MotionMixer: MLP-based 3D Human Body Pose Forecasting

Arij Bouazizi[1,2*], **Adrian Holzbock**[2], **Ulrich Kressel**[1], **Klaus Dietmayer**[2] and **Vasileios Belagiannis**[3 †]
[1]Mercedes-Benz AG, Stuttgart, Germany
[2]Ulm University, Ulm, Germany
[3]Otto von Guericke University Magdeburg, Magdeburg, Germany
{arij.bouazizi, ulrich.kressel}@mercedes-benz.com, {adrian.holzbock, klaus.dietmayer}@uni-ulm.de, vasileios.belagiannis@ovgu.de

Abstract

In this work, we present *MotionMixer*, an efficient 3D human body pose forecasting model based solely on multi-layer perceptrons (MLPs). *MotionMixer* learns the spatial-temporal 3D body pose dependencies by sequentially mixing both modalities. Given a stacked sequence of 3D body poses, a spatial-MLP extracts fine-grained spatial dependencies of the body joints. The interaction of the body joints over time is then modelled by a temporal MLP. The spatial-temporal mixed features are finally aggregated and decoded to obtain the future motion. To calibrate the influence of each time step in the pose sequence, we make use of squeeze-and-excitation (SE) blocks. We evaluate our approach on Human3.6M, AMASS, and 3DPW datasets using the standard evaluation protocols. For all evaluations, we demonstrate state-of-the-art performance, while having a model with a smaller number of parameters. Our code is available at: https://github.com/MotionMLP/MotionMixer.

1 Introduction

Forecasting 3D human motion is at the core of many different applications ranging from virtual reality to autonomous driving [Wiederer *et al.*, 2020] and robotics [Gui *et al.*, 2018]. Fundamentally, the task of human motion prediction is defined as the prediction of future body poses from past ones. To model the human spatial-temporal dynamics, classic approaches adopted Hidden Markov Models [Kulić *et al.*, 2012] or Gaussian Processes [Wang *et al.*, 2007]. Despite the tangible progress in predicting periodic motion, these approaches impose strong assumptions on body pose, leading to performance degradation. Furthermore, it is difficult for these methods to forecast reliable 3D poses because of the complicated bio-mechanical kinematics. Even manually imposing expert knowledge does not help prior-based approaches to generalise to new environments and subjects.

Figure 1: Long-term predictions of *MotionMixer* for the actions *Directions* and *Posing* of the Human3.6M dataset. The first and the third line indicate the ground-truth 3D human motion. The frames on the left are the observations. The right part, shown in pink is the long-term motion prediction. One every three frames are shown. The model predictions accurately match the ground-truth body poses.

Recently, the availability of large-scale datasets, e.g. Human3.6M [Ionescu *et al.*, 2013], AMASS [Mahmood *et al.*, 2019] or 3DPW [von Marcard *et al.*, 2018], the development of human pose estimation algorithms [Belagiannis *et al.*, 2014; Bouazizi *et al.*, 2021] and the advent of deep learning methods pushed the evolution towards forecasting future 3D poses with less priors. Several learning-based approaches were proposed to tackle the problem of 3D human motion prediction. Methods like [Fragkiadaki *et al.*, 2015; Martinez *et al.*, 2017; Tang *et al.*, 2018] build upon the success of recurrent neural networks (RNNs) to better model the temporal correlation between the human body joints. Nev-

*Contact Author
†Most of this work was done while Vasileios Belagiannis was with Ulm University.

ertheless, the use of RNNs come with many drawbacks, including vanishing or exploding gradients [Fragkiadaki *et al.*, 2015], first frame discontinuity, problems in processing longer input sequences as well as shorter forecast horizons leading to stationary predictions or frozen motion [Martinez *et al.*, 2017]. More recent works [Li *et al.*, 2018; Sofianos *et al.*, 2021] propose to replace RNN models with dedicated temporal convolutional architectures. In contrast to RNNs, Convolutional Neural Networks (CNNs) maintain a hierarchical structure, enabling them to capture both spatial and temporal correlations effectively. Lately, graph convolutional networks (GCNs) have received increasing attention. Several works [Mao *et al.*, 2019; Mao *et al.*, 2020; Sofianos *et al.*, 2021; Liu *et al.*, 2021] attempted to utilize GCNs to learn fine-grained spatial relationships among joints. [Mao *et al.*, 2019] for instance, encoded the joints history in the frequency domain and proposed a GCN with learnable connectivity to predict the future motion. Although effective, these methods still need structural priors [Aksan *et al.*, 2020] or frequency transformation [Mao *et al.*, 2019] to address the inherent spatial-temporal dependency of the human motion.

While RNNs, CNNs and GCNs led to a significant performance gain in forecasting 3D human body poses, the existing methods are unnecessarily complex. In this paper, we present *MotionMixer*, the first model using exclusively MLPs to address the inherent problems of human motion. Equipped with a simple, yet effective architecture, the model aims to learn the spatial-temporal dependencies of the human body pose. Inspired by [Tolstikhin *et al.*, 2021], we propose two types of layers: one with MLPs applied independently to time steps (i.e. "mixing" the temporal information) and another with MLPs applied across body poses (i.e. "mixing" the spatial information). The interchangeable spatial and temporal mixing operations allow the model to access current and past information directly and capture both the structural and the temporal dependencies explicitly.

Our contributions are summarized as follows: 1) We propose to jointly model the spatial locations of the body joints and their temporal dependency with a spatial-temporal MLP. To the best of our knowledge, *MotionMixer* is the first 3D body pose forecasting approach based solely on MLPs. 2) We design an efficient architecture, that significantly reduces the computational cost of the pose forecasting model. 3) An extensive evaluation on three challenging large-scale datasets demonstrates state-of-the-art performance for short-term and long-term motion prediction.

2 Related Work

Recurrent-based Motion Prediction. 3D human motion prediction with RNNs has been widely studied in the last years. [Fragkiadaki *et al.*, 2015] proposed a recurrent encoder-decoder model, which incorporates nonlinear encoder and decoder networks before and after recurrent layers. A curriculum learning strategy was adopted to prevent error accumulation during the training. To better model the long-term temporal dependency, [Martinez *et al.*, 2017] incorporated a residual connection between the RNN units. To make reliable future predictions, [Tang *et al.*, 2018] proposed a motion context modeling by summarizing the historical human motion with respect to the current prediction within a recurrent prediction framework. Though these methods have also incorporated different modules into RNNs, exploring a recurrent-free backbone to address human motion prediction tasks is rarely studied.

Convolutional-based Motion Prediction. Thanks to their hierarchical structure and their effectiveness in capturing spatial and temporal correlations, there has been recently a great interest in integrating CNNs in human motion prediction. [Li *et al.*, 2018] for instance, encoded the history motion into a long-term hidden variable, which is used with a decoder to predict the future sequence. The decoder itself also has an encoder-decoder structure, with a short-term encoder and a long-term decoder. [Sofianos *et al.*, 2021] proposed a space-time-separable GCN, where the space-time graph connectivity is factored into space and time affinity matrices. [Mao *et al.*, 2019] designed a fully connected GCN to adaptively learn the spatial connectivity of the human skeletons and converted the joint trajectory to the frequency domain to handle the temporal information. [Dang *et al.*, 2021] proposed a multi-scale residual graph network with descending and ascending GCNs to extract features in both fine-to-coarse and coarse-to-fine manners. Despite the advantages in capturing long-range temporal correlations, the quite high computational cost of convolution-based approaches remains a bottleneck. Unlike these approaches, we propose an MLP-based model with lower computational complexity, that better exploits the spatial-temporal dependencies of the body pose.

Attention- and MLP-based Architectures. Inspired by the success of the self-attention mechanism [Vaswani *et al.*, 2017] in natural language processing, many works have explored its application in human motion prediction. [Mao *et al.*, 2020] proposed to capture the similarity between the current motion context and the historical motion sub-sequences in the frequency domain with the attention mechanism. [Aksan *et al.*, 2020] proposed to autoregressively learn spatial-temporal representations with decoupled temporal and spatial self-attention. The key role of self-attention is to re-weight the relative importance of each pose in the sequence with respect to all other poses. This resulted in a high computational memory overhead with increasing number of history poses. On the other end of the spectrum, there have been new works that support replacing self-attention with MLPs. MLP-Mixer [Tolstikhin *et al.*, 2021] for instance, relaxed the quadratically increasing computational memory by replacing the self-attention module with a two-layer MLP. The idea behind the Mixer architecture is to learn to separate the per-location operations and cross-location operations, allowing communication between different image patches. With a design based solely on MLPs, the model was originally developed for visual recognition tasks. However, unlike image classification, where only spatial correlations exist, there exist complicated spatial-temporal dynamics in human motion. In this work, we delve deeper and propose a new architecture based on MLPs to learn the spatial-temporal dependencies of the human body. Unlike the above-discussed approaches, we show that MLPs are effective in learning human dynamics.

3 Method

We define the human body motion as a sequence of $T_h + T_f$ consecutive frames, where each frame parameterizes the angles or 3D coordinates of the human body joints. Let $\mathbf{X}_{1:T_h} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_t, \dots \mathbf{x}_{T_h}\} \in \mathbb{R}^{3\times J\times T_h}$ be the historical motion sequence until the current time step T_h, with the 3D body pose $\mathbf{x}_t \in \mathbb{R}^{3\times J}$ and J is the number of body keypoints. Our goal is to learn the mapping that bridges the history sequence $\mathbf{X}_{1:T_h}$ to the future sequence $\mathbf{X}_{T_h+1:T_h+T_f} = \{\mathbf{x}_{T_h+1}, \mathbf{x}_{T_h+2}, \dots, \mathbf{x}_{T_h+t}, \dots \mathbf{x}_{T_h+T_f}\} \in \mathbb{R}^{3\times J\times T_f}$ with a pure MLP-based network. Below, we provide the details of our network architecture.

3.1 MotionMixer

MotionMixer is a sequence to sequence model with mainly three modules: pose embedding, spatial-temporal mixing, and pose prediction, as illustrated in Fig. 2. The pose embedding and the spatial-temporal mixing are coupled together to encode the spatial-temporal dependencies of the human body joints. The pose prediction module, which consists of two fully-connected layers decodes the future 3D motion. Given the historical sequence, each pose is first embedded by a fully-connected layer and given to repeated N *STMixer* blocks, each of which includes two MLPs with skip connections. The interaction of the body joints over time is modelled by two mixing operations within a single spatial-temporal MLP. The spatial-mixing allows the interplay between the spatial location of the joints, whereas the temporal-mixing allows the long-range interactions of the observed motion. In the pose prediction module, the outputs of the mixing are finally aggregated into a global vector and fed to an MLP to forecast the future motion. Below, we describe each module in detail.

Pose Embedding

Given the observed motion sequence $\mathbf{X}_{1:T_h}$, the skeleton of each time step $\mathbf{x}_t$ is flattened into a vector of length $K = 3 \times J$. This yields a two-dimensional tensor $\mathbf{X}_{1:T_h} \in \mathbb{R}^{T_h\times K}$ with one temporal dimension T_h and one spatial dimension K. For simplicity, we omit the subscript T_h, thus replacing $\mathbf{X}_{1:T_h}$ with $\mathbf{X}$. The flattened sequence $\mathbf{X}$ is then processed by a learnable embedding, which linearly projects each body skeleton $\mathbf{x}_t \in \mathbb{R}^K$ through a single fully-connected layer to the hidden dimension C. We refer to the output of the learnable pose embedding $\mathbf{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_t, \dots \mathbf{y}_{T_h}\} \in \mathbb{R}^{C\times T_h}$ as:

$$\mathbf{Y} = \mathbf{W}_0\mathbf{X} + \mathbf{b}_0, \tag{1}$$

where $\mathbf{W}_0 \in \mathbb{R}^{C\times T_h\times K}$ and $\mathbf{b}_0 \in \mathbb{R}^{C\times T_h}$ are weights of the fully-connected layer.

Spatial-Temporal Mixing

The proposed mixing module is motivated by the fact that the human spatial-temporal dynamics are contiguous. The spatial-temporal mixing stacks N *STMixer* blocks of identical size and structure. Each block, as shown in Fig. 2 includes two types of MLP layers: spatial-MLP and temporal-MLP, each followed by a squeeze-and-excitation block [Hu *et al.*, 2018]. The spatial-mixing aims to learn fine-grained spatial dependencies between the body joints by acting on the columns of the pose embedding $\mathbf{Y}$. Each column encodes the spatial information of one timestep. The spatial-mixing operation can be written as follows:

$$\hat{\mathbf{Y}} = \mathbf{Y} + \mathbf{W}_2\sigma(\mathbf{W}_1\text{LN}(\mathbf{Y})), \tag{2}$$

where $\mathbf{W}_1 \in \mathbb{R}^{C\times C}$, $\mathbf{W}_2 \in \mathbb{R}^{C\times C}$, $\sigma(\cdot)$ is a GELU activation function [Hendrycks and Gimpel, 2016] and LN$(\cdot)$ denotes the layer normalization [Ba *et al.*, 2016]. Driven by the fact that the human body joints contribute unequally to the forecasted motion, mixing the columns of $\mathbf{Y}$ allows the communication between different spatial pose embeddings. To enable the interchanging between the spatial and temporal domains, the spatial-mixed skeleton features $\hat{\mathbf{Y}}$ are transposed and fed to the temporal-mixing MLP. The temporal-mixing MLP, by acting on rows of $\hat{\mathbf{Y}}$ aims to learn the temporal correlation of the spatial-mixed features. The temporal-mixing operation is shared across all rows, which encode the temporal information of one body joint. Formally, this can be written as follows:

$$\widetilde{\mathbf{Y}} = \hat{\mathbf{Y}} + (\mathbf{W}_4\sigma(\mathbf{W}_3\text{LN}(\hat{\mathbf{Y}}^\top)))^\top, \tag{3}$$

where $\mathbf{W}_3 \in \mathbb{R}^{T_h\times T_h}$, $\mathbf{W}_4 \in \mathbb{R}^{T_h\times T_h}$. Each linear operator of the temporal-mixing assigns each time step as a linear combination of all frames where the linear weights depend on the frame's location. As such, the temporal information can be maintained in each mixing step, allowing the model to capture long-term dependencies by applying long-range interactions between frames.

In the motion prediction, each time step has a different contribution that is not known in advance. We introduce a squeeze-and-excitation (SE) block [Hu *et al.*, 2018] into *STMixer* to automatically regulate the input importance. The SE block, as shown in Fig. 2 is added after each mixing operation helping the network to re-weight the influence of each time step. Formally, this is defined as:

$$\begin{aligned}\hat{\mathbf{Y}} &= \mathbf{Y} + \delta(\mathbf{W}_s\sigma_R(\mathbf{W}_e(\mathbf{W}_2\sigma(\mathbf{W}_1\text{LN}(\mathbf{Y}))))), \\ \widetilde{\mathbf{Y}} &= \hat{\mathbf{Y}} + \delta(\mathbf{W}_s\sigma_R(\mathbf{W}_e(\mathbf{W}_4\sigma(\mathbf{W}_3\text{LN}(\hat{\mathbf{Y}}^\top)))^\top),\end{aligned} \tag{4}$$

where $\delta(\cdot)$ and $\sigma_R(\cdot)$ are respectively the Softmax and ReLU activation functions. The weights $\mathbf{W_s} \in \mathbb{R}^{s\times e}$ and $\mathbf{W_e} \in \mathbb{R}^{e\times s}$ are shared across the spatial and temporal mixing units. After the mixing operation, an MLP-based pose prediction learns to generate the future human motion.

Pose Prediction

Let $\widetilde{\mathbf{Y}}$ be the output features of the spatial-temporal mixing. An MLP-based decoder further propagates the mixed features for future pose forecasting. Each $\widetilde{\mathbf{y}}_t$ of $\widetilde{\mathbf{Y}}$ is projected to a vector of length T_f based on a two-layer non-linear feed-forward network. The computation of the predicted body poses $\hat{\mathbf{X}}_{T_h+1:T_h+T_f}$ is described as:

$$\hat{\mathbf{X}}_{T_h+1:T_h+T_f} = \mathbf{W}_{p2}(\sigma_R(\mathbf{W}_{p1}(\widetilde{\mathbf{Y}}) + \mathbf{b}_{p1})) + \mathbf{b}_{p2}, \tag{5}$$

where $\mathbf{W}_{p1} \in \mathbb{R}^{C\times T_f}$, $\mathbf{W}_{p2} \in \mathbb{R}^{3\times J\times T_f}$ and $\mathbf{b}_{p2} \in \mathbb{R}^{3\times J}$, $\mathbf{b}_{p2} \in \mathbb{R}^C$ are the weights of the fully-connected layers.

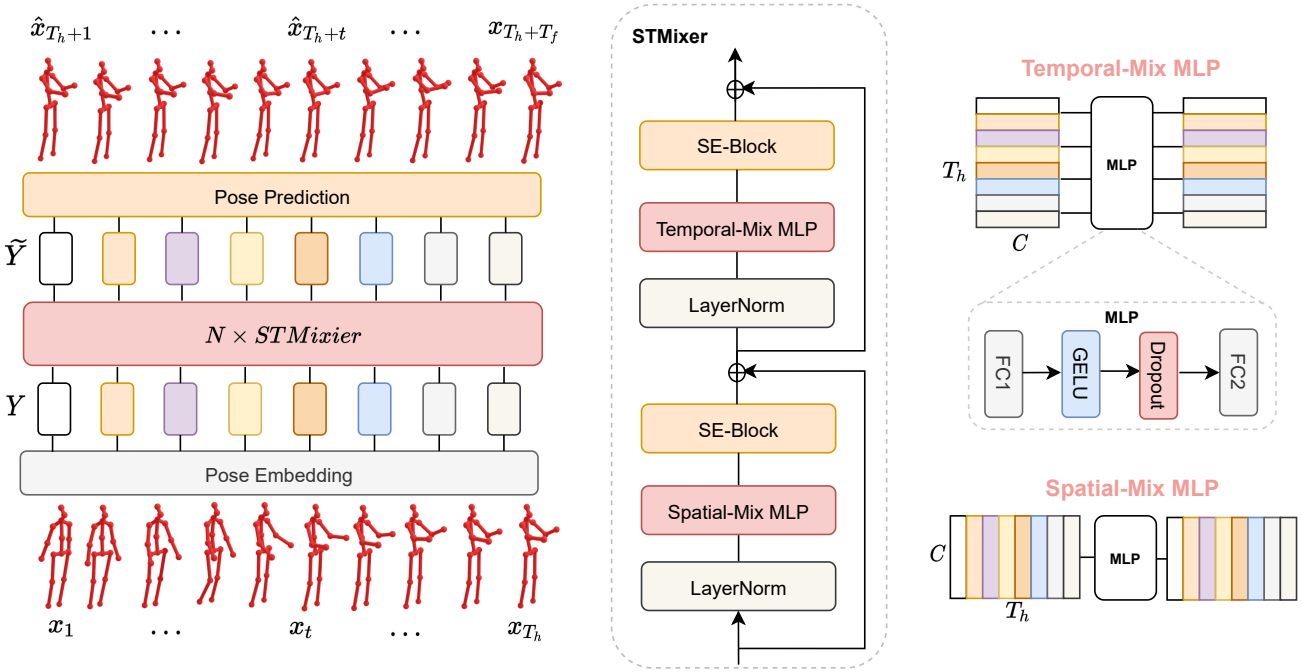

Figure 2: **Overview of the proposed MotionMixer**. It mainly consists of three modules: pose embedding, spatial-temporal mixing, and pose prediction. First, the pose embedding module linearly projects each of the past 3D body poses through a single fully-connected layer to a hidden dimension C. The learned features are then fed to N *STMixer* layers. Equipped with a spatial-MLP, a temporal-MLP and a squeeze-and-excitation (SE) block, *STMixer* aims to learn fine-grained spatial-temporal dependencies of the human motion. The mixing blocks are shown on the right. In each layer, we depict how our framework aggregates information via spatial-temporal mixing. An MLP-based body pose prediction is then applied to the mixed features to forecast the future human motion.

3.2 Training

Each joint is represented by the position displacement between two adjacent frames. We train our model by predicting future displacements, which are then added to the most recent pose to get the full-body pose sequence. The loss in terms of Mean Per Joint Position Error (MPJPE) is given by:

$$\mathcal{L}_{3D} = \frac{1}{J \times T_f} \sum_{j=1}^{J} \sum_{t=T_h+1}^{T_h+T_f} \| \Delta\hat{\mathbf{x}}_{t,j} - \Delta\mathbf{x}_{t,j} \|_2, \quad (6)$$

with $\Delta\hat{\mathbf{x}}_{t,j}$ and $\Delta\mathbf{x}_{t,j}$ denoting the predicted and ground-truth displacement of a joint j between two adjacent frames. $||\cdot||_2$ indicates the ℓ_2 norm. For the angle-based representation, the loss between the predicted joint angles and the ground truth in the exponential map representation is given by:

$$\mathcal{L}_{MAE} = \frac{1}{J \times T_f} \sum_{j=1}^{J} \sum_{t=T_h+1}^{T_h+T_f} \| \hat{\mathbf{x}}_{t,j} - \mathbf{x}_{t,j} \|_2, \quad (7)$$

where $\hat{\mathbf{x}}_{t,j}$ denotes the predicted angle of the joint j at frame t and $\mathbf{x}_{t,j}$ the corresponding ground-truth.

4 Experimental Evaluation

We evaluate our model on three large-scale public benchmarks. Below, we first introduce the datasets, the evaluation metrics and the baselines we compare with. We then present our results using 3D coordinates and joint angles.

4.1 Datasets and Metrics

Human3.6M. [Ionescu *et al.*, 2013] consists of 7 actors performing 15 different actions. The original data is transformed from exponential map to 3D joint coordinates. We consider 22 joints for forecasting the 3D body poses and 16 for the angle-based prediction. Following [Sofianos *et al.*, 2021; Mao *et al.*, 2020], we use the subject (S11) for validation, (S5) for testing, and the rest of the subjects for training.

AMASS. [Mahmood *et al.*, 2019] is a recently published dataset, which consists of 40 subjects performing the action of *walking*. Following [Sofianos *et al.*, 2021; Mao *et al.*, 2021], we select 8 datasets for training, 4 for validation, and one (BMLrub) as the test set. For each body pose, we consider 18 joints.

3DPW. The 3D Pose in the Wild dataset [von Marcard *et al.*, 2018] consists of video sequences acquired by a moving phone camera. Overall, it contains 51,000 frames of indoor and outdoor actions captured at 30Hz. We use the official test set to test the generalization of a model trained on AMASS.

Metrics. Following the standard evaluation protocol [Li *et al.*, 2018; Mao *et al.*, 2020; Sofianos *et al.*, 2021], we report the euclidean distance between the predicted and ground-truth joint angles. Due to the inherent ambiguity of the Euler-angle representation [Mao *et al.*, 2019; Mao *et al.*, 2020], we further report results in terms of 3D error. We make use of the Mean Per Joint Position Error (MPJPE) in millimeters. We provide the results at the particular frame, as well as

	Walking					Eating					Smoking					Discussion				
milliseconds	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
Res. sup [Martinez *et al.*, 2017]	23.2	40.9	61.0	66.1	79.1	16.8	31.5	53.5	61.7	98.0	18.9	34.7	57.5	65.4	102.1	25.7	47.8	80.0	91.3	131.8
convSeq2Seq [Li *et al.*, 2018]	17.7	33.5	56.3	63.6	82.3	11.0	22.4	40.7	48.4	87.1	11.6	22.8	41.3	48.9	81.7	17.1	34.5	64.8	77.6	129.3
LTD-10-25 [Mao *et al.*, 2019]	12.3	23.2	39.4	44.4	60.9	7.8	16.3	31.3	38.6	75.8	8.2	16.8	32.8	39.5	72.1	11.9	25.9	55.1	68.1	118.5
RNN-GCN [Mao *et al.*, 2020]	10.0	19.5	34.2	39.8	58.1	6.4	14.0	28.7	36.2	75.7	7.0	14.9	29.9	36.4	69.5	10.2	23.4	52.1	65.4	119.8
MSRGCN [Dang *et al.*, 2021]	12.1	22.6	38.6	45.2	63.0	8.3	17.0	33.0	40.4	77.1	8.0	16.2	31.3	38.1	71.6	11.9	26.7	57.0	69.7	117.6
MultiAttention [Mao *et al.*, 2021]	9.9	19.3	33.7	39.0	57.1	7.9	17.5	37.4	45.2	73.7	7.0	14.3	25.4	29.0	68.7	8.6	22.8	51.0	64.0	117.5
STSGCN [Sofianos *et al.*, 2021] †	10.7	16.8	29.1	38.2	51.8	6.7	11.3	22.6	31.6	52.5	7.1	11.6	22.3	30.6	50.1	9.7	16.7	33.4	45.0	78.8
GAGCN [Zhong *et al.*, 2022] †	10.3	16.1	28.8	32.4	51.1	6.4	11.5	21.7	25.2	51.4	7.1	11.8	21.7	24.3	48.7	9.7	17.1	31.4	38.9	76.9
Ours	10.8	22.4	36.5	42.4	59.9	7.7	14.0	27.3	36.1	76.6	7.1	14.0	29.1	36.8	68.5	10.2	22.5	51.0	64.1	117.4
Ours †	**7.3**	**12.9**	**23.5**	**28.6**	**49.2**	**4.3**	**8.3**	**16.9**	**20.9**	**47.4**	**4.7**	**8.8**	**17.3**	**21.4**	**45.4**	**6.4**	**13.1**	**28.6**	**35.5**	**78.0**
	Directions					Greeting					Phoning					Posing				
milliseconds	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
Res. sup [Martinez *et al.*, 2017]	21.6	41.3	72.1	84.1	129.1	31.2	58.4	96.3	108.8	153.9	21.1	38.9	66.0	76.4	126.4	29.3	56.1	98.3	114.3	183.2
convSeq2Seq [Li *et al.*, 2018]	13.5	29.0	57.6	69.7	115.8	22.0	45.0	82.0	96.0	147.3	13.5	26.6	49.9	59.9	114.0	16.9	36.7	75.7	92.9	187.4
LTD-10-25 [Mao *et al.*, 2019]	8.8	20.3	46.5	58.0	105.5	16.2	34.2	68.7	82.6	136.8	9.8	19.9	40.8	50.8	105.1	12.2	27.5	63.1	79.9	174.8
RNN-GCN [Mao *et al.*, 2020]	7.4	18.4	44.5	56.5	106.5	13.7	30.1	63.8	78.1	138.8	8.6	18.3	39.0	49.2	105.0	10.2	24.2	58.5	75.8	178.2
MSRGCN [Dang *et al.*, 2021]	8.6	19.6	43.2	53.8	100.6	16.4	36.9	77.3	93.3	-	10.1	20.7	41.5	51.2	-	12.8	29.4	66.9	85.0	-
MultiAttention [Mao *et al.*, 2021]	11.3	22.9	50.6	62.6	105.7	12.9	26.6	68.2	85.4	136.7	11.2	19.6	37.7	44.1	104.6	9.8	23.7	62.2	78.7	172.9
STSGCN [Sofianos *et al.*, 2021] †	7.4	13.5	29.2	40.9	71.0	12.4	21.7	42.1	54.5	91.6	8.2	13.7	26.8	36.6	66.1	9.9	18.0	38.2	52.6	106.4
GAGCN [Zhong *et al.*, 2022] †	7.3	12.8	30.3	34.5	69.9	11.8	20.1	40.5	48.4	87.7	8.8	13.5	25.5	28.7	66.0	10.1	17.0	35.5	45.1	99.1
Ours	8.3	18.1	43.8	53.4	105.4	12.8	33.4	62.3	82.2	136.5	10.0	20.1	37.4	51.1	104.4	11.7	23.3	62.4	79.5	174.9
Ours †	**4.4**	**9.7**	**22.5**	**29.2**	**66.5**	**8.8**	**17.7**	**36.9**	**46.2**	**93.6**	**5.6**	**10.7**	**21.9**	**27.8**	**63.4**	**6.0**	**13.1**	**30.2**	**40.1**	**99.7**
	Purchases					Sitting					Sitting Down					Taking Photo				
milliseconds	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
Res. sup [Martinez *et al.*, 2017]	28.7	52.4	86.9	100.7	154.0	23.8	44.7	78.0	91.2	152.6	31.7	58.3	96.7	112.0	187.4	21.9	41.4	74.0	87.6	153.9
convSeq2Seq [Li *et al.*, 2018]	20.3	41.8	76.5	89.9	151.5	13.5	27.0	52.0	63.1	120.7	20.7	40.6	70.4	82.7	150.3	12.7	26.0	52.1	63.6	128.1
LTD-10-25 [Mao *et al.*, 2019]	15.2	32.9	64.9	78.1	134.9	10.4	21.9	46.6	58.3	118.7	17.1	34.2	63.6	76.4	143.8	9.6	20.3	43.3	54.3	115.9
RNN-GCN [Mao *et al.*, 2020]	13.0	29.2	60.4	73.9	135.9	9.3	20.1	44.3	56.0	138.8	14.9	30.7	59.1	72.0	143.6	8.3	18.4	40.7	51.5	115.9
MSRGCN [Dang *et al.*, 2021]	14.7	32.4	66.1	79.6	-	10.5	21.9	46.2	57.8	-	16.1	31.6	62.4	76.8	-	9.8	21.0	44.5	56.3	-
MultiAttention [Mao *et al.*, 2021]	18.1	36.8	58.4	67.9	133.1	9.9	24.3	53.8	66.3	115.0	10.4	26.6	54.6	66.3	141.8	5.9	14.8	38.0	49.4	115.2
STSGCN [Sofianos *et al.*, 2021] †	11.9	21.3	41.9	54.8	93.5	9.1	15.1	29.8	39.8	75.3	14.4	23.7	41.9	53.8	94.3	8.1	14.1	29.7	41.9	76.9
GAGCN [Zhong *et al.*, 2022] †	11.9	20.7	41.8	47.6	85.1	9.3	14.4	29.6	38.5	71.1	14.1	24.8	40.0	47.4	84.1	8.5	13.9	28.8	35.1	70.0
Ours	14.6	31.3	62.8	76.1	135.1	10.0	20.9	43.7	54.5	115.7	12.0	31.4	61.4	74.5	141.1	9.0	18.9	41.0	51.6	114.6
Ours †	**8.4**	**16.9**	**34.1**	**42.7**	**88.7**	**6.5**	**11.8**	**23.6**	**29.8**	**68.9**	**10.9**	**18.8**	**35.1**	**42.6**	**89.3**	**5.5**	**10.4**	**22.1**	**27.9**	**66.6**
	Waiting					Walking Dog					Walking Together					Average				
milliseconds	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
Res. sup [Martinez *et al.*, 2017]	23.8	44.2	75.8	87.7	135.4	36.4	64.8	99.1	110.6	164.5	20.4	37.1	59.4	67.3	98.2	25.0	46.2	77.0	88.3	136.6
convSeq2Seq [Li *et al.*, 2018]	14.6	29.7	58.1	69.7	117.7	27.7	53.6	90.7	103.3	162.4	15.3	30.4	53.1	61.2	87.4	16.6	33.3	61.4	72.7	124.2
LTD-10-25 [Mao *et al.*, 2019]	12.3	23.2	39.4	44.4	108.3	7.8	16.3	31.3	38.6	146.4	8.2	16.8	32.8	39.5	65.7	11.9	25.9	55.1	68.1	112.4
RNN-GCN [Mao *et al.*, 2020]	8.7	19.2	43.4	54.9	108.2	20.1	40.3	73.3	86.3	146.9	8.9	18.4	35.1	41.9	64.9	10.4	22.6	47.1	58.3	112.1
MSRGCN [Dang *et al.*, 2021]	10.6	23.0	48.2	59.2	-	20.6	42.8	80.3	93.3	-	10.5	20.9	37.4	43.8	65.9	12.1	25.5	51.6	62.9	114.2
MultiAttention [Mao *et al.*, 2021]	9.0	22.5	55.7	71.1	105.1	29.5	54.8	100.3	119.0	141.4	8.0	17.6	33.2	42.0	63.2	11.0	23.6	49.2	60.0	110.1
STSGCN [Sofianos *et al.*, 2021] †	8.6	14.7	29.6	40.7	72.0	17.6	29.3	52.6	66.4	102.6	8.6	14.3	26.5	35.1	51.1	10.1	17.1	33.1	38.3	75.6
GAGCN [Zhong *et al.*, 2022] †	8.5	14.1	29.8	33.8	69.3	17.0	28.8	50.1	59.4	91.3	-	-	-	-	-	10.1	16.9	32.5	38.5	72.9
Ours	10.2	21.1	45.2	56.4	107.7	20.5	42.8	75.6	87.8	142.2	10.5	20.6	38.7	43.5	65.4	11.0	23.6	47.8	59.3	111.0
Ours †	**5.4**	**10.9**	**23.2**	**30.0**	**68.2**	**13.4**	**24.6**	**45.2**	**54.1**	**99.6**	**5.9**	**11.3**	**22.2**	**27.4**	**50.4**	**9.0**	**13.2**	**26.9**	**33.6**	**71.6**

Table 1: Performance comparison between different methods in terms of short-term and long-term pose prediction via mean per joint position error for each activity from the Human3.6M dataset. We provide the error results for the particular frame as well as the average over all frames. (†) indicates methods that compute the average error over all frames. All other approaches evaluate at the particular frame, where the error is measured between the predictions and ground truth at each frame. The best performance is highlighted in boldface.

the average over all frames following [Sofianos *et al.*, 2021; Zhong *et al.*, 2022]. For the particular frame evaluation, the MPJPE is measured between the predicted pose sequence and the corresponding ground truth at each frame, whereas for the average frame evaluation, the errors in all previous frames w.r.t. a considered one are computed and then averaged.

4.2 Implementation Details

MotionMixer contains three *STMixer* blocks with $C = 60$ channels. In each MLP block, a dropout layer with a rate of 0.1 is added to prevent overfitting. We use Adam [Kingma and Ba, 2014] as the optimizer. During training, the learning rate is set to 10^{-2} and decayed by a factor of 0.1 every 10 epochs. We train our model for 50 epochs with a batch size of 50 for Human3.6M and 256 for AMASS.

4.3 Baselines

Following previous works [Martinez *et al.*, 2017; Li *et al.*, 2018; Mao *et al.*, 2019], we quantitatively evaluate our proposed model on all kinds of actions against the state-of-the-art for 400ms short-term (*i.e.*, 10 frames) and 1000ms long-term (*i.e.*, 25 frames) predictions. We include nine methods with recurrent [Martinez *et al.*, 2017; Mao *et al.*, 2020], convolutional [Li *et al.*, 2018; Tang *et al.*, 2018], graph-convolutional [Mao *et al.*, 2019; Sofianos *et al.*, 2021; Dang *et al.*, 2021; Zhong *et al.*, 2022], and attention-based architectures [Mao *et al.*, 2021] in our comparison .

4.4 Results

Human3.6M. In Tab. 1, we provide the results for each activity of the Human3.6M dataset using 3D body poses. *MotionMixer* outperforms all previous methods for short-term and long-term prediction in the average-frame evaluation. In particular, we outperform the current best-performing approach [Zhong *et al.*, 2022] by at least 1.3% over all time horizons. The gain over the second-best approach [Sofianos *et al.*, 2021] ranges from 13% in the case of 400ms, up to 5% for 1000ms. In the particular frame evaluation, we reach very competitive results to the state-of-the-art approaches. Only [Mao *et al.*, 2021] outperforms *MotionMixer* in long-term prediction. Our method, however, yields larger improvements on activities with more complex dynamics such as *WalkingDog* or *Purchase*. Also on highly aperiodic actions like *Posing*, our model still produces accurate predictions.

	Average 3D								Average MAE							
milliseconds	80	160	320	400	560	720	880	1000	80	160	320	400	560	720	880	1000
Res. sup. [Martinez *et al.*, 2017]	25.0	46.2	61.4	88.3	106.3	119.4	130.0	136.6	0.36	0.67	1.02	1.15	-	-	-	-
convSeq2Seq [Li *et al.*, 2018]	16.6	33.3	77.0	72.7	90.7	104.7	116.7	124.2	0.38	0.68	1.01	1.13	1.35	1.50	1.69	1.82
MHU [Tang *et al.*, 2018]	-	-	-	-	-	-	-	-	0.39	0.68	1.01	1.13	1.14	1.28	1.46	1.57
LTD-10-25 [Mao *et al.*, 2019]	11.2	23.4	47.9	58.9	78.3	93.3	106.0	114.0	0.32	0.55	0.91	1.04	1.26	1.44	1.59	1.68
RNN-GCN [Mao *et al.*, 2020]	10.4	22.6	47.1	58.3	77.3	91.8	104.1	112.1	0.31	0.55	0.90	1.04	1.25	1.42	1.56	1.65
Motion-Attention [Mao *et al.*, 2021]	11.0	23.6	49.1	60.0	75.9	90.4	102.5	110.1	0.27	0.51	0.81	0.93	1.12	1.27	1.46	1.57
STSGCN [Sofianos *et al.*, 2021](†)	10.1	17.1	33.1	38.3	50.8	60.1	68.9	75.6	0.24	0.39	0.59	0.66	0.79	0.92	1.00	1.09
GAGCN [Zhong *et al.*, 2022](†)	10.1	16.9	32.5	38.5	50.0	-	-	72.9	0.24	0.38	**0.54**	0.65	**0.74**	-	-	**1.02**
Ours	11.0	23.6	47.8	59.3	77.8	91.4	106.0	111.0	0.29	0.54	0.81	0.94	1.20	1.30	1.40	1.57
Ours (†)	**6.9**	**13.2**	**26.9**	**33.6**	**46.1**	**56.5**	**65.7**	**71.6**	**0.20**	**0.34**	0.55	**0.63**	0.78	**0.91**	**0.99**	1.08

Table 2: Average short-term and long-term 3D and mean angle prediction errors over all actions of Human3.6M. We provide the error results for the particular frame as well as the average over all frames. (†) indicates methods that compute the average error over all frames. All other approaches evaluate at the particular frame, where the error is measured between the predictions and ground truth at each frame. The best performance is highlighted in boldface.

	AMASS-BMLrub								3DPW							
milliseconds	80	160	320	400	560	720	880	1000	80	160	320	400	560	720	880	1000
convSeq2Seq [Li *et al.*, 2018]	20.6	36.9	59.7	67.6	79.0	87.0	91.5	93.5	18.8	32.9	52.0	58.8	69.4	77.0	83.6	87.8
LTD-10-25 [Mao *et al.*, 2019]	11.0	20.7	37.8	45.3	57.2	65.7	71.3	75.2	12.6	23.2	39.7	46.6	57.9	65.8	71.5	75.5
RNN-GCN [Mao *et al.*, 2020]	11.3	20.7	35.7	42.0	51.7	58.6	63.4	67.2	12.6	23.1	39.0	45.4	56.0	63.6	69.7	73.7
Motion-Attention [Mao *et al.*, 2021]	11.0	20.3	35.0	41.2	50.7	57.4	61.9	65.8	12.4	22.6	38.1	44.4	54.7	62.1	67.9	71.8
STSGCN [Sofianos *et al.*, 2021]	10.0	12.5	21.8	24.5	31.9	38.1	42.7	45.5	8.6	12.8	21.0	24.5	30.4	35.7	39.6	42.3
Ours	**6.6**	**10.3**	**18.0**	**21.9**	**28.8**	**33.6**	**38.8**	**41.6**	**7.4**	**11.4**	**19.3**	**22.8**	**29.3**	**34.6**	**39.0**	**42.1**

Table 3: Short-term and long-term prediction of 3D body poses on AMASS-BMLrub (left) and 3DPW (right). All results are in millimeters. The best performance is highlighted in boldface.

Fig. 1 illustrates the future predictions of the *Posing* action. The predicted skeletons accurately match the ground-truth body poses. This demonstrates the effectiveness of spatial-temporal mixing in learning fine-grained motion patterns. In Tab. 2, we additionally provide the results over all actions using respectively the 3D body poses and the joint angles. Despite the inherent ambiguity of the angle-based representation, our method outperforms the compared methods in short-term and yields the lowest angle error of 0.2 at 80ms. In long-term prediction, we reach comparable results with previous approaches.

AMASS & 3DPW. The results of short-term and long-term prediction in 3D on AMASS and 3DPW are shown in Tab. 3. *MotionMixer* gets the best average error at all short-term forecast times on the AMASS dataset. For long-term prediction, our method consistently outperforms all previous approaches. The performance gain ranges from 4% for 1000ms up to 36% for 80ms, which further shows the benefits of the proposed spatial-temporal mixing. We further test the generalization of a model trained on AMASS on 3DPW. Without any fine-tuning, our approach outperforms previous approaches at different forecast times and can therefore better generalize to complex outdoor environments.

4.5 Ablation Studies

Model Architecture. We first study the influence of individual components of the proposed method through different ablation studies. Specifically, we report the impact of the number of layers N on the motion prediction. In Fig. 3, we show the average prediction errors at different forecast times on the Human3.6M dataset. *MotionMixer* yields the best average error with $N = 3$. Stacking more than three layers did not empirically improve the performance. To verify the impact of predicting the pose displacement instead of the 3D body pose, we train our model directly with 3D joints. With a $5mm$ performance gain, our model takes advantage of the pose displacement representation. This is reasonable, since such transformation may help the network focus more on motion patterns rather than the appearance of the body pose, hence, generalizing better to new environments and subjects.

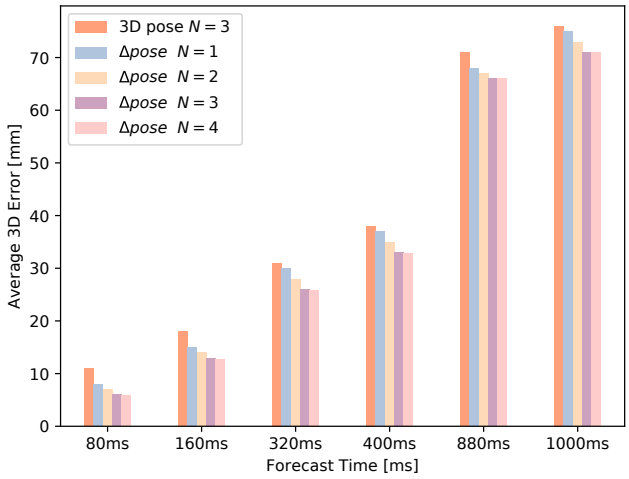

Figure 3: Comparison of average 3D error in mm over all actions of the Human3.6M dataset at different prediction times.

Spatial-Temporal Mixing. To demonstrate the effect of the *STMixer*, we train the *Spatial-Mix MLP* and the *Temporal-Mix MLP* independently for short-term and long-term prediction and report the results in Tab. 4. By removing the temporal or the spatial mixing, the error increases at 1000ms by 4% and 6%, respectively. The best results are achieved when simultaneously mixing the body pose in time and space. This

is expected since the human spatial-temporal dynamics are interleaved. We also empirically evaluate the effect of the squeeze-and-excitation (SE) blocks, which shed new light on LSTMs hidden-state weighting by giving higher importance to influential time steps. With a 2mm performance gain over all time horizons, the SE-blocks help the network re-calibrate the influence of each pose in the sequence and predict more accurate motion patterns.

	Human3.6M							
milliseconds	80	160	320	400	560	720	880	1000
Spatial-Mix MLP	12.1	17.5	32.6	37.4	51.6	61.9	70.1	77.9
Temporal-Mix MLP	10.5	15.2	30.5	36.8	49.5	59.3	68.6	74.0
STMixer w/o SE-Block	8.5	14.5	29.1	35.5	48.3	58.6	67.8	73.2
STMixer	**6.9**	**12.2**	**26.9**	**33.6**	**46.1**	**56.5**	**65.7**	**71.6**

Table 4: Influence of different parts of the *STMixer* on the performance. "SE-Block" denotes the squeeze-and-excitation blocks. The best results are shown in bold.

Computational Complexity. We also evaluate the trade-off between the model's computational cost and performance. The results are shown in Tab. 5. We report the number of parameters and an estimate of the floating operations FLOPs to predict 25 frames (1000ms). We compare our model with the current best approaches. In comparison to [Mao *et al.*, 2021], *MotionMixer* reaches nearly the same performance with only 1.4% of the parameters. We outperform [Sofianos *et al.*, 2021] and [Dang *et al.*, 2021] respectively by 4% and 3%, while using only 40% and 0.5% of the parameters.

Model	Parameters	$\approx$ FLOPs	Average 3D
Motion-Attention [Mao *et al.*, 2021]	3.4M	-	110.1
MSRGCN [Dang *et al.*, 2021]	6.3M	192.4M	114.2
STSGCN [Sofianos *et al.*, 2021] †	57.5k	7.1M	75.6
Ours $N = 1$	12.2k	1.5M	117.3
Ours $N = 2$	18.2k	1.8M	115.5
Ours $N = 3$	30.2k	2.1M	111.0
Ours $N = 1$ †	12.2k	1.5M	75.6
Ours $N = 2$ †	18.2k	1.8M	74.8
Ours $N = 3$ †	30.2k	2.1M	71.6

Table 5: Computational complexity analysis. (†) indicates results with the average error over all frames.

4.6 Limitations

In addition to the qualitative results in Fig.1, we examine some failure cases of *MotionMixer*. Fig. 5 illustrates an example of the predicted skeletons for the *WalkDog* action. As can be seen, the last three frames do not match the ground-truth poses. This failure is also common for previous methods [Mao *et al.*, 2021; Sofianos *et al.*, 2021] since various actions in Human3.6M are performed in different arts in the training dataset. In addition, the human motion is highly uncertain. A sequence of past poses may imply various possible futures. Thus, predicting the inter-joint and inter-frame dependencies in long-term becomes even more complex.

5 Conclusion

In this work, we presented an MLP-based pose forecasting approach that effectively exploits the spatial-temporal dependencies of the 3D human body pose. By learning to mix features across the spatial and temporal domains, our method improved the state-of-the-art for short-term and long-term forecasting on three large-scale benchmark datasets. Enhanced by squeeze-and-excitation (SE) blocks, which aim to calibrate the influence of each time step in the pose sequence, our model has much less parameters than current best-performing approaches.

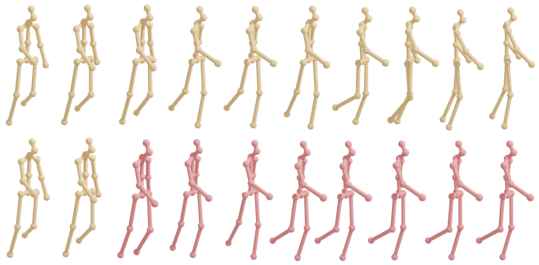

Figure 4: Example of the failure cases of our 3D pose forecasting approach. The first line indicates the ground-truth 3D human motion. The frames on the left are the observations. The right part, shown in pink is the predicted future motion.

Acknowledgments

The research leading to these results is funded by the German Federal Ministry for Economic Affairs and Energy within the project "KI Delta Learning" (Förderkennzeichen 19A19013A). The authors would like to thank the consortium for the successful cooperation.

References

[Aksan *et al.*, 2020] Emre Aksan, Peng Cao, Manuel Kaufmann, and Otmar Hilliges. A spatio-temporal transformer for 3d human motion prediction. *arXiv preprint arXiv:2004.08692*, 2020.

[Ba *et al.*, 2016] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[Belagiannis *et al.*, 2014] Vasileios Belagiannis, Christian Amann, Nassir Navab, and Slobodan Ilic. Holistic human pose estimation with regression forests. In *International Conference on Articulated Motion and Deformable Objects*, pages 20–30. Springer, 2014.

[Bouazizi *et al.*, 2021] Arij Bouazizi, Julian Wiederer, Ulrich Kressel, and Vasileios Belagiannis. Self-supervised 3d human pose estimation with multiple-view geometry. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8. IEEE, 2021.

[Dang *et al.*, 2021] Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Msr-gcn: Multi-scale residual graph convolution networks for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11467–11476, 2021.

[Fragkiadaki *et al.*, 2015] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4346–4354, 2015.

[Gui *et al.*, 2018] Liang-Yan Gui, Kevin Zhang, Yu-Xiong Wang, Xiaodan Liang, José MF Moura, and Manuela Veloso. Teaching robots to predict human motion. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 562–567. IEEE, 2018.

[Hendrycks and Gimpel, 2016] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[Hu *et al.*, 2018] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[Ionescu *et al.*, 2013] 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 transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Kulić *et al.*, 2012] Dana Kulić, Christian Ott, Dongheui Lee, Junichi Ishikawa, and Yoshihiko Nakamura. Incremental learning of full body motion primitives and their sequencing through human motion observation. *The International Journal of Robotics Research*, 31(3):330–345, 2012.

[Li *et al.*, 2018] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional sequence to sequence model for human dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5226–5234, 2018.

[Liu *et al.*, 2021] Zhenguang Liu, Pengxiang Su, Shuang Wu, Xuanjing Shen, Haipeng Chen, Yanbin Hao, and Meng Wang. Motion prediction using trajectory cues. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13299–13308, 2021.

[Mahmood *et al.*, 2019] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5442–5451, 2019.

[Mao *et al.*, 2019] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9489–9497, 2019.

[Mao *et al.*, 2020] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *European Conference on Computer Vision*, pages 474–489. Springer, 2020.

[Mao *et al.*, 2021] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Multi-level motion attention for human motion prediction. *International Journal of Computer Vision*, pages 1–23, 2021.

[Martinez *et al.*, 2017] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2891–2900, 2017.

[Sofianos *et al.*, 2021] Theodoros Sofianos, Alessio Sampieri, Luca Franco, and Fabio Galasso. Space-time-separable graph convolutional network for pose forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11209–11218, 2021.

[Tang *et al.*, 2018] Yongyi Tang, Lin Ma, Wei Liu, and Wei-Shi Zheng. Long-term human motion prediction by modeling motion context and enhancing motion dynamics. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 935–941. International Joint Conferences on Artificial Intelligence Organization, 7 2018.

[Tolstikhin *et al.*, 2021] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Peter Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[von Marcard *et al.*, 2018] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 601–617, 2018.

[Wang *et al.*, 2007] Jack M Wang, David J Fleet, and Aaron Hertzmann. Gaussian process dynamical models for human motion. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):283–298, 2007.

[Wiederer *et al.*, 2020] Julian Wiederer, Arij Bouazizi, Ulrich Kressel, and Vasileios Belagiannis. Traffic control gesture recognition for autonomous vehicles. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10676–10683. IEEE, 2020.

[Zhong *et al.*, 2022] Chongyang Zhong, Lei Hu, Zihao Zhang, Yongjing Ye, and Shihong Xia. Spatial-temporal gating-adjacency gcn for human motion prediction. *arXiv preprint arXiv:2203.01474*, 2022.