International Journal on Semantic Web and Information Systems
Volume 21 • Issue 1 • January-December 2025

Web-Based Semantic Framework for Enhanced Human Motion Prediction With MSTIA-Net

Yanzheng He
Wenzhou University, China

Pengjun Wang
https://orcid.org/0000-0002-1461-3719
Wenzhou University, China

Xiaochun Guan
Wenzhou University, China

Han Li
Wenzhou University, China

ABSTRACT

Researchers in human motion prediction have focused on mathematical modeling of the human skeletal structure, often overlooking the spatio-temporal characteristics of human pose sequences. To address this, we propose the multi-scale spatio-temporal information aggregation net (MSTIA-Net), which includes two key modules: the graph convolutional spatio-temporal information aggregation (GCSTIA) module and the windowed discrete cosine transform (WDCT) temporal encoding module. GCSTIA extracts and integrates multi-scale temporal and spatial features of human motion sequences, while WDCT removes high-frequency noise and compresses data. model's efficacy is demonstrated on three datasets: Human3.6, CMU, and 3DPW, achieving performance improvements of 2.4%, 4.1%, and 1.7%, respectively.

KEYWORDS

Human Motion Prediction, ST-GCN, Spatio-Temporal Information Aggregation, WDCT

INTRODUCTION

The applications of human motion prediction (HMP) are becoming increasingly prevalent in our daily lives. This can be observed in various fields, including human–computer interaction (Drosos et al., 2024; Hu et al., 2024), autonomous driving (K. Wang et al., 2023), and intelligent sports (Yeh et al., 2023). The data currently employed for HMP are based on a range of factors, including 2D skeleton points, 3D skeleton points, electromyographic signals, and so forth. In our experiments, the principal rationale for employing 3D skeleton-based human data is that they more closely align with the actual, real-life movements of humans in 3D space. In the nascent stages of HMP, traditional mathematical models were typically employed, including nonlinear Markov models (Lehrmann et al., 2014), Gaussian process dynamical models (Wang et al., 2007), and restricted Boltzmann machines (Taylor et al., 2006). The ongoing developments in deep learning have led to considerable advances in the field of HMP. Given that human motion is a temporal series, recurrent neural networks initially

DOI: 10.4018/IJSWIS.368221

emerged as a method for HMP (Pavllo et al., 2020). However, due to the intrinsic characteristics of recurrent neural networks, they are prone to accumulating model errors. In light of these developments, attention was drawn to convolutional neural networks (CNNs; Cui et al., 2021). In these approaches, the temporal and skeletal spatial dimensions of human motion sequence data are approximated as 2D data within an image for convolution processing. However, the convolutional kernels employed in convolution operations are unable to effectively extract global information from each joint in the human skeleton. This results in incomplete information extraction by CNNs in HMP. Currently, extensive research demonstrates the strong applicability of GCNs in HMP (X. Wang et al., 2023; Wang et al., 2024; Zand et al., 2023). GCNs treat human joints as graph nodes and the physical connections between joints as edges, extracting information from the human skeleton to aid in motion prediction.

However, these methods often focus only on either the skeletal structure or the time series, failing to fully explore the complementarity of temporal and spatial information in skeletal data. Furthermore, the methodologies employed in the modeling of human motion sequences are becoming increasingly diverse with a particular focus on the modeling of the skeleton (Ren et al., 2023; Wang et al., 2024). The majority of contemporary GCN-based approaches to forecasting human movement tend to treat the human skeleton as a discrete entity divorced from the temporal dynamics of movement sequences. This approach overlooks the inherent complementarity between the skeleton and the time series, which are inextricably linked in the context of human movement. The graph convolutional spatiotemporal information aggregation (GCSTIA) module proposed in this paper can effectively integrate information from both temporal and spatial domains, leveraging link-weighted computations to capture relationships between temporal and spatial dimensions.

A further common issue is that human motion data is typically collected using a variety of sensors, which may introduce a range of types of noise into the acquired data. These may include internal noise from the sensors themselves or environmental noise from the surrounding capture environment. While these noises do not necessarily corrupt the original motion information, they pose challenges for deep learning models. The discrete cosine transform (DCT) converts human movement sequences from the time domain to the frequency domain, thereby removing noise through the application of a high-frequency filtering process (Sun & Chowdhary, 2023). As the recorded human action sequences are discrete in timing, the DCT is susceptible to the boundaries between each frame of the human action. To address this issue, this paper proposes the use of windowed discrete cosine transform (WDCT) to reduce the effect of boundaries on the time-frequency transformation. In summary, the main contributions of this paper are as follows:

- We novelly adopt a WDCT time encoding method that effectively encodes discrete time series, achieving denoising by compressing the encoded information.
- We introduce a GCSTIA where multiscale pooling in the temporal domain enhances the model's understanding of human motion trends. This module separately extracts temporal and spatial domain features of human motion, followed by pooling-weighted fusion of these two parts of information.
- We conduct extensive experiments on various standard datasets for HMP to evaluate the effectiveness of the proposed methods. Experimental results demonstrate significant performance improvements across different datasets.

RELATED WORK

HMP

The traditional approach to HMP entails the manual construction of human motion features in conjunction with machine learning techniques (Lehrmann et al., 2014; Starke et al., 2019; Taylor et al.,

2006; Wang et al., 2007). However, these methods not only encounter difficulties in modeling complex human movement sequences but also require a considerable input of human effort to generate the requisite movement features. Deep learning-based methods, which are capable of learning spatiotemporal representations of human motion sequences and utilizing end-to-end mechanisms, have demonstrated considerable success in the domain of HMP. Pavllo et al. (2020) put forth the QuaterNet approach, which employs quaternions for rotation parameterization. This methodology is designed to circumvent bone stretching and invalid configurations (Cui et al., 2021). Corona et al. (2020) explore a context-aware motion prediction architecture using a semantic graph representation. Gopalakrishnan et al. (2019) employ motion derivative features and a novel multi-objective loss function to achieve excellent performance in long-term motion synthesis. Scholkopf et al. (xxx) put forth a nonlinear generative model for human motion data, employing an undirected model with binary latent variables and real-valued "visible" variables representing joint angles (Pavllo et al., 2020). Cui et al. (2021) put forth a temporal convolutional generative adversarial network as a means of accurately predicting future poses. Wang et al. (2023) were the first to establish a theoretical connection between MLP mixers and GCNs. Zand et al. (2023) put forth the concept of ResChunk, which entails learning residuals between successive target sequence chunks. Ren et al. (2023) present a method that simultaneously extracts spatiotemporal information and decomposes the task into multiple steps. However, these approaches typically do not integrate features across temporal and spatial domains. Our objective is to aggregate features across these domains to achieve more accurate predictions.

Spatiotemporal Information Modeling

Current popular methods for HMP focus on the extraction of multiscale features of the human skeleton (Li et al., 2022; Mascaró et al., 2024; Wang et al., 2024). Wang and Qureshi (2023) present AnyPose, a lightweight continuous-time neural architecture. Sofianos et al. (2021) propose STS-GCN, which allows cross talk between motion and spatial correlations. Saadatnejad et al. (2023) present a temporal cascaded diffusion model to better handle long-term prediction horizons. These methods typically focus on either the spatial or temporal characteristics of human motion sequences. Despite their effectiveness, these methods often fail to fully account for temporal information or the interrelationship between temporal and spatial domains, leading to suboptimal model learning. Recently, residual learning structures based on discrete sine transform time coding have been adopted to enhance skeletal topology extraction by transforming time-domain information into frequency (Guo et al., 2023; Mao et al., 2019). While discrete sine transform effectively models the temporal relationships between each joint, it struggles to handle abrupt transitions between the time stamps of each joint. To address this, we employ a novel temporal coding method, WDCT, which incorporates windowing functions to achieve smoother transitions between timestamps.

METHOD

Problem Formulation

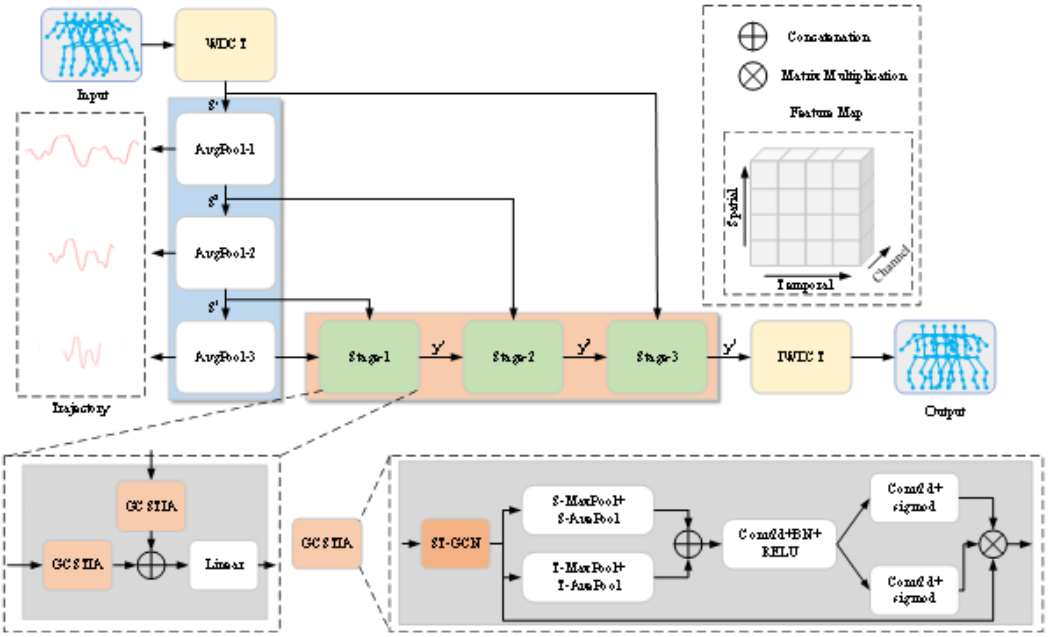
Assuming $S_{1:T} = \{S_1, \ldots, S_T\}$ represents a human pose sequence, where each frame $S_t \in \mathbb{R}^{J \times D}$ denotes the posture of the human at time t, including J joints, each represented in D dimensions. Typically, D can represent the 3D coordinates of joints, 2D coordinates, or quaternion representations of orientations, among others. Thus, the joint j at time step t can be represented as $S_t^j \in \mathbb{R}^{J \times D}$.

In the task of HMP, the objective is to predict the future p frames of human motion sequence $y = \{S_{T+1:T+p}\}$ given the observed T frames of human motion sequence. Therefore, our model needs to have $S_{T+1:T+p} = \theta(\{S_{1:T}\})$ mapping relationships. Below, we will provide information about our network structure in detail.

As shown in Figure 1, the multiscale spatiotemporal information aggregation net (MSTIA-Net) contains three main parts: the WDCT temporal codec module, the multiscale temporal dependency

extraction module, and the multiscale spatiotemporal feature intermingling module. In the following subsections, we describe the details of our proposed method.

Figure 1. A multi-intermediate state incremental graph convolution framework for human motion prediction (HMP), spanning N periods



Note. WDCT = windowed discrete cosine transform; IWDCT = inverse windowed discrete cosine transform; GCSTIA = graph convolutional spatiotemporal information aggregation; ST-GCN = XX; RELU = XX; BN = XX; Conv2d = XX; AvgPool = XX; T-MaxPool = XX; S-MaxPool = XX; T-AvePool = XX; S-AvePool = XX. Each period takes the output of the previous period as its input. For the first period, the input is composed of the past sequence and the last pose filled in.

Temporal Codec Module

In order to facilitate the transformation of temporal domain information to and from frequency domain, we propose the incorporation of a time domain codec module. The columns of matrix $S_{1:T}$ represent specific time points of human poses, while the rows of $S_{1:T}$ describe features of each joint. Assuming represents the motion trajectory of the *k*-th joint across *T* frames.

WDCT Temporal Encoding Structure

In light of the invariance principles of motion under Euclidean geometric transformations and interactions with intelligent agents (Xu et al., 2023), we put forth the proposition of adopting a trajectory representation based on the WDCT. The principal objective of this representation is to transform temporal sequences into the frequency domain, thus eliminating high-frequency noise and facilitating a more accurate capture of the smoothness of human motion.

In comparison to DCT, WDCT represents an enhancement of the process by introducing window functions, which facilitate the manipulation of the dataset. The application of appropriate window functions (such as Hamming or Hanning windows) to each joint trajectory prior to the application of DCT enables the smoothing of trajectory boundaries and the reduction of boundary effects (e.g. the Gibbs phenomenon). This approach not only suppresses noise in the frequency domain, thereby

improving the signal-to-noise ratio, but also enhances the accuracy and reliability of motion pattern representation. In particular, the use of window functions effectively suppresses side lobes in the spectrum, enhances spectral resolution, and significantly improves motion prediction performance without significantly increasing computational complexity. In particular, for a given trajectory, its corresponding *l*-th WDCT coefficient is computed by the following formula:

$$C_{k,l} = \sqrt{\frac{2}{T}} \sum_{t=1}^{T} S_{k,t} \boldsymbol{F}(\bullet) \cos\left(\frac{\pi}{2T}(2t-1)(l-1)\right) \tag{1}$$

where $\boldsymbol{F}(\bullet)$ represents window function. In specific experiments, we retained all $l \in \{1,2,\ldots,T\}$. However, for datasets with significant noise, we may ignore higher values, which means mitigating the influence of high-frequency motion. We can compute WDCT coefficients using Equation 1 and then model the temporal information of each joint using the WDCT system.

Inverse Windowed Discrete Cosine Transform Temporal Decoding Structure

Moreover, we can restore the original joint temporal information through the inverse windowed discrete cosine transform (IWDCT) operation. The specific formula is as follows:

$$s_{k,t} = \sqrt{\frac{2}{T}} \sum_{t=1}^{T} C_{k,l} \boldsymbol{F}(\bullet) \cos\left(\frac{\pi}{2T}(2t-1)(l-1)\right) \tag{2}$$

where $t \in \{1,2,\ldots,T\}$. It is worthy of note that if all WDCT coefficients are employed, the transformed data is lossless. As previously stated, by truncating high frequencies, some of the jitter effects caused by sensor sampling can be mitigated.

Multiscale Temporal Dependency Extraction Module

In order to simultaneously capture both local and global information pertaining to human pose sequences, we propose the implementation of a multiscale, temporal dependency extraction module. As shown in Figure 1, the multiscale spatiotemporal information fusion module consists of pooling layers. Pooling operations are used to generate descriptive features of the input data $S = \{S^1, S^2, S^3, \ldots, S^L\}$. Where *L* represents the layer of the module. The use of multiple averaging pooling layers in series can effectively extract trend components from raw data, making it easier to grasp the overall trends in human motion. We use three-layer average pooling with a convolution kernel of three, a step size of two and a padding of one to obtain the results S^1, S^2, S^3, S^4.

$$S^1 = s \tag{3}$$

$$S^{z+1} = AvgPool(S^z),\ z \in [1, 3] \tag{4}$$

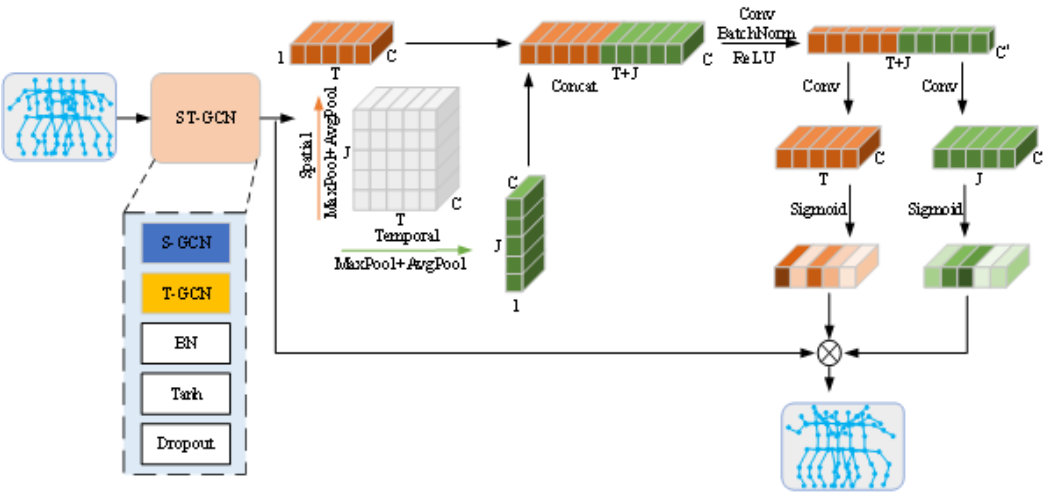
Multiscale temporal dependency extraction module effectively captures trend features, with trend information concentrating more at higher layers and richer detail information at lower layers.

Multiscale Spatiotemporal Feature Intermingling Module

The multiscale temporal dependency extraction module facilitates the formation of multiscale temporal dependencies. In order to optimize the utilization of these dependencies, it is necessary to implement a multiscale spatiotemporal feature intermingling module to prune and fuse the features in order to predict the output of stage $y = \{y^1, y^2, y^3\}$. The length of y^i is the sum of the lengths of S^i and y^{I-1}. As illustrated in Figure 1, each stage incorporates GCSTIA, which serves the function of

extracting data pertaining to the human skeleton and aggregating spatiotemporal information about it, as demonstrated in Figure 2.

Figure 2. Graph convolutional spatiotemporal information aggregation (GCSTIA) feature mapping diagram



Note. ST-GCN = XX; S-GCN = XX; T-CGN = XX; ReLU = XX; BN = XX; Conv = XX; BatchNorm = XX.

If we assume that the human skeletal structure can be represented as an undirected graph, then the extraction of temporal and spatial information from human motion using a graph CNN involves two main operations: graph convolution on the temporal dimension and graph convolution on the spatial dimension. Let $A \in \mathbb{R}^{J\times J}$ denote the adjacency matrix, $W_t^i, W_s^i \in \mathbb{R}^{D^i\times D^{i+1}}$ the trainable parameter matrix at layer i-th, $H^i \in \mathbb{R}^{T\times J\times D^i}$ the input at the current layer i, and D^i the number of channels at the current layer i. The specific operation for graph convolution on the temporal dimension is as follows:

$$\boldsymbol{H}^i = \sigma\left(A H^i W_t^i\right) \tag{5}$$

Following the application of graph convolution to the temporal dimension, the output of the current layer must be transposed in order to undergo graph convolution on the spatial dimension. The precise operation is as follows:

$$H^{i+1} = \sigma\left(A \boldsymbol{H}_{transpose}^i W_s^i\right) \tag{6}$$

where $\boldsymbol{H}^i_{\text{transpose}}$ is the result of the $\boldsymbol{H}^i$ after swapping the temporal and spatial dimensions, and σ is the nonlinear activation function. After the ST-GCN, we need to aggregate the temporal and spatial features. This paper uses pooling operations with kernels $(T, 1)$ and $(1, J)$ to encode each channel along the temporal and spatial dimensions respectively. The output of the c-th channel at time step t can be expressed as:

$$z_c^t(t)_{ave} = \frac{1}{J}\sum_{0\le q<J} S_c(q,t) \tag{7}$$

$$z_c^t(t)_{max} = Max\{S_c(0,t), S_c(1,t), \ldots, S_c(J,t)\} \tag{8}$$

$$z_c^t(t) = z_c^t(t)_{ave} + z_c^t(t)_{max} \tag{9}$$

Similarly, the output of the c-th channel at the j-th joint can be expressed as:

$$z_c^j(j)_{ave} = \frac{1}{T}\sum_{0\le k<T} S_c(j,k) \tag{10}$$

$$z_c^j(j)_{max} = Max\{S_c(j,0), S_c(j,1), \ldots, S_c(j,T)\} \tag{11}$$

$$z_c^j(j) = z_c^j(j)_{ave} + z_c^j(j)_{max} \tag{12}$$

The two aforementioned operations aggregate features along the temporal and spatial dimensions of human motion, respectively, resulting in a single-directional feature perception map. This feature perception map allows our model to capture long-range dependencies in one direction while maintaining positional information in the other direction. For subsequent aggregation of temporal and spatial features, we first concatenate the feature maps from both directions and then feed them into a shared convolutional transformation function.

$$f = \sigma\left(F_1([z^j, z^t])\right) \tag{13}$$

where $[\bullet, \bullet]$ denotes the concatenation operation of the two feature maps, F_1 denotes the operation of convolution and batchnorm, $f \in \mathbb{R}^{C'\times(J+T)}$ represents the intermediate feature map encoded along the temporal and spatial dimensions, and C' denotes the scaled number of channels.

Then, we split f along the concatenation direction into two separate tensors $f^t \in \mathbb{R}^{C'\times T}$ and $f^j \in \mathbb{R}^{C'\times J}$. Additionally, we use two 1×1 convolutional transformations F_T and F_J to transform f_t and f_j back to the original channel number C, respectively.

$$g^T = \delta\left(F_T(f^t)\right) \tag{14}$$

$$g^J = \delta\left(F_J(f^j)\right) \tag{15}$$

where δ denotes the sigmoid function. Finally, we use the resulting g^T and g^J as attention weights to balance the output obtained by the ST-GCN, thereby achieving the aggregation of features between the temporal and spatial dimensions. The output of the c-th channel can be expressed as follows:

$$y_c(q,k) = s_c(q,k) \times g_c^T(q) \times g_c^J(k) \tag{16}$$

EXPERIMENTS

Datasets

Human3.6M Dataset (Ionescu et al., 2013)

The Human3.6M dataset is currently the most widely used one for human motion applications. It comprises data from seven actors (S1, S5, S6, S7, S8, S9, and S11). We use S5 and S11 as the test set and validation set, respectively, while the remaining actors form the training set. Each actor

performs 15 types of actions. Each action includes 32 joints, but due to 10 redundant joints, we use only 22 nonredundant joints.

3D Poses in the Wild Dataset (Von Marcard et al., 2018)

3D Poses in the Wild (3DPW) is a dataset for 3D human pose estimation, and the actions in it are more complex compared to other public datasets. Each of its actions includes 26 joints, but since three joints are redundant, we use only the 23 nonredundant joints.

Carnegie Mellon University Motion Capture Dataset (Dang et al., 2021)

Carnegie Mellon University Motion Capture (CMU-MoCap) is a human motion capture dataset created by the Graphics Lab at Carnegie Mellon University. Each action includes 38 joints, and there are a total of eight human motion categories.

Experimental Setup

Metric

The Mean Per Joint Position Error is a widely adopted evaluation metric in HMP. It is used to compare the similarity between two human motion sequences. The specific formula is as follows:

$$\mathscr{L}_{MPJPE} = \frac{1}{J \times T} \sum_{t=1}^{T} \sum_{j=1}^{J} \left\| \hat{s}(j,t) - s(j,t) \right\| \tag{17}$$

Model Setup

The depth of pooling in our model is three for long-term forecasting and two for short-term forecasting. We used Adam as our optimizer. Initially, the learning rate was set to 0.005 and gradually decreased to 0.0005 as the training progressed. The model was trained under the PyTorch framework with a batch size of 12 and for 30 epochs. The hardware used for training included an NVIDIA RTX 3080 graphics processing unit and an Intel Xeon Silver 4310 central processing unit.

Comparative Results

A comparison is presented between our method and Traj-GCN, DMGNN, MSR-GCN, PGBIG, and SPGSN on the Human3.6M, 3DPW, and CMU-MoCap datasets. In contrast to the fixed human joint adjacency matrix, Traj-GCN employs a trainable parameter matrix. DMGNN gradually extracts human features by successively dividing the human structure into multiple subgraphs. MSR-GCN proposes a multiscale structure for the extraction of human skeleton features. PGBIG employs a multi-stage training approach with the objective of achieving increasingly accurate prediction results. SPGSN divides the human body into upper and lower parts in order to separately extract different human features. To ensure the comparability of the results, the epoch was set to 30 and the same random seed was used for all experiments.

Human3.6M

Table 1 shows the total number of trainable parameters for each model. Table 2 presents a quantitative comparison of our method with other methods of short-term prediction on the Human3.6M dataset. Table 3 presents a comparison of long-term prediction at 560 ms and 1000 ms. As can be observed from the tables, our proposed method demonstrates superior performance at the majority of timestamps. In comparison to the baseline methods, we achieve improvements of 3.4% and 1.4% in short-term and long-term prediction, respectively. As illustrated in Figure 3, our model is capable of more accurately capturing the overall trend of human motion. As the timestamp increases, there is a notable divergence from the ground truth among other models.

Table 1. Each baseline parameter

Baseline	Traj-GCN	MSR-GCN	DMGNN	SPGSN	PGBIG	MSTIA-Net
Paraments	9.7M	6.3M	6.2M	4.4M	1.7M	3.8M

Note. Traj-GCN = XX; MSR-GCN = XX; DMGNN = XX; SPGSN = XX; PGPIB = XX; MSTIA-Net = XX.

Table 2. Comparison of short-term predicted mean per joint position error on the Human3.6M dataset showing the predicted results for the next 80 ms,160 ms,320 ms,400 ms of motion

Motion	Walking				Eating				Smoking			
Millisecond	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms
Traj-GCN	12.3	23.0	39.8	46.1	8.4	16.9	33.2	40.7	7.9	16.2	31.9	38.9
MSR-GCN	12.2	22.7	38.6	45.2	8.4	17.1	33.0	40.4	8.0	16.3	31.3	38.2
DMGNN	17.3	30.7	54.6	65.2	11.0	21.4	36.2	43.9	9.0	17.6	32.1	40.3
SPGSN	12.5	22.7	38.4	44.3	8.5	16.9	32.8	40.1	8.1	16.0	30.6	37.1
PGBIG	11.2	22.1	37.7	43.2	6.8	14.8	30.0	37.4	7.4	15.9	31.3	38.3
MSTIA-Net	**10.6**	**20.9**	**37.0**	**43.0**	**6.6**	**14.4**	**29.4**	**37.0**	**7.1**	**14.9**	**29.7**	**36.6**
Motion	Discussion				Greeting				Posing			
Millisecond	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms
Traj-GCN	12.5	27.4	58.5	71.7	18.7	38.7	77.8	93.4	13.7	29.9	66.6	84.0
MSR-GCN	12.0	26.8	57.1	69.7	16.5	37.0	77.3	93.3	12.8	29.4	67.0	85.0
DMGNN	17.3	34.8	61.0	69.8	23.3	50.3	107.3	132.1	15.3	29.3	71.5	96.7
SPGSN	12.4	26.8	57.3	70.3	18.2	37.9	76.9	92.6	13.4	29.0	64.5	81.3
PGBIG	10.2	23.7	52.1	65.4	15.2	34.1	71.6	87.1	10.7	25.7	60.0	76.6
MSTIA-Net	**10.0**	**23.0**	**51.3**	**64.4**	**13.5**	**29.8**	**63.2**	**77.4**	**9.6**	**23.2**	**56.5**	**73.1**
Motion	Purchase				Sitting				Taking a Photo			
Millisecond	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms
Traj-GCN	15.6	32.8	65.7	79.2	10.6	21.9	46.3	57.9	9.9	20.9	45.0	56.6
MSR-GCN	14.6	32.8	65.2	78.2	10.2	20.3	43.7	53.6	9.1	20.1	44.6	55.7
DMGNN	21.4	38.7	75.7	92.7	11.9	25.1	44.6	50.2	13.6	29.0	45.0	56.6
SPGSN	15.2	31.8	64.3	77.6	10.6	21.6	45.1	56.5	10.3	21.3	45.0	56.5
PGBIG	**13.0**	29.6	62.1	75.9	9.2	19.7	43.6	55.3	9.1	19.3	42.4	53.8
MSTIA-Net	13.1	**29.5**	**60.7**	**74.4**	**9.0**	**19.5**	**42.6**	**53.8**	**8.2**	**18.2**	**40.4**	**51.6**
Motion	Waiting				Walking a Dog				Average			
Millisecond	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms
Traj-GCN	11.3	24.0	50.0	61.5	21.5	42.5	79.4	92.3	12.7	26.0	52.3	63.5
MSR-GCN	10.7	23.0	48.3	59.2	20.7	42.9	80.4	93.3	12.1	25.6	51.6	62.9
DMGNN	12.2	47.1	93.3	160.1	47.1	93.3	160.1	171.2	17.0	33.6	65.9	79.7
SPGSN	11.2	23.0	47.4	58.3	21.6	42.3	76.6	88.6	12.3	25.4	50.8	61.8

continued on following page

Table 2. Continued

Motion	Walking				Eating				Smoking			
PGBIG	9.4	20.8	44.3	55.0	**19.2**	39.9	74.6	87.0	10.9	23.4	48.3	59.2
MSTIA-Net	**8.5**	**18.7**	**42.3**	**53.7**	19.9	**39.6**	**71.6**	**84.9**	**10.4**	**22.5**	**46.7**	**57.9**

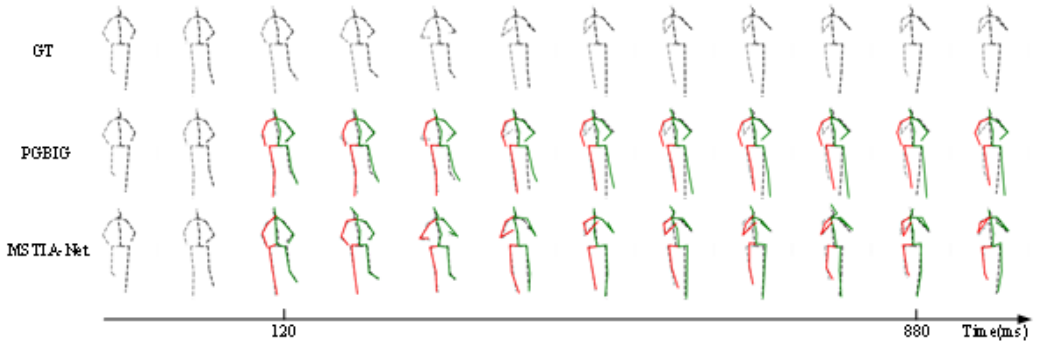
Note. Traj-GCN = XX; MSR-GCN = XX; DMGNN = XX; SPGSN = XX; PGPIB = XX; MSTIA-Net = XX. The most favorable outcomes are bold, while less optimal results are underlined.

Table 3. Comparison of long-term predicted mean per joint position error on the Human3.6M dataset showing the results of future motion predictions for 560 ms and 1000 ms

Motion	Waking		Eating		Smoking		Discussion		Greeting	
Millisecond	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms
Traj-GCN	54.1	59.8	53.4	77.8	50.7	72.6	91.6	121.5	115.4	148.8
MSR-GCN	52.7	63.0	52.5	77.1	49.4	71.6	88.6	117.6	116.2	147.2
DMGNN	73.4	95.8	58.1	86.7	50.9	72.2	**81.9**	138.3	152.5	157.7
SPGSN	52.1	60.3	52.6	77.1	48.5	69.8	90.6	121.2	115.1	148.0
PGBIG	48.1	56.4	51.7	76.3	47.2	**69.6**	87.1	117.1	110.1	142.4
MSTIA-Net	**47.6**	**56.3**	**51.1**	**74.8**	**46.9**	70.9	87.0	**116.8**	**101.2**	**135.1**
Motion	Posing		Sitting		Waiting		Walking a Dog		Average	
Millisecond	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms
Traj-GCN	114.5	173.0	80.9	114.0	116.3	180.6	141.8	185.4	81.6	114.3
MSR-GCN	116.3	174.3	78.2	120.0	76.3	106.3	111.9	148.2	81.1	114.2
DMGNN	163.9	310.1	75.5	115.4	85.5	113.7	183.2	210.2	93.6	127.6
SPGSN	110.4	166.7	76.8	118.9	76.1	106.0	105.0	141.1	79.9	112.7
PGBIG	107.0	164.5	76.7	118.8	72.7	104.2	105.2	140.1	77.7	110.8
MSTIA-Net	**106.1**	**163.7**	**74.5**	**113.2**	**72.0**	**103.9**	**101.7**	**139.9**	**75.9**	**110.2**

Note. Traj-GCN = XX; MSR-GCN = XX; DMGNN = XX; SPGSN = XX; PGPIB = XX; MSTIA-Net = XX. The most favorable outcomes are bold, while less optimal results are underlined.

Figure 3. Motion prediction results on "eating" motion classes from the Human3.6M dataset



Note. GT = XX; PGBIG = XX; MSTIA-NET = XX. The red solid line represents the predicted value on the left side of the body. The green solid line represents the predicted value on the right side of the body. The grey dashed line represents the ground truth.

CMU-MoCap and 3DPW

Tables 4 and 5 present a comparison of the performance of the baseline methods and our proposed method on the CMU-MoCap and 3DPW datasets, respectively. As some methods do not address long-term prediction on CMU-MoCap, we similarly refrain from discussing it here. As evidenced by the tables, our method achieves notable performance enhancements in long-term prediction on the challenging 3DPW dataset. This improvement can be attributed to our model's capacity to extract the overall motion trend through multilayer pooling operations, which facilitates the prediction of final results for complex actions.

Table 4. Performance comparison of individual actions on the Carnegie Mellon University motion capture (CMU-MoCap) dataset

Motion	Basketball				Jumping				Directing Traffic			
Millisecond	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms
Traj-GCN	11.7	21.3	41.0	50.8	17.2	32.4	60.1	72.6	6.9	13.7	30.3	40.0
MSR-GCN	10.3	18.9	37.7	47.0	15.0	28.7	**55.9**	69.1	5.9	12.1	28.4	38.0
DMGNN	15.6	28.7	59.0	73.1	32.0	54.3	96.7	119.9	10.2	20.9	41.6	52.3
SPGSN	12.6	23.2	45.2	56.5	18.2	34.2	64.2	77.7	7.5	14.6	31.3	40.7
PGBIG	10.1	18.5	36.9	46.0	14.6	28.6	57.0	70.2	5.3	10.2	24.3	32.3
MSTIA-Net	**9.8**	**17.9**	**36.6**	**44.3**	**13.7**	**27.9**	56.3	**68.2**	**5.1**	**9.9**	**23.9**	**31.7**
Motion	Soccer				Washing a Window				Average			
Millisecond	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms
Traj-GCN	13.3	24.0	43.8	53.2	6.0	11.6	24.8	31.6	9.9	18.0	33.6	41.0
MSR-GCN	**10.8**	**19.5**	**37.0**	**46.4**	5.5	11.1	25.1	32.5	8.7	15.8	30.6	38.0
DMGNN	14.9	25.3	52.2	65.4	7.9	14.7	33.3	44.2	14.1	24.4	45.9	55.4
SPGSN	15.2	27.8	51.5	62.9	6.7	13.0	27.4	34.7	11.3	20.8	39.1	47.8

continued on following page

Table 4. Continued

Motion	Basketball				Jumping				Directing Traffic			
PGBIG	11.9	22.0	42.0	51.5	5.0	9.6	21.9	28.6	8.3	15.9	30.8	38.1
MSTIA-Net	11.0	20.4	38.2	47.1	**4.9**	**9.5**	**21.0**	**28.1**	**8.1**	**14.9**	**29.3**	**36.9**

Note. Traj-GCN = XX; MSR-GCN = XX; DMGNN = XX; SPGSN = XX; PGPIB = XX; MSTIA-Net = XX.

Table 5. Performance comparison on the 3D poses in the wild (3DPW) dataset

Millisecond	400ms	600ms	800ms	1000ms
Traj-GCN	66.8	93.6	107.6	114.8
MSR-GCN	65.0	93.8	108.2	116.3
DMGNN	70.4	94.1	109.7	123.9
SPGSN	64.5	91.6	104.0	111.1
PGBIG	58.2	79.8	94.4	104.1
MSTIA-Net	**58.1**	**78.2**	**89.8**	**97.8**

Note. Traj-GCN = XX; MSR-GCN = XX; DMGNN = XX; SPGSN = XX; PGPIB = XX; MSTIA-Net = XX.

Ablation Analysis

To further analyze the reliability of our proposed method, we conducted ablation experiments using the Human3.6M dataset.

Effect of Different Time Encoding Modules

In order to identify the optimal window function, the mapping relationship inherent to the WDCT calculation formula was modified. Specifically, the window was replaced with sin, Hamming, Hanning and Blackman. The results in Table 6 demonstrate that the sine window yields the best performance, providing the optimal smoothing effect between each joint's timestamps. As evidenced by the data presented in the table, the utilization of the window function has resulted in a discernible enhancement in the performance of the models. Conversely, the window function has demonstrated its capacity to mitigate the abrupt transitions between discrete timestamps.

Table 6. Performance comparison of different time coding functions

Encoder	Window	320ms	400ms	1000ms
DCT		47.5	58.7	111.6
WDCT	sin	**46.7**	**57.9**	**110.2**
WDCT	Hamming	46.8	58.0	110.7
WDCT	Hanning	47.1	58.5	110.7
WDCT	Blackman	47.5	58.7	111.5

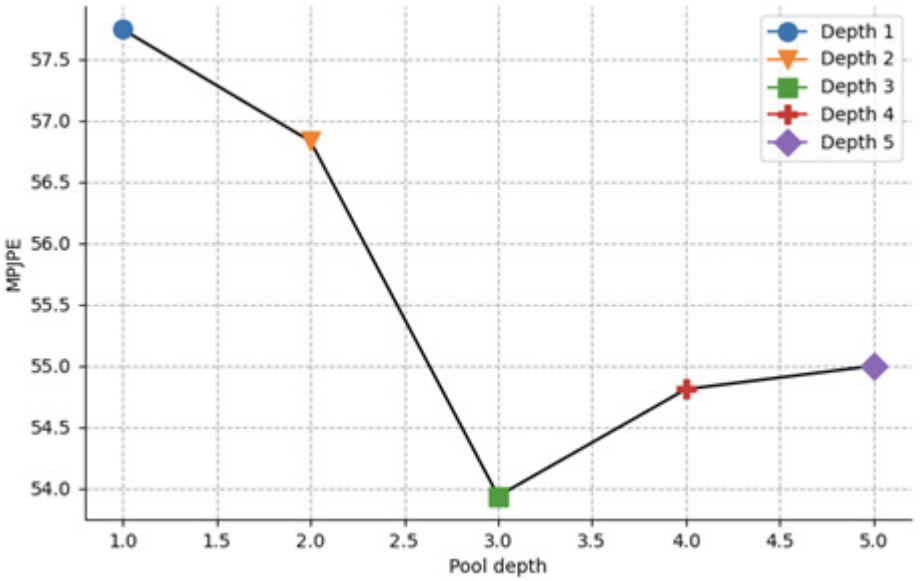
Note. DCT = XX; WDCT = XX.

Effect of Pooling Depth

In order to ascertain the optimal pooling depth for the accurate and efficient extraction of human motion features, a series of experiments were conducted utilizing different pooling depths. As

illustrated in Figure 4, when the pooling depth is less than three layers, the model demonstrates a gradual improvement in performance as the pooling depth increases. However, when the pooling depth exceeds three layers, the model loses local information due to over-pooling, which hinders the recovery of detailed information from subsequent layers. Nevertheless, the precise pooling depth must be determined based on the specific human action sequences under investigation.

Figure 4. Ablations on depth of pool

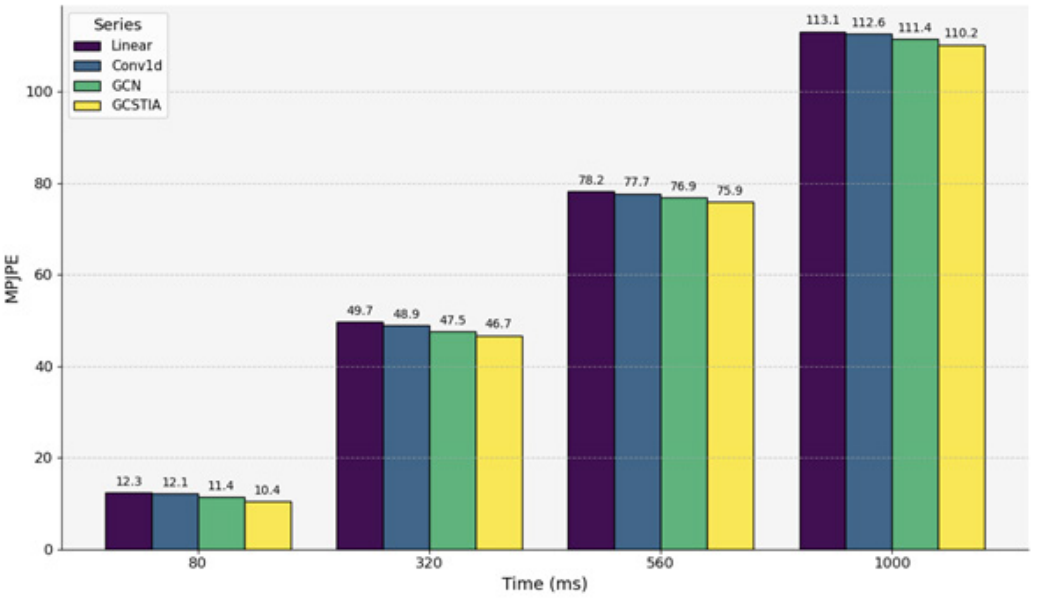


Note. MPJPE = XX.

Effect of GCSTIA

To demonstrate the capacity of the GCSTIA module to aggregate spatiotemporal data, we conducted ablation experiments using linear, convolutional, and graph convolutional layers in place of GCSTIA. As illustrated in Figure 5, the linear layer, convolutional layer, graph convolution, and capacity of GCSTIA for the aggregation of spatiotemporal information are demonstrated in a successive order. Furthermore, the spatiotemporal information aggregation ability of these structures remains largely unaltered with the alteration of the timestamp.

Figure 5. Ablations on architecture



Note. Conv1d = XX; GCN = XX; GCSTIA = XX; MPJPE = XX.

CONCLUSION

This paper proposes a novel spatiotemporal model for HMP, which employs a WDCT to filter out high-frequency noise in human motion sequences. In contrast to previous research, the proposed approach employs a two-stage process, whereby temporal and spatial information are extracted and subsequently integrated through convolution, thereby facilitating the capture of more useful information from motion data. The interplay between spatial joints and temporal patterns is enhanced, and modal bias is mitigated by assigning weights to the shared spatiotemporal information for human action sequences. Evaluated on the Human3.6M, CMU-MoCap, and 3DPW motion prediction datasets, MSTIA-Net achieves performance gains of 2.4%, 4.1%, and 1.7%, respectively. In future work, we will further investigate other discrete temporal encoding methods, such as wavelet filtering, along with lower overhead spatiotemporal information fusion strategies. These approaches aim to facilitate easier adaptation of MSTIA-Net for HMP on low-power devices.

COMPETING INTERESTS

The authors of this publication declare there are no competing interests.

DATA AVAILABILITY

The data used to support the findings of this study are included within the article.

All datasets used in this study are publicly available and do not contain any private or sensitive personal information.

FUNDING

This research was supported by the Key R&D Program of Zhejiang (2022C03114) and the Wenzhou basic public welfare scientific research project (G20220034, S2023015).

AUTHOR NOTE

Correspondance concerning this article should be addressed to Pengjun Wang, College of Electrical and Electronic Engineering, Wenzhou University, Wenzhou, Zhejiang, 325035, China. Email: wangpengjun@wzu.edu.cn

PROCESS DATES

January 17, 2025
Received: November 7, 2024, Revision: December 1, 2024, Accepted: December 4, 2024

REFERENCES

Corona, E., Pumarola, A., Alenya, G., & Moreno-Noguer, F. (2020). Context-aware human motion prediction. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6992-7001.

Cui, Q., Sun, H., Kong, Y., Zhang, X., & Li, Y. (2021). Efficient human motion prediction using temporal convolutional generative adversarial network. *Information Sciences*, *545*, 427–447. DOI: 10.1016/j.ins.2020.08.123

Dang, L., Nie, Y., Long, C., Zhang, Q., & Li, G. (2021). Msr-gcn: Multi-scale residual graph convolution networks for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11467-11476. DOI: 10.1109/ICCV48922.2021.01127

Drosos, I., Sarkar, A., Xu, X., Negreanu, C., Rintel, S., & Tankelevitch, L. (2024). "It's like a rubber duck that talks back": Understanding generative AI-assisted data analysis workflows through a participatory prompting study. In *Proceedings of the 3rd Annual Meeting of the Symposium on Human-Computer Interaction for Work*, 1-21.

Gopalakrishnan, A., Mali, A., Kifer, D., Giles, L., & Ororbia, A. G. (2019). A neural temporal model for human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12116-12125. DOI: 10.1109/CVPR.2019.01239

Guo, W., Du, Y., Shen, X., Lepetit, V., Alameda-Pineda, X., & Moreno-Noguer, F. (2023). Back to mlp: A simple baseline for human motion prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 4809-4819. DOI: 10.1109/WACV56688.2023.00479

Hu, F., Zhang, L., Yang, X., & Zhang, W. A. (2024). EEG-based driver fatigue detection using spatio-temporal fusion network with brain region partitioning strategy. *IEEE Transactions on Intelligent Transportation Systems*, *25*(8), 9618–9630. DOI: 10.1109/TITS.2023.3348517

Ionescu, C., Papava, D., Olaru, V., & Sminchisescu, C. (2013). Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(7), 1325–1339. DOI: 10.1109/TPAMI.2013.248 PMID: 26353306

Lehrmann, A. M., Gehler, P. V., & Nowozin, S. (2014). Efficient nonlinear Markov models for human motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1314-1321. DOI: 10.1109/CVPR.2014.171

Li, M., Chen, S., Zhang, Z., Xie, L., Tian, Q., & Zhang, Y. (2022). Skeleton-parted graph scattering networks for 3D human motion prediction. In *European Conference on Computer Vision*, 18-36. Cham: Springer Nature Switzerland. DOI: 10.1007/978-3-031-20068-7_2

Mao, W., Liu, M., Salzmann, M., & Li, H. (2019). Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9489-9497. DOI: 10.1109/ICCV.2019.00958

Mascaró, E. V., Ahn, H., & Lee, D. (2024). A unified masked autoencoder with patchified skeletons for motion synthesis. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(6), 5261–5269. DOI: 10.1609/aaai.v38i6.28333

Pavllo, D., Feichtenhofer, C., Auli, M., & Grangier, D. (2020). Modeling human motion with quaternion-based neural networks. *International Journal of Computer Vision*, *128*(4), 855–872. DOI: 10.1007/s11263-019-01245-6

Ren, H., Shi, Y., & Liang, K. (2023). Multi-graph convolution network for pose forecasting. *arXiv preprint arXiv:2304.04956*.

Saadatnejad, S., Rasekh, A., Mofayezi, M., Medghalchi, Y., Rajabzadeh, S., Mordan, T., & Alahi, A. (2023). A generic diffusion-based approach for 3D human pose prediction in the wild. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 8246-8253. IEEE. DOI: 10.1109/ICRA48891.2023.10160399

Sofianos, T., Sampieri, A., Franco, L., & Galasso, F. (2021). Space-time-separable graph convolutional network for pose forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11209-11218. DOI: 10.1109/ICCV48922.2021.01102

Starke, S., Zhang, H., Komura, T., & Saito, J. (2019). Neural state machine for character-scene interactions. *ACM Transactions on Graphics*, *38*(6), 178. DOI: 10.1145/3355089.3356505

Sun, J., & Chowdhary, G. (2023). Towards accurate human motion prediction via iterative refinement. *ArXiv preprint arXiv:2305.04443*.

Taylor, G. W., Hinton, G. E., & Roweis, S. (2006). Modeling human motion using binary latent variables. *Advances in Neural Information Processing Systems*, 19.

Von Marcard, T., Henschel, R., Black, M. J., Rosenhahn, B., & Pons-Moll, G. (2018). Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 601-617. https://api.semanticscholar.org/CorpusID:51883209

Wang, J. M., Fleet, D. J., & Hertzmann, A. (2007). Gaussian process dynamical models for human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(2), 283–298. DOI: 10.1109/TPAMI.2007.1167 PMID: 18084059

Wang, K., Li, Z., Yu, T., & Sakaguchi, K. (2023). Smart mobility digital twin for automated driving: Design and proof-of-concept. In *2023 IEEE 97th Vehicular Technology Conference (VTC2023-Spring)*, 1-6.

Wang, X., Cui, Q., Chen, C., & Liu, M. (2024). Gcnext: Towards the unity of graph convolutions for human motion prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(6), 5642–5650. DOI: 10.1609/aaai.v38i6.28375

Wang, X., Cui, Q., Chen, C., Zhao, S., & Liu, M. (2023). Graph-guided mlp-mixer for skeleton-based human motion prediction. In *Proceedings of the 5th ACM International Conference on Multimedia in Asia*, 1-7. DOI: 10.1145/3595916.3626387

Wang, Z., & Qureshi, A. H. (2023). AnyPose: Anytime 3D human pose forecasting via neural ordinary differential equations. *arXiv preprint arXiv:2309.04840*.

Xu, C., Tan, R. T., Tan, Y., Chen, S., Wang, Y. G., Wang, X., & Wang, Y. (2023). Eqmotion: Equivariant multi-agent motion prediction with invariant interaction reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1410-1420. DOI: 10.1109/CVPR52729.2023.00142

Yeh, W. H., Lin, P. H., Su, Y. A., Cheng, W. H., & Ku, L. W. (2023). MAAIG: Motion analysis and instruction generation. In *Proceedings of the 5th ACM International Conference on Multimedia in Asia Workshops*, 1-5.

Zand, M., Etemad, A., & Greenspan, M. (2023). Multiscale residual learning of graph convolutional sequence chunks for human motion prediction. *arXiv preprint arXiv:2308.16801*.