MotionMixer: MLP-based 3D Human Body Pose Forecasting

Arij Bouazizi\textsuperscript{1,2*}, Adrian Holzbock\textsuperscript{2}, Ulrich Kressel\textsuperscript{1}, Klaus Dietmayer\textsuperscript{2} and Vasileios Belagiannis\textsuperscript{3}†

\textsuperscript{1}Mercedes-Benz AG, Stuttgart, Germany
\textsuperscript{2}Ulm University, Ulm, Germany
\textsuperscript{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 \textit{MotionMixer}, an efficient 3D human body pose forecasting model based solely on multi-layer perceptrons (MLPs). \textit{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 \cite{wiederer2020forecasting} and robotics \cite{gui2018}. 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 \cite{kuli2012} or Gaussian Processes \cite{wang2007}. 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.

\*Contact Author

\textsuperscript{†}Most of this work was done while Vasileios Belagiannis was with Ulm University.

Figure 1: Long-term predictions of \textit{MotionMixer} for the actions \textit{Directions} and \textit{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 \cite{ionescu2013}, AMASS \cite{mahmood2019} or 3DPW \cite{vonmarcard2018}, the development of human pose estimation algorithms \cite{belagiannis2014, bouazizi2021} 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 \cite{fragkiadaki2015, martinez2017, tang2018} build upon the success of recurrent neural networks (RNNs) to better model the temporal correlation between the human body joints. Nev-
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. 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 $X_{1:T_h} = \{x_1, x_2, \ldots, x_t, \ldots, x_{T_h}\} \in \mathbb{R}^{3 \times J \times T_h}$ be the historical motion sequence until the current time step $T_h$, with the 3D body pose $x_t \in \mathbb{R}^{3 \times J}$ and $J$ is the number of body keypoints. Our goal is to learn the mapping that bridges the historical motion sequence until the current time step $T_h$, thus replacing the forward network. The computation of the predicted body pose embedding $Y$ is projected to the future sequence $\hat{X}_{T_h+1:T_h+T_f} = \{\hat{x}_{T_h+1}, \hat{x}_{T_h+2}, \ldots, \hat{x}_{T_h+1}, \hat{x}_{T_h+2}, \ldots, \hat{x}_{T_h+1}, \hat{x}_{T_h+2}, \ldots, \hat{x}_{T_h+1}\} \in \mathbb{R}^{3 \times J \times T_f}$ 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 is encoded by MLPs and fed to the temporal-mixing 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 finitely 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 $X_{1:T_h}$, the skeleton of each time step $x_t$ is flattened into a vector of length $K = 3 \times J$. This yields a two-dimensional tensor $X_{1:T_h} \in \mathbb{R}^{T_h \times K}$ with one temporal dimension $T_h$ and one spatial dimension $K$. For simplicity, we omit the subscript $T_h$, thus replacing $X_{1:T_h}$ with $X$. The flattened sequence $X$ is then processed by a learnable embedding, which linearly projects each body skeleton $x_t \in \mathbb{R}^K$ through a single fully-connected layer to the hidden dimension $C$. We refer to the output of the learnable pose embedding $Y = \{y_1, \ldots, y_1, \ldots, y_{T_h}\} \in \mathbb{R}^{C \times T_h}$ as:

$$Y = W_0X + b_0,$$

where $W_0 \in \mathbb{R}^{C \times T_h \times K}$ and $b_0 \in \mathbb{R}^{C \times T_h}$ 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 $Y$. Each column encodes the spatial information of one timestep. The spatial-mixing operation can be written as follows:

$$\hat{Y} = Y + W_2\sigma(W_1LN(Y)),$$

where $W_1 \in \mathbb{R}^{C \times C}, W_2 \in \mathbb{R}^{C \times C}, \sigma(\cdot)$ 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 $Y$ allows the communication between different spatial pose embeddings. To enable the interchanging between the spatial and temporal domains, the spatial-mixed skeleton features $\hat{Y}$ are transposed and fed to the temporal-mixing MLP. The temporal-mixing MLP, by acting on rows of $\hat{Y}$ 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:

$$\bar{Y} = \hat{Y} + (W_4\sigma(W_3LN(\hat{Y}^T)))^T,$$

where $W_3 \in \mathbb{R}^{T_h \times T_h}, W_4 \in \mathbb{R}^{T_h \times T_h}$. 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:

$$\tilde{Y} = Y + \delta(W_5\sigma(R(\hat{Y}^T)))^T,$$

where $\delta(\cdot)$ and $\sigma(R(\cdot))$ are respectively the Softmax and ReLU activation functions. The weights $W_{5e} \in \mathbb{R}^{e \times e}$ and $W_{5e} \in \mathbb{R}^{e \times e}$ 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 $\bar{Y}$ be the output features of the spatial-temporal mixing. An MLP-based decoder further propagates the mixed features for future pose forecasting. Each $\tilde{y}_i$ of $\bar{Y}$ 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 $\hat{X}_{T_h+1:T_h+T_f}$ is described as:

$$\hat{X}_{T_h+1:T_h+T_f} = \tilde{W}_{p2}(\sigma(\tilde{W}_{p1}(\bar{Y}) + b_{p1})) + b_{p2},$$

where $\tilde{W}_{p1} \in \mathbb{R}^{C \times T_f}, \tilde{W}_{p2} \in \mathbb{R}^{3 \times J \times T_f}$ and $b_{p2} \in \mathbb{R}^{3 \times J}$, $b_{p2} \in \mathbb{R}^{C}$ 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:

$$L_{3D} = \frac{1}{J \times T_f} \sum_{j=1}^{J} \sum_{t=T_h+1}^{T_f} \| \Delta \hat{x}_{t,j} - \Delta x_{t,j} \|_2,$$

with $\Delta \hat{x}_{t,j}$ and $\Delta x_{t,j}$ denoting the predicted and ground-truth displacement of a joint $j$ between two adjacent frames. $\| : \|_2$ indicates the $\ell_2$ 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:

$$L_{MAE} = \frac{1}{J \times T_f} \sum_{j=1}^{J} \sum_{t=T_h+1}^{T_f} \| \hat{x}_{t,j} - x_{t,j} \|_2,$$

where $\hat{x}_{t,j}$ denotes the predicted angle of the joint $j$ at frame $t$ and $x_{t,j}$ 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...
the average over all frames following [Sofianos et al., 2022] as the optimizer. During training, the learning rate is set to 0.1 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 526 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.; 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.
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.

<table>
<thead>
<tr>
<th>Miliseconds</th>
<th>AMASS-BMLrub</th>
<th>3DPW</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>80 160 320 400 560 720 880 1000</td>
<td>80 160 320 400 560 720 880 1000</td>
</tr>
<tr>
<td>Res. sup. [Martinez et al., 2017]</td>
<td>20.6 36.9 59.7 67.6 79.0 87.0 91.5 93.5</td>
<td>18.8 32.9 52.0 58.8 69.4 77.0 83.6 87.8</td>
</tr>
<tr>
<td>convSeq2Seq [Li et al., 2018]</td>
<td>11.0 20.7 37.8 45.3 57.2 65.7 71.3 75.2</td>
<td>12.6 23.2 39.7 46.6 57.9 65.8 71.5 75.5</td>
</tr>
<tr>
<td>LTD-10-25 [Mao et al., 2019]</td>
<td>11.0 20.7 37.8 45.3 57.2 65.7 71.3 75.2</td>
<td>12.6 23.2 39.7 46.6 57.9 65.8 71.5 75.5</td>
</tr>
<tr>
<td>RNN-GCN [Mao et al., 2020]</td>
<td>11.3 20.7 35.7 42.0 51.7 58.6 63.4 67.2</td>
<td>12.6 23.1 39.0 45.4 56.0 63.6 69.7 73.7</td>
</tr>
<tr>
<td>Motion-Attention [Mao et al., 2021]</td>
<td>11.0 20.3 35.0 41.2 50.7 57.4 61.9 65.8</td>
<td>12.4 22.6 38.1 44.4 54.7 62.1 67.9 71.8</td>
</tr>
<tr>
<td>STSGCN <a href="i">Soofianos et al., 2021</a></td>
<td>10.0 12.5 21.8 24.5 31.9 38.1 42.7 45.5</td>
<td>8.6 12.8 21.0 24.5 30.4 35.7 39.9 42.3</td>
</tr>
<tr>
<td>Ours</td>
<td>6.6 10.3 18.0 21.9 28.8 33.6 38.8 41.6</td>
<td>7.4 11.4 19.3 22.9 29.3 34.6 39.8 42.1</td>
</tr>
</tbody>
</table>

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 shows 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.
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.

### 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.

#### 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


