Rethinking Human Motion Prediction with Symplectic Integral

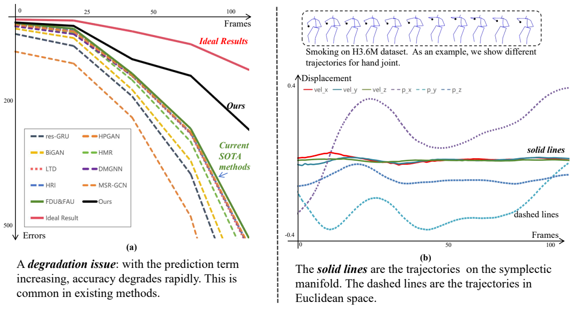
Haipeng Chen[1], Kedi Lyu[1]✉, Zhenguang Liu[2]✉, Yifang Yin[3]✉, Xun Yang [4], Yingda Lyu [5]
[1]College of Computer Science and Technology, Jilin University, Changchun, China
[2]Zhejiang University, Hangzhou, China
[3]Institute for Infocomm Research (I^{2}R), A*STAR, Singapore
[4]University of Science and Technology of China, Hefei, China
[5]Public Computer Education and Research Center, Jilin University, Changchun, China
lvkd19@mails.jlu.edu.cn, {chenhp, ydlv}@jlu.edu.cn, liuzhenguang2008@gmail.com,
yin_yifang@i2r.a-star.edu.sg, xyang21@ustc.edu.cn

Abstract

Long-term and accurate forecasting is the long-standing pursuit of the human motion prediction task. Existing methods typically suffer from dramatic degradation in prediction accuracy with increasing prediction horizon. It comes down to two reasons: 1) Insufficient ***numerical stability*** *caused by unforeseen high noise and complex feature relationships in the data, and 2) Inadequate* ***modeling stability*** *caused by unreasonable step sizes and undesirable parameter updates in the prediction. In this paper, we design a novel and symplectic integral-inspired framework named symplectic integral neural network (SINN), which engages symplectic trajectories to optimize the pose representation and employs a stable symplectic operator to alternately model the dynamic context. Specifically, we design a Symplectic Representation Encoder that performs on enhanced human pose representation to obtain trajectories on the symplectic manifold, ensuring numerical stability based on Hamiltonian mechanics and symplectic spatial splitting algorithm. We further present the Symplectic Temporal Aggregation module, which splits the long-term prediction into multiple accurate short-term predictions generated by a symplectic operator to secure modeling stability. Moreover, our approach is model-agnostic and can be efficiently integrated with different physical dynamics models. The experimental results demonstrate that our method achieves the new state-of-the-art, outperforming existing methods by 20.1% on Human3.6M, 16.7% on CUM Mocap, and 10.2% on 3DPW.*

1. Introduction

Anticipating future human motions based on historical observations presents a crucial and formidable issue in the realm of computer vision, particularly in scenarios such as *collision avoidance* or *handshake interactions*. Owing to its paramount significance, machine learning-driven human motion prediction has witnessed a surge in research efforts, yielding a diverse range of practical applications encompassing human-computer interaction, motion synthesis, and autonomous driving [4, 24, 37, 44, 46, 51]



A ***degradation issue***: with the prediction term increasing, accuracy degrades rapidly. This is common in existing methods.

The ***solid lines*** are the trajectories on the symplectic manifold. The dashed lines are the trajectories in Euclidean space.

Figure 1. Illustration of human motion prediction challenges. (a) Existing methods suffer from significant long-term performance degradation. (b) Conventional motion representation is not numerically stable.

Drawing on the advancements in neural networks, early methods (e.g., [3, 7, 8, 16, 34, 43]) have employed recurrent neural networks (RNNs) [36] such as Long Short-Term Memory (LSTM) [14] and Gated Recurrent Unit (GRU) [5] to realise human motion prediction as a vanilla temporal prediction issue. Despite the impressive performance, the extension of such methods to complex pose sequences still remains challenging due to the intricate kinematic and anatomical nature of human motions. With all this in mind, researchers have endeavored to incorporate *prior knowledge* (kinematics and anatomy) to strengthen existing solutions, such as utilizing graph Graph convolutional networks (GCN) [22, 23, 41, 47, 50] to capture spatial depen-

✉ Corresponding author.

dencies among the joints and human poses. Some works [1, 6, 17, 26] attempt to exploit generative adversarial networks (GANs) [11] to predict future poses by adversarial *training strategy*. [1] presents a common practice to achieve diverse possible predictions. [26] utilizes the GANs to simulate path integrals and predict future motion profiles.

Upon scrutinizing and experimenting on the released implementations of previous methods [16, 17, 22, 29, 31, 41], we have empirically observed that incorporating motion priors and pertinent training strategies, whether implicitly or explicitly, enhances the performance of motion prediction. Moreover, as shown in Figure 1 (a), existing methods suffer from dramatic degradation with increasing prediction horizon. We attribute this issue is mainly due to the following reasons. i) *insufficient* ***numerical stability***: human motion exhibits a stochastic nature, which is reflected in the data through high noise levels and complex distributional characteristics (see Figure 1 (b)). The larger the prediction field, the higher the likelihood of capturing excessive noise. The high dimension of pose sequences makes it challenging to accurately estimate features and relationships in the data, and this challenge increases with larger prediction horizons. ii) *inadequate* ***modeling stability***: different network architectures have distinct performance bottlenecks when it comes to handling temporal sequence problems. Pursuing long-term temporal forecasting without proper layout can lead to misinterpretations within the network, giving rise to parametric models with fluctuations that complicate the training process and network design.

To tackle the above challenges, we propose a symplectic integral neural network (SINN) as illustrated in Figure 2. It consists of two major components, namely a *symplectic representation encoder* (SRE) and a *Symplectic Temporal Aggregation* module (STA). These components mitigate the degradation in prediction performance over time together based on the symplectic integral.

The objective of SRE is to construct *numerical stability* by converting motion from traditional 3D Euclidean space into a Hamiltonian mechanical system, which is composed of kinetic and potential energy (*i.e.*, location, and velocity). We explicitly model the physical dynamics of joints with 3D velocity, which is utilized as the control actions to produce the joint trajectories. Moreover, we observe that it is beneficial to transform the original high-dimensional pose sequence into low-dimensional per-joint and per-direction sliced trajectories to be processed by the neural network parallelly (see Figure 2). This is because that learning the interactions between joints from the high-dimensional motion data itself poses a significant challenge on conventional methods. We thus utilize the symplectic spatial splitting algorithm to ameliorate this issue. Specifically, the dimension of our model's output is reduced to $T \times 2$ trajectories on the symplectic manifold, where T is the number of frames to be predicted, each composed of a velocity magnitude and a direction angle *w.r.t.* the hyperplanes in a 3D coordinate system. To summarize, our proposed SRE generates *numerically stable* motion representations with the following advantages. First, leveraging velocity to model the dynamics of human motions can filter out unwanted noise and irregular fluctuations. Second, reducing the data dimension not only prevents entering scenarios of network overfitting but also promotes learning more robust and significant features and dependencies. Finally, our predicted low-dimensional trajectories, which consist of the control actions only, can be integrated into human pose efficiently and effectively through a user-defined dynamics model without additional overhead for network learning.

The mission of STA is to guarantee *modeling stability* by a stable symplectic operator based on the symplectic temporal splitting algorithm. Specifically, a model is first trained to perform relatively short-term motion prediction, replicas of which are next temporally concatenated and fine-tuned for long-term prediction. The model and its replicas are shared weights, so our proposed method is parameter-efficient without increasing the original model size. Based on these, to ensure the stability of the symplectic structure, we employ pre-trained networks as the *symplectic operators* to predict the outcome in different sub-steps (*i.e.*, short-term steps) with a stable parameter structure. Additionally, we design a GAN-based network, which utilizes adversarial training to distinguish between the generated and real motion trajectories. This adversarial loss is, for the first time, applied to human motion prediction to enhance long-term prediction accuracy. Notably, this pre-trained symplectic operator can effectively control the perceptual field so that the network will not be affected by the perceptual field due to changes in the prediction term. The symplectic integral-based deep neural network is model-agnostic, which is also able to improve the long-term prediction capability of other existing motion prediction models.

Contributions. To summarize, our key contributions are as follows: 1) We propose a novel framework that is model-agnostic based on symplectic integral for human motion prediction tasks. 2) We present a symplectic representation encoder and a symplectic temporal aggregation module to separately enhance pose representation and model dynamic context stably and effectually for long-term and accurate motion prediction. 3) Our method achieves state-of-the-art results in predicting human motion on three challenging benchmark datasets, Human3.6M, CMU Mocap, and 3DWP. To the best of our knowledge, we are the first to combine symplectic integral with deep learning and explicitly model the dynamic controls of joints for human motion prediction. To facilitate future research, our source code is released at https://github.com/adamlyu789/SINN.

2. Related Work

Articulated Poses Representation. Researchers recognize the importance of investigating different schemes for parameterized human poses. These pose representation schemes can be broadly classified into two categories, namely *physical representations* and *mathematical representations*. For the physical schemes [12, 26, 38, 42, 50], human poses are represented as hierarchical body parts reflecting the body structures and the kinematic relations. Further, joint velocity and acceleration are engaged to model motion dynamics. In MPT [27], the human pose is presented as the joint trajectory. For the mathematical schemes [2, 25, 30, 35, 45], human motion sequences are mapped to different mathematical spaces and abstracted into different distributions, with the goal of facilitating the learning process. QuaterNet [35] represents rotations with quaternions. HMR [25] proposes to explicitly encode anatomical constraints by modeling their skeletons with a Lie algebra representation. These approaches based on modified pose representation have achieved laudable results. Therefore, we attempt to propose a more valid representation that can alleviate the inherent problem of human poses and achieve efficient long-term prediction performance.

Motion Prediction. Early approaches utilized nonlinear Markov models [19], Gaussian Process dynamic models [40], and Restricted Boltzmann Machines [20] to tackle this problem. With the advancement of deep learning, numerous methods based on deep neural networks have emerged and demonstrated remarkable outcomes. RNNs [36] have been playing important roles in modeling and predicting human motion due to their outstanding performance in dealing with temporal issues. A first class of works relies on the combination of different types of RNNs structures [8, 10, 34, 36, 48] to achieve predictions. Primitively, LSTM units are adopted in prediction, yielding classical networks such as LSTM-3LR [34] and ERD [8]. Further, DAE [10] combines ERD with a dropout auto-encoder to model temporal and spatial structures. To overcome accumulated errors and discontinuity across frames, res-GRU [34] is presented, which also incorporates the velocities of joints in motion representation. Another class of works proposes improvised RNN architectures. For instance, Liu *et al.* [25, 28] design Hierarchical recurrent networks (HMR) to model global and local motion contexts for long-term prediction. Chopin *et al.* [6] and Zhao *et al.* [49] proposed a method for human posture prediction based on adversarial generative networks (GANs) [11], which involves a discriminator and a generator. GANs [17, 26] facilitate the learning of complex distributions through this adversarial training strategy. Thereby, we attempt to utilize GANs to construct our network to alleviate the complex distribution among human motion data.

3. Our Approach

Problem formulation. Given an observed pose sequence $P = [p_1, p_2, \dots, p_T]$, we aim to predict its future pose sequence $\hat{P} = [p_{T+1}, p_{T+2}, \dots, p_{T+N}]$, where T and N represent the number of observed and future frames, respectively. Formally, we seek to learn a parameterized function $\mathcal{F}(\cdot)$ that extrapolates from the observed pose sequence to predict the future $\hat{P}_{T+1:T+N} = \mathcal{F}(P_{1:T})$.

To the best of our knowledge, the majority of existing techniques depict human motions using the 3D coordinates of body joints. However, the dynamics and the interactions between joints are not explicitly captured by this representation, but rather expected to be learned by the model from the complex, high-dimensional motion data. Inspired by symplectic integral, we propose to disentangle a human motion sequence into distinct trajectories, each corresponding to the tangent plane of an individual joint on the symplectic manifold (hereinafter referred to as ”*symplectic trajectories*”). Formally, let ζ_i^t denote the symplectic trajectories of the i-th joint at frame t. ζ comprises coordinates where α, β correspond to the generalized notions of location and momentum, respectively. Thereby, the Hamiltonian $H(\alpha, \beta, t)$, which defines the energy function of this dynamical system, can be decomposed as:

$$H = K + V \tag{1}$$

where $K(\alpha)$ is the kinetic energy, a function of the generalized momentum coordinates, and $V(\beta)$ is the potential energy, a function of the generalized position coordinates. Thus, we model the Hamiltonian of each joint through a learnable network function of the generalized momentum and positions, which *inherently takes into account* the dynamics and constraints between joints to some extent.

Moreover, it is important to acknowledge that accurately learning the interactions between joints from the high-dimensional motion data itself poses a significant challenge. Incorrect joint relations can be possibly learned due to data limitation or model overfitting. Here we reduce the dimension of motion data by a factor of J (*i.e.*, the number of joints) by modeling joint trajectories individually. We empirically observed that this dimension reduction effectively reduces the risk of model overfitting where more robust correlations can be learned from the observed data. Compared to the performance gain, the information loss is negligible.

Method overview. The overall architecture of our proposed framework is illustrated in Figure 2. Our framework consists of two key components: a Symplectic Representation Encoder (SRE) (Sec.3.1) and a Symplectic Temporal Aggregation module (STA) (Sec.3.2). Specifically, we *first* utilize SRE to generate a novel low-dimensional symplectic trajectory ζ to ensure numerical stability. *Next*, these trajectories are learned by the symplectic operator in STA,

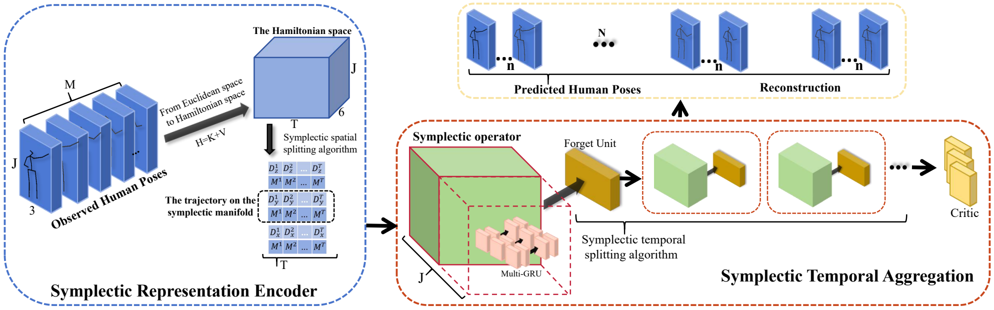

The Hamiltonian space
M
J
T
6
3
From Euclidean space to Hamiltonian space
H=K+V
Observed Human Poses
Symplectic spatial splitting algorithm
D_z^1 D_z^2 ... D_z^T
M^1 M^2 ... M^T
The trajectory on the symplectic manifold
D_y^1 D_y^2 ... D_y^T
M^1 M^2 ... M^T
D_x^1 D_x^2 ... D_x^T
M^1 M^2 ... M^T
T
Symplectic Representation Encoder
N
n
n
n
Predicted Human Poses
Reconstruction
Symplectic operator
Forget Unit
Critic
Symplectic temporal splitting algorithm
Multi-GRU
J
Symplectic Temporal Aggregation

Figure 2. The proposed Symplectic Integral Neural Network (SINN) architecture.

which is implemented with a robust symplectic structure that is resistant to perturbations. The symplectic operator is used only for accurate short-term prediction. *Finally*, we leverage integrator in STA to achieve reliable long-term predictions through the aggregation of multiple accurate short-term forecasts. The following sections elucidate our proposed SRE and STA in detail.

3.1. Symplectic Representation Encoder

As noted previously, conventional motion representation suffers from two drawbacks: (i) Dependence solely on raw, high-dimensional position information struggles to capture the physical dynamics and characteristic relationships of human motions; (ii) The presence of complex human structures and movements introduces high noise levels and intricate data distributions. These factors compromise the robustness of conventional representations of human motions, imposing a significant challenge on the network's learning capacity. Motivated by these observations and insights, we present a novel symplectic representation encoder (SRE), which addresses the above challenges through *Hamiltonian representation* and *symplectic spatial splitting algorithm*. The objective of SRE is to establish numerical stability in the human motion data.

Hamiltonian representation. The typical input to human motion prediction models is an $T \times J \times 3$ human pose sequence P consisting of T frames, J joints, and 3D coordinates of each joint. Here we propose to represent each joint in the Hamiltonian mechanical system, denoted as $\zeta =< \alpha, \beta >$ where ζ is composed of the location α (*i.e.*, the 3D joint location of the last observed pose) and the momentum β (*i.e.*, the frame-wise joint displacement viewed as joint instantaneous velocity). A naive representation of the 3D joint velocity can be a three-dimensional tuple consisting of the magnitude projected onto the x, y, and z axes. However, we empirically observe that this naive representation tends to suffer from information loss. To solve this issue, we present a new 3D joint velocity representation by introducing a little redundancy. Specifically, the representation is a four-dimensional tuple, which consists of one velocity magnitude $\mathcal{M}$ and three direction angles $\mathcal{D}_{x,y,z}$ with respect to the coordinate hyperplanes xy-plane, yz-plane, and xz-plane. As shown in Figure 1, the displacements in the Hamiltonian system exhibit stability and similarity. As a result, our representation can be considered as displacements in a multidimensional space, which not only improves the numerical stability of data but also minimizes the loss of raw information. In this way, the momentum β can be expressed as:

$$\beta = (\mathcal{M}, \mathcal{D}_x, \mathcal{D}_y, \mathcal{D}_z) \tag{2}$$

Symplectic spatial splitting algorithm. Based on the findings on the negative effect of high dimension as shown in Figure 1, we design a symplectic spatial splitting algorithm for human motion prediction. To be specific, our proposed representation captures the physical dynamics of each joint along each axis. Then, we propose to predict the symplectic trajectories $\zeta_{x,y,z}^{1:T}$ along the three directions (x, y, z) parallelly. The process is given by:

$$\zeta^{1:T} \xrightarrow{Reduce} [\zeta_x^{1:T}, \zeta_y^{1:T}, \zeta_z^{1:T}] \tag{3}$$

As the location α is the initial quantity, the future sequence of which can be reconstructed from the current location and the predicted future velocities, the output dimension of our model is further reduced to $T \times 2$. Taking the x-axis direction as an example,

$$\zeta_x^{1:T} = (\mathcal{M}^{1:T}, \mathcal{D}_x^{1:T}) \tag{4}$$

Pose integration via velocity controls. Let $\hat{\zeta}_i^{1:T} = (\hat{\mathcal{M}}_i^{1:T}, \hat{\mathcal{D}}_i^{1:T})$ denote the predicted velocity along direction i where $i \in x, y, z$. We first reconstruct the 3D velocity $\hat{\zeta}^{1:T} = (\hat{\mathcal{M}}^{1:T}, \hat{\mathcal{D}}_x^{1:T}, \hat{\mathcal{D}}_y^{1:T}, \hat{\mathcal{D}}_z^{1:T})$ by computing $\hat{\mathcal{M}}^{1:T}$ as

$$\hat{\mathcal{M}}^{1:T} = (\hat{\mathcal{M}}_x^{1:T} + \hat{\mathcal{M}}_y^{1:T} + \hat{\mathcal{M}}_z^{1:T})/3 \tag{5}$$

Next, starting from an initial joint location α in the last observed pose, we reconstruct the joint future locations based on $\hat{\zeta}^{1:T}$, and further recover the pose $P^{1:T}$ consisting of J joints by stacking back individual joint trajectories. This whole reconstruction process is denoted as

$P^{1:T} = Rec(\hat{\zeta}_x^{1:T}, \hat{\zeta}_y^{1:T}, \hat{\zeta}_z^{1:T}, \alpha)$ where we revert the predicted 3D velocity to the raw position without information loss. Next, we will introduce how to accurately predict the velocities $\hat{\zeta}^{1:T}$ based on our proposed model.

3.2. Symplectic Temporal Aggregation

While the SRE reduces the high dimension of human poses and high noise in data, it is still highly challenging to achieve accurate long-term prediction results. Therefore, we need to pursue not only numerical stability but also stable and effective parametric models to maintain long-term and accurate predictions. Generally, Euler's integrator starts from the initial state z_0 at time t_0 and estimates the function $z(t)$ at uniformly spaced time points $t_n = t_0 + n\Delta t$ with the recursive expression

$$z_{n+1} = z_n + \Delta t f(z_n, t_n) \tag{6}$$

However, using Euler's method could easily lead to unstable solutions unless the time-step is chosen to be very small [18]. Here we can consider f_θ to be the parametric equation of our network. z_0 is the last observed joint location, while $z_1, \ldots, z_n$ form the future joint trajectory, which can be generated based on our choice of the integrator $Rec(\cdot)$ as introduced in the previous section. Subsequently, a more general form is given as,

$$\{z_i\}_{i=1}^{T} = Integrator(z_0, f_\theta, \{t_i\}_{i=1}^{T}) \tag{7}$$

Thus, our goal here is to optimize our model f_θ based on the dynamics control (*i.e.*, velocity) of the joint trajectory, which is equivalent to minimize the MSE loss between the ground-truth trajectory $\{\zeta_i\}_{i=0}^{T}$ and the predicted trajectory $\{\hat{\zeta}_i\}_{i=0}^{T}$ in conventional methods. Inspired by the symplectic integral, we designed a symplectic temporal aggregation (STA) module as shown in Figure. 2. It consists of two components: a *symplectic temporal splitting* algorithm and a *Temporal aggregation and finetune* module.

Symplectic temporal splitting algorithm. The main idea of the symplectic integral is that numerical integrations are first performed in each subinterval, which are next summed to determine an approximation of the whole integral. Thus, we employ a symplectic temporal splitting algorithm (STS) for the purpose of long-term prediction by splitting the long-term prediction into multiple accurate short-term predictions. The process is given by:

$$\mathcal{STS}(\{t_i\}_{i=1}^{T}) = \bigcup_{n=0}^{T/t-1} \{t_i\}_{i=n*t+1}^{(n+1)*t} \tag{8}$$

where $n = 0, 1, 2, \ldots, T/t - 1$, t is short-term duration. The exact length of t is determined by the implementation of the symplectic operator. As shown in Eq. 7, the aim of **symplectic operator** is to obtain a stable parametric submodel $\{f_{so}^i, t_i\}_{i=n*t+1}^{(n+1)*t}$. The symplectic operator takes the observed symplectic trajectories ζ as input and predicts the future trajectories $\hat{\zeta} = \mathcal{SO}(\zeta)$. *Importantly*, to maintain the stable symplectic nature, we design the symplectic operator as a pre-trained network to ensure stability in computations. Since this approach is model-agnostic, we tentatively design a succinct model to verify this structure. This pre-trained network uses multiple GRUs to perform prediction, as illustrated in Figure 2. This pre-trained symplectic operator can effectively control the perceptual field of the network, so that the network will not be affected by the perceptual field due to changes in the prediction term.

Temporal aggregation and finetune. Our model has been trained to perform relatively short-term motion predictions. Now we concatenate and fine-tune its replica in a temporal fashion to perform long-term predictions.

$$\{f_\theta, t_i\}_{i=1}^{T} = [\{f_{so}^i, t_i\}_{i=1}^{t}, \ldots, \{f_{so}^i, t_i\}_{i=n*t+1}^{(n+1)*t}] \tag{9}$$

When information is transmitted in a long-term temporal sequence, we design a *forget unit* which can keep inputs real-time by dropping earlier n frames in the input sequence and supplementing the newly generated n frames to the sequence. Considering the complex distribution of motion data, we therefore employ a GAN architecture, featuring a symplectic operator as a generator and a three-layer fully connected network as a critic to finetune the long-term prediction by adversarial training. Notably, GAN does not increase the parameters of the model by adding the critic.

3.3. Training Objectives

Our training process takes place in symplectic space, so the objective is to minimize the error between predicted $\hat{\zeta}$ and target ζ. There are mainly three loss functions utilized in our method: *generator loss* L_g, *critic loss* L_c, and *symplectic operator loss*. For short-term prediction, to supervise the output of the symplectic operator (t frames), we employ symplectic operator loss L_{so}.

$$L_{so} = \sum_{i=1}^{t} ||\zeta^i - \hat{\zeta}^i||_2 \tag{10}$$

We first pre-train the symplectic operator based on L_{so}, next we optimize the end-to-end SINN (T frames) based on L_g + L_c. The generator loss L_g is given by

$$L_g = \frac{1}{T} \sum_{i=1}^{T} |\zeta^i - \hat{\zeta}^i|^2 \tag{11}$$

The critic loss L_c is defined as:

$$L_c = \mathbf{ED}(\zeta^{1:T}) - \mathbf{ED}(\hat{\zeta}^{1:T}) \tag{12}$$

where $ED(\cdot)$ is the mean critic score.

Time (ms)	80	160	320	400	1,000	80	160	320	400	1,000	80	160	320	400	1,000	80	160	320	400	1,000
	Directions					Greeting					Phoning					Posing				
res-GRU [34]	36.4	56.6	80.3	98.1	126.3	36.8	73.3	138.2	155.6	189.5	24.3	42.3	72.6	82.3	124.2	26.7	52.4	129.5	159.4	181.7
HPGAN [1]	80.9	101.3	148.6	168.8	234.6	81.5	118.8	178.4	200.1	258.6	78.8	100.3	152.7	179.0	244.2	75.5	107.4	168.3	178.0	250.1
BiGAN [17]	22.0	37.5	58.9	72.0	114.7	24.6	45.8	89.9	103.0	148.1	17.0	29.7	54.1	62.1	112.0	16.8	35.0	86.4	105.6	187.0
HMR [25]	23.3	25.0	47.2	61.5	116.9	12.9	31.9	55.6	82.5	123.2	12.5	21.3	39.3	58.6	112.8	13.6	23.5	62.5	114.1	143.6
LTD [32]	9.2	20.6	46.9	58.8	105.8	16.7	33.9	67.5	81.6	140.2	10.2	20.2	40.9	50.9	105.1	12.5	27.5	62.5	79.6	171.7
DMGNN [21]	12.3	23.8	46.2	**55.5**	90.3	14.0	29.8	74.0	89.1	140.2	10.2	14.0	32.8	40.0	104.1	9.2	23.5	65.0	82.8	170.2
HRI [33]	7.4	18.4	44.5	56.5	106.5	13.7	30.1	63.8	78.1	138.8	8.6	18.3	39.0	49.2	105.0	10.2	24.4	58.5	75.8	178.2
MSR-GCN [22]	8.6	19.7	43.3	53.8	-	16.5	37.0	77.3	93.4	-	10.1	20.7	41.5	51.3	-	12.8	29.4	67.0	85.0	-
FDU&FAU [9]	6.6	16.4	39.6	50.1	97.2	13.0	30.7	63.1	78.2	141.8	7.8	17.2	37.5	47.3	96.7	**7.5**	19.3	47.1	62.0	149.5
Ours	**5.4**	**10.3**	**21.6**	**27.3**	**62.0**	**12.7**	**23.2**	**45.3**	**55.5**	**106.7**	**6.2**	**11.7**	**30.4**	**37.2**	**82.0**	8.7	**16.0**	**32.7**	**41.2**	**94.6**
	Waiting					Eating					Smoking					Discussion				
res-GRU [34]	20.5	39.8	78.2	90.3	120.1	17.5	34.3	71.1	87.5	117.6	22.4	39.9	80.2	92.5	119.2	25.8	43.4	83.5	95.8	129.1
HPGAN [1]	70.1	89.6	98.2	121.0	145.2	64.1	78.4	99.9	113.7	136.2	67.2	88.6	100.1	123.9	140.4	71.4	91.3	105.2	129.7	150.4
BiGAN [17]	17.5	31.3	53.9	61.4	128.5	13.6	26.1	51.4	63.1	84.1	11.0	21.0	33.1	38.2	88.1	19.2	39.0	67.7	75.3	122.5
HMR [25]	17.2	31.4	53.5	61.1	99.0	13.2	26.0	51.1	62.6	74.0	10.3	20.5	33.0	37.2	69.1	19.0	38.8	67.3	75.0	121.5
LTD [32]	10.5	21.6	45.9	57.1	106.9	7.7	15.8	30.5	37.6	74.1	8.4	16.8	32.5	39.5	73.6	12.2	25.8	53.9	66.7	118.6
DMGNN [21]	12.2	24.1	60.0	77.5	128.0	11.0	21.4	36.1	43.9	57.0	9.0	17.6	25.1	40.3	-	17.3	34.8	61.0	70.0	-
HRI [33]	8.7	19.2	43.4	54.9	108.2	8.7	18.7	39.5	47.1	57.0	7.0	14.9	29.9	36.4	69.5	10.2	23.4	52.1	65.4	119.8
MSR-GCN [22]	10.7	23.1	48.3	59.2	-	8.4	17.1	33.0	40.0	-	8.0	16.3	31.3	38.2	-	12.0	26.8	57.1	70.0	-
FDU&FAU [9]	8.2	18.4	41.3	52.1	101.2	6.3	13.7	29.1	36.3	71.1	**5.1**	9.1	21.3	29.9	59.3	7.4	17.1	42.9	50.4	149.5
Ours	**7.2**	**13.8**	**28.7**	**36.2**	**80.1**	**5.7**	**10.1**	**19.2**	**23.7**	**52.2**	5.3	**9.0**	**19.3**	**24.4**	**58.0**	**7.3**	**14.2**	**30.0**	**37.4**	**79.9**
	Purchases					Sitting					Walking					Takingphoto				
res-GRU [34]	38.5	70.1	101.0	102.3	131.2	34.1	53.2	110.4	115.0	150.1	29.5	60.4	118.1	138.5	165.3	23.1	47.0	92.3	110.1	149.2
HPGAN [1]	42.4	88.9	95.0	120.2	170.2	36.3	60.0	120.0	123.1	168.2	65.2	98.1	148.3	168.2	199.9	38.0	49.3	79.9	83.8	160.4
BiGAN [17]	29.0	54.1	82.2	92.4	139.0	19.9	41.0	76.3	88.2	120.5	17.8	36.4	74.4	90.2	188.9	14.2	27.1	53.5	66.1	128.0
HMR [25]	15.3	30.6	64.7	73.9	122.7	12.6	25.6	44.7	60.7	118.4	12.8	24.5	45.2	85.1	101.9	7.9	19.0	31.5	57.3	108.5
LTD [32]	15.5	32.3	64.9	78.1	135.9	10.7	24.6	50.6	62.0	115.7	12.6	23.6	39.4	44.5	60.9	9.9	20.5	43.8	55.2	120.2
DMGNN [21]	21.4	38.7	75.7	92.7	-	11.9	25.1	44.6	50.2	-	9.6	21.8	56.9	71.9	-	13.6	29.0	46.0	58.8	-
HRI [33]	13.0	29.2	60.4	73.9	134.2	9.3	20.1	44.3	56.0	115.9	10.0	19.5	34.2	39.8	58.1	8.3	18.4	40.7	51.5	115.9
MSR-GCN [22]	14.8	32.4	66.1	79.6	-	10.3	22.0	46.3	57.8	-	8.9	14.9	29.0	33.1	-	9.9	21.0	44.6	56.3	-
FDU&FAU [9]	11.8	27.2	56.4	63.9	130.7	8.7	18.9	42.1	53.2	114.5	8.8	16.9	31.5	37.0	**50.3**	8.1	18.0	39.2	50.6	116.1
Ours	**9.2**	**17.3**	**34.5**	**42.6**	**87.5**	**5.8**	**11.0**	**22.7**	**40.4**	**68.3**	**8.4**	**14.5**	**28.2**	**33.0**	65.4	**5.8**	**11.0**	**22.9**	**41.4**	**70.8**
	Sittingdown					Walkingdog					Walkingtogether					Average				
res-GRU [34]	28.6	55.2	85.6	115.8	180.0	59.8	78.6	152.3	178.3	200.1	25.4	53.2	89.8	99.6	183.4	29.9	53.3	98.8	114.7	151.1
HPGAN [1]	39.9	65.9	92.1	130.0	200.2	83.1	92.1	170.0	198.4	238.8	68.6	79.9	95.3	108.4	188.1	64.2	87.3	122.1	143.0	192.3
BiGAN [17]	17.0	34.8	66.5	76.9	152.0	41.2	78.3	116.2	130.1	210.5	14.8	30.1	54.2	65.1	150.2	19.7	37.8	67.9	79.3	138.3
HMR [25]	9.6	18.6	41.1	57.7	148.3	38.2	63.6	109.3	125.6	190.0	12.2	25.2	46.2	50.2	134.1	15.4	28.4	52.8	70.9	119.6
LTD [32]	17.0	33.4	61.6	74.4	144.1	22.9	43.5	74.5	86.4	142.2	10.8	21.7	39.6	47.0	69.6	12.5	25.5	46.3	61.3	112.3
DMGNN [21]	15.0	32.9	77.1	93.0	-	47.1	93.3	160.1	171.2	-	14.3	26.7	50.1	63.2	-	15.2	30.4	60.7	73.3	-
HRI [33]	14.9	30.7	59.1	72.0	143.6	20.1	40.3	73.3	86.3	142.2	8.0	**15.4**	35.1	41.6	64.9	10.5	22.7	47.9	59.0	110.9
MSR-GCN [22]	16.1	31.6	62.5	76.8	-	20.6	42.9	80.4	93.3	-	10.6	20.9	37.4	43.9	-	11.9	25.1	51.0	62.1	-
FDU&FAU [9]	13.9	25.6	54.2	67.2	145.3	14.5	32.7	63.8	76.0	123.1	7.4	15.2	30.0	36.4	**58.7**	9.0	19.8	42.6	52.7	107.0
Ours	**7.6**	**13.8**	**27.8**	**35.1**	**84.6**	**13.7**	**25.2**	**49.0**	**59.8**	**114.5**	**7.3**	**14.8**	**29.5**	**35.8**	76.4	**7.8**	**14.4**	**29.0**	**37.0**	**80.1**

Table 1. Comparisons of prediction on H3.6M. Results at 80 ms, 160 ms, 320 ms, 400 ms, and 1,000 ms in the future are shown. The best results are highlighted in bold.

4. Experiments

4.1. Experimental Settings

Datasets. We evaluate our proposed method on three benchmark datasets, namely Human3.6m (H3.6M), CMU Motion Capture (CMU Mocap), and 3D Poses in the Wild (3DPW). The **H3.6M** dataset [15] contains 3.6 million human images recorded by a Vicon motion capture system. 7 subjects perform 15 different classes of actions. Following the evaluation protocol of previous works [13, 16], duplicate points in the human pose are removed and downsampled to 25 FPS (frames per second). Subject 5 is utilized as the test set. The **CMU Mocap** dataset is released by researchers from Carnegie Mellon University. 12 infrared cameras are utilized to capture human poses. Following previous works [32], we adopt the same training/test splits. The **3DPW** dataset [39] is proposed primarily for wild scenes that are recorded by a handheld smartphone camera or IMU. It contains 60 video sequences with more than 51,000 indoor or outdoor poses.

Evaluation metric. We evaluate our proposed approach on H3.6M, CMU Mocap, and 3DPW datasets by measuring the mean per joint position error (MPJPE) after the alignment of the root joint. In these experiments, we consider three kinds of prediction: short-term prediction (less than 400 ms), long-term prediction (400-1,000 ms), and longer-term prediction (2,000 ms). Experimental results at 80 ms, 160 ms, 320 ms, 400 ms, and 1,000 ms in the future are shown for comparisons.

Implementation details. We utilize Python 3.8 to implement our method. The experiments are performed with an NVIDIA 3090 GPU. For the symplectic operator, the length of the observed sequence is 25 frames (1,000 ms) and the length of the predicted sequence is 5 frames (200 ms). The symplectic operator is multiple GRUs with 256 units each. The GAN consists of two components, namely 256-unit GRUs and a three-layer FCN with 256 units (*i.e.*, the critic). The Adam Optimizer with a 0.001 learning rate is used. Note that different datasets and different actions are trained independently in our method.

4.2. Comparisons with State-of-the-art Approaches

We compare our proposed approach against the following human motion prediction methods: res-GRU [13], HPGAN [1], BiGAN [17], HMR [25], LTD [32], DMGNN [21], HRI [33], MSR-GCN [22], and FDU&FAU [9]. These methods are representative state-of-the-art (SOTA) methods covering a diverse set of technologies. In the following section, we analyze their prediction performance with comprehensive quantitative and qualitative comparisons. The results on H3.6M, CMU Mocap, and 3DPW are reported in Tables 1, 2, and 3, respectively. The empirical results demonstrate that our method significantly outperforms current SOTA approaches on all three datasets, which reveals the feasibility of tackling the human motion prediction problem with the symplectic integral.

Starting from the experimental results on the **H3.6M** dataset, the current SOTA methods obtain short- and long-term results of 52.7 and 107.0, respectively. Our SINN further pushes forward the performance boundary to 37.0 (↑ 15.7) and 80.1 (↑ 19.9), which is a quantum leap in performance. Moreover, we observe that existing methods tend to work well on regular motions (*e.g., walking*). However, they do not perform well on non-regular and complex motions (*e.g., discussion*). The long-term predicted average performance for *discussion* is at 129.8. Our method improves the performance to 79.9 (↑ 49.8). We believe our method benefits from the proposed motion representation for the following two reasons. *First*, the symplectic representation encoder provides a stable data system for prediction, which aids prediction, and in our experiments, we achieved such promising results with just a concise model. *Second*, our approach does not introduce excessive human-defined skeletal constraints, which might lead to human bias and limit the generalization ability of the model.

For long-term prediction, we observe that the prediction errors of all methods increase as the predicted sequence gets longer. This confirms that long-term prediction is more challenging. Quantitative results in Table 1 suggest that our approach outperforms existing methods in long-term motion prediction with a 20.1% accuracy increase. We believe that this mainly benefits from our proposed SINN because that 1) The symplectic operator exploits the strengths of prediction networks in short-term forecasting, it gains 30% increase. This ensures stable implementation of long-term integration; and 2) our proposed STA involves performing a definite integration of a function over a given interval, which ensures that the mean squared error over the entire interval is minimized. With respect to human motion prediction tasks, human motion is complex and rife with uncertainties, and can be considered a non-smooth and even steeply varying function from a data standpoint. Therefore, symplectic integration is particularly meaningful for handling human motion prediction tasks.

Time (ms)	80	160	320	400	560	1000
DMGNN [21]	13.6	24.1	47.0	58.8	77.4	112.6
MSR-GCN [22]	8.1	15.2	30.6	38.6	53.7	83.0
FDU&FAU [9]	6.4	13.9	27.9	36.0	50.1	75.4
Ours	**5.8**	**11.5**	**21.8**	**30.3**	**42.2**	**62.8**

Table 2. Comparisons of average prediction errors on CMU Mocap at 80 ms, 160 ms, 320 ms, 400 ms, 560 ms, and 1,000 ms.

Time (ms)	200	400	600	800	1000
DMGNN [21]	37.3	67.8	94.5	109.7	123.6
MSR-GCN [22]	37.8	71.3	93.9	110.8	121.5
FDU&FAU [9]	26.1	54.2	72.3	87.2	94.5
Ours	**19.9**	**46.5**	**65.8**	**80.3**	**88.2**

Table 3. Comparisons of average prediction errors on 3DPW at 200 ms, 400 ms, 600 ms, 800 ms, and 1,000 ms.

The results on the other two datasets show a similar pattern as that on the H3.6M dataset. The CMU Mocap dataset has a much higher frame rate than the H3.6M dataset. After downsampling to 25 frames, Table 2 shows that our method still performs better than existing methods, which validates the robustness of our method. A comparison of the average results shows that our method improves the performance by 16.7%. Finally, the 3DPW dataset is more challenging as it collects human motion data in the wild with a handheld smartphone, as shown in Table 3. Our method sets the new SOTA performance and achieves a 10.2% accuracy improvement over the current SOTA approach. The empirical results on three different datasets verify the efficacy of the proposed method.

Visualization results. We visualize the results of our method and current state-of-the-art methods HRI [33], MSR-GCN [22] and FDU&FAU [9] in Figure 3, where the results for ”Smoking” activity on H3.6M dataset. Existing methods for highly stochastic actions often suffer from a significant issue where the predicted motions tend to converge to static and motionless states. Our proposed SINN overcomes this problem by providing richer and smoother individualized motion contexts, which are easier to model compared to traditional pose representations. Figure 4 further visualizes the results of current state-of-the-art MSR-GCN [22], FDU&FAU [9], and our method for the longer-term prediction (2,000 ms). Prominently, most existing methods are shown to have plausible predictions within 1,000 ms. However, they usually converge to stationary or unnatural poses in the longer-term prediction (2,000 ms). Our method is able to maintain a natural motion and trend in the longer-term prediction.

4.3. Ablation Studies

To quantitatively analyze the effect of different components, we compare our method to two variations by removing the Symplectic Representation Encoder (SRE) or by removing the Symplectic Temporal Aggregation (STA) from our pro-

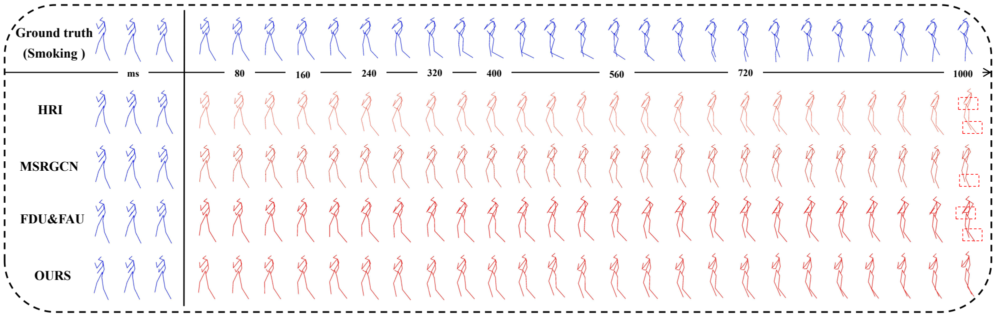


Figure 3. Result visualizations on H3.6M dataset up to 1,000 ms.

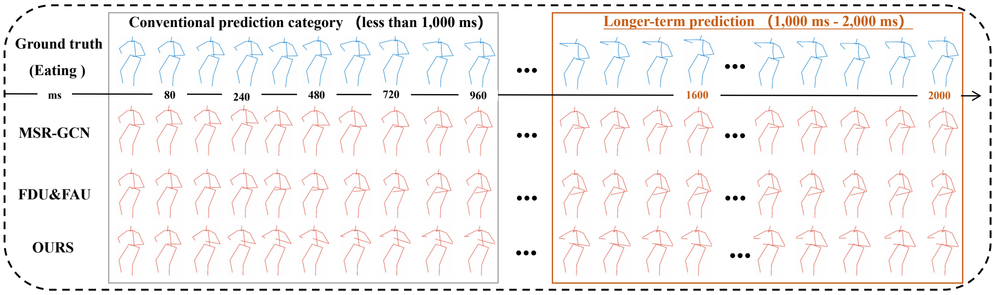

Conventional prediction category (less than 1,000 ms)
Longer-term prediction (1,000 ms - 2,000 ms)
Ground truth
(Eating)
ms
80
240
480
720
960
1600
2000
MSR-GCN
FDU&FAU
OURS

Figure 4. Result visualizations on H3.6M dataset for longer-term prediction (2,000 ms).

SRE	STA	80	160	320	400	1000
			Eating on H3.6M			
✓	×	10.2	19.5	35.6	45.2	65.1
×	✓	18.9	34.2	45.1	55.3	80.5
✓	✓	5.7	10.1	19.2	23.7	52.2
			Baketball on CMU Mocap			
✓	×	11.2	19.5	35.6	48.2	95.1
×	✓	19.6	41.0	58.1	85.3	108.5
✓	✓	10.0	18.3	34.5	45.2	88.1

Table 4. Ablation studies on the impact of different components.

	Current methods					Current methods + SINN				
Time (ms)	80	160	320	400	1,000	80	160	320	400	1,000
res-GRU [34]	17.5	34.3	71.1	87.5	117.6	16.5	38.8	68.0	85.1	114.3
BiGAN [17]	13.6	26.1	51.4	63.1	84.1	12.5	24.5	48.2	59.8	82.5
LTD [32]	7.7	15.8	30.5	37.6	74.1	7.1	14.9	31.8	36.0	72.9
DMGNN [21]	11.0	21.4	36.1	43.9	57.0	10.1	20.0	35.2	42.3	55.5
MSR-GCN [22]	8.4	17.1	33.0	40.0	-	8.0	16.5	32.2	35.7	-
FDU&FAU [9]	6.3	13.7	29.1	36.3	59.3	6.0	12.5	27.2	35.7	57.1

Table 5. Studies on model-agnostic learning.

posed method. The empirical results are reported in Table 4. From the results, we observe that 1) the SRE can largely improve the prediction power of the method. 2) the SRE representation is more helpful for short-term predictions. and 3) SINN can effectively improve the method and contribute to both the long-term and short-term predictions.

4.4. Studies on Model-agnostic Learning

Our proposed SINN is a model-agnostic framework, which can be easily integrated with existing methods. To demonstrate its effectiveness, we conduct experiments by utilizing various well-established methods such as res-GRU [34], BiGAN [17], LTD [32], DMGNN [21], MSR-GCN [22], and FDU&FAU [9]. These methods represent different types of networks including RNNs [34], GCNs [21], and GANs [17], thereby covering a wide range of network classes. The results, as shown in Table 5, clearly indicate that SINN significantly improves the prediction accuracy of these existing methods. This empirical evidence demonstrates the usefulness and extendability of our proposed SINN to other motion prediction approaches.

5. Conclusion

In this paper, we propose a novel human motion prediction approach that models human motion based on symplectic integral. We try to tackle the motion prediction problem from two aspects: (1) represent 3D human poses on symplectic representation to gain numerical stability, and (2) propose a network named symplectic integral network to construct the stable structure for modeling context. Our method provides a new perspective for this research area and contributes a reproducible research toward more accurate and longer human motion prediction.

6. Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62372402), and the Key R&D Program of Zhejiang Province (No. 2023C01217).

References

[1] Emad Barsoum, John R. Kender, and Zicheng Liu. HP-GAN: probabilistic 3d human motion prediction via GAN. In *CVPR Workshops 2018*, pages 1418–1427, 2018. 2, 6, 7

[2] Yujun Cai, Lin Huang, Yiwei Wang, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Xu Yang, Yiheng Zhu, Xiaohui Shen, Ding Liu, Jing Liu, and Nadia Magnenat-Thalmann. Learning progressive joint propagation for human motion prediction. In *ECCV*, pages 226–242. Springer, 2020. 3

[3] Haoming Chen, Runyang Feng, Sifan Wu, Hao Xu, Fengcheng Zhou, and Zhenguang Liu. 2d human pose estimation: a survey. *Multim. Syst.*, 29(5):3115–3138, 2023. 1

[4] Wenheng Chen, He Wang, Yi Yuan, Tianjia Shao, and Kun Zhou. Dynamic future net: Diversified human motion generation. In *MM 2020*, pages 2131–2139, 2020. 1

[5] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734, 2014. 1

[6] Baptiste Chopin, Naima Otberdout, Mohamed Daoudi, and Angela Bartolo. 3-d skeleton-based human motion prediction with manifold-aware GAN. *IEEE Trans. Biom. Behav. Identity Sci.*, 5(3):321–333, 2023. 2, 3

[7] Enric Corona, Albert Pumarola, Guillem Alenyà, and Francesc Moreno-Noguer. Context-aware human motion prediction. In *CVPR 2020,*, pages 6990–6999, 2020. 1

[8] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik. Recurrent network models for human dynamics. In *ICCV*, 2015. 1, 3

[9] Xuehao Gao, Shaoyi Du, Yang Wu, and Yang Yang. Decompose more and aggregate better: Two closer looks at frequency representation learning for human motion prediction. In *CVPR 2023,*, pages 6451–6460, 2023. 6, 7, 8

[10] Partha Ghosh, Jie Song, Emre Aksan, and Otmar Hilliges. Learning human motion models for long-term predictions. In *3DV*, pages 458–466, 2017. 3

[11] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial networks. *CoRR*, abs/1406.2661, 2014. 2, 3

[12] Xiao Guo and Jongmoo Choi. Human motion prediction via learning local structure representations and temporal dependencies. In *AAAI 2019*, pages 2580–2587, 2019. 3

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6, 7

[14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997. 1

[15] 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 TPAMI*, 36 (7):1325–39, 2014. 6

[16] Ashesh Jain, Amir Roshan Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *CVPR 2016*, pages 5308–5317, 2016. 1, 2, 6

[17] Jogendra Nath Kundu, Maharshi Gor, and R Venkatesh Babu. Bihmp-gan: Bidirectional 3d human motion prediction gan. In *AAAI 2019*, pages 8553–8560, 2019. 2, 3, 6, 7, 8

[18] J. D. Lambert. Numerical methods for ordinary differential systems: The initial value problem. 1991. 5

[19] Andreas M. Lehrmann, Peter V. Gehler, and Sebastian Nowozin. Efficient nonlinear markov models for human motion. In *CVPR*, pages 1314–1321, 2014. 3

[20] Peng Lei and Sinisa Todorovic. Modeling human-skeleton motion patterns using conditional deep boltzmann machine. In *ICPR 2016*, pages 1845–1850. IEEE, 2016. 3

[21] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In *CVPR, 2020*, pages 211–220, 2020. 6, 7, 8

[22] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Multiscale spatio-temporal graph neural networks for 3d skeleton-based motion prediction. *IEEE Trans. Image Process.*, 30:7760–7775, 2021. 1, 2, 6, 7, 8

[23] Maosen Li, Siheng Chen, Zijing Zhang, Lingxi Xie, Qi Tian, and Ya Zhang. Skeleton-parted graph scattering networks for 3d human motion prediction. In *ECCV 2022*, pages 18–36, 2022. 1

[24] Peng Li, Xiaofei Pei, Zhenfu Chen, Xingzhen Zhou, and Jie Xu. Human-like motion planning of autonomous vehicle based on probabilistic trajectory prediction. *Appl. Soft Comput.*, 118:108499, 2022. 1

[25] Zhenguang Liu, Shuang Wu, Shuyuan Jin, Qi Liu, Shijian Lu, Roger Zimmermann, and Li Cheng. Towards natural and accurate future motion prediction of humans and animals. In *CVPR,2019*, pages 10004–10012, 2019. 3, 6, 7

[26] Zhenguang Liu, Kedi Lyu, Shuang Wu, Haipeng Chen, Yanbin Hao, and Shouling Ji. Aggregated multi-gans for controlled 3d human motion prediction. In *AAAI, 2021*, pages 2225–2232, 2021. 2, 3

[27] Zhenguang Liu, Pengxiang Su, Shuang Wu, Xuanjing Shen, Haipeng Chen, Yanbin Hao, and Meng Wang. Motion prediction using trajectory cues. In *ICCV,2021*, pages 13279–13288, 2021. 3

[28] Zhenguang Liu, Shuang Wu, Shuyuan Jin, Shouling Ji, Qi Liu, Shijian Lu, and Li Cheng. Investigating pose representations and motion contexts modeling for 3d motion prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, pages 1–16, 2022. 3

[29] Zhenguang Liu, Roger Zimmermann, and Li Cheng. Special issue on human-centric intelligent multimedia understanding. *Multim. Syst.*, 29(2):457–458, 2023. 2

[30] Kedi Lyu, Zhenguang Liu, Shuang Wu, Haipeng Chen, Xuhong Zhang, and Yuyu Yin. Learning human motion prediction via stochastic differential equations. In *MM, 2021*, pages 4976–4984, 2021. 3

[31] Kedi Lyu, Haipeng Chen, Zhenguang Liu, Beiqi Zhang, and Ruili Wang. 3d human motion prediction: A survey. *Neurocomputing*, 489:345–365, 2022. 2

[32] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *ICCV*, pages 9488–9496, 2019. 6, 7, 8

[33] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *ECCV, 2020*, pages 474–489, 2020. 6, 7

[34] J. Martinez, M. Black, and J. Romero. On human motion prediction using recurrent neural networks. In *CVPR*, 2017. 1, 3, 6, 8

[35] Dario Pavllo, David Grangier, and Michael Auli. Quaternet: A quaternion-based recurrent model for human motion. In *BMVC 2018*, page 299, 2018. 3

[36] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back propagating errors. *Nature*, 323(6088):533–536, 1986. 1, 3

[37] André Salomão, Flávio Andaló, Gabriel Souza Prim, Milton Luiz Horn Vieira, and Nicolas Canale Romeiro. Case studies of motion capture as a tool for human-computer interaction research in the areas of design and animation. In *Human-Computer Interaction. Theoretical Approaches and Design Methods*, pages 302–311, 2022. 1

[38] Pengxiang Su, Zhenguang Liu, Shuang Wu, Lei Zhu, Yifang Yin, and Xuanjing Shen. Motion prediction via joint dependency modeling in phase space. In *MM, 2021*, pages 713–721. ACM, 2021. 3

[39] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV 2018*, pages 614–631, 2018. 6

[40] Jack M. Wang, David J. Fleet, and Aaron Hertzmann. Gaussian process dynamical models. In *NIPS*, pages 1441–1448, 2005. 3

[41] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI 2018*, pages 7444–7452, 2018. 1, 2

[42] Xun Yang, Meng Wang, and Dacheng Tao. Person re-identification with metric learning using privileged information. *IEEE Transactions on Image Processing*, 27(2):791–805, 2017. 3

[43] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. Deconfounded video moment retrieval with causal intervention. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1–10, 2021. 1

[44] Xun Yang, Shanshan Wang, Jian Dong, Jianfeng Dong, Meng Wang, and Tat-Seng Chua. Video moment retrieval with cross-modal neural architecture search. *IEEE Transactions on Image Processing*, 31:1204–1216, 2022. 1

[45] Jiawei Yao and Yingxin Lai. Dynamicbev: Leveraging dynamic queries and temporal context for 3d object detection. *CoRR*, abs/2310.05989, 2023. 3

[46] Jiawei Yao, Chuming Li, Keqiang Sun, Yingjie Cai, Hao Li, Wanli Ouyang, and Hongsheng Li. Ndc-scene: Boost monocular 3d semantic scene completion in normalized device coordinates space. In *ICCV 2023*, pages 9421–9431, 2023. 1

[47] Jiawei Yao, Tong Wu, and Xiaofeng Zhang. Improving depth gradient continuity in transformers: A comparative study on monocular depth estimation with cnn. *arXiv preprint arXiv:2308.08333*, 2023. 1

[48] Yifang Yin, Sheng Zhang, Yicheng Zhang, Yi Zhang, and Shili Xiang. Context-aware aircraft trajectory prediction with diffusion models. In *ITSC*, pages 5312–5317, 2023. 3

[49] M. Zhao, H. Tang, P. Xie, and et al. Bidirectional transformer gan for long-term human motion prediction. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2023. 3

[50] Honghong Zhou, Caili Guo, Hao Zhang, and Yanjun Wang. Learning multiscale correlations for human motion prediction. In *ICDL 2021*, pages 1–7, 2021. 1, 3

[51] Zixiang Zhou and Baoyuan Wang. UDE: A unified driving engine for human motion generation. In *CVPR 2023*, pages 5632–5641, 2023. 1