Decompose More and Aggregate Better: Two Closer Looks at Frequency Representation Learning for Human Motion Prediction

Xuehao Gao[1], Shaoyi Du[1], Yang Wu[2], Yang Yang[1,*]

[1]Xi'an Jiaotong University, [2]Tencent AI Lab

{gaoxuehao.xjtu, dushaoyi}@gmail.com, dylanywu@tencent.com

Abstract

Encouraged by the effectiveness of encoding temporal dynamics within the frequency domain, recent human motion prediction systems prefer to first convert the motion representation from the original pose space into the frequency space. In this paper, we introduce two closer looks at effective frequency representation learning for robust motion prediction and summarize them as: decompose more and aggregate better. Motivated by these two insights, we develop two powerful units that factorize the frequency representation learning task with a novel decomposition-aggregation two-stage strategy: (1) frequency decomposition unit unweaves multi-view frequency representations from an input body motion by embedding its frequency features into multiple spaces; (2) feature aggregation unit deploys a series of intra-space and inter-space feature aggregation layers to collect comprehensive frequency representations from these spaces for robust human motion prediction. As evaluated on large-scale datasets, we develop a strong baseline model for the human motion prediction task that outperforms state-of-the-art methods by large margins: 8%∼12% on Human3.6M, 3%∼7% on CMU MoCap, and 7%∼10% on 3DPW.

1. Introduction

3D skeleton-based human motion prediction system forecasts future poses given a past motion. It helps machines understand human behavior and plan their own responses, which is crucial in many real-world applications, including intelligent surveillance [11, 40], human-machine interaction [16, 17] and autonomous driving [18, 32]. The core challenge behind this task lies in developing a powerful mapping function that effectively bridges past body motion to the future [23, 25, 27, 30, 33].

*Corresponding author.

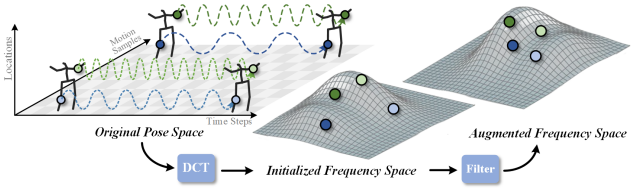

Figure 1. Diverse frequency distributions of body poses. As for a human action, the differences in temporal smoothness between different body joints and motion samples enlarge the representation gap in its frequency space.

Earlier prediction algorithms tend to extract motion patterns from the original pose space [7, 9, 10, 22, 34, 43]. Due to the subject-specific nature of pose space, their embedding representations intertwine body motion information and structure information jointly. In this case, they encapsulate an inductive bias on general human stature and thus suffer from limited robustness against body shape perturbation. Inspired by the effectiveness of encoding temporal smoothness in frequency domain, frequency space encourages human motion prediction systems to focus on trajectory-related cues [1, 42]. As an initial attempt, Mao et al. [28] propose to convert the motion representation from the pose space into the frequency space with discrete cosine transform (DCT). Following this insight, recent methods widely use DCT as a routine operation in the data preprocessing stage and extract feature embeddings from the single frequency space initialized by the DCT [21, 24, 26, 38]. In this context, the frequency features extracted from past body motions dominate the future motion prediction. A further investigation into developing a powerful frequency representation learning framework for robust human motion prediction remains fundamental yet under-explored.

As sketched in Figure 1, diverse frequency distributions of body motions lie in intra-sample and inter-sample levels: (1) *intra-sample difference.* Since the human skeleton

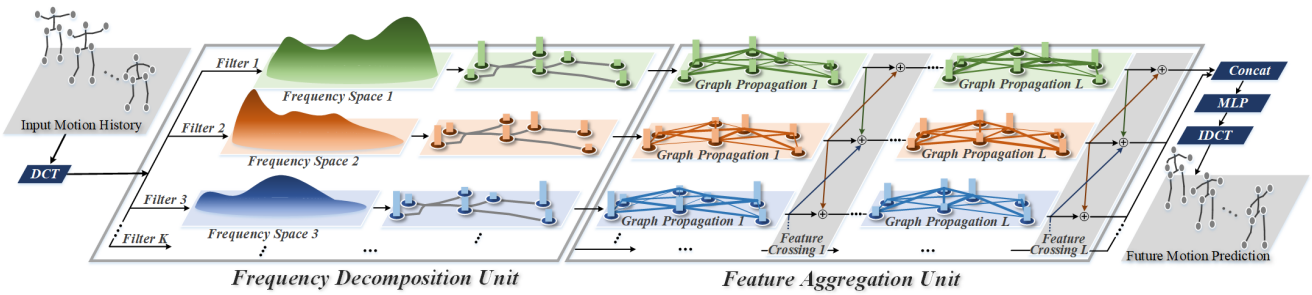

Figure 2. Network Architecture. we factorize the frequency representation learning into two-stage decomposition-aggregation scheme: Frequency Decomposition Unit (FDU) extracts multi-view frequency features from an input body motion by embedding its frequency representations into K spaces; Feature Aggregation Unit (FAU) deploys L intra-space and inter-space feature aggregation layers to collect comprehensive frequency representations for robust human motion prediction.

is a non-rigid articulated structure, different body joints exhibit different frequency appearances in their motion trajectories; (2) *inter-sample difference.* Different personal motion styles in the same activity brings subtle intra-class bias to different data samples, enlarging the frequency representation gap between human motion samples. These diverse frequency distributions make human motion prediction systems prone to be incapable of governing the input body trajectories with unseen frequency variations. It prompts us to develop multi-view augmentation learning into a promising solution for robust human motion prediction. Instead of extracting features from a single frequency space initialized by the DCT, we first introduce an input body motion into multiple frequency spaces to enrich its spectral encoding. Then, we collect richer multi-view frequency representations from these spaces for robust human motion prediction.

Specifically, as illustrated in Figure 2, we factorize the frequency representation learning into two sequential stages: (1) *Frequency Decomposition Unit* (FDU) unweaves finer frequency representations from an input body motion by tuning each body joint trajectory with multiple versatile filters. By embedding the frequency representation into multiple feature spaces, FDU explores multi-view frequency representations on input body poses; (2) *Feature Aggregation Unit* (FAU) first deploys a series of adaptive graph filters within each frequency space and then interleaves feature-crossing layers to promote message exchange between spaces. These intra-space and inter-space information aggregations benefit FAU in extracting comprehensive body features for robust body motion prediction. Integrating both FDU and FAU components, we reformulate the frequency representation learning into a novel and powerful decomposition-aggregation scheme.

The main contributions of this paper are summarized into the following:

- We propose a frequency decomposition unit (FDU) that develops multiple versatile filters to embed each body joint trajectory into multiple frequency spaces. By exploring multi-view frequency representations on an input body motion, FDU enriches its encodings in the spectral domain.

- Pairing with FDU, we design a feature aggregation unit (FAU) that deploys a series of intra-space and inter-space feature aggregation layers to extract comprehensive representations from multiple frequency spaces. By promoting message propagation within and between different spaces, FAU collects richer multi-view body features for robust motion prediction.

- Integrating FDU with FAU, we develop a powerful motion prediction system that factorizes the frequency representation learning into a decomposition-aggregation scheme. As verified on three datasets, it significantly outperforms state-of-the-art methods in short-term and long-term motion predictions.

2. Related Work

2.1. Human Motion Prediction

Traditional human motion prediction methods adopt shallow state models, such as Gaussian Processes [39], Hidden Markov Models [19], Restricted Boltzmann Machine [35], to propagate the state of neural cells for pose representation learning. Notably, these methods impose strong assumptions such as Gaussian distribution on the body dynamics, suffering from the potential generalization limitation [4, 14]. Recently, since the topology of human skeleton is a nature graph, some feed-forward networks introduce graph convolution layers to discover motion patterns across space and time, such as DMGNN [22], MSR-GCN [7], STSGCN [34], PGBIG [24], and GAGCN [43]. Motivated by the effectiveness of self-attention mechanism [36] in long-range dependency modeling, some methods adopt Transformer-based backbones to encourage wide-range receptive fields and make distant neighbors reachable, such as MRT [38], ST-Trans [2], PJP-Trans [6], and POTR [29].

However, since most of them focus on learning motion patterns from the original and direct pose space, their embedding features intertwine the body joint trajectory cues and body shape structure cues jointly. In this case, they have an inductive bias on general human stature and suffer from limited robustness against body shape perturbation.

2.2. Frequency Representation Learning

Inspired by the effectiveness of encoding temporal smoothness in the frequency domain, frequency representation learning purifies trajectory-related information from body poses and compacting them into a more abstract representation [1, 3, 20, 41]. As an initial attempt, Mao et al. [28] propose to convert the motion representation from the original pose space into the frequency space with discrete cosine transform (DCT). Following this insight, recent works tend to extract features from the frequency space and widely use DCT as a routine operation in the data preprocessing stage [6, 8, 12, 21, 24, 38]. However, they focus on learning embeddings from the single frequency space initialized by the DCT. Besides, they rarely rethink to develop a dedicated and powerful feature encoder for frequency-specific representation learning. In this context, a closer look at powerful feature extraction for robust human motion prediction remains fundamental and under-explored. In this work, we propose a novel scheme that factorizes frequency representation learning into a decomposition-aggregation strategy. We hope it will develop into a strong baseline and inspire more investigations and explorations in the community.

3. Methodology

In this section, we first introduce a problem formulation and its related notations for the human motion prediction task. Then, we briefly analyze the scheme-wise difference between the conventional frequency representation learning strategy and ours. Finally, we elaborate on the technical details of the key components proposed in our scheme.

3.1. Problem Formulation

Human motion prediction system aims at forecasting the future body poses from given motion history. Mathematically, let $\mathbf{X} \in \mathbb{R}^{D\times T\times J}$ denotes an input body motion sequence at past T time steps in the D-dimensional pose space, where a skeleton at each frame contains J body joints and here D is 3. Considering $\mathbf{X}$ as a set of N body joint trajectories, $\mathbf{x}$ represents the T-frame motion trajectory of arbitrary one of N joints, where $\mathbf{x} \in \mathbf{X}$ and $N = J \times D$. The main challenge of developing a powerful motion prediction system lies in formulating an effective predictor $\mathcal{F}_{\text{pred}}$ that maps $\mathbf{X}$ to future poses $\mathbf{X}'$ in next T' frames $\mathbf{X}' = \mathcal{F}_{\text{pred}}(\mathbf{X})$ to approximate its ground-truth $\overline{\mathbf{X}}'$.

3.2. Paradigm Review

Conventional Scheme. Previous attempts on frequency representation learning for motion prediction tend to extract the embedding from the single frequency space initialized by the DCT. Mathematically, the generic formulation they followed can be summarized as:

$$\mathcal{F}_{\text{pred}}(\mathbf{X}) = \mathcal{F}_{\text{IDCT}}\Big(\mathcal{F}_{\text{enc}}\big(\mathcal{F}_{\text{DCT}}(\mathbf{X})\big)\Big), \tag{1}$$

where $\mathcal{F}_{\text{DCT}}$ denotes a DCT operation that converts body representations from the original pose space into the frequency domain. $\mathcal{F}_{\text{enc}}$ denotes feature encoding (e.g., graph neural networks). $\mathcal{F}_{\text{IDCT}}$ is an inverse DCT that recovers the pose space.

Proposed Scheme. In this work, we introduce a novel scheme that factorizes the frequency representation learning into a decomposition-aggregation strategy as:

$$\begin{aligned} &\mathcal{F}_{\text{pred}}(\mathbf{X}) = \mathcal{F}_{\text{IDCT}}\big(\mathcal{F}_{\text{enc}}(\widehat{\mathbf{X}}_1, \widehat{\mathbf{X}}_2, \cdots, \widehat{\mathbf{X}}_K)\big) \\ &\text{where } \widehat{\mathbf{X}}_k = \mathcal{F}^k_{\text{filt}}\big(\mathcal{F}_{\text{DCT}}(\mathbf{X})\big) \end{aligned} \tag{2}$$

In the first decomposition stage, K different filters $(\mathcal{F}^1_{\text{filt}}, \cdots, \mathcal{F}^K_{\text{filt}})$ unweave multi-view frequency frequency representations from an input body motion by embedding its frequency features into K different spaces. In the following aggregation stage, we propose a powerful encoder $\mathcal{F}_{\text{enc}}$ that extracts comprehensive body features from these spaces with a serious of intra-space and inter-space information aggregations. In the following sections, we introduce these stages in detail.

3.3. Decompose More: *Frequency Decomposition Unit*

Taking the historical motion $\mathbf{X}$ as the input, we first apply the discrete cosine transform (DCT) along the time axis to convert its temporal dynamics from the original pose space into the frequency space as $\widetilde{\mathbf{X}} = \mathcal{F}_{\text{DCT}}(\mathbf{X})$. Accordingly, $\mathcal{F}_{\text{DCT}}$ compacts the body joint trajectory $\mathbf{x}$ into an abstract frequency representation $\widetilde{\mathbf{x}}$. Instead of extracting feature embeddings from the single frequency space initialized by the DCT, we propose a frequency decomposition unit (FDU) to enrich the spectral encoding of a body motion with introducing $\widetilde{\mathbf{X}}$ into multiple frequency spaces.

As proven in the digital signal processing theory [5, 31], the *frequency resolution* of an input signal affects its spectral appearances, such as temporal smoothness. It inspires us to explore multi-view representations on body motions by encoding $\widetilde{\mathbf{X}}$ within different frequency resolution contexts. Furthermore, according to the theoretical analyses on frequency resolution, its influencing factors lie in two aspects: sampling window size, and sampling interval. Therefore, with different sampling window and interval configurations for different filtered signals, we introduce $\widetilde{\mathbf{X}}$ into multiple frequency resolution contexts and encode its multi-view features.

Mathematically, as shown in Eq.2, FDU develops K context-sensitive filters $(\mathcal{F}_{\text{filt}}^1, \cdots, \mathcal{F}_{\text{filt}}^K)$ on $\widetilde{\mathbf{X}}$ and embeds it into K-view representations $(\widehat{\mathbf{X}}_1, \widehat{\mathbf{X}}_2, \cdots, \widehat{\mathbf{X}}_K)$. In general, each filter $\mathcal{F}_{\text{filt}}$ deploys a trainable filtering signal $\mathbf{f}$ on $\widetilde{\mathbf{X}}$ and performs discrete signal convolution between its each node frequency feature $\widetilde{\mathbf{x}}$ and $\mathbf{f}$ as:

$$\begin{aligned} \widehat{\mathbf{X}} &= \mathcal{F}_{\text{filt}}(\widetilde{\mathbf{X}}) \\ \text{where } \ \widehat{\mathbf{x}} &= \mathbf{f} * \widetilde{\mathbf{x}}, \ \widehat{\mathbf{x}} \in \widehat{\mathbf{X}}. \end{aligned} \tag{3}$$

Specially, different $\mathcal{F}_{\text{filt}}$ develop the generic filtering formulation in Eq. 3 into different specific instantiations by choosing different sampling window w and sampling interval i for $\mathbf{f}$. Taking $\mathcal{F}_{\text{filt}}^k$ as an example, it embeds $\widetilde{\mathbf{x}}$ into $\widehat{\mathbf{x}}_k$ and sampling window and interval of $\mathbf{f}_k$ is denoted as w_k and i_k, respectively. The specific filtering operation between $\mathbf{f}_k \in \mathbb{R}^{w_k}$ and $\widetilde{\mathbf{x}} \in \mathbb{R}^T$ at time t is defined as:

$$\widehat{\mathbf{x}}_k(t) = (\mathbf{f}_k * \widetilde{\mathbf{x}})(t) = \sum_{m=1}^{w_k} \widetilde{\mathbf{x}}(t + m i_k)\mathbf{f}_k(m) \tag{4}$$

As an adaptive filter, the elements of $\mathbf{f}$ are learnable , which can be optimized by the BP algorithm. We provide multiple choices of window size w and sampling interval i as $w \in \{w_1, w_2, \ldots, w_W\}$, $i \in \{i_1, i_2, \ldots, i_I\}$. Supposing that there are totally $K = W \times I$ configuration combinations of (w, i), FDU introduces $\widetilde{\mathbf{X}}$ into K different resolution contexts. With deploying an adaptive filter within each context, FDU unweaves K-view frequency representations $\widehat{\mathbf{X}}_1, \widehat{\mathbf{X}}_2, \cdots, \widehat{\mathbf{X}}_K$ from $\widetilde{\mathbf{X}}$, developing the initial frequency space into multiple ones.

3.4. Aggregate Better: *Feature Aggregation Unit*

After embedding $\widetilde{\mathbf{X}}$ into K spaces, the issue of how to extract their features naturally arises. We propose a powerful feature aggregation unit (FAU) to extract comprehensive body features by promoting message propagation within and between these spaces. As a powerful feature encoder $\mathcal{F}_{\text{enc}}$, FAU interleaves L intra-space aggregation layers and L inter-space aggregation layers on $\widehat{\mathbf{X}}_1, \widehat{\mathbf{X}}_2, \cdots, \widehat{\mathbf{X}}_K$ to collect a richer feature $\mathbb{X}$ for robust motion prediction as:

$$\mathbb{X} = \mathcal{F}_{\text{enc}}(\widehat{\mathbf{X}}_1, \widehat{\mathbf{X}}_2, \cdots, \widehat{\mathbf{X}}_K). \tag{5}$$

Specifically, since different frequency spaces reflect different body node correlations, each intra-space aggregation layer develops K adaptive graph propagation filters to promote information flows between body nodes within K spaces. Then, inter-space aggregation layers encourage information interchanges between spaces by feature crossing. Taking $\widehat{\mathbf{X}}_k$ as an example, we introduce its operations in the l-th intra-space aggregation layer and l-th inter-space aggregation layer. Denoting its input at the l-th layer as $\widehat{\mathbf{X}}_k^{(l)}$, we first deploy an adaptive graph filter $\mathbf{A}_k^{(l)} \in \mathbb{R}^{N \times N}$ on it and perform a graph convolution between $\widehat{\mathbf{X}}_k^{(l)}$ and $\mathbf{A}_k^{(l)}$ as:

$$\widehat{\mathbf{X}}_k^{(l+1)} = \sigma(\mathbf{A}_k^{(l)} \widehat{\mathbf{X}}_k^{(l)} \Theta_k^{(l)}), \tag{6}$$

where $\sigma(\cdot)$ is an activation function, and $\Theta_k^{(l)}$ is a trainable weight matrix at layer l. In the following l-th inter-space information aggregation layer, we encourage feature crossing between neighboring spaces as:

$$\widehat{\mathbf{X}}_k^{(l+1)} = \begin{cases} \widehat{\mathbf{X}}_1^{(l+1)} + \widehat{\mathbf{X}}_2^{(l+1)} & \text{if } \ k = 1, \\ \widehat{\mathbf{X}}_{k-1}^{(l+1)} + \widehat{\mathbf{X}}_k^{(l+1)} + \widehat{\mathbf{X}}_{k+1}^{(l+1)} & \text{if } \ 1 < k < K, \\ \widehat{\mathbf{X}}_{K-1}^{(l+1)} + \widehat{\mathbf{X}}_K^{(l+1)} & \text{if } \ k = K. \end{cases} \tag{7}$$

Stacking these L intra-space and inter-space feature aggregation layers sequentially, FAU updates $\widehat{\mathbf{X}}_1, \widehat{\mathbf{X}}_2, \cdots, \widehat{\mathbf{X}}_K$ into $\widehat{\mathbf{X}}_1^L, \widehat{\mathbf{X}}_2^L, \cdots, \widehat{\mathbf{X}}_K^L$ and then concatenates them as:

$$\mathbb{X} = [\widehat{\mathbf{X}}_1^L, \widehat{\mathbf{X}}_2^L, \cdots, \widehat{\mathbf{X}}_K^L]. \tag{8}$$

To predict the future human motion in next T' time steps, we first deploy a MLP with one hidden layer to compact $\mathbb{X}$ into a T'-dimensional frequency space. Then, an inverse DCT (IDCT) converts $\mathbb{X}$ back to the pose space as:

$$\mathbf{X}' = \mathcal{F}_{\text{IDCT}}(\mathbb{X}). \tag{9}$$

3.5. Loss Function

During training, we consider ℓ_2 loss to minimize the distance between the predicted 3D motion $\mathbf{X}'$ and its ground-truth $\overline{\mathbf{X}}'$. Hence, the loss function is defined as:

$$\mathcal{L} = \frac{1}{T'N} \left\| \mathbf{X}' - \overline{\mathbf{X}}' \right\|_2, \tag{10}$$

All the trainable parameters in components are optimized end-to-end, including all node filters $\mathbf{f}$ and graph filters $\mathbf{A}$.

4. Discussion

In this section, we give in-depth analysis of our proposed decomposition-aggregation scheme. It can be intuitively interpreted as enforcing a multi-view augmentation in frequency domain. Different (w, i) pairs of filter $\mathbf{f}$ affect the spectral encoding of input $\mathbf{X}$, augmenting versatile frequency spaces for multi-view representation learning. Specifically, as for the case $w = i = 1$, it can be viewed as incorporating the original frequency feature space initialized by the DCT into ours. As varied in the experiments (section 5.3), the proposed decomposition-aggregation scheme benefits from enriching the spectral diversity of an input body motion, making it less prone to overfitting on the limited motion samples.

Table 1. Comparisons of short-term prediction on H3.6M. Results at 80ms, 160ms, 320ms, 400ms in the future are shown. The best results are highlighted in bold, and the second best are marked by underline.

scenarios	walking				eating				smoking				discussion			
millisecond	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms
DMGNN [22]	17.3	30.7	54.6	65.2	11.0	21.4	36.2	43.9	9.0	17.6	32.1	40.3	17.3	34.8	61.0	69.8
MSR-GCN [7]	12.2	22.7	38.6	45.2	8.4	17.1	33.0	40.4	8.0	16.3	31.3	38.2	12.0	26.8	57.1	69.7
PGBIG [24]	10.2	19.8	34.5	40.3	7.0	15.1	30.6	38.1	6.6	14.1	28.2	34.7	10.0	23.8	53.6	66.7
SPGSN [21]	10.1	19.4	34.8	41.5	7.1	14.9	30.5	37.9	6.7	13.8	28.0	34.6	10.4	23.8	53.6	67.1
Ours	**8.8**	**16.9**	**31.5**	**37.0**	**6.3**	**13.7**	**29.1**	**36.3**	**5.1**	**9.1**	**21.3**	**29.9**	**7.4**	**17.1**	**42.9**	**50.4**
scenarios	directions				greeting				phoning				posing			
millisecond	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms
DMGNN [22]	13.1	24.6	64.7	81.9	23.3	50.3	107.3	132.1	12.5	25.8	48.1	58.3	15.3	29.3	71.5	96.7
MSR-GCN [7]	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
PGBIG [24]	7.2	17.6	40.9	51.5	15.2	34.1	71.6	87.1	8.3	18.3	38.7	48.4	10.7	25.7	60.0	76.6
SPGSN [21]	7.4	17.1	39.8	50.3	14.6	32.6	70.6	86.4	8.7	18.3	38.7	48.5	10.7	25.3	59.9	76.5
Ours	**6.6**	**16.4**	**39.6**	**50.1**	**13.0**	**30.7**	**63.1**	**78.2**	**7.8**	**17.2**	**37.5**	**47.3**	**7.5**	**19.3**	**47.1**	**62.0**
scenarios	purchases				sitting				sittingdown				takingphoto			
millisecond	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms
DMGNN [22]	21.4	38.7	75.7	92.7	11.9	25.1	44.6	50.2	15.0	32.9	77.1	93.0	13.6	29.0	46.0	58.8
MSR-GCN [7]	14.8	32.4	66.1	79.6	10.5	22.0	46.3	57.8	16.1	31.6	62.5	76.8	9.9	21.0	44.6	56.3
PGBIG [24]	12.5	28.7	60.1	73.3	8.8	19.2	42.4	53.8	**13.9**	27.9	57.4	71.5	8.4	18.9	42.0	53.3
SPGSN [21]	12.8	28.6	61.0	74.4	9.3	19.4	42.3	53.6	14.2	27.7	56.8	70.7	8.7	18.9	41.5	52.7
Ours	**11.8**	**27.2**	**56.4**	**63.9**	**8.7**	**18.9**	**42.1**	**53.2**	**13.9**	**25.6**	**54.2**	**67.2**	**8.1**	**18.0**	**39.2**	**50.6**
scenarios	waiting				walkingdog				walkingtogether				average			
millisecond	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms
DMGNN [22]	12.2	24.2	59.6	77.5	47.1	93.3	160.1	171.2	14.3	26.7	50.1	63.2	17.0	33.6	65.9	79.7
MSR-GCN [7]	10.7	23.1	48.3	59.2	20.7	42.9	80.4	93.3	10.6	20.9	37.4	43.9	12.1	25.6	51.6	62.9
PGBIG [24]	8.9	20.1	43.6	54.3	18.8	39.3	73.7	86.4	8.7	18.6	34.4	41.0	10.3	22.7	47.4	58.5
SPGSN [21]	9.2	19.8	43.1	54.1	18.2	37.3	71.3	84.2	8.9	18.2	33.8	40.9	10.4	22.3	47.1	58.3
Ours	**8.2**	**18.4**	**41.3**	**52.1**	**14.5**	**32.7**	**63.8**	**76.0**	**7.4**	**15.2**	**30.0**	**36.4**	**9.3**	**19.7**	**41.0**	**51.1**

5. Experiments

5.1. Datasets and Model Configuration

Human 3.6m (H3.6M). H3.6M dataset [13] collects 15 types of human actions performed by 7 subjects. A pose at each time step consists of 32 body joints. Following the common-used setups [21, 24], we only use 22 key joints and downsample the frame rate from 50 fps to 25 fps. Following the recommended official evaluation protocol, the model is trained on 6 subjects and tested on the 5-th subject.

CMU Motion Capture (CMU Mocap). CMU MoCap * collects 8 general types of human actions. A pose at each time step is represented by 38 body joints. Following the common-used setups [21, 24], we only use 25 joints and downsample the frame rate to 25 fps. The division of training and testing data samples is also consistent with the recommended official evaluation protocol.

3D Pose in the Wild (3DPW). 3DPW [37] collects 51k frames with 3D human poses, including general indoor and outdoor activities. A pose at each time step contains 26 body joints, and we use 23 of them. The evaluation protocols we adopted are following the official suggestion.

Implementation Details. To validate the proposed method, we report its performances on both short-term (80∼400 ms) and long-term (560∼1000 ms) motion prediction on H3.6M, CMU Mocap and 3DPW datasets. We give 400-milliseconds history (T=10) as input and predict the future human motions in future 1 seconds (T'=25). In the FDU, we choose five window sizes, and they range from 1 to T with step of 2 (i.e., $w \in \{1, 3, 5, 7, 9\}$). As for each w, its choices of sampling interval i range from 1 to 5 with step of 2 (i.e., $i \in \{1, 3, 5\}$). Therefore, there are totally $K = 15$ filters with different configuration combinations of (w, i). In the FAU, we stack 12 graph propagation layers and 12 feature-crossing layers for intra-space and inter-space feature aggregation ($i.e.$, $L = 12$). Their feature dimensions are both 64. Finally, we implement the prediction system with PyTorch 1.4 on one NVIDIA RTX-3090Ti GPU. We use Adam optimizer [15] to train it with setting batch size as 32 and epoch number 100. The initial learning rate is 0.001 with a 0.96 decay for every two epochs.

*http://mocap.cs.cmu.edu/

5.2. Evaluation Metrics and Baselines

Evaluation Metrics. Following the standard evaluation metric commonly used in previous methods [21, 24, 26], we adopt Mean Per Joint Position Error (MPJPE) as a metric for quantitatively evaluating 3D human motion prediction. Specifically, MPJPE report the average Euclidean distance between the predicted joints and target ones over all N nodes. We report the MPJPE performance at different time steps (milliseconds) in millimeters.

Comparison Baselines. We compare the our prediction system with many state-of-the-art methods, including DMGNN [22], MSR-GCN [7], PGBIG [24], SPGSN [21]. In the following section, we analyze their prediction performances with comprehensive quantitative and qualitative comparisons, including short-term, long-term, and few-sample predictions.

Table 2. Comparisons of long-term prediction on H3.6M. Results at 560ms and 1000ms in the future are shown.

scenarios	walking		eating		smoking		discussion		directions		greeting		phoning		posing	
millisecond	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms
DMGNN [22]	73.4	95.8	58.1	86.7	50.9	72.2	81.9	138.3	110.1	115.8	152.5	157.7	78.9	98.6	163.9	310.1
MSR-GCN [7]	52.7	63.0	52.5	77.1	49.5	71.6	88.6	117.6	71.2	100.6	116.3	147.2	68.3	104.4	116.3	174.3
PGBIG [24]	48.1	56.4	51.1	76.0	46.5	69.5	87.1	118.2	69.3	100.4	110.2	143.5	65.9	102.7	106.1	164.8
SPGSN [21]	46.9	53.6	49.8	73.4	46.7	68.6	89.7	118.6	70.1	100.5	111.0	143.2	66.7	102.5	110.3	165.4
Ours	**45.2**	**50.3**	**49.0**	**71.1**	**40.6**	**59.3**	**59.5**	**92.3**	**68.1**	**97.2**	**109.4**	**141.8**	**65.1**	**96.7**	**93.3**	**149.5**
scenarios	purchases		sitting		sittingdown		takingphoto		waiting		walkingdog		walkingtogether		average	
millisecond	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms
DMGNN [22]	118.6	153.8	60.1	104.9	122.1	168.8	91.6	120.7	106.0	136.7	194.0	182.3	83.4	115.9	103.0	137.2
MSR-GCN [7]	101.6	139.2	78.2	120.0	102.8	155.5	77.9	121.9	76.3	106.3	111.9	148.2	52.9	65.9	81.1	114.2
PGBIG [24]	95.3	133.3	74.4	116.1	96.7	147.8	74.3	118.6	72.2	103.4	104.7	139.8	51.9	64.3	76.9	110.3
SPGSN [21]	96.5	133.9	75.0	116.2	98.9	149.9	75.6	118.2	73.5	103.6	102.4	138.0	49.8	60.9	77.4	109.6
Ours	**94.8**	**130.7**	**72.3**	**114.5**	**94.3**	**145.3**	**72.2**	**116.1**	**70.0**	**101.2**	**94.6**	**123.1**	**47.9**	**58.7**	**67.2**	**100.3**

Table 3. Comparisons of average prediction errors on CMU Mocap at 80ms, 160ms, 320ms, 400ms, 560ms, and 1000ms.

millisecond	80ms	160ms	320ms	400ms	560ms	1000ms
DMGNN [22]	13.6	24.1	47.0	58.8	77.4	112.6
MSR-GCN [7]	8.1	15.2	30.6	38.6	53.7	83.0
PGBIG [24]	7.6	14.3	29.0	36.6	50.9	80.1
SPGSN [21]	8.3	14.8	28.6	37.0	51.2	77.8
Ours	**6.4**	**13.9**	**27.9**	**36.0**	**50.1**	**75.4**

Table 4. Comparisons of average prediction errors on 3DPW at 200ms, 400ms, 600ms, 800ms, and 1000ms.

millisecond	200ms	400ms	600ms	800ms	1000ms
DMGNN [22]	37.3	67.8	94.5	109.7	123.6
MSR-GCN [7]	37.8	71.3	93.9	110.8	121.5
PGBIG [24]	29.3	58.3	79.8	94.4	104.1
SPGSN [21]	32.9	64.5	91.6	104.0	111.1
Ours	**26.1**	**54.2**	**72.3**	**87.2**	**94.5**

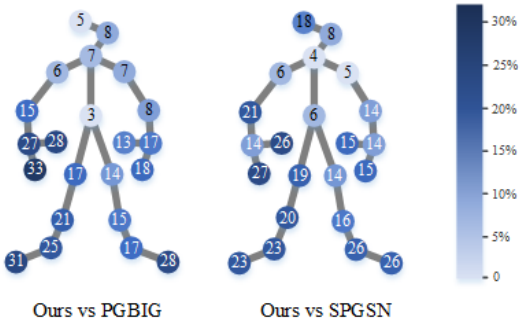

Figure 3. Performance gains on each body joint (H3.6M). Considering PGBIG and SPGSN as two baselines, we report our prediction accuracy improvements on each joint in predicting "posing".

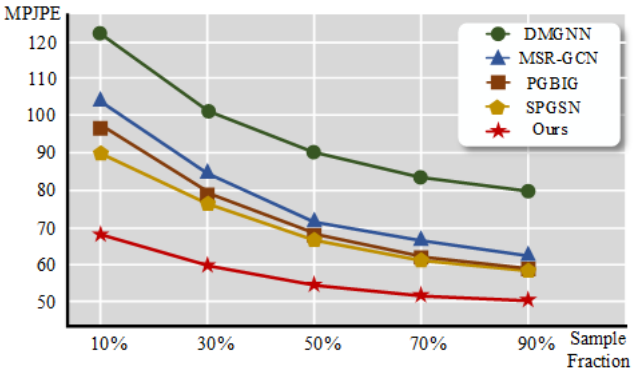

Figure 4. Prediction performances with fewer training samples on H3.6M.

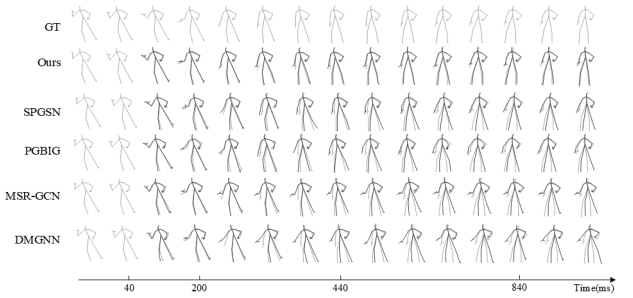

Figure 5. Visualization comparison on a H3.6M data sample for both short-term and long-term prediction.

5.3. Quantitative Comparison

Short-term and Long-term Prediction. We report the MPJPE results at different time steps to evaluate the performances of short-term and long-term motion prediction. As verified in Table 1 and Table 2, we develop a strong baseline that outperforms all state-of-the-art methods on the H3.6M dataset for both short-term and long-term prediction. Besides, Table 3 and Table 4 indicate that it also shows consistent superiority on the CMU Mocap and 3DPW datasets. To further investigate the key factor behind our prediction performance gains on the "posing", we consider PGBIG and SPGSN as two baselines and plot the prediction accuracy improvements of our method on each body joint in Figure 3. As can be seen, we achieve higher performance gains on limbs, including hands and feet. We conjecture that, compared with central body joints (e.g., abdomen), more distant joints (e.g., feet and hands) have more diversified motion frequency patterns. Our prediction system has the advantage of extracting effective frequency representations from these body joints and thus substantially facilitates robust human motion prediction.

Prediction with Fewer Samples. To investigate the robustness of our prediction system on a limited number of

Table 5. The prediction and efficiency performances of FDU with different number of filters.

Number of Filters			# Params (M)	MPJPEs			
K	w	i		80ms	160ms	320ms	400ms
1	{1}	{1}	3.1	14.6	26.0	51.3	62.7
5	{1,3,5,7,9}	{1}	5.9	11.7	22.6	46.8	56.9
9	{1,3,5}	{1,3,5}	10.9	9.9	20.4	43.7	53.9
15	{1,3,5,7,9}	{1,3,5}	15.3	9.3	19.7	41.0	51.1
25	{1,3,5,7,9}	{1,3,5,7,9}	29.6	**9.1**	**19.5**	**39.7**	**51.0**

Table 6. The prediction performances of FDU with different permutations of filters.

Permutation of Filters			MPJPEs			
setup	w	i	80ms	160ms	320ms	400ms
I	{1,3,5,7,9}	{1,3,5}	9.3	19.7	**41.0**	**51.1**
II	{1,5,3,9,7}	{1,3,5}	9.4	**19.6**	41.1	51.2
III	{1,9,5,7,3}	{1,3,5}	**9.2**	19.8	41.3	51.3
IV	{1,3,5,7,9}	{3,5,1}	9.4	19.7	41.2	**51.1**
V	{1,3,5,7,9}	{5,1,3}	9.3	19.8	**41.0**	**51.1**

samples, we first retrain these prediction methods by randomly sampling a fraction of the H3.6M training dataset and then evaluate their prediction performance on the original test dataset. As verified in Figure 4, our method significantly outperforms these baseline methods when training with 10%, 30%, 50% data samples. It indicates that the multi-view frequency representation augmentation enriches the spectral diversity of body motions and prevents the prediction system from overfitting on limited training samples.

5.4. Qualitative Comparison

We visualize the predicted human motion samples of our method, SPGSN, PGBIG, MSR-GCN, and DMGNN on the H3.6M dataset. As shown in Figure 5, compared with these baselines, our system clearly enhances the accuracy of long-term motion prediction without leading to divergent or freezing predictions. These results indicate that we develop a powerful and robust human motion prediction system for long-term prediction.

5.5. Ablation Study

We thoroughly analyze the individual components and their configurations in the final architecture. Unless stated, the performances reported in the following section are average MPJPE results on the H3.6M dataset.

5.5.1 Component Studies on FDU

Effects of the number of filters. The intention that motivates us to tune the number of filters is twofold: (I) verify the effectiveness of multiple filters; (II) investigate the optimal number of multiple filters. Therefore, as shown in Table 5, we provide five different configuration choices for filters in FDU (i.e., $K = 1, 5, 9, 15, 25$). Taking $K = 25$ as an example, we first chooses five different w. Then, five different i are chosed with each w. Particularly, in the case

Table 7. The prediction performances of FAU with different configurations of graph propagation layers.

Graph Propagation			MPJPE			
inter-space-shared	intra-space-shared	L	80ms	160ms	320ms	400ms
✗	✗	4	11.4	21.6	45.2	54.7
		12	9.3	**19.7**	**41.0**	**51.1**
		20	**9.1**	19.9	42.3	53.4
	✔	4	16.3	25.7	53.6	61.2
		12	14.1	22.3	50.2	59.3
		20	13.6	21.0	48.9	57.8
✔	✗	4	17.9	26.1	54.9	63.0
		12	16.3	25.5	52.8	61.1
		20	15.4	26.7	55.7	64.6
	✔	4	19.9	28.3	58.9	69.1
		12	18.3	27.6	59.1	70.3
		20	17.6	29.0	59.7	72.1

of $K = 1$, we focus on extracting features from single frequency space initialized by the DCT without multi-view frequency augmentations. Considering $K = 1$ as a baseline, enriching the frequency representation with multiple filters (i.e., $K>1$) brings clear performance gains, verifying the effectiveness of the FDU component. Then, we further investigate the optimal number of multiple filters. Table 5 reports that using 25 filters ($K = 25$) achieves the best prediction performance. However, compared with $K = 15$, the performance gains are limited given additional computational cost brought by more filters. Therefore, we choose $K = 15$ to balance the prediction performance with computational efficiency.

Effects of the permutation of filters. As presented in section 3.4, FAU develops a series of feature-crossing layers to aggregate the features between spaces. In this case, different deployment orders of filters in FDU will result in different permutations of inter-space feature aggregations in FAU. Therefore, we tune the deployment order of filters to investigate whether their different permutations will affect the final prediction performance. As shown in Table 6, we provide 5 different deployment orders for 15 filters (setup I $\sim$ V). Specifically, these setups can be divided into two comparison groups: setup I II, and III investigate the effects of different permutations of w; setup I, IV, and V investigate the effects of different permutations of i. As verified in Table 6, our prediction performance is insensitive to the different deployment orders between multiple filters. Therefore, for simplicity, we choose setup I as the default configuration in the final model deployment.

5.5.2 Component Studies on FAU

Effects of intra-space graph propagation. The intention that motivates us to tune the configuration of intra-space graph propagation layers is twofold: (1) investigate the effect of developing adaptive graph propagation layers into intra-space-shared or inter-space-shared ones; (2) choose the optimal number of graph propagation layers. To this end, as shown in Table 7, we first consider all the graph propagation layers from two aspects: with the inter-space-

Table 8. The prediction performances of FAU with different configurations of feature-crossing layers.

Feature Crossing		MPJPE			
Crossing Branches	Crossing Interval	80ms	160ms	320ms	400ms
1	1	12.8	23.5	46.1	58.2
2	1	9.8	20.4	42.6	53.7
	2	10.3	20.9	43.0	54.2
	3	10.2	20.9	43.2	54.4
3	1	9.3	**19.7**	**41.0**	**51.1**
	2	**9.2**	20.0	41.2	51.7
	3	9.4	20.4	41.3	51.5
4	1	9.3	**19.7**	41.1	51.3
	2	9.4	19.9	41.2	51.5
	3	9.5	20.1	41.3	51.6

shared setup, a graph propagation filter is shared among K feature spaces; with the intra-space-shared setup, a graph propagation filter is shared among L layers in one feature space. Then, we further provide L with 3 choices to explore the optimal number of graph propagation layers. Table 7 verifies that different spaces and different layers both reflect different node correlations. Therefore, we deploy $K \times L$ adaptive graph propagation layers to promote intra-space feature aggregation within K spaces. Furthermore, as Table 7 suggested, we adopt $L = 12$ as the optimal number of graph propagation layers.

Effects of inter-space feature crossing. As presented in section 3.4, we deploy a feature-crossing layer behind each intra-space graph propagation layer to collect the information from three branches and promote inter-space feature aggregation. Our intentions behind tuning the configurations of these feature-crossing layers lie in two aspects: (1) investigate the effects of the breadth of feature aggregation; (2) investigate the effects of the depth of feature aggregation. To this end, as shown in Table 8, we first change the breadth of feature aggregation by tuning the number of branches aggregated by each feature-crossing layer from 1 to 4. Particularly, we consider the one-branch feature crossing as a baseline, since it deploys a residual connection within a space without introducing inter-space information flows. Then, we further change the depth of feature aggregation by tuning the interval between two feature-crossing layers from 1 to 3. Following the suggestion in Table 8, we deploy three-branch one-interval feature-crossing layers in FAU, as described in the default model configuration.

5.5.3 Node Feature and Correlation Visualization

We visualize the node features and correlations to analyze their responses in different spaces and layers. We separately normalize the values of node features and correlations in different spaces and layers and plot them in Figure 6. The visualizations verify two key observations: (1) Different frequency spaces reflect different body joint features, enriching the spectral diversity of body motions for robust human motion prediction; (2) Different spaces and different layers in the same space reflect different node correlations, bringing better flexibility to inter-space and intra-space feature aggregation.

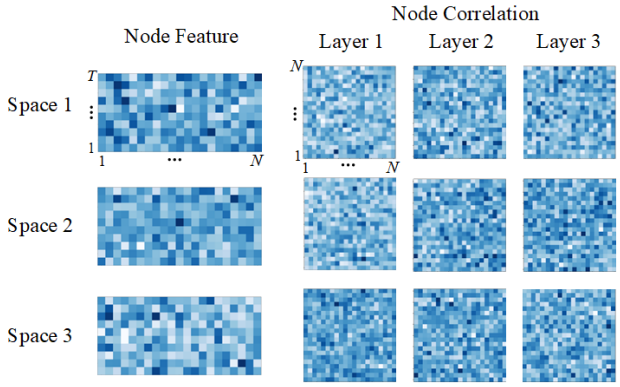

Figure 6. Visualization of node features and correlations in different spaces and layers. The darker color represents the higher response.

6. Limitation and Future Work

In this section, we analyze the limitation of our approach to inspire its further development. We consider our scheme in current version is a static model that has fixed computation flows and parameters at the inference stage. For example, the number of filters (i.e., K), the number of graph propagation layers (i.e., L), and their parameters are fixed across different input motion samples. In the future work, we will develop it into dynamic ones that can adapt its network structure and parameters to different input motion samples, leading to notable advantages in terms of accuracy, computation efficiency, adaptability, etc.

7. Conclusion

In this work, we introduce two closer looks at effective frequency representation learning for human motion prediction. We develop two powerful components that factorize the frequency representation learning into a decomposition-aggregation scheme. First, the frequency decomposition unit explores multi-view frequency representations on an input body motion to enrich its spectral encodings. Then, the feature aggregation unit promotes intra-space and inter-space message propagation to collect comprehensive body features for robust motion prediction. As evaluated on three datasets, our model outperforms state-of-the-art prediction methods by large margins. The strength of this decomposition-aggregation scheme suggests that, despite a recent surge in interest, frequency representation learning in body motions remains under-explored.

Acknowledgements. This work was supported by the National Key Research and Development Program of China under Grant No.2018AAA0102500. Xuehao Gao sincerely thanks He Wang for his feedback on the draft.

References

[1] Ijaz Akhter, Yaser Sheikh, Sohaib Khan, and Takeo Kanade. Nonrigid structure from motion in trajectory space. In *NeurIPS*, pages 41–48, 2008. 1, 3

[2] Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. A spatio-temporal transformer for 3d human motion prediction. In *3DV*, pages 565–574, 2021. 2

[3] Rasel Ahmed Bhuiyan, Md. Amiruzzaman, Nadeem Ahmed, and Md. Rashedul Islam. Efficient frequency domain feature extraction model using EPS and LDA for human activity recognition. In *ICKII*, pages 344–347, 2020. 3

[4] Matthias Blohm, Glorianna Jagfeld, Ekta Sood, Xiang Yu, and Ngoc Thang Vu. Comparing attention-based convolutional and recurrent neural networks: Success and limitations in machine reading comprehension. In *CoNLL*, pages 108–118, 2018. 2

[5] J. M. Boss, K. S. Cujia, J. Zopes, and C. L. Degen. Quantum sensing with arbitrary frequency resolution. *Science*, 356(6340):837–840, 2017. 3

[6] 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, 2020. 2, 3

[7] Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. MSR-GCN: multi-scale residual graph convolution networks for human motion prediction. In *ICCV*, pages 11447–11456, 2021. 1, 2, 5, 6

[8] Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Diverse human motion prediction via gumbel-softmax sampling from an auxiliary space. In *ACM MM*, 2022. 3

[9] Xuehao Gao, Yang Yang, and Shaoyi Du. Contrastive self-supervised learning for skeleton action recognition. In *NeurIPS 2020 Workshop on Pre-registration in Machine Learning*, volume 148, pages 51–61, 2020. 1

[10] Xuehao Gao, Yang Yang, Yimeng Zhang, Maosen Li, Jin-Gang Yu, and Shaoyi Du. Efficient spatio-temporal contrastive learning for skeleton-based 3d action recognition. *IEEE Trans. Multimedia*, 2021. 1

[11] Utkarsh Gaur, Yingying Zhu, Bi Song, and Amit K. Roy-Chowdhury. A "string of feature graphs" model for recognition of complex activities in natural videos. In *ICCV*, 2011. 1

[12] Wen Guo, Yuming Du, Xi Shen, Vincent Lepetit, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Back to MLP: A simple baseline for human motion prediction. In *WACV*, 2023. 3

[13] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(7):1325–1339, 2014. 5

[14] Angel Kennedy and Cara MacNish. An investigation of the state formation and transition limitations for prediction problems in recurrent neural networks. In *ACSC*, volume 74, pages 137–145, 2008. 2

[15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[16] Hema Swetha Koppula and Ashutosh Saxena. Anticipating human activities for reactive robotic response. In *IROS*, page 2071, 2013. 1

[17] Hema Swetha Koppula and Ashutosh Saxena. Anticipating human activities using object affordances for reactive robotic response. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(1):14–29, 2016. 1

[18] Vasileios Lefkopoulos, Marcel Menner, Alexander Domahidi, and Melanie N. Zeilinger. Interaction-aware motion prediction for autonomous driving: A multiple model kalman filtering scheme. *IEEE Robotics Autom. Lett.*, 6(1):80–87, 2021. 1

[19] Andreas M. Lehrmann, Peter V. Gehler, and Sebastian Nowozin. Efficient nonlinear markov models for human motion. In *CVPR*, pages 1314–1321, 2014. 2

[20] Baihua Li and Horst Holstein. Recognition of human periodic motion - A frequency domain approach. In *ICPR*, pages 311–314, 2002. 3

[21] 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. 1, 3, 5, 6

[22] 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*, pages 211–220, 2020. 1, 2, 5, 6

[23] Hengbo Ma, Jiachen Li, Ramtin Hosseini, Masayoshi Tomizuka, and Chiho Choi. Multi-objective diverse human motion prediction with knowledge distillation. In *CVPR*, pages 8151–8161, 2022. 1

[24] Tiezheng Ma, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Progressively generating better initial guesses towards next stages for high-quality human motion prediction. In *CVPR*, pages 6437–6446, 2022. 1, 2, 3, 5, 6

[25] Takahiro Maeda and Norimichi Ukita. Motionaug: Augmentation with physical correction for human motion prediction. In *CVPR*, pages 6417–6426, 2022. 1

[26] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *ECCV*, pages 474–489, 2020. 1, 5

[27] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. Weakly-supervised action transition learning for stochastic human motion prediction. In *CVPR*, pages 8141–8150, 2022. 1

[28] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *ICCV*, pages 9488–9496, 2019. 1, 3

[29] Ángel Martínez-González, Michael Villamizar, and Jean-Marc Odobez. Pose transformers (POTR): human motion prediction with non-autoregressive transformers. In *ICCVW*, pages 2276–2284, 2021. 2

[30] Hee-Seung Moon and Jiwon Seo. Fast user adaptation for human motion prediction in physical human-robot interaction. *IEEE Robotics Autom. Lett.*, 7(1):120–127, 2022. 1

[31] tuomas paatero and matti karjalainen. kautz filters and generalized frequency resolution: theory and audio applications. *Journal of the Audio Engineering Society*, 51(1/2):27–44, 2003. 3

[32] Brian Paden, Michal Cáp, Sze Zheng Yong, Dmitry S. Yershov, and Emilio Frazzoli. A survey of motion planning and control techniques for self-driving urban vehicles. *IEEE*

Trans. Intell. Veh., 1(1):33–55, 2016. 1

[33] Xiangbo Shu, Liyan Zhang, Guo-Jun Qi, Wei Liu, and Jinhui Tang. Spatiotemporal co-attention recurrent neural networks for human-skeleton motion prediction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(6):3300–3315, 2022. 1

[34] Theodoros Sofianos, Alessio Sampieri, Luca Franco, and Fabio Galasso. Space-time-separable graph convolutional network for pose forecasting. In *ICCV*, pages 11189–11198, 2021. 1, 2

[35] Graham W. Taylor, Geoffrey E. Hinton, and Sam T. Roweis. Modeling human motion using binary latent variables. In *NeurIPS*, pages 1345–1352, 2006. 2

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 2

[37] 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*, volume 11214, pages 614–631, 2018. 5

[38] Jiashun Wang, Huazhe Xu, Medhini Narasimhan, and Xiaolong Wang. Multi-person 3d motion prediction with multi-range transformers. In *NeurIPS*, pages 6036–6049, 2021. 1, 2, 3

[39] Jack M. Wang, David J. Fleet, and Aaron Hertzmann. Gaussian process dynamical models. In *NeurIPS*, pages 1441–1448, 2005. 2

[40] Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, and Nicu Sebe. Learning deep representations of appearance and motion for anomalous event detection. In *BMVC*, 2015. 1

[41] Jiachen Xu, Min Wang, Jingyu Gong, Wentao Liu, Chen Qian, Yuan Xie, and Lizhuang Ma. Exploring versatile prior for human motion via motion frequency guidance. In *3DV*, pages 606–616, 2021. 3

[42] Guang Yang, Wu Liu, Xinchen Liu, Xiaoyan Gu, Juan Cao, and Jintao Li. Delving into the frequency: Temporally consistent human motion transfer in the fourier space. In *ACM MM*, pages 1156–1166, 2022. 1

[43] Chongyang Zhong, Lei Hu, Zihao Zhang, Yongjing Ye, and Shihong Xia. Spatio-temporal gating-adjacency gcn for human motion prediction. In *CVPR*, pages 6447–6456, 2022. 1, 2