measure balanced and unbalanced static postures. However, these approaches do not measure stability during motion or over time, rather independent instances of errors only. In Figure 1, we show how an unnatural lean in the predicted pose may appear stable but overtime will lead to a loss of balance in simulation.

To overcome these limitations, we evaluate the physical plausibility of 3D HPEs through physics simulation. Physically plausible poses should obey the laws of physics, while physical simulators must replicate friction, gravity, and collisions. We hypothesize a link between pose plausibility and simulation stability: more stable simulations likely indicate more plausible poses. We measure the physical plausibility on simulated bodies using two new metrics: CoM distance, the center-of-mass trajectory distance with the kinematics reference, and Pose Stability Duration, the time at which a pose can be simulated before reaching an irrecoverable failure. These two measures describe the temporal stability of the simulated body undergoing motion, and unlike earlier physics-based metrics, we encapsulate dynamics and collisions without relying on hand-crafted heuristics. We also note, our simulated-based metrics can be applied to any off-the-shelf 3D HPE result without ground truth data.

Our contributions are two physical simulation-based metrics, CoM distance and Pose Stability Duration, for evaluating the physical plausibility of 3D HPEs and a comprehensive analysis of publicly available state-of-the-art methods and baselines. All simulation and evaluation code have been made publicly available [1].

[1] https://github.com/MichiganCOG/Simulation_Physical_Plausibility

2 Related Work

3D Human Pose Estimation

3D HPE is performed using monocular [15, 22, 30] or multi-view [12, 21, 24, 26, 33, 40] camera inputs. Monocular camera-based solutions are cost-effective, however, they present unique challenges due to their inherent depth ambiguity. Multi-view methods resolve these ambiguities through aggregation of multiple camera but require known camera projection matrices. Within these two classes, single-stage approaches approximate the 3D pose by estimating parameters of a statistical body model [15, 17, 38] and two-stage methods [21, 22, 41] employ a 2D detector to predict a 2D pose and “lift” it into 3D. The most commonly used metrics for quantitative evaluation, MPJPE, MPVE, and PCK, essentially measure the average displacement between predicted joints and corresponding ground truth joints. However, they fall short in their ability to quantitatively evaluate plausible postures or motion. Our proposed metrics CoM distance and Pose Stability Duration address this through physical simulation of the predicted pose.

Physics-aware 3D Human Pose Estimation

The prior 3D HPE methods are kinematics-based, only modeling joint location and disregard underlying forces and environmental factors. Failure to account for dynamics results in errors such as floating, foot skating, and unrealistic postures [27]. To address these implausibilities, some works have proposed ground contact metrics to measure ground penetration [28, 39] and foot sliding (or skating) [27, 39]. But these metrics use handcrafted heuristics and are most reliable on a calibrated floor plane. To measure stability of poses, other works [28, 32] incorporate terms from biomechanics literature [8, 9, 37] to discern balanced stationary postures. However, these functions are only valid for stationary poses and not a pose undergoing motion because it does not account for the velocity of the center of gravity [9]. Despite these advances, these metrics do not analyze the temporal impacts of physical implausibility or a body in motion. For a more comprehensive evaluation, we look to physical simulation as a test bed for plausible motion assessment under the effects of gravity, frictional forces, and collisions.

3 Method

We propose a new evaluation method for assessing the physical plausibility of 3D human pose estimates (HPE); we show an overview of our approach in Figure 2. From a given video, we predict 3D human poses using an off-the-shelf 3D HPE model and estimate a reference kinematic trajectory. Following, we apply these kinematics to a simulated body and optimize its motion under physical effects of the simulation environment. Finally, we measure the stability of the simulation using our proposed metrics to quantitatively evaluate the physical plausibility of the predicted 3D poses. In Supplemental Material, we discuss qualitative analysis of estimated ground contacts forces.

3.1 Kinematic Initialization

Given a video v of a person performing some action, we predict a sequence of 3D skeletal poses $\mathbf{X} \in \mathbb{R}^{T \times J \times 3}$ using an off-the-shelf 3D HPE method, for J joints across T frames.

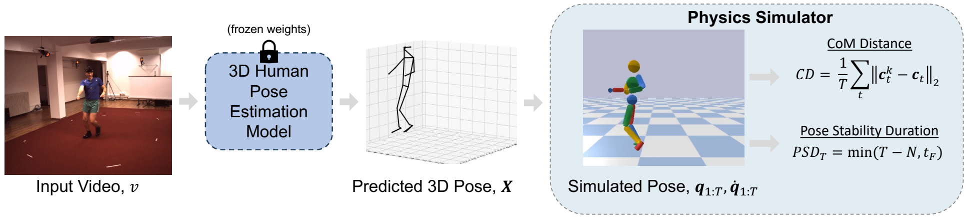

(frozen weights)
3D Human Pose Estimation Model
Input Video, v
Predicted 3D Pose, $\boldsymbol{X}$
Physics Simulator
CoM Distance
$CD = \frac{1}{T}\sum_t \|\boldsymbol{c}_t^k - \boldsymbol{c}_t\|_2$
Pose Stability Duration
$PSD_T = \min(T - N, t_F)$
Simulated Pose, $\boldsymbol{q}_{1:T}, \dot{\boldsymbol{q}}_{1:T}$

Figure 2: From a video v, we estimate 3D human poses $\mathbf{X}$ from an off-the-shelf 3D HPE model. Next, we initialize kinematic joint targets $\mathbf{q}^k_{1:T}$ on a simulated body and optimize it to mimic the reference motion under simulated environmental effects. We measure the plausibility of this optimized output using CoM distance and Pose Stability Duration.

Models generating a mesh, such as SMPL [17], can use a pre-trained regressor to estimate a 3D skeletal pose. We apply minimal pre-processing to reduce noise in X, first a median filter ($w = 15$ frames) and then constraining bone lengths to their mean values.

To apply the pose to a simulated body, we initialize a kinematic representation using a set of kinematic poses $\mathbf{q}^k_{1:T}$ from X. Each kinematic pose $\mathbf{q} = (q_1, q_2, ..., q_D)$ is a concatenation of all joint angles, parameterized as quaternions, for D degrees-of-freedom (DOF) on the simulated body. The kinematic pose is generally optimized from an inverse kinematics problem, however the ill-posed nature of this optimization often produces implausible and substantially different poses from the predictions in X. Similar to [23], we define a kinematic tree and use change of basis rotations, from the root to the end effectors, to approximate the joint angles in $\mathbf{q}$. This assumes both the skeletal pose and the simulated body have compatible kinematic trees. The resulting output, $\mathbf{q}^k_{1:T}$, is a direct kinematic representation of the 3D pose sequence without alteration from an inverse kinematics solver.

The simulated body is a humanoid with 28 DOFs, and 12 controllable joints. This includes eight spherical joints with 3 DOFs each and four revolute joints (knees and elbows) with 1 DOF each. The root node (pelvis) is excluded as a controllable joint, since applying forces directly to the root of the kinematic tree is not realistic and lacks physical meaning [3]. Each joint is paired with a Stable PD controller [31], which takes as input target joint angles, joint velocities, and gains (k_p,k_d) to output a torque on each joint. At each time step, the torque inputs are computed as

$$\tau = k_p(\mathbf{q}^k - \mathbf{q}) + k_d(\dot{\mathbf{q}}^k - \dot{\mathbf{q}}). \tag{1}$$

Here, $\mathbf{q}^k$ represents the target kinematic pose while $\mathbf{q}$ is the current kinematic pose, and joint velocities $\dot{\mathbf{q}}^k_{1:T}$ are estimated using first-order finite differences. The proportional and derivative gains k_p and k_d are fixed across all experiments.

3.2 Trajectory Optimization

We perform trajectory optimization [1, 2, 5] to optimize the kinematic pose targets so that the simulated body mimics the action in video. Given that physical simulators are highly advanced and generally non-differentiable, we use the derivative-free Covariance Matrix Adaptation Evolution Strategy (CMA-ES) [7] algorithm. We run the simulator at 1kHz with kinematic targets set at 25Hz, optimizing for 200 iterations and a population size of 100, on eight overlapping windows of length 0.5s. Rather than optimize all time steps simultaneously, we reduce the search space by representing the targets as Euler angles in a cubic B-spline and optimize the knots. We minimize the following cost functions: The center-of-mass loss ,

$$L_{COM} = \sum_t \left(\mathbf{c}_t^k - \mathbf{c}_t \right)^2, \tag{2}$$

the distance between the COM of the kinematic target, $\mathbf{c}^k$, and the current COM, $\mathbf{c}$, from the simulated body. The center-of-mass velocity loss,

$$L_{COMv} = \sum_t \left(\dot{\mathbf{c}}_t^k - \dot{\mathbf{c}}_t \right)^2, \tag{3}$$

the distance between COM velocities of the kinematic target and the current velocity. The root orientation loss,

$$L_{orn} = \sum_t \arccos \left(\left| \left\langle \mathbf{q}_{root}^k, \mathbf{q}_{root} \right\rangle \right| \right), \tag{4}$$

constraining the root orientation of the body to align with the kinematic reference. The kinematic pose loss,

$$L_{pose} = \sum_t W * \left(\mathbf{q}_t^k - \mathbf{q}_t \right)^2, \tag{5}$$

and the kinematic pose velocity loss,

$$L_{vel} = \sum_t W * \left(\dot{\mathbf{q}}_t^k - \dot{\mathbf{q}}_t \right)^2, \tag{6}$$

minimizing the joint angle and joint velocity differences. We minimize these two losses by parameterizing the joint angles as Euler angles and W is the joint weighting array applied element-wise to each joint. Next, is the joint acceleration loss,

$$L_{acc} = \sum_t ||\ddot{\mathbf{q}}_t||^2. \tag{7}$$

to encourage a smoother trajectory and avoid jittery motions. And finally, we introduce a contact loss,

$$L_{feet} = \sum_t ||\mathbf{p}_t^k - \mathbf{p}_t||_1, \tag{8}$$

to encourage alignment of foot contact states with those of the input pose. We use joint values in X to estimate ground truth foot contacts $\mathbf{p}^k$, discussed briefly in Supplemental Material. The foot contact states in simulation $\mathbf{p}$ are determined using a small height threshold (0.0005) between the ground plane and feet. We observe that this constraint helps to maintain balance when shifting weight between feet. The total optimization objective function is a linear combination of the aforementioned losses with the following weights $w_{COM} = 20$, $w_{COMv} = 0.5$, $w_{orn} = 1.0$, $w_{pose} = 1.0$, $w_{vel} = 5e^{-3}$, $w_{acc} = 1e^{-10}$, and $w_{feet} = 1.5$.

3.3 Simulation-based Metrics

Finally, we measure the stability of the optimized simulated body motion using our proposed metrics CoM distance and Pose Stability Duration to correlate to physical plausibility.

Metric 1: CoM distance The CoM distance is straightforward, computing the L2 distance between the kinematic COM trajectory and the final optimized COM,

$$\mathrm{CD} = \frac{1}{T}\sum_t \|\mathbf{c}_t^k - \mathbf{c}_t\|_2, \tag{9}$$

in millimeters (mm). Comparing the final trajectory with the kinematic reference indicates how well the 3D pose estimate X can be simulated. A low value suggests that the reference pose is plausible and easy for the simulated body to follow, while a high value indicates significant deviations from the reference.

Metric 2: Pose Stability Duration We introduce Pose Stability Duration to understand the time at which instability in simulation occurs. First, we identify two states of motion from the predicted pose: stationary and non-stationary. Values $|\dot{\mathbf{c}}_t^k|$ below a threshold (250mm/s) are considered stationary, while values above that threshold are considered non-stationary.

For the stationary state, biomechanics literature [8, 9, 37] states that a stationary pose is considered balanced if its center of gravity (CoG) falls within its base-of-support (BoS), *i.e.* the convex hull of all ground contacts. We compute the BoS on the simulated body, and denote the number of instances where the center of gravity is beyond the convex hull with N. A slight loss of balance in the stationary case is recoverable. For the non-stationary state, a body in motion may have its CoG lie outside of its BoS, invalidating the previous assumption [9]. Within the simulator we note a loss of balance in this state as the time t_F when a body part, other than the foot, from the simulated body touches the ground. A loss of balance in the non-stationary case is not recoverable. We note that prior knowledge of the action (*e.g.* cartwheels, push-ups) may be required for accurate computation. We defined this metric as

$$\mathrm{PSD}_T = \min(T - N, t_F), \tag{10}$$

where T is the total number of frames in each sequence. By examining both stationary and non-stationary poses, we assess the maximum number of simulated frames before a catastrophic failure, *i.e.* an unrecoverable deviation from the reference motion.

4 Experiments

4.1 Evaluation Dataset

All experiments are performed on Human3.6M [11], a video dataset capturing human actions using four cameras and a motion capture system where full camera projection matrices are assumed known. We use the validation subjects *S9, S11*, and the same subset of actions as prior work [28]: *Directions, Discussion, Greeting, Posing, Purchases, Photo, Waiting, WalkDog, WalkTogether*, and *Walking*. Videos are downsampled from 50fps to 25fps on 100 frames for computational efficiency in simulation.

4.2 Evaluation Models

We evaluate physical plausibility using a monocular model, PoseFormer [41], a physics-aware model, NeuralPhysCap [29], and a multi-view baseline. PoseFormer is a spatio-temporal transformer that produces a 3D human pose from 2D pose detections. PoseFormer outputs a pose in the camera space, hence to estimate a global pose we minimize the 2D reprojection loss through a differentiable projection function. NeuralPhysCap proposes a differentiable framework to generate physically plausible poses through contact estimation, force estimation, and physics-aware optimization. We chose both PoseFormer and NeuralPhysCap for their publicly available code and to cover two kinematic and physics-aware aspects of monocular 3D HPE models. Last, we include a multi-view triangulated baseline that generates global 3D poses by applying RANSAC triangulation on 2D wholebody MSCOCO detections [13, 16] from all available camera views. NeuralPhysCap generates a 16-joint skeleton and the multi-view baseline produces a 23-joint skeleton, including heel and toe joints, differing from the 17-joint H36M skeleton. However, we only compute MPJPE on the mutual joints. We provide additional details on pose formats in Supplemental Material.

4.3 Evaluation Metrics

We evaluate models using kinematics-based and physics-based metrics. First, we use MPJPE, measured in mm, to compute the pose error relative to the root location. Then, MPJPE-2D to evaluate the 2D image alignment between the projected ground truth and projected 3D prediction. And MPJPE-G to account for the global 3D position error in the world space. For physics-based metrics, in addition to our proposed metrics, we include Footskate (FS%) [27] and ground penetration (GP) [28, 39]. FS% measures the percentage of frames where the foot moves more than 2cm while in contact with the ground. And GP computes the average distance to the ground for joints below the ground plane. We discuss details on estimating floor height and ground contact states in Supplemental Material.

5 Results

In our main results in Table 1, we introduce GT 3D, the ground truth surface markers from H36M, as a soft upperbound for the 3D pose. The first three columns focus on MPJPE-derived metrics which emphasize spatial alignment of poses, while the latter columns evaluate physical plausibility. PoseFormer [41] has the lowest MPJPE, 42.5mm, and MPJPE-2D, 10.6 pixels. The baseline has the lowest MPJPE-G with 57.2mm because it utilizes multiple views and known camera projection matrices to regress a global pose. NeuralPhysCap has the highest amount of error, likely due to its miscalculation of the camera intrinsic parameters. This can be seen in the 2D spatial alignment in Figures 3 and 4, however we observe less impact in our proposed plausibility metrics.

Method	MPJPE ↓	MPJPE-G ↓	MPJPE-2D ↓	FS (%) ↓	GP ↓	CD (Ours) ↓	PSD$_{100}$ (Ours) ↑
GT 3D	-	-	-	**0.0**	1.08	33.7	63.1
NeuralPhysCap [29]	81.6	439.6	36.3	29.7	2.62	29.6	62.3
PoseFormer [41]	**42.5**	299.4	**10.6**	4.4	0.30	36.2	64.8
Baseline	55.6	**57.2**	12.0	2.6	**0.26**	**27.3**	**70.6**

Table 1: Results on the validation subset of the Human3.6M dataset, comparing kinematics and physics-based plausibility metrics.

In the next four columns we examine physical plausibility. The baseline produces the second lowest foot skate, 2.6% and the lowest GP error, 0.26mm, demonstrating a consistent estimation of pose with ground plane. While GT 3D has no foot skate error, it surprisingly has the second highest ground penetration. But this variance in ground penetration depth could be due to the inherent margin of error from motion capture surface markers. Comparatively with our metrics, the baseline is demonstrated to be the most physically plausible with the lowest CD, 27.3mm, and the highest PSD_{100}, 70.6 frames. Unlike the prior metrics we observe that NeuralPhysCap performs closer to the other methods within simulation, with 29.6mm and 62.3 frames, respectively. This indicates some invariance to strict 2D spatial alignment of the pose prediction and an increased focus on the plausibility of its motion. In Table 2, we examine the per-class performance of Pose Stability Duration. The lowest performing classes: *Greeting, Purchases, Waiting, and WalkDog*, contain crouching and bending over movements which increases the difficulty of optimizing balanced movements on the simulated body.

Method	Dir.	Disc.	Greet	Photo	Pose	Purch.	Wait	WalkD.	WalkT.	Walk	Avg.
GT 3D	**91.6**	**83.8**	10.5	**94.5**	**82.7**	37.3	37.6	36.9	**80.0**	76.1	63.1
NeuralPhysCap [29]	86.5	82.3	73.1	89.6	62.7	35.8	**70.4**	52.1	26.2	44.3	62.3
PoseFormer [41]	69.8	77.8	**81.0**	82.8	80.5	13.3	42.2	**55.8**	66.1	**78.3**	64.8
Baseline	88.2	82.4	66.0	87.6	72.9	**72.4**	53.5	27.7	77.8	77.2	**70.6**

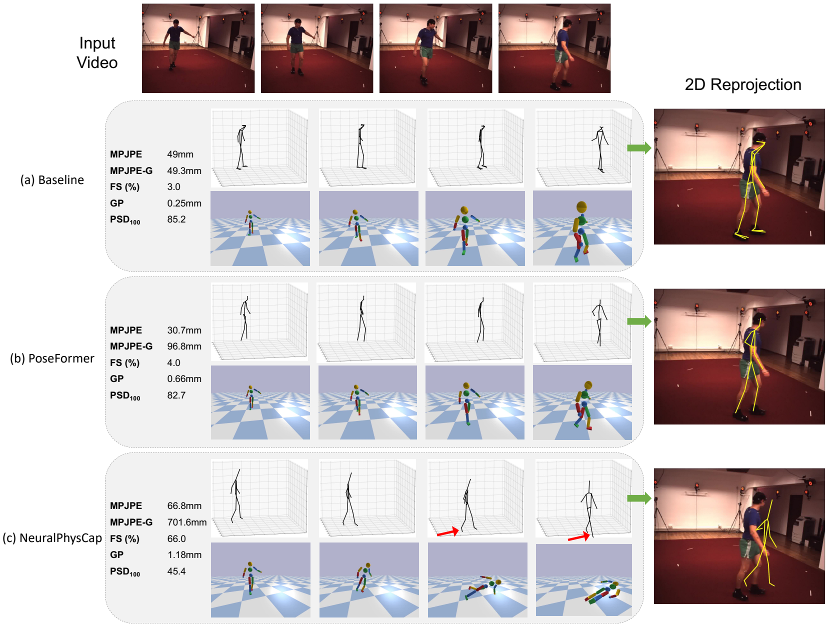
Table 2: We break down the per-class performance for Pose Stability Duration (PSD_{100}).

We provide a visual example in Figure 3. The baseline (a) and PoseFormer (b) show comparable performance on physical plausibility metrics and produce similar simulation results. While, NeuralPhysCap fails early due to inaccurate camera intrinsic and ground plane estimation, more notable in this example due to the large amount of motion for this action. The red arrows show where the 3D pose attempts to penetrate the ground plane and on the right column, we see the 2D misalignment in the image plane

2D Image Alignment In Figure 4, we provide an example of NeuralPhysCap to show how spatially misaligned poses can still generate physically plausible poses. Although the kinematic metrics errors are higher than average, 110.4mm for MPJPE and 50.0mm for MPJPE-2D, physical plausibility according to our metrics CD, 0.381, and PSD_{100}, 71.9, are on par with Table 1. Conversely, FS, 32%, is very high while and GP is 0.07mm. NeuralPhysCap explicitly corrects poses for physical implausibilities, so for relatively little motion it makes sense that the output pose is plausible, despite poor spatial alignment. This suggests that we are assessing the plausibility of the pose and not strict pose alignment, which can improved with more accurate camera parameter estimation.

6 Discussion

In this work, we propose two simulation-based metrics, CoM distance and Pose Stability Duration, to measure the physical plausibility of 3D HPE using physical simulation. CoM distance captures how closely the 3D pose estimate can be simulated when used as a kinematic reference. And Pose Stability Duration records the time the pose can be simulated before reaching an irrecoverable failure. While prior approaches capture independent instances of physical implausibilities, they do not demonstrate the ability to understand the progression of these instabilities. Our metrics incorporates the temporal aspects of stability while also integrating simulated physical properties. Moreover, we demonstrate some invariance to strict spatial alignment of 3D poses, highlighting instead the feasibility as shown by 2D image alignment disparities. Our experiments shows consistency with other kinematics and physics-based metrics when determining the most physically plausible output.



(a) Baseline

MPJPE	49mm
MPJPE-G	49.3mm
FS (%)	3.0
GP	0.25mm
PSD_{100}	85.2

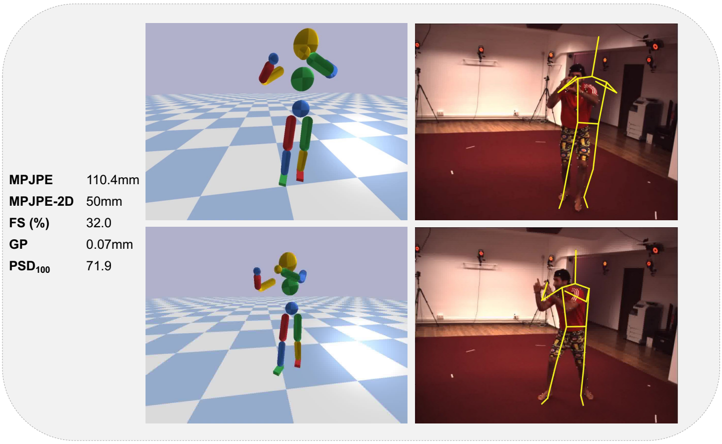
(b) PoseFormer

MPJPE	30.7mm
MPJPE-G	96.8mm
FS (%)	4.0
GP	0.66mm
PSD_{100}	82.7

(c) NeuralPhysCap

MPJPE	66.8mm
MPJPE-G	701.6mm
FS (%)	66.0
GP	1.18mm
PSD_{100}	45.4

Figure 3: For the *S11 - WalkTogether* example, we show the 3D pose, optimized simulation, and 2D reprojection. Inaccurate camera and ground plane assumptions in (c) causes the motion to fail early on as the simulated body tries to step through the ground plane (red arrows).



MPJPE	110.4mm
MPJPE-2D	50mm
FS (%)	32.0
GP	0.07mm
PSD_{100}	71.9

Figure 4: For *S9 - Photo 1* example, we show a misaligned 2D re-projected (right column) can still produce physically plausible simulation (left column) from NeuralPhysCap [29]. While this example displays higher MPJPE-2D=50.0mm, we measure Pose Stability Duration to be on par with other methods, $PSD_{100} = 71.9$.

Limitations This work is impacted by the convergence of the trajectory optimization. If the optimization falls into a local minimum, it may deviate from the reference motion which we mitigate through our constraints. Additionally, we retarget all humans to the same simulated body, which can introduce modeling errors when the shape and size vary drastically between subjects. Future work may resolve this by modifying the limbs of the simulated body to reflect the approximate shape and size attributes of the detected humans.

7 Acknowledgements

This work has been supported by NIH Award 5 R01 HL-146619-05.

References

[1] Mazen Al Borno, Martin De Lasa, and Aaron Hertzmann. Trajectory optimization for full-body movements with complex contacts. *IEEE transactions on visualization and computer graphics*, 19(8):1405–1414, 2012.

[2] Mazen Al Borno, Ludovic Righetti, Michael J Black, Scott L Delp, Eugene Fiume, and Javier Romero. Robust physics-based motion retargeting with realistic body shapes. In *Computer Graphics Forum*, volume 37, pages 81–92. Wiley Online Library, 2018.

[3] Marcus A Brubaker, Leonid Sigal, and David J Fleet. Estimating contact dynamics. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2389–2396. IEEE, 2009.

[4] Mihai Fieraru, Mihai Zanfir, Silviu Cristian Pirlea, Vlad Olaru, and Cristian Sminchisescu. Aifit: Automatic 3d human-interpretable feedback models for fitness training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9919–9928, 2021.

[5] Erik Gärtner, Mykhaylo Andriluka, Hongyi Xu, and Cristian Sminchisescu. Trajectory optimization for physics-based reconstruction of 3d human pose from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13106–13115, 2022.

[6] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020.

[7] Nikolaus Hansen. The cma evolution strategy: a comparing review. *Towards a new evolutionary computation: Advances in the estimation of distribution algorithms*, pages 75–102, 2006.

[8] At L Hof. The equations of motion for a standing human reveal three mechanisms for balance. *Journal of biomechanics*, 40(2):451–457, 2007.

[9] At L Hof, MGJ Gazendam, and WE Sinke. The condition for dynamic stability. *Journal of biomechanics*, 38(1):1–8, 2005.

[10] Ludovic Hoyet, Rachel McDonnell, and Carol O'Sullivan. Push it real: Perceiving causality in virtual interactions. *ACM Transactions on Graphics (TOG)*, 31(4):1–9, 2012.

[11] 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.

[12] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7718–7727, 2019.

[13] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[14] 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.

[15] 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.

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[17] 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.

[18] Nathan Louis, Jason J Corso, Tylan N Templin, Travis D Eliason, and Daniel P Nicolella. Learning to estimate external forces of human motion in video. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3540–3548, 2022.

[19] Jean-Sébastien Monzani, Paolo Baerlocher, Ronan Boulic, and Daniel Thalmann. Using an intermediate skeleton and inverse kinematics for motion retargeting. In *Computer Graphics Forum*, volume 19, pages 11–19. Wiley Online Library, 2000.

[20] Fotini Patrona, Anargyros Chatzitofis, Dimitrios Zarpalas, and Petros Daras. Motion analysis: Action detection, recognition and evaluation based on motion capture data. *Pattern Recognition*, 76:612–622, 2018.

[21] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Harvesting multiple views for marker-less 3d human pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6988–6997, 2017.

[22] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[23] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel Van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions On Graphics (TOG)*, 37(4):1–14, 2018.

[24] Haibo Qiu, Chunyu Wang, Jingdong Wang, Naiyan Wang, and Wenjun Zeng. Cross view fusion for 3d human pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4342–4351, 2019.

[25] Paul SA Reitsma and Nancy S Pollard. Perceptual metrics for character animation: sensitivity to errors in ballistic motion. In *ACM SIGGRAPH 2003 Papers*, pages 537–542. 2003.

[26] Edoardo Remelli, Shangchen Han, Sina Honari, Pascal Fua, and Robert Wang. Lightweight multi-view 3d pose estimation through camera-disentangled representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6040–6049, 2020.

[27] Davis Rempe, Leonidas J Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. Contact and human dynamics from monocular video. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 71–87. Springer, 2020.

[28] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. Physcap: Physically plausible monocular 3d motion capture in real time. *ACM Transactions on Graphics (ToG)*, 39(6):1–16, 2020.

[29] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, Patrick Pérez, and Christian Theobalt. Neural monocular 3d human motion capture with physical awareness. *ACM Transactions on Graphics (ToG)*, 40(4):1–15, 2021.

[30] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Black Michael J., and Tao Mei. Monocular, One-stage, Regression of Multiple 3D People. In *ICCV*, 2021.

[31] Jie Tan, Karen Liu, and Greg Turk. Stable proportional-derivative controllers. *IEEE Computer Graphics and Applications*, 31(4):34–44, 2011.

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

[33] Hanyue Tu, Chunyu Wang, and Wenjun Zeng. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 197–212. Springer, 2020.

[34] Shubham Tulsiani, Saurabh Gupta, David F Fouhey, Alexei A Efros, and Jitendra Malik. Factoring shape, pose, and layout from the 2d image of a 3d scene. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 302–310, 2018.

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

[36] Zhenzhen Weng and Serena Yeung. Holistic 3d human and scene mesh estimation from single view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 334–343, 2021.

[37] David A Winter. Human balance and posture control during standing and walking. *Gait & posture*, 3(4):193–214, 1995.

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

[39] Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason Saragih. Simpoe: Simulated character control for 3d human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7159–7169, 2021.

[40] 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: 13153–13164, 2021.

[41] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11656–11665, 2021.