PIMNet: Physics-infused Neural Network for Human Motion Prediction

Zhibo Zhang[1*], Yanjun Zhu[2*], Rahul Rai[3], and David Doermann[2]

***Abstract*—Human motion prediction (HMP) predicts future human pose sequences given the past ones. HMP has recently attracted attention in computer vision and the robotics domain as it helps machines understand human behavior, plan target actions and optimize interaction strategies. Existing methods for HMP are based on either purely physics-based models or statistical models. However, each of these methods has its shortcomings. The physics-based techniques are complex and error-prone, while the statistical methods require a large amount of data and lack physical consistency. To overcome their limitations, we propose a physics-infused neural network (PIMNet), which combines both physics-based and statistical methods. We first computed the contact forces and joint torques for each pose using the physics-based human dynamical model. Then they are fed into an Encoder-Decoder machine learning architecture to predict future ones. In this way, PIMNet simultaneously obtains computational efficiency and physical consistency. Extensive experimental results on Human 3.6M show that the proposed PIMNet could accurately predict human motion in both short-term and long-term scope. It achieves better or comparable prediction accuracy than the state-of-the-art, even using a basic LSTM as the machine learning model.**

***Index Terms*—Human Detection and Tracking; Deep Learning Methods; AI-Based Methods**

I. Introduction

HUMANS exhibit a tremendous ability to make accurate short-term predictions based on current environmental information [1]. Human motion prediction (HMP) is one of the most critical aspects of the ability. It is generally acknowledged that predicting human motion is a challenging task due to the complexity of the human musculoskeletal system. The future movement of humans has various possibilities. However, based on specific prior information, the range of outcomes can be restricted to a manageable degree of variation [2]. Research on human motion prediction helps machines understand human behavior, plan target actions, and optimize interaction strategies. HMP techniques are widely used in many applications in computer vision, medical, and robotics, such as human tracking [3], [4], human behavior monitoring [5], and human-robot interaction [6]–[8]. There are two kinds of approach to modeling human motion: physics-based models underpinned by laws of physics and statistical models underpinned by data.

Traditional physics-based modeling of human motion hinges on the existing knowledge of human behavior derived from dynamical information and biomechanical information. Physics plays a vital role in characterizing, describing and predicting human motion [9]. A meticulously designed physics model could provide a reliable prediction. However, it is not easy to accurately model human motion due to the complexity of human biomechanical models and the agnostic environment. Moreover, high-fidelity physics-based human motion models suffer from extremely high computational complexity, which fails to employ those models in real-time applications.

Statistical motion models are often expressed as mathematical equations or functions that describe human motion using a limited number of parameters and their associated probability distributions [10]. The statistical motion model is favorable for human motion representation as it does not require a complex hand-crafted modeling effort. As long as motion data are available, any human motion model can theoretically be fitted [10]. Especially with the widespread success of deep learning, an increasing number of statistical techniques have proven promising performance in human motion modeling. However, the black-box behavior nature of the statistical motion model can also be a double-edged sword, as it can be undesirably physics agnostic [11], [12]. Failure to strictly obey physics principles raises reliability concerns about their use in modeling human motion. Another primary concern is that deep learning models are notoriously data-hungry. The quantity and quality of the data directly impact the model's performance. However, accurate motion capture data is costly to obtain. For example, there are only 210 sequences in the H3.6M dataset, one of the most significant benchmark datasets for HMP. Kinetics, a public benchmark for video action recognition, contains 650,000 video clips. In addition, statistical models are often poor in generalizing beyond their initial training set. Furthermore, it is unrealistic and costly to have a dataset covering all human movements. Therefore, the pure statistical model may not be the best solution for HMP under current conditions.

To overcome the limitations mentioned above for both approaches, we propose a hybrid model, termed **PIMNet**, which combines physics-based and statistical methods in a unified way. Instead of developing the high-fidelity physics model from scratch, we propose using a simplified low-fidelity physics-based motion model. We then combine a low-fidelity physics-based motion model with an off-the-shelf

Manuscript received: February, 23, 2022; Revised May, 10, 2022; Accepted June, 12, 2022.

This paper was recommended for publication by Editor Gentiane Venture upon evaluation of the Associate Editor and Reviewers comments. This work was supported by Defense Advanced Research Projects Agency (DARPA)

[1]Manufacturing and Design Lab (MADLab), Department of Mechanical and Aerospace Engineering, University at Buffalo, Buffalo, NY zzhang38@buffalo.edu

[2]Artificial Intelligence Innovation Lab (A2IL), Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY

[3]Corresponding author. Geometric Reasoning and Artificial Intelligence Lab (GRAIL), Clemson University, Greenville, SC, USA rrai@clemson.edu

* Contributed equally

Digital Object Identifier (DOI): see top of this page.

deep neural (statistical) model such as LSTM to arrive at a hybrid HMP model. When predicting motion, the proposed hybrid model takes into account historical motion data and considers force information (from the physics model) that is the root cause of motion. Thus, PIMNet as a physics-infused machine learning model obtains computational efficiency and physical consistency. The contribution of this work lies in three aspects: (1) we propose a low-fidelity human motion model to extract physical motion features; (2) we propose a novel hybrid mechanism for human motion prediction that takes into account both physics-based model and statistical model; (3) extensive experiments on public benchmarks demonstrate the effectiveness of the proposed hybrid model over the pure machine learning model and the traditional physics-based model.

The remainder of this paper is organized as follows. Section II introduces related work. Section III outlines the architecture of the proposed hybrid model, including a physics-based dynamic model and deep learning prediction model. Section IV describes the experimental setups and the metrics used to evaluate our model's performance. Training and testing results are also discussed in Section IV. Section V ends with concluding remarks.

II. Literature Review

A. Physics-based human motion model

All movements originate from the action of both internal and external forces. Thus, Newton's laws of motion provide a clear relationship between the applied force and the resultant change in movement. Many researchers have explored how to model human motion using physics-based models. Wren and Pentland proposed a dynamic model coupled with human behavior to track the precise movement of the upper body [13]. Metaxas and Terzopoulos combined a 3D dynamic model with continuous non-linear Kalman filtering to estimate human motion [14]. Kakadiaris and Metaxas presented a force-based method for tracking human movements, which mitigates the difficulties arising from occlusion among parts of the body [15]. In Bissacco and Soatto's work, a switching, linear dynamical system is utilized to model the motion and contact forces [16]. Additionally, the space-time constraint method is applied to minimize the energy of motion, which results in modeling human movement [17], [18]. In addition, control strategies for physics simulation are also used to model human motion [19], [20]. In summary, all of these models fall into three categories: modeling the kinematics and dynamics of humans, modeling an optimization problem, and modeling based on a control strategy. However, the physics-based approaches mentioned above cannot predict human motion with both accuracy and computational efficiency at the same time because of the limitations of modeling the complex environment and the muscle skeleton system of humans.

B. Statistical based human motion model

Statistical models are favorable because they are compact and only work on the collected motion data. The development of statistical models of human motion has gone through two stages: conventional statistical models and deep learning models. In the traditional statistical model, HMM [21], sequential Monte Carlo [22], particle filter [23], Gaussian process [10], and restricted Boltzmann machine [24], [25] are used to model and predict human movements. However, due to the performance limitations of these conventional approaches, the prediction results cannot meet our desirable requirements.

With the rapid emergence of deep learning in recent years, many researchers have been modeling human motion using deep learning-based methods. These methods fall into three categories: RNN-based model, GNN-based (graph neural network), and GAN-based (generative adversarial network). Because of the success of sequence-to-sequence prediction, RNN-based models could be desirable candidates to model human motion. Fragkiadaki et al. proposed two RNN-based models, that is, ERD (Encoder-Recurrent-Decoder) and LSTM-3LR (LSTM with three layers), which involve a nonlinear encoder-decoder before and after recurrent layers [2]. In [26], the authors combine the spatio-temporal graphs with RNN, enabling a promising long-term human motion prediction. Based on previous work, Martinez et al. improved the SRNN model with three changes that boosted the prediction performance [27]. Since the skeleton tree can be regarded as a graph, it is reasonable to involve GNN in the human motion prediction. In [28], the authors exploited a graph neural network (GNN) to capture temporal and spatial information about human joints using the graph of temporally connected kinematic trees. Li et al. combined two types of links (actional links and structural links) into a generalized skeleton graph and proposed an actional-structural GNN to make motion prediction [29]. Furthermore, in his subsequent work, a multiscale GNN is developed to leverage human skeleton structure awareness resulting in better prediction [30]. GANs have resulted in extraordinary achievements in image generation. In human motion prediction, GANs can learn a probability density function of future human movement conditioned on previous states. Barsoum et al. proposed HP-GAN based on Wasserstein GAN [31]. Gui et al. incorporated local geometric structure constraints in GANs to make the model geometrically meaningful [32].

However, all approaches mentioned above share the same property: they do not consider the dynamics which cause the motion. Therefore, the results of the prediction are often physically implausible. Furthermore, similar to other data-driven models, the quality of the data directly affects the prediction results.

C. Hybrid Model

Several attempts have been made to integrate physics-based models with statistic models [10]. However, these approaches are based on conventional statistical methods that cannot handle complex problems like human motion prediction. In recent years, researchers have started to apply the hybrid idea in multiple domains using machine learning [33] [34]. Several works also show decent performance in the human-robotics domain [35] [36] [37]. By accounting for physical information and statistical priors simultaneously, our proposed model instills physical realism into deep learning-based motion models and

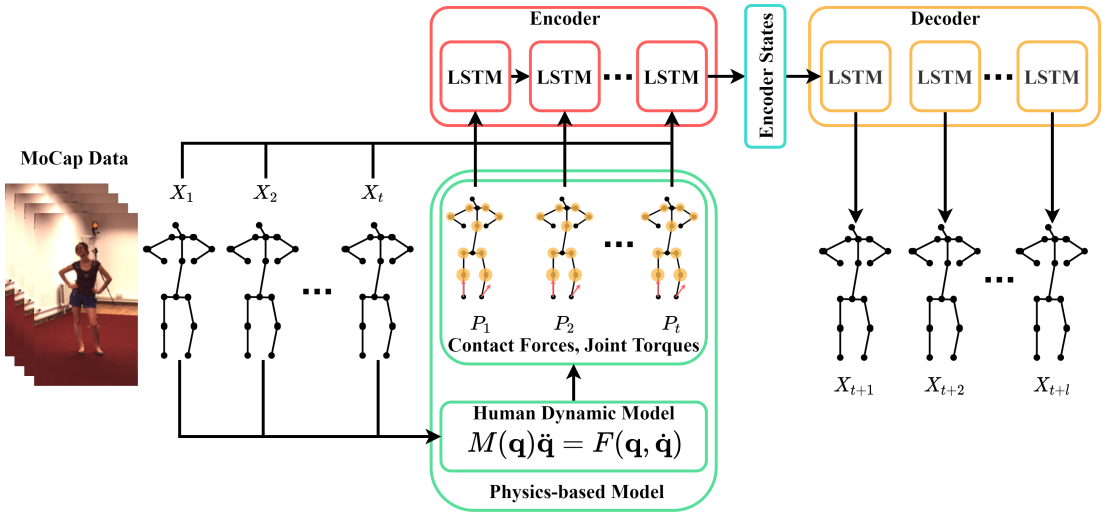

Fig. 1. PIMNet architecture

boosts the physics-based model's performance with lower computational cost.

III. Methodology

A. Physics Dynamic Model

Our primary goal is to predict human motion based on MoCap (Motion Capture) data. In addition to the statistical prediction network, a feasible physics-based model of humans is required. One of the natural approaches is to model full-body dynamics, which has been explored in many domains, such as robotic, biomechanical, and computer animation. However, a full-body high-fidelity dynamic model for humans is highly complex. Thus, we developed and utilized a low-fidelity model in our hybrid modeling framework, simplifying several aspects of the human body and forces. Next, we discuss the entire modeling procedure in detail.

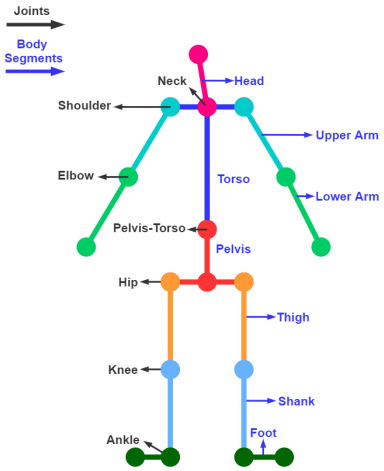

Fig. 2. Human skeleton model

TABLE I
Joints in the Human Body Model

Joint	Neck	Shoulder	Elbow	Hip	Knee	Ankle	Pelvis-Torso
DOF	3	3	1	3	1	2	3
Joint Type	Ball	Ball	Hinge	Ball	Hinge	Saddle	Ball
# of Joints	1	2	2	2	2	2	1

As shown in Figure 2, our human body model contains 13 segments that are regarded as rigid bodies. They are connected to their parent segments by 1-DOF (degree-of-freedom), 2-DOF, or 3-DOF joints (as shown in Table I). The pelvis is considered the root segment "linked" to the world, coordinate by a 6-DOF "joint" that contains the global position and orientation information. Our model uses the Quaternion to represent relative joint angles because of its advantages in dealing with rotations. Then the generalized coordinates, $\mathbf{q} \in \mathbb{R}^N$ (N is the total DOFs), consist of the root position, orientation (6-DOF), and all joint angles. Thus, a Newtonian dynamics equation can be used to represent the full-body movement:

$$M(\mathbf{q})\ddot{\mathbf{q}} = F(\mathbf{q}, \dot{\mathbf{q}}) + C(\mathbf{q}, \dot{\mathbf{q}}) \tag{1}$$

where M is the generalized mass matrix, F is a vector of generalized forces acting on the human body, C denotes all other terms to enforce joint constraints, $\dot{\mathbf{q}}$ and $\ddot{\mathbf{q}}$ represent the vector of joint angle velocities and joint angle accelerations. The desired goal of the aforementioned dynamic model is to extract the joint forces (torques) and contact forces when given joint states $\mathbf{q}$, $\dot{\mathbf{q}}$, and $\ddot{\mathbf{q}}$. The joint states can be directly obtained from MoCap data. Next, we explain the generalized force term.

1) Generalized Forces: The generalized force term can be divided into two parts: internal torques F_e from the joint muscles and external forces F_i from the environment, which can be expressed as:

$$F(\mathbf{q}, \dot{\mathbf{q}}) = F_i(\mathbf{q}, \dot{\mathbf{q}}) + F_e(\mathbf{q}, \dot{\mathbf{q}}) \tag{2}$$

Then we can further decompose the external forces into those due to gravity f_g, contact with the environment f_c, and unexplained components f_r:

$$F_e(\mathbf{q}, \dot{\mathbf{q}}) = J_f(\mathbf{q})(\mathbf{f}_g + \mathbf{f}_c + \mathbf{f}_r) \tag{3}$$

in which $\mathbf{f}_g$ and $\mathbf{f}_c$ are vectors represented in the body frame, comprising three forces and three torques for each of the body segments. And $\mathbf{f}_r$, which contains 6 DOFs, is the vector of residue forces acting on the pelvis segment to explain the residue effects of the environment, such as wind drag. J_f is the Jacobian matrix that maps the forces on body segments to generalized forces. It is easy to model the gravity forces $\mathbf{f}_g$ by applying gravity to the CoM (center of mass) of each segment.

We used a continuous contact model, which has been applied in the work of Liu et al. [38] and follows the method proposed by Brubaker et al. [39]. We assume that the contact forces come from the interaction between the body segments (e.g., foot, hand) and the fixed surface. Our contact model is based on a normal-direction damper-spring model and a tangential-direction friction damper model modulated by two sigmoidal functions. The contact force $\mathbf{f}_c$ is expressed as:

$$\mathbf{f}_c = S(-60d)S(5f_n)(f_n\mathbf{n}_s + \mathbf{f}_t) \tag{4}$$

where $S(x) = 0.5(1 + \tanh(x))$ is the sigmoid function, d is the shortest distance from the contact point to the surface, f_n is the magnitude of the normal force coming from the linear spring, $\mathbf{n}_s$ is the unit normal of the surface (in this research, $\mathbf{n}_s = [0, 0, 1]$ is a constant vector), $\mathbf{f}_t$ is the tangential force of the frictional damper. The magnitudes of the normal force and the tangential force are expressed as follows.

$$f_n = -k(d - 1) - c_n\mathbf{v}^T\mathbf{n}_s \tag{5}$$

$$\mathbf{f}_t = -c_t(\mathbf{v}^T - \mathbf{n}_s^T\mathbf{v}\mathbf{n}_s) \tag{6}$$

in which, k is the stiffness of the spring, c_n, c_t denote the normal and tangential damping coefficient, and $\mathbf{v}$ is the velocity vector of the contact point. To simplify the expression, $\theta = [k, c_n, c_t]$ is used to denote all contact parameters. Until now, we have all terms except $\mathbf{f}_r$.

The most straightforward way to estimate the joint torques and contact parameters is to minimize the difference between the MoCap data and the motion generated by the simulation. Nevertheless, this approach is computationally complex due to noise and the existence of local minima. Therefore, we consider employing a residue force $\mathbf{f}_r$ to explain the difference between the dynamic model and the MoCap data. This residue force does not exist in reality. However, with the help of the residue force, the physics model can be greatly simplified. The residual force is only used to model the noise not accounted for by the contact model.

In sum, by substituting Eq. (2), (3), (4), (5), and (6) to (1), we derive the human full-body, dynamic model:

$$M(\mathbf{q})\ddot{\mathbf{q}} = F_i(\mathbf{q}, \dot{\mathbf{q}}) + J_f(\mathbf{q})(\mathbf{f}_g + \mathbf{f}_c(\mathbf{q}, \dot{\mathbf{q}}, \theta) + \mathbf{f}_r) + C(\mathbf{q}, \dot{\mathbf{q}}) \tag{7}$$

2) Parameter optimization: An optimization problem is then formulated to estimate the contact forces and the internal joint torque. The objective of the optimization problem is to search for the surface parameters θ that minimize the magnitude of the residue forces. By reorganizing the Equation (7), and expanding the internal torque $F_i(\mathbf{q}, \dot{\mathbf{q}}) = J_\tau(\mathbf{q})\tau$, we get:

$$[J_\tau, J_f]\begin{bmatrix}\tau \\ \mathbf{f}_r\end{bmatrix} = M(\mathbf{q})\ddot{\mathbf{q}} - J_f(\mathbf{q})(\mathbf{f}_g + \mathbf{f}_c(\mathbf{q}, \dot{\mathbf{q}}, \theta)) + C(\mathbf{q}, \dot{\mathbf{q}}) \tag{8}$$

where J_τ is the Jacobian matrix that maps the joint torques into the vector of generalized forces, and τ is the vector of the internal torques actuated by muscles. $\mathbf{f}_r$ and τ, are function of surface parameters θ. The overall optimization problem can then be formulated as follows.

$$\begin{aligned} \min_{\theta} \quad & \sum_{i=1}^{6} ||\mathbf{f}_r^{(i)}(\theta(k, c_n, c_t))||^2 \\ \text{s.t.} \quad & 1 \le k \le 20 \\ & 0.1 \le c_n, c_t \le 20 \end{aligned} \tag{9}$$

The formulated optimization problem is a single-objective constrained optimization problem. Therefore, a gradient-based optimizer can be used to find the minimal value of $\mathbf{f}_r$ and the corresponding surface parameters θ. Then, the internal torques τ can be calculated by substituting θ into Eq. (8).

To sum up, there are three steps in the human dynamic model to estimate the internal joint torques and contact forces: (1) Calculate the angle velocity $\dot{\mathbf{q}}$ and angle acceleration $\ddot{\mathbf{q}}$ by using forward differences; (2) Build the human body model by defining the biomechanical information; (3) Estimate the internal joint torques and contact forces by minimizing residue forces.

B. PIMNet

The previous subsection outlined the details of the simplified human full-body dynamic model that extracts pertinent physical information (contact forces and joint torques). Our primary goal is to predict human motion based on a hybrid model. Our hybrid model is a deep neural network-based model that integrates MoCap data and physical information (from a low-fidelity physics model) to enable HMP. In this section, we introduce the notations and problem formulation and outline the details of the proposed hybrid model.

1) Notation and problem formulation: Given T prior time series human MoCap data in terms of the Quaternion, that is, $X_i = [X^1, X^2, ..., X^T] \in \mathbb{R}^{T\times N\times Q}$, where N is the number of joints and Q = 4 denotes the dimension of the Quaternion, we want to predict L future poses $\hat{X}_o = [\hat{X}^{t+1}, \hat{X}^{t+2}, ..., \hat{X}^{T+L}] \in \mathbb{R}^{L\times N\times Q}$. Additionally, based on the physics-based model $X_p^{(t)} = \mathcal{P}(X^{(t)})$ described in the previous section, we get the physics inputs $X_p = [X_p^1, X_p^2, ..., X_p^T] \in \mathbb{R}^{L\times M\times S}$, where M is the total number of joints plus the number of contact forces, $S = 3$ is the dimension of the forces (torques). Our objective is to design a model f that narrows the difference between the predicted results $\hat{X}_o$ and the ground truth $X_o = [X^{t+1}, X^{t+2}, ..., X^{T+L}] \in \mathbb{R}^{L\times N\times Q}$, which is formulated as follows.

$$\hat{X}_o = f(X_i, X_p) \tag{10}$$

2) PIMNet Architecture: As shown in Figure 1, PIMNet mainly consists of two components: a physics-infused encoder and a temporal decoder.

The encoder is essentially RNN, which can encode the input sequence as feature representations in machine translation [40]. For human motion prediction, the physics-infused encoder can be used to learn a mapping from X_i^t, X_p^t to h^t at the time step t, which can be written as:

$$h^t = f_e(h_{t-1}, X_i^t, X_p^t) \tag{11}$$

where $h^t \in \mathbb{R}^H$ is the embedded hidden state, H is the size of the hidden state, and f_e is a non-linear activation function that could be an LSTM or GRU (gated recurrent unit). We use basic LSTM as f_e to explore long-term dependencies in this work. The use of LSTM brings us the merits of avoiding the problem of vanishing gradients and capturing long-term dependencies of time series.

To predict the future motion series, we applied another LSTM network to decode the hidden states h. On the contrary, the hidden state vector is regarded as the input of the LSTM network. And the decoder network can be expressed as:

$$\hat{X}^t = f_d(h^t) \tag{12}$$

in which f_d is an LSTM-based decoding function. At this point, our proposed model f is capable of generating the future motion based on prior MoCap data and physical information, which can be denoted as:

$$\hat{X}_o = f(X_i, X_p) = f_d(f_e(X_i, X_p; W_e); W_d) \tag{13}$$

where W_e and W_d are the trainable weights within the encoder and decoder.

To train our prediction model, we consider the l_2 loss. Then, for N training samples, the cost function is:

$$\begin{aligned} \mathcal{L} &= \frac{1}{N}\sum_{n=1}^{N} ||(X_o)_n - (\hat{X}_o)_n||^2 \\ &= \frac{1}{N}\sum_{n=1}^{N} ||(X_o)_n - f_n(X_i, X_p; W_e, W_d)||^2 \end{aligned} \tag{14}$$

Using backpropagation, we find optimal values of W_e and W_d that minimize the cost function $\mathcal{L}$. In our experiment, Adam is utilized as the optimizer.

IV. Experiments and Results

A. Dataset and preprocessing

We evaluate our model on the Human3.6M (H3.6M) dataset [41]. H3.6M is the largest MoCap dataset for 3D human motion analysis. It contains 3.6 million different 3D articulated poses captured from 11 different professional actors across 17 different scenarios (activities). A Vicon motion capture system supports H3.6M and provides accurate 3D human joint locations at the global coordinate. Our experiments follow the same data preprocessing step presented in [2], [26], [27]. In the original dataset, 32 joints are recorded; however, some are overlapping or constant. We eliminate redundant joint information. The motion sampling rate is reduced from 50 to 25 fps. Furthermore, we adopt the joint angles as the exponential map [26]. Before training, all features are normalized to $[-1, 1]$.

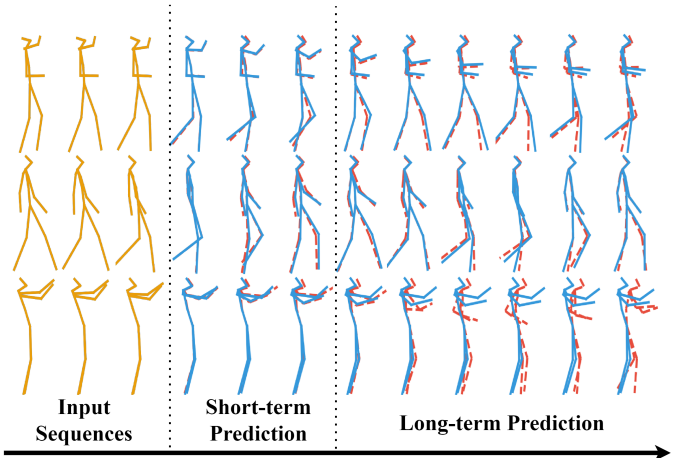

Fig. 3. Example results. From top to bottom, we show the results of three activities" Eating," "Walking," and "Discussion." The yellow line is historical input data, the red dash line is ground truth, and the blue line is predicted results.

B. Training procedure

The following equations are used to convert the exponential map representation into the Quaternion representation that is used in the physics-based model to derive the joint torques and contact forces:

$$\mathbf{e} = [e_1, e_2, e_3] \tag{15}$$

$$\theta = \sqrt{e_1^2 + e_2^2 + e_3^2} \tag{16}$$

$$\mathbf{q} = [q_0, q_1, q_2, q_3] = [\cos\frac{\theta}{2}, \frac{e_1 \sin\frac{\theta}{2}}{\theta}, \frac{e_2 \sin\frac{\theta}{2}}{\theta}, \frac{e_3 \sin\frac{\theta}{2}}{\theta}] \tag{17}$$

where, $\mathbf{e} = [e_1, e_2, e_3]$ is the exponential map, $\mathbf{q} = [q_0, q_1, q_2, q_3]$ is the corresponding Quaternion, θ is the rotation angle.

Our test is carried out on subject 5 (S5) in H3.6M, and the rest of the data serves as the training set. The Euclidean distance between the predicted and ground truth joint angles is used as an error metric to compare the performance of different methods. We compared our approach with several recent works: ERD [2], LSTM-3LR [2], SRNN [26], Res-GRU [27], Traj-GCN [42], and DMGNN [30].

In our model, 1024 LSTM cells are used in the physics-infused encoder, and 1024 cells are used in the temporal decoder. The learning rate of the Adam optimizer is 0.001, with a 0.9 decay rate for every 10,000 steps. The batch size is set to 16.

C. Computational resource

The physics-based dynamic model is developed in MATLAB. The training and testing procedures are executed using Tensorflow on a server with Linux Centos 7.5.x operating system, Intel Xeon Gold 6230 processor (40 cores @2.10GHz), 32GB RAM, and NVIDIA Tesla V100 GPU.

TABLE II
MEAN ANGLE ERRORS OF DIFFERENT METHODS FOR SHORT-TERM PREDICTION ON DIFFERENT ACTIONS OF H3.6M.

Method	Walking				Eating				Smoking				Discussion			
milliseconds	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
ERD [2]	0.93	1.18	1.59	1.78	1.27	1.45	1.66	1.80	1.66	1.95	2.35	2.42	2.27	2.47	2.68	2.76
LSTM-3LR [2]	0.77	1.00	1.29	1.47	0.89	1.09	1.35	1.46	1.34	1.65	2.04	2.16	1.88	2.12	2.25	2.23
SRNN [26]	0.81	0.94	1.16	1.30	0.97	1.14	1.35	1.46	1.45	1.68	1.94	2.08	1.22	1.49	4.83	1.93
Res-sup. [27]	0.28	0.49	0.72	0.81	0.23	0.39	0.62	0.76	0.33	0.39	0.62	0.76	0.31	0.68	1.01	1.09
Traj-GCN [42]	**0.18**	0.32	**0.49**	0.56	**0.16**	0.29	0.50	0.62	0.22	0.41	0.84	0.79	**0.20**	0.51	0.79	0.86
DMGNN [30]	**0.18**	**0.31**	**0.49**	0.58	0.17	**0.30**	0.49	0.59	**0.21**	**0.39**	0.81	0.77	0.26	0.65	0.92	0.99
PIMNet	0.22	0.35	0.50	**0.55**	0.20	0.31	**0.49**	**0.59**	**0.21**	**0.39**	**0.66**	**0.70**	0.23	**0.50**	**0.71**	**0.79**

Method	Direction				Greeting				Phoning				Posing				Purchases			
millisecond	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
Res-sup. [27]	0.26	0.47	0.72	0.84	0.75	1.17	1.74	1.83	**0.23**	**0.43**	0.69	0.82	0.36	0.71	1.22	1.48	0.51	0.97	1.07	1.16
Traj-GCN [42]	0.26	0.45	0.71	0.79	0.36	**0.60**	0.95	1.13	0.53	1.02	1.35	1.48	0.19	0.44	1.01	1.24	0.43	0.65	1.05	1.13
DMGNN [30]	0.25	0.44	0.65	0.71	0.36	0.61	0.94	1.12	0.52	0.97	1.29	1.43	0.2	0.46	1.06	1.34	0.41	0.61	1.05	1.14
PIMNet	**0.23**	**0.40**	**0.64**	**0.68**	**0.33**	**0.60**	**0.89**	**0.96**	0.37	**0.43**	**0.66**	**0.79**	**0.22**	**0.43**	**0.96**	**1.11**	**0.39**	**0.60**	**1.00**	**1.06**

Method	Taking Photo				Waiting				Walking Dog				Walking Together				Average			
millisecond	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
Res-sup. [27]	0.24	0.51	0.9	1.05	0.28	0.53	1.02	1.14	0.56	0.91	1.26	1.4	0.31	0.58	0.87	0.91	0.36	0.65	0.99	1.11
Traj-GCN [42]	**0.14**	0.34	0.58	0.7	0.23	0.50	0.91	1.14	0.46	0.79	1.12	1.29	**0.15**	0.34	0.52	0.57	0.27	0.51	0.83	0.95
DMGNN [30]	0.15	0.34	0.58	0.71	0.22	**0.49**	0.88	1.10	**0.42**	**0.72**	1.16	1.34	**0.15**	**0.33**	**0.50**	0.57	0.27	0.51	0.83	0.95
PIMNet	**0.14**	**0.32**	**0.55**	**0.68**	**0.20**	0.50	**0.85**	**1.07**	0.44	**0.72**	**1.10**	**1.23**	**0.15**	0.34	**0.50**	**0.53**	**0.26**	**0.45**	**0.73**	**0.83**

D. Results and Analysis

Qualitative Analysis: First, we present qualitative evaluations of our results, as shown in Fig. 3. In Fig. 3, the yellow line represents historical input sequences, the red line represents ground truth, and the blue line represents predicted results from PIMNet. Comparing the ground truth with our predicted sequences shows that our model provided accurate predictions for "Eating" and "Walking" in both the short and long term. However, our model fails to predict the movements in the long-term prediction of "Discussion." This failure occurs because the actor only moves minimally in the input sequence. Therefore, the physics-based model can only provide limited prediction information. Predicting transition scenarios from static to dynamic is a challenging task. This could be a task for future work.

Results on H3.6M: We use all activities from H3.6M to evaluate the performance of our model, except for "Sitting" and "Sitting Down." Within these two activities, actors continue sitting either on a chair or on the ground, which is not aligned with the assumption of the physics-based model that feet are the only contact points. Figure 4 shows the mean angle error for all scenarios in the overall prediction of human motion. As shown in Figure 4, our model outperforms the comparison methods. Our model only has minor improvements in the short-term prediction compared to the latest approaches (Traj-GCN and DMGNN). However, in long-term prediction, our model always performs the best. Additionally, the gap between our PIMNet and other models increases with increasing prediction time. The improved performance of PIMNet can be attributed to the infusion of physical information from the physics-based model, making our hybrid model retain physics-plausible features.

We also present the short-term prediction results for each scenario in Table II. As the table shows, our model cannot beat the comparison models in short-term prediction (2-4 frames). However, our model shows better results in the relatively more extended forecast. Our model has an impressive improvement, especially in long-term prediction, as shown in Table III.

TABLE III
MEAN ANGLE ERROR OF DIFFERENT METHODS FOR LONG-TERM PREDICTION ON FOUR ACTIONS

Motion	Walking		Eating		Smoking		Discussion	
milliseconds	560	1k	560	1k	560	1k	560	1k
ERD [2]	2.00	2.38	2.36	2.41	3.68	3.82	3.47	2.92
LSTM-3LR [2]	1.81	2.2	2.49	2.82	3.24	3.42	2.48	2.93
SRNN [26]	1.90	2.13	2.28	2.58	3.21	3.23	2.39	2.43
Res-sup. [27]	0.93	1.03	0.95	1.08	1.25	1.5	1.43	1.69
Traj-GCN [42]	0.65	0.67	0.76	1.12	0.87	1.57	1.33	1.70
DMGNN [30]	0.66	0.75	0.74	1.14	0.83	1.52	1.33	1.45
PIMNet	**0.60**	**0.66**	**0.68**	**0.96**	**0.76**	**1.33**	**0.90**	**1.35**

Furthermore, our proposed model delivers better results in the case of nonperiodic movements. The deep learning-based models predict the results based on the prior data. The stronger the periodicity of the data, the better the prediction performance of the deep learning-based model. However, informed by the physics-based model, our proposed hybrid model reduces dependence on prior training data and shows more robust performance without periodicity in the data.

Abalation Analysis: To evaluate the contribution of the physics-based model. We carried out another numerical experiment that keeps the encoder-decoder prediction model but removes the physics-based model. The results of the ablation experiment are shown as the solid black line (ED) in Figure 4. Ablation analysis shows the efficacy of incorporating physical information from physics-based models. Figure 4 highlights that the proposed model without the infusion of physical information leads to inferior results (the accuracy is close to the Res-sup model). Performance comparable to the Res-sup model is understandable since our deep learning-based prediction model is similar to most encoder-decoder models.

V. CONCLUDING REMARKS

This paper introduces a hybrid model called PIMNet to model human motion. The proposed model combines a physics-based model and a machine learning-based model. With the help of the simplified human full-body dynamic model, our

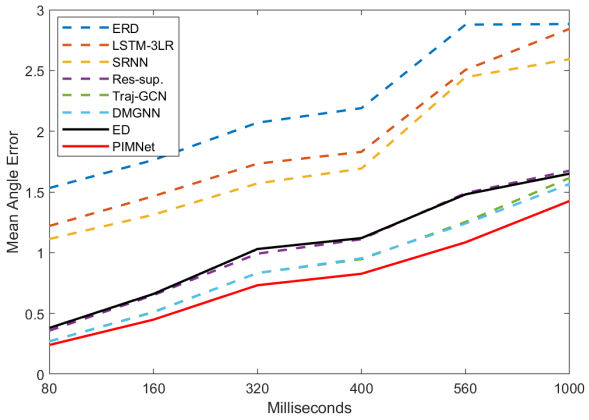

Fig. 4. Performance comparison between different methods in terms of mean angle error for all scenarios

LSTM-based machine learning model can accurately predict human motion in the short-term and long-term. By comparing the performance of our proposed model with several state-of-the-art approaches, we conclude that our physics-infused hybrid model could be beneficial for HMP tasks.

There are several directions for future exploration: (1) Develop a better physics model which involves kinetic information to infuse more comprehensive physical details. (2) Use a sequential machine learning model with more robust predictive capabilities. (3) Propose a wiser fusion method to integrate the physics-based and machine learning models.

ACKNOWLEDGEMENT

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Agreement No. HR00111890037. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA).

REFERENCES

[1] D. Vernon, C. Von Hofsten, and L. Fadiga, *A roadmap for cognitive development in humanoid robots*. Springer Science & Business Media, 2011, vol. 11.

[2] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, "Recurrent network models for human dynamics," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4346–4354.

[3] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 961–971.

[4] H. Gong, J. Sim, M. Likhachev, and J. Shi, "Multi-hypothesis motion planning for visual object tracking," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 619–626.

[5] J. A. Ambrósio and A. Kecskeméthy, "Multibody dynamics of biomechanical models for human motion via optimization," in *Multibody Dynamics*. Springer, 2007, pp. 245–272.

[6] D.-A. Huang and K. M. Kitani, "Action-reaction: Forecasting the dynamics of human interaction," in *European Conference on Computer Vision*. Springer, 2014, pp. 489–504.

[7] L.-Y. Gui, K. Zhang, Y.-X. Wang, X. Liang, J. M. Moura, and M. Veloso, "Teaching robots to predict human motion," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 562–567.

[8] H. S. Koppula and A. Saxena, "Anticipating human activities using object affordances for reactive robotic response," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 14–29, 2015.

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

[10] X. Wei, J. Min, and J. Chai, "Physically valid statistical models for human motion generation," *ACM Transactions on Graphics (TOG)*, vol. 30, no. 3, pp. 1–10, 2011.

[11] Z. Zhang, R. Rai, S. Chowdhury, and D. Doermann, "Midphynet: Memorized infusion of decomposed physics in neural networks to model dynamic systems," *Neurocomputing*, vol. 428, pp. 116–129, 2021.

[12] A. Sharma, Z. Zhang, and R. Rai, "The interpretive model of manufacturing: a theoretical framework and research agenda for machine learning in manufacturing," *International Journal of Production Research*, vol. 59, no. 16, pp. 4960–4994, 2021. [Online]. Available: https://doi.org/10.1080/00207543.2021.1930234

[13] C. R. Wren and A. P. Pentland, "Dynamic models of human motion," in *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 1998, pp. 22–27.

[14] D. Metaxas and D. Terzopoulos, "Shape and nonrigid motion estimation through physics-based synthesis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 6, pp. 580–591, 1993.

[15] I. A. Kakadiaris and D. Metaxas, "Model-based estimation of 3d human motion with occlusion based on active multi-viewpoint selection," in *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 1996, pp. 81–87.

[16] A. Bissacco and S. Soatto, "Hybrid dynamical models of human motion for the recognition of human gaits," *International journal of computer vision*, vol. 85, no. 1, pp. 101–114, 2009.

[17] A. Witkin and M. Kass, "Spacetime constraints," *ACM Siggraph Computer Graphics*, vol. 22, no. 4, pp. 159–168, 1988.

[18] A. Safonova, J. K. Hodgins, and N. S. Pollard, "Synthesizing physically realistic human motion in low-dimensional, behavior-specific spaces," *ACM Transactions on Graphics (ToG)*, vol. 23, no. 3, pp. 514–521, 2004.

[19] J. K. Hodgins, W. L. Wooten, D. C. Brogan, and J. F. O'Brien, "Animating human athletics," in *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, 1995, pp. 71–78.

[20] K. Yin, K. Loken, and M. Van de Panne, "Simbicon: Simple biped locomotion control," *ACM Transactions on Graphics (TOG)*, vol. 26, no. 3, pp. 105–es, 2007.

[21] D. Gehrig, H. Kuehne, A. Woerner, and T. Schultz, "Hmm-based human motion recognition with optical flow data," in *2009 9th IEEE-RAS International Conference on Humanoid Robots*. IEEE, 2009, pp. 425–430.

[22] M. A. Brubaker, D. J. Fleet, and A. Hertzmann, "Physics-based person tracking using the anthropomorphic walker," *International journal of computer vision*, vol. 87, no. 1-2, p. 140, 2010.

[23] J. Cui, Y. Liu, Y. Xu, H. Zhao, and H. Zha, "Tracking generic human motion via fusion of low-and high-dimensional approaches," *IEEE transactions on systems, man, and cybernetics: systems*, vol. 43, no. 4, pp. 996–1002, 2013.

[24] I. Sutskever, G. E. Hinton, and G. W. Taylor, "The recurrent temporal restricted boltzmann machine," in *Advances in neural information processing systems*, 2009, pp. 1601–1608.

[25] G. W. Taylor and G. E. Hinton, "Factored conditional restricted boltzmann machines for modeling motion style," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 1025–1032.

[26] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-rnn: Deep learning on spatio-temporal graphs," in *Proceedings of the ieee conference on computer vision and pattern recognition*, 2016, pp. 5308–5317.

[27] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2891–2900.

[28] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[29] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3595–3603.

[30] M. Li, S. Chen, Y. Zhao, Y. Zhang, Y. Wang, and Q. Tian, "Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 214–223.

[31] E. Barsoum, J. Kender, and Z. Liu, "Hp-gan: Probabilistic 3d human motion prediction via gan," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 1418–1427.

[32] L.-Y. Gui, Y.-X. Wang, X. Liang, and J. M. Moura, "Adversarial geometry-aware human motion prediction," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 786–803.

[33] R. Rai and C. K. Sahu, "Driven by data or derived through physics? a review of hybrid physics guided machine learning techniques with cyber-physical system (cps) focus," *IEEE Access*, vol. 8, pp. 71 050–71 073, 2020.

[34] C. Meng, S. Seo, D. Cao, S. Griesemer, and Y. Liu, "When physics meets machine learning: A survey of physics-informed machine learning," *arXiv preprint arXiv:2203.16797*, 2022.

[35] K. Xie, T. Wang, U. Iqbal, Y. Guo, S. Fidler, and F. Shkurti, "Physics-based human motion estimation and synthesis from videos," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 532–11 541.

[36] S. Shimada, V. Golyanik, W. Xu, and C. Theobalt, "Physcap: Physically plausible monocular 3d motion capture in real time," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, pp. 1–16, 2020.

[37] X. Du, R. Vasudevan, and M. Johnson-Roberson, "Bio-lstm: A biomechanically inspired recurrent neural network for 3-d pedestrian pose and gait prediction," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1501–1508, 2019.

[38] C. K. Liu, A. Hertzmann, and Z. Popović, "Learning physics-based motion style with nonlinear inverse optimization," *ACM Transactions on Graphics (TOG)*, vol. 24, no. 3, pp. 1071–1081, 2005.

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

[40] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *EMNLP*, 2014.

[41] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1325–1339, jul 2014.

[42] W. Mao, M. Liu, M. Salzmann, and H. Li, "Learning trajectory dependencies for human motion prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9489–9497.