Physics-based Human Motion Estimation and Synthesis from Videos: Supplementary

Kevin Xie[1,2], Tingwu Wang[1,2], Umar Iqbal[2]
Yunrong Guo[2], Sanja Fidler[1,2], Florian Shkurti[1]
[1]University of Toronto and Vector Institute, [2]Nvidia
kevincxie@cs.toronto.edu

1. Detailed Loss Formulation

Here we detail the losses used in our optimization method. To obtain the physics losses we first must obtain velocity and acceleration of our character. We use a finite difference scheme that corresponds to implicit integration.

$$\dot{q}_t \approx (q_{t+1} - q_t)/\Delta t$$

$$\ddot{q}_t \approx (\dot{q}_{t+1} - \dot{q}_t)/\Delta t$$

For the contact loss, we compute the contact variables $c_{t,i}$ using k_1 and k_2 parameters. Here k_1 controls the stiffness of the contact. The higher it is, the closer the soft contact loss approaches a hard step function corresponding to contact complementarity constraint. Then k_2 is simply an offset that ensures that $c_{t,i} = 0$ when $f_{t,i} = 0$. We employ 2 additional penalties that keep the contact forces physical. First we penalize contact forces from violating the friction cone constraint. We set the friction constant $\mu = 1.0$ (which is a generous overestimate representing rubber on rubber contact) and calculate the deviation from this.

$$L_{friction}(t) = w_\mu \sum_i^{n_c} \max\left(\frac{||f_i^c(t)_\parallel||^2}{||f_i^c(t)_\perp||^2} - \mu, 0\right)$$

Here $\parallel$ indicates the component of force tangential to the contact surface and $\perp$ the normal force. The second one is to prevent overly excessive contact forces that are unreasonable for natural human motion. We set this to 8 times the force required for an evenly balanced standing motion. Since there are 8 foot contact points, each contact point would hence be restricted to exert no more than the whole body weight on its own. This means each foot can generate total contact force of 4 times body weight which is similar to highly dynamic dance motions.

We base L_{pose} on losses common to many body shape estimation works [?].

$$L_{pose} = L_{prior} + L_{pose2d} + L_{pose3d}$$

L_{prior} is the per-frame SMPL prior and is the log prob. of a Gaussian Mixture Model over pose and L2 regularization of the body shape parameter β.

$$L_{prior} = w_\beta ||\beta||_2^2 + w_{GMM} \sum_t \log p_{GMM}({\theta^{joints}}_t)$$

L_{pose2d} is error in pixel space, re-projecting our motion using true camera projection matrix P and uses robust loss ρ [?].

$$L_{pose2d} = w_{2d}\rho(Pp - x^{pe,2D})$$

L_{pose3d} measures local keypoint 3d error where global root position is subtracted out to obtain relative keypoint positions p_{rel}. Here R is the camera extrinsic rotation.

$$L_{pose3d} = w_{3d}||Rp_{rel} - sp_{rel}^{pe}||^2 + w_{scale}(s-1)^2$$

Since scale of the original 3d pose estimation is inherently ambiguous, the scale parameter s is jointly optimized with the motion which accounts for this ambiguity. The actual scale of the character in our optimization will be adjusted through the β shape parameters and informed through contact geometry and motion (scale typically does not diverge too much from 1). In our case the pose estimator we use also emits a score representing the confidence in its estimation (ranging from 0 to 1). In this case, we also weigh the pose estimation losses per joint by this confidence.

We give weights for each of the losses in Table 1

2. Rigid Body Human Body Model

We construct the body model out of geometric primitives as shown in Figure 1. Mass and inertial properties are calculated assuming constant density of $1000kg/m^3$.

Sizes of primitives are heuristically set and are (differentiably) scaled in proportion to the lengths of the corresponding bones of the skeleton resulting from the SMPL body shape params β. Specifically, we scale box and sphere primitives corresponding to foot, torso and head uniformly in all 3d dimensions in proportion to the distance to the

Name	Value
$w_{dynamics}$	50
w_e	200
$w_{\dot{e}}$	50
k_1	10
w_μ	1
w_{pen}	100
w_{2d}	1e-3
w_{3d}	0.5
w_{scale}	1e-3
w_β	5e-3
w_{GMM}	2.5e-3
$w_{\ddot{p}}$	0.15
$w_{\ddot{\theta}}$	1e-4

Table 1. Table of constants used and their values.

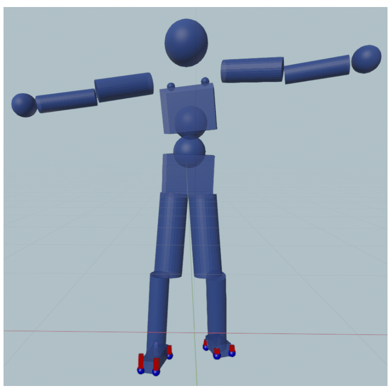

Figure 1. **Body Model.** Example geometry of our human body model.

Method	Feet	Body	Body-Align 1
[43] Physics (MTC)	508.7	499.8	421.9
[43] Physics (Our PE)	345.2	382.0	310.8
Ours (Kinematic)	251.0	190.1	**114.6**
Ours (Physics)	**82.4**	**101.1**	156.0

Table 2. Comparison with [43] on HumanEva dataset. Errors are mean over time and measured in millimeters.

next child joint, whereas we scale the cylinders representing limbs only in the length-wise direction and maintain a constant thickness. Admittedly, this does not fully capture variation in human body shapes, as it does not distinguish between characters with similar skeletons that differ in body mass.

3. Additional Comparison

We adopt the same experimental setting presented in [43]. Specifically we evaluate on the same 15 short (2 second) sequences extracted from the walking clips in the HumanEva [45] dataset. Using the camera extrinsics given in the dataset, we found that the ground had a slight z offset in the clips. We estimated a conservative 6cm offset for all clips and applied our method with this elevated ground plane. In [43], MTC is used as the pose estimator for the initial motion. For fairer comparison, we also adapt a version that uses the outputs of our pose estimator (Our PE) instead. Note that unlike our method, [43] does not optimize the shape of the body during optimization which can lead to large errors especially in the depth of the root.

We present our results using the metrics reported in [43] in Table 2. Here “Feet” measures the global position error of the 2 feet joints. This especially highlights foot floating and sliding artifacts. The “Body” metric measures whole body global position error.

The metric “Body-Align 1” measures the average error between the poses after aligning the 1st frame root position. This metric is not very robust as it is quite sensitive to the root position on the first frame. For example, if one motion started with an error in the root position on the first frame but later in the motion recovers from this error, it will be penalized for the rest of the motion as the first frame is erroneously compensated for.

Our method greatly outperforms [43] in every pose accuracy metric.

References

[1] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. Structured prediction helps 3d human motion modelling. In *ICCV*, 2019.

[2] Kevin Bergamin, Simon Clavet, Daniel Holden, and James Richard Forbes. Drecon: Data-driven responsive control of physics-based characters. *ACM Trans. Graph.*, 38(6), Nov. 2019.

[3] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science. Springer International Publishing, Oct. 2016.

[4] M. A. Brubaker, L. Sigal, and D. J. Fleet. Estimating contact dynamics. In *IEEE 12th International Conference on Computer Vision*, pages 2389–2396, 2009.

[5] Simon Clavet. Motion matching and the road to next-gen animation. *Proceedings of GDC*, 2016.

[6] E. Daneshmand, M. Khadiv, F. Grimminger, and L. Righetti. Variable horizon mpc with swing foot dynamics for bipedal walking control. *IEEE Robotics and Automation Letters*, 6(2):2349–2356, 2021.

[7] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV ’15, page 4346–4354, USA, 2015. IEEE Computer Society.

[8] P. Ghosh, J. Song, E. Aksan, and Otmar Hilliges. Learning human motion models for long-term predictions. *2017 International Conference on 3D Vision (3DV)*, pages 458–466, 2017.

[9] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision, 2020.

[10] I. Habibie, Daniel Holden, Jonathan Schwarz, Joe Yearsley, and T. Komura. A recurrent variational autoencoder for human motion synthesis. In *BMVC*, 2017.

[11] Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. Moglow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Trans. Graph.*, 39(6), Nov. 2020.

[12] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *TPAMI*, 36(7):1325–1339, 2014.

[13] Umar Iqbal, Pavlo Molchanov, Thomas Breuel, Juergen Gall, and Jan Kautz. Hand pose estimation via 2.5D latent heatmap regression. In *ECCV*, 2018.

[14] Umar Iqbal, Pavlo Molchanov, and Jan Kautz. Weakly-supervised 3d human pose learning via multi-view images in the wild. In *CVPR*, 2020.

[15] Umar Iqbal, Kevin Xie, Yunrong Guo, Jan Kautz, and Pavlo Molchanov. KAMA: 3d keypoint aware body mesh articulation. In *ArXiv*, 2021.

[16] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018.

[17] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Computer Vision and Pattern Recognition (CVPR)*, 2019.

[18] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[19] Lucas Kovar, Michael Gleicher, and Frédéric Pighin. Motion graphs. *ACM Trans. Graph.*, 21(3):473–482, July 2002.

[20] Taesoo Kwon, Yoonsang Lee, and Michiel Van De Panne. Fast and flexible multilegged locomotion using learned centroidal dynamics. *ToG*, 39(4):46–1, 2020.

[21] Yongjoon Lee, Kevin Wampler, Gilbert Bernstein, Jovan Popović, and Zoran Popović. Motion fields for interactive character locomotion. *ACM Trans. Graph.*, 29(6), Dec. 2010.

[22] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *CVPR*, 2021.

[23] Jiaman Li, Yihang Yin, H. Chu, Y. Zhou, Tingwu Wang, S. Fidler, and H. Li. Learning to generate diverse dance motions with transformer. *ArXiv*, abs/2008.08171, 2020.

[24] Zongmian Li, Jiri Sedlar, Justin Carpentier, Ivan Laptev, Nicolas Mansard, and Josef Sivic. Estimating 3d motion and forces of person-object interactions from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[25] 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 *ECCV*, 2014.

[26] Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel Van De Panne. Character controllers using motion vaes. *ACM Trans. Graph.*, 39(4), July 2020.

[27] C. Liu and Sumit Jain. A quick tutorial on multibody dynamics. 2012.

[28] Libin Liu, KangKang Yin, and Baining Guo. Improving sampling-based motion control. In *Computer Graphics Forum*, volume 34, pages 415–423. Wiley Online Library, 2015.

[29] Libin Liu, KangKang Yin, Michiel van de Panne, Tianjia Shao, and Weiwei Xu. Sampling-based contact-rich motion control. *ACM Transactions on Graphics (TOG)*, 29(4):128, 2010.

[30] Miao Liu, Dexin Yang, Yan Zhang, Zhaopeng Cui, James M. Rehg, and Siyu Tang. 4d human body capture from egocentric video via 3d scene grounding, 2020.

[31] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *SIGGRAPH Asia*, 34(6):248:1–248:16, 2015.

[32] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *ICCV*, 2019.

[33] Julieta Martinez, Michael J. Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[34] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. Xnect. *ACM Transactions on Graphics*, 39(4), Jul 2020.

[35] Igor Mordatch, Kendall Lowrey, Galen Andrew, Zoran Popovic, and Emanuel V. Todorov. Interactive control of diverse complex characters with neural networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[36] Igor Mordatch, Zoran Popović, and Emanuel Todorov. Contact-invariant optimization for hand manipulation. In *SIGGRAPH*, pages 137–144, 2012.

[37] Igor Mordatch, Emanuel Todorov, and Zoran Popović. Discovery of complex behaviors through contact-invariant optimization. *TOG*, 31(4):1–8, 2012.

[38] Igor Mordatch, Jack M Wang, Emanuel Todorov, and Vladlen Koltun. Animating human lower limbs using contact-invariant optimization. In *Siggraph Asia*, 2013.

[39] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, New York, NY, USA, second edition, 2006.

[40] Zherong Pan, Bo Ren, and Dinesh Manocha. Gpu-based contact-aware trajectory optimization using a smooth force model. In *Proceedings of the 18th Annual ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '19, New York, NY, USA, 2019. Association for Computing Machinery.

[41] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. Deepmimic: Example-guided deep reinforce-

ment learning of physics-based character skills. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018.

[42] Paul S. A. Reitsma and Nancy S. Pollard. Evaluating motion graphs for character animation. *ACM Trans. Graph.*, 26(4), Oct. 2007.

[43] Davis Rempe, Leonidas J. Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. Contact and human dynamics from monocular video. In *ECCV*, 2020. 2

[44] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. Physcap: Physically plausible monocular 3d motion capture in real time. *ToG*, 39(6), dec 2020.

[45] Leonid Sigal, Alexandru O. Balan, and Michael J. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 87(1):4–27, 2010. 2

[46] Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *ACM Trans. Graph.*, 38(6), Nov. 2019.

[47] Sebastian Starke, Yiwei Zhao, Taku Komura, and Kazi Zaman. Local motion phases for learning multi-contact character movements. *ACM Trans. Graph.*, 39(4), July 2020.

[48] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019.

[49] 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 *ECCV*, 2018.

[50] M. Vondrak, L. Sigal, and O. C. Jenkins. Physical simulation for probabilistic motion tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[51] Tingwu Wang, Yunrong Guo, Maria Shugrina, and Sanja Fidler. Unicon: Universal neural controller for physics-based character motion, 2020.

[52] Alexander W Winkler, C Dario Bellicoso, Marco Hutter, and Jonas Buchli. Gait and trajectory optimization for legged systems through phase-based end-effector parameterization. *IEEE Robotics and Automation Letters*, 3(3):1560–1567, 2018.

[53] Jungdam Won, Deepak Gopinath, and Jessica Hodgins. A scalable approach to control diverse behaviors for physically simulated characters. *ACM Transactions on Graphics (TOG)*, 39(4):33–1, 2020.

[54] Xinchen Yan, Akash Rastogi, Ruben Villegas, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, Ersin Yumer, and Honglak Lee. Mt-vae: Learning motion transformations to generate multimodal human dynamics. In *European Conference on Computer Vision*, pages 276–293. Springer, 2018.

[55] Ze Yang, Siva Manivasagam, Ming Liang, Bin Yang, Wei-Chiu Ma, and Raquel Urtasun. Recovering and simulating pedestrians in the wild, 2020.

[56] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *ECCV*, 2020.

[57] Ye Yuan and Kris Kitani. Residual force control for agile human behavior imitation and extended motion synthesis. In *NeurIPs*, 2020.

[58] Jason Y. Zhang, Panna Felsen, Angjoo Kanazawa, and Jitendra Malik. Predicting 3d human dynamics from video. In *ICCV*, 2019.

[59] Yan Zhang, Michael J. Black, and Siyu Tang. We are more than our joints: Predicting how 3d bodies move. In *CVPR*, 2021.

[60] Tiancheng Zhi, Christoph Lassner, Tony Tung, Carsten Stoll, Srinivasa G. Narasimhan, and Minh Vo. Texmesh: Reconstructing detailed human texture and geometry from rgb-d video. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 492–509. Springer International Publishing, 2020.

[61] Yi Zhou, Zimo Li, Shuangjiu Xiao, Chong He, Zeng Huang, and Hao Li. Auto-conditioned recurrent networks for extended complex human motion synthesis. In *International Conference on Learning Representations*, 2018.