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

1 Supplemental Material

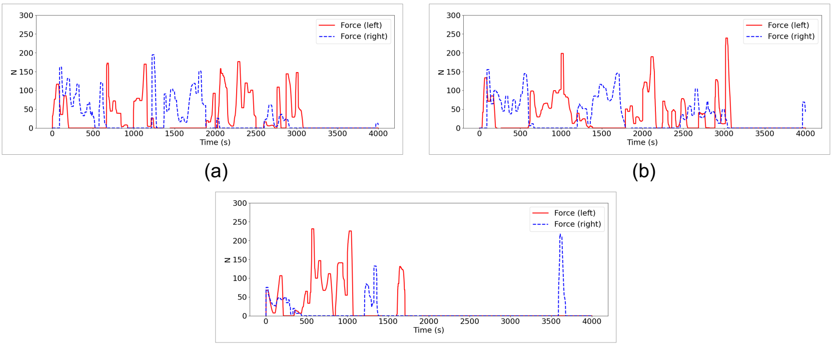
1.1 Contact Force Estimation

We demonstrate the ability to generate plausible ground contact forces from within the simulator in Figure 1 for the *S11 - WalkTogether 1* sequence. Because there are no ground truth contacts or forces, this is useful for subjective analysis. The ground contact states are estimated from the ground truth MoCap markers and the contact forces are approximated within the physical simulator by summing the normal and lateral forces when the foot joint makes contact with the ground plane. For the walking motion, we expect alternating peak magnitudes for each foot and we see that the Baseline and PoseFormer show this in the estimated forces and ground contacts. While the NeuralPhysCap example fails very early on.

We determine the height of the ground plane from the average $\lfloor 0.05 * T \rfloor$ lowest joint locations, roughly 5% of the T frames. The ground plane is assumed to be normal to the initial pose location. And to estimate ground contact states on each sequence, we use the estimated ground plane and follow the same heuristics as Rempe *et al*. [1] employing a height threshold of 5cm and velocity threshold of 2cm/s on the foot joints to identify ground contact.

1.2 Skeleton Pose Formats

In Figure 2, we layout the joints supported by each skeletal pose. PoseFormer uses the H36M skeleton which is composed of markers on top of the skin rather than body joint centers. The Baseline and NeuralPhysCap use skeleton formats from human annotated datasets, *e.g*. MSCOCO, with annotated joint centers. All of the skeleton poses are mapped to the same simulated body using in-common upperbody and lowerbody joints. The joint angles for the simulated body are neck, chest, shoulders (2), pelvis, elbows (2), hips (2), knees (2), ankles (2). There are a total of 12 controllable joints on the simulated body, the pelvis is not controlled.



(a)

(b)

(c)

Figure 1: Generated ground contact forces on *S11 - WalkTogether 1* from Human3.6m. The following results are from (a) Baseline (b) PoseFormer (c) NeuralPhysCap

We define a kinematic tree and use change of basis rotations, from the root (pelvis) to the end effectors (hands and feet), to approximate all joint angles. First, we define the origin basis for the coordinate system $A = [\mathbf{x}, \mathbf{y}, \mathbf{z}]$. Then, we find the orientation of the pelvis $B = [\mathbf{x}, \mathbf{y}, \mathbf{z}]$ from the positions of the thorax, pelvis, and right hip using the right hand rule. And finally, the root orientation for the pelvis is defined as the rotation between A and B. We repeat this for the next pair of joints (*e.g.* pelvis and chest), traversing outwards to the end effectors for the remaining joints. If toe and heel joints are not detected, the orientation of the ankle for the kinematic initialization is unknown. Instead, we initialize the ankle with a neutral pose and impose no constraints during the optimization process.

Impact of toe and heel joints

Our main results may suggest that the baseline outperforms other methods because of additional toe and heel joints. To identify the impact of these joints, we run the baseline with only 17 joints, removing toes and heels from the kinematic initialization. Results are shown in Tables 1 and 2. In Table 1, we observe comparable plausibility metrics, but with noticeable improvements for GP and PSD_{100}. These gains instead suggest that while the toes and heels provide more information about the orientation of the foot, it adds additional variance into the pose estimation. The observed physical plausibility improves when omitting the toe and heel joints. In Table 2, we note similar per class performance to the baseline counterpart, with the most increases coming from the lower-performing classes that contain significant crouching or bending over movements.

Method	FS (%)	GP ↓	CD ↓	PSD_{100} ↑
Baseline-17	1.6	**0.11**	28.9	**74.7**

Table 1: We show results on our validation subset on Human3.6M dataset. Baseline-17 removes the toe and heel keypoints. Results are comparable to Table 1 in the main text.

	Joint Name	Baseline	PoseFormer	NeuralPhysCap
Head	Head top		x	x
	Nose	x	x	
	L Eye	x		
	L Ear	x		
	R Eye	x		
	R Ear	x		
	Neck		x	x
Upperbody	Pelvis		x	
	Thorax		x	
	L Shoulder	x	x	x
	L Elbow	x	x	x
	L Wrist	x	x	x
	R Shoulder	x	x	x
	R Elbow	x	x	x
	R Wrist	x	x	x
Lowerbody	L Hip	x	x	x
	L Knee	x	x	x
	L Ankle	x	x	x
	R Hip	x	x	x
	R Knee	x	x	x
	R Ankle	x	x	x
Feet	L Big Toe	x		x
	L Small Toe	x		
	L Heel	x		
	R Big Toe	x		x
	R Small Toe	x		
	R Heel	x		

Figure 2: The skeleton pose formats and the shared joints between the different HPE methods. If the pelvis joint is not detected, it is estimated using the midpoint of the left and right hip joints. We apply a similar approach to approximate the positions of the neck and thorax.

1.3 Simulation details

The CMA-ES algorithm is executed on 40 cpus for 9-10 hours on 100 frames in each video sequence. In a sliding fashion, two adjacent time windows are optimized for 200 iterations. The result of the first window initializes the starting point of the next two subsequent time windows. We select the hyper-parameters in the cost function through manual fine-tuning to minimize the CD trajectory distance.

1.4 Qualitative Examples

In Figures 3 and 4, we show qualitative examples on two higher performing class, *Walking* and *Waiting*. We run both of these examples on our multi-view Baseline, where we note low MPJPE-G scores and no ground penetration. The consistent ground estimation, accurate multi-view estimation of the limbs, and the linear movement result in much more stable motion of the simulated body. In Figures 5 and 6, we show qualitative examples on two of the lower performing classes, *Purchases* and *WalkDog*. We run both examples on the PoseFormer architecture. While MPJPE is low, the *Purchases* sequence has higher ground penetration, suggesting an inconsistent estimation of the floor leading to instability. While

Method	Dir.	Disc.	Greet	Photo	Pose	Purch.	Wait	WalkD.	WalkT.	Walk	Avg.
Baseline-17	84.3	75.7	77.2	**98.5**	75.2	82.4	37.8	55.4	**83.3**	77.0	**74.7**

Table 2: Here we show the per-class performance for the PSD$_{100}$ (Higher is better) metric on Baseline-17. Baseline-17 removes the toe and heel keypoints. Results are comparable to Table 2 in the main text.

we have observed that this is negligible with a stationary pose, the simulated body falls forward when bending over. The *WalkDog* sequence is more stable, but struggles when the simulated body does a turnaround, possibly due to sub-optimal optimization on ankle joint angle.

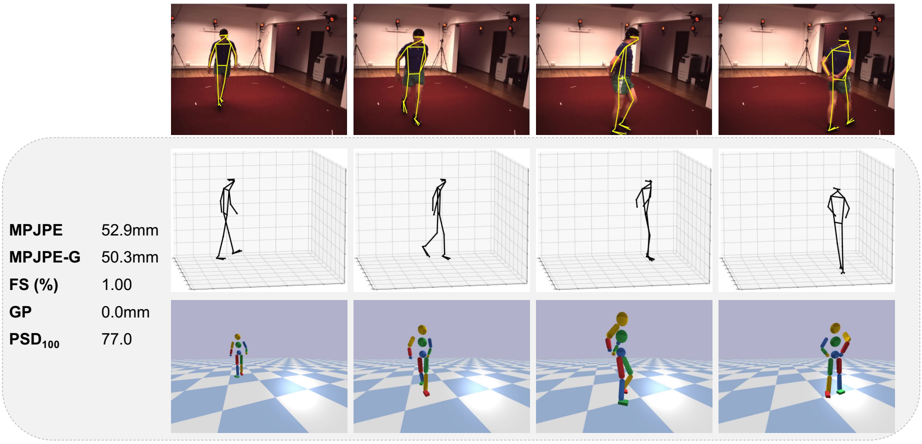
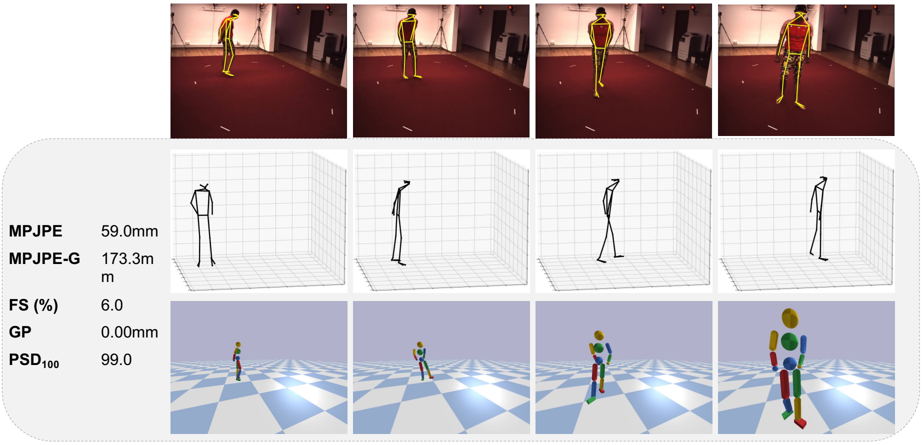
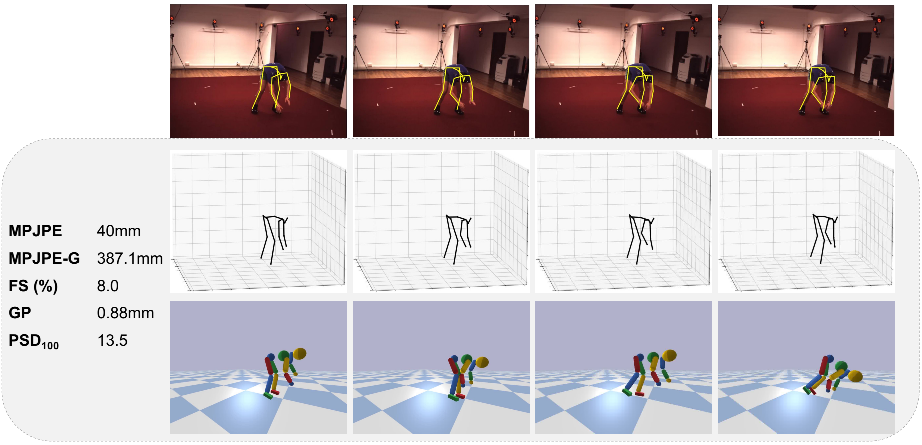


Figure 3: Results on the Baseline for the *S11* - *Walking 1* sequence.



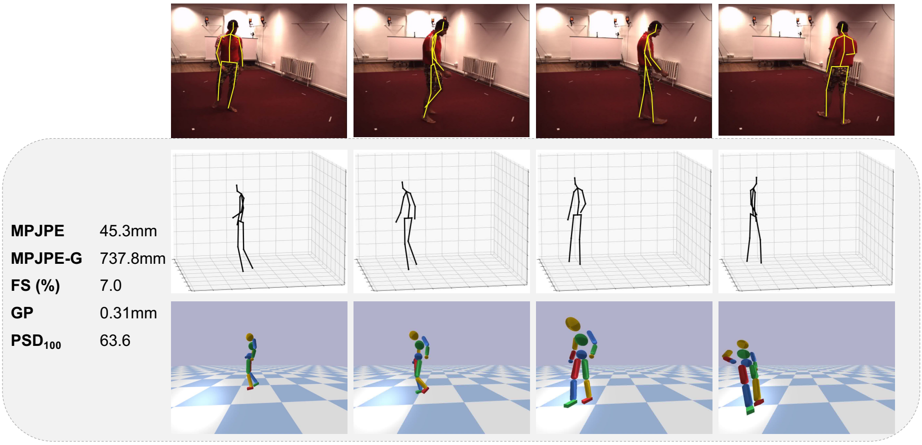
MPJPE	59.0mm
MPJPE-G	173.3mm
FS (%)	6.0
GP	0.00mm
PSD$_{100}$	99.0

Figure 4: Results on the Baseline for the *S9* - *Waiting 1* sequence.



MPJPE	40mm
MPJPE-G	387.1mm
FS (%)	8.0
GP	0.88mm
PSD$_{100}$	13.5

Figure 5: Results on PoseFormer for the *S11* - *Purchases 1* sequence.



MPJPE	45.3mm
MPJPE-G	737.8mm
FS (%)	7.0
GP	0.31mm
PSD$_{100}$	63.6

Figure 6: Results on PoseFormer for the *S9* - *WalkDog 1* sequence.

References

[1] 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.