Intention-based Long-Term Human Motion Anticipation

Julian Tanke*, Chintan Zaveri*, Juergen Gall
University of Bonn
{tanke|gall}@iai.uni-bonn.de

Abstract

Recently, a few works have been proposed to model the uncertainty of the future human motion. These works do not forecast a single sequence but multiple sequences for the same observation. While these works focused on increasing the diversity, this work focuses on keeping a high quality of the forecast sequences even for very long time horizons of up to 30 seconds. In order to achieve this goal, we propose to forecast the intention of the person ahead of time. This has the advantage that the generated human motion remains goal oriented and that the motion transitions between two actions are smooth and highly realistic. We furthermore propose a new quality score for evaluation that correlates better with human perception than other metrics. The results and a user study show that our approach forecasts multiple sequences that are more plausible compared to the state-of-the-art.

1. Introduction

Anticipating human motion is highly relevant for many interactive activities such as sports, manufacturing, or navigation [25] and significant progress has been made in forecasting human motion [8, 9, 10, 11, 15, 17, 23, 26, 35]. Most progress has been made in anticipating motion over a short time horizon of around half a second. However, these methods fail when anticipating longer time horizons as they either produce unrealistic poses or the motion freezes. Another issue that occurs when the time horizon gets larger is the fact that there are more than one future sequence that are plausible for a single observed sequence of human motion as it is shown in Figure 2. Going from a short time horizon of less than one second to a larger time horizon of a few seconds therefore imposes the following challenges: (a) How can we model the uncertainty of the future? (b) How can we ensure that the motion remains plausible? (c) How can we measure the quality of methods that generate more than one sequence?

Handling the uncertainty of the future has been so far only addressed in very few recent works [4, 28, 37] for human motion anticipation. These approaches are able to forecast diverse sequences from the same observation, but the quality of the sequences decreases for longer time horizons beyond 1 second. In this work, we also propose a network that generates multiple sequences as shown in Figure 2, but our goal is to generate more plausible sequences for time horizons of multiple seconds. In order to achieve this goal, we not only model the human motion but also the intention of the person as illustrated in Figure 1. In fact, human motion anticipation depends on two factors, namely the past motion and the intention. The latter, which is ignored by existing works, is very important for longer sequences since a motion without a goal is perceived as random and unrealistic. We therefore model the intention as discrete actions and propose to forecast the intention as well as the human motion. The key aspect is that our model forecasts the intention ahead of time and that the forecast human motion is conditioned on the past motion and on the forecast intention as shown in Figure 1.

It, however, remains an open issue how methods that generate multiple sequences are best compared. Recent works suggest to evaluate both the quality of the generated motion as well as the sample diversity. While diversity is commonly measured by using the average pairwise distance between multiple generated predictions [4, 37], measuring the quality is still an open problem. In [37], for instance, multiple sequences are forecast but only the error of the sequence with the lowest error is reported. Such measures are misleading since they evaluate only one forecast sequence while the other sequences can be implausible. In fact, we show in the supplementary material that this measure can be easily fooled by a simple but unrealistic baseline approach, yielding competitive results on clearly unrealistic motion. In [4], pre-trained skeleton-based action classifiers are used to compute the inception score and a quality score over all generated sequences. While the inception score is an indicator for plausibility it is highly depended on the model. The authors did not make the models publicly available, making an evaluation very difficult. Normalized Power Spectrum Similarity [10] (NPSS) evaluates sequences in the power spectrum to account for frequency shifts that cannot be cap-

*equal contribution

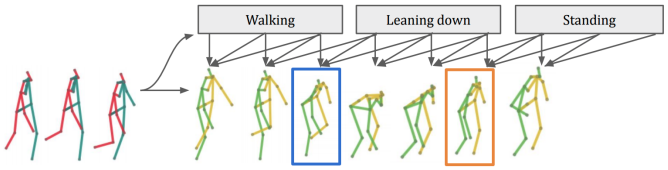


Figure 1: Intention-based human motion anticipation. Given a human motion input sequence (red-blue skeletons), our method forecasts the intention of the person ahead of time (top row) and the human motion (green-yellow skeletons) conditioned on the previous motion and the future intention. This allows not only long-term forecasting but also realistic transitions between different actions. For example, the blue and orange boxes show how the motion already prepares for the next action *leaning down* or *standing*, respectively.

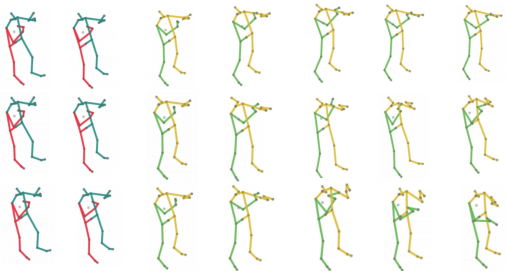


Figure 2: Our approach forecasts multiple sequences of plausible future human motion for long time horizons. Each row shows a different prediction of three seconds made by our model, given the same input sequence (*Discussion* from Human3.6M [14]). The red-blue skeletons represent the ground-truth input while the green-yellow skeletons are model predictions. During the first second, the model generates fairly consistent human poses but it starts to generate diverse but realistic human motion after 1 second. The qualitative results are best viewed in the supplementary video.

tured by MSE. However, NPSS is uni-modal as it compares the motion to a single ground-truth sequence. We therefore propose a new complementary similarity score that measures the normalized directional motion similarity between motion snippets of forecast and real motions that have the same semantic meaning. The measure has the advantage that it takes the multi-modality of human motion into account and that it correlates much better with human perception than NPSS.

Our contribution is therefore two-fold:

- We propose a novel quality score for long-term human motion anticipation that measures the plausibility of multiple generated sequences and that correlates better with human perception than other metrics.
- We propose a novel approach for human motion forecasting that forecasts the intention of a person ahead of time and that is capable of generating multiple plausible future sequences for long time horizons.

2. Related Work

Human Motion Anticipation: In recent years, deep neural networks [12, 6, 9] have been used to synthesize and anticipate human motion. Auto-regressive methods [9, 23, 38] model first-order motion derivatives using the sequence-to-sequence model [32] popularized in machine translation. QuaterNet [27] replaces the exponential map representation with quaternions, which do not suffer from common 3D rotational problems such as gimbal locks. Furthermore, the authors show that the model can generate cyclic motion for very long time horizons when frame-wise user control is provided, similar to [16, 18, 19, 31]. A similar approach is utilized in Hierarchical Motion Recurrent networks [20] and Structured Prediction [2] where novel RNN structures are proposed which better represent skeletal structures. Graph-convolutional neural networks [22] can be utilized to learn human motion in trajectory space, using Discrete Cosine Transform, rather than in pose space. Highly competitive results are achieved by recent attention-based models [21]. The idea of utilizing discrete representations for human poses was first proposed in [33] where a conditional restricted Boltzmann machine (RBM) is used as a generative model for synthesizing or filling missing pose data. While RBMs or Deep Belief Networks learn a binary representation of the data, they are nowadays outperformed by other approaches that learn continuous hidden states such as RNNs. Recently, an adversarial generative grammar model was proposed in [28] for future prediction where stochastic production rules are learned jointly with its latent non-terminal representations. By selecting various production rules during inference, many different forecast outputs can be generated. However, our experiments show that the model does not forecast long term natural human motion. Recently, models based on adversarial training gained some attention: Convolutional sequence-to-sequence models [17] utilize a convolutional encoder-decoder structure with the adversarial loss to prevent overfitting. The adversarial geometry-aware encoder-decoder [11] utilizes two adversarial losses: one to tackle motion discontinuity, which is a common problem in previous models, and one to ensure that realistic motion is generated. On top of that, the geodesic instead of the Euclidean distance is used as reconstruction loss. MotionGAN [29] frames human motion anticipation as an inpainting problem. Wang et al. [35] combine an adversarial loss with rein-

forcement learning to forecast realistic poses. Early works on multi-modal human motion anticipation utilize stochastic conditional variational autoencoders [30, 7]. Recently, novel sampling methods [4, 37] for conditional variational autoencoders were proposed for multi-modal human motion anticipation. While Mix-and-Match [4] randomly perturbs the hidden state to increase stochasticity, DLow [37] maps a random variable to a latent code. It employs a two-stage approach by first learning a conditional variational autoencoder and then the mapping.

Human Motion Evaluation: Evaluating complex multivariate time series with a high degree of stochasticity, such as human motion, remains a challenging research problem. The simplest approaches calculate the Euclidean distance [12, 15, 23] to a target sequence independently for each time step, which works well for very short time horizons ($< 0.5s$). However, frame-wise distances completely ignore motion dynamics and forecasting only the last pose results in competitive results [23]. To address these challenges, frequency-based metrics have been proposed. Frequency-based methods such as NPSS [10] incorporate motion information, but they accumulate it over the entire sequence. On top of that, distances in the frequency domain are difficult to interpret and make it hard to pinpoint when a motion can still be considered as realistic or not. In [4] the inception score [13] is adapted by training a model on skeleton data to evaluate the quality of the generated sequences. Complementary, a binary classifier is trained for quality assessment. However, both models are not publicly available, making comparisons difficult. While the work [37] uses the average pairwise distance to measure diversity as [4], it only evaluates the best generated sequence using the quality metrics from [12, 15, 23].

3. Stochastic Human Motion Anticipation from Intention

In this work, we address the task of forecasting human motion. This means that we observe 3d human skeletons for t frames, which are denoted by $\mathbf{x}_1^t = (x_1, \dots, x_t) \in \mathbb{R}^{t \times \mathfrak{d}}$ and where $\mathfrak{d}$ is the feature dimension that represents the human pose, and our goal is to forecast plausible future pose sequences $\hat{\mathbf{x}}_{t+1}^T \sim p(\mathbf{x}_{t+1}^T | \mathbf{x}_1^t)$ where $\hat{\mathbf{x}}_{t+1}^T = (\hat{x}_{t+1}, \dots, \hat{x}_T) \in \mathbb{R}^{(T-t) \times \mathfrak{d}}$ and $p(\mathbf{x}_{t+1}^T | \mathbf{x}_1^t)$ is the distribution of all plausible future sequences given the observed human motion.

As it is illustrated in Figure 2, our approach does not predict a single sequence but aims to learn the distribution $p(\mathbf{x}_{t+1}^T | \mathbf{x}_1^t)$ such that we can generate multiple plausible future sequences

$$\hat{\mathcal{X}}_{t+1}^T = \{\hat{\mathbf{x}}_{t+1}^T : \hat{\mathbf{x}}_{t+1}^T \sim p(\mathbf{x}_{t+1}^T | \mathbf{x}_1^t)\}. \quad (1)$$

While we introduce in Section 4 a new quality score that evaluates the plausibility of the set $\hat{\mathcal{X}}_{t+1}^T$ and that correlates very well with human perception, we first discuss the novel approach that forecasts (1).

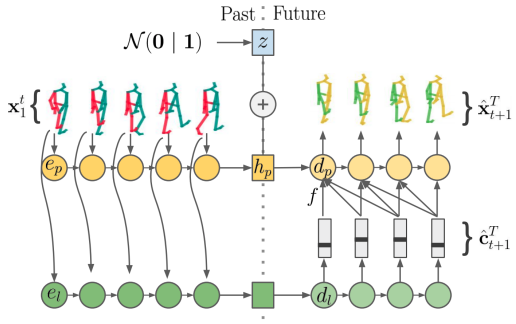


Figure 3: Overview of our method. The blue-red skeletons are the observed human poses while the yellow-green skeletons are forecast future human poses. The network forecasts the human poses at two levels: at the pose level (yellow) and at an intention level (green). During inference, the network forecasts the intention labels ahead in time which guide then the generation of the future poses. By conditioning the pose decoder d_p in addition on z, multiple plausible sequences can be generated for a single sequence of observed human poses.

Although the recent works [4, 28, 37] are able to forecast diverse sequences, the quality of the sequences decreases for longer time horizons beyond 1 second as we show in the user study reported in Table 4. This is expected since the methods model human motion but not the intention of the person. The latter, however, is very important for longer sequences since a motion without a goal is perceived as random and unrealistic.

We therefore propose an approach that generates multiple future sequences that remain plausible even for longer time horizons of 4 seconds. In order to achieve this goal, our network not only forecasts human poses, but also the intention as shown in Figure 1 and 3. An important aspect of our network is that it forecasts the intention $\hat{\mathbf{c}}_{t+1}^T$ ahead in time, which then guides the generated poses

$$\hat{\mathbf{x}}_{t+1}^T \sim p(\mathbf{x}_{t+1}^T | \mathbf{x}_1^t, \hat{\mathbf{c}}_{t+1}^T) \quad (2)$$

and ensures plausible motion transitions when the intention changes. We describe the module of the network that forecasts the intention in Section 3.1 and the module that forecasts the human motion conditioned on the intention in Section 3.2.

3.1. Intention Anticipation

We model the intention by a categorical representation $c_t \in C$ where C is set of possible intention classes. While we forecast the intention ahead in time as shown in Figure 1, we estimate it for each future frame $\hat{\mathbf{c}}_{t+1}^T$. In Section 3.3, we describe how C can be obtained in an unsupervised way.

To anticipate future intent, we use a recurrent encoder-decoder where the recurrent encoder e_l takes as input a se-

quence of observed human motion $\mathbf{x}_1^t$ and the recurrent decoder d_l forecasts the future intentions $\hat{\mathbf{c}}_{t+1}^T$:

$$\hat{\mathbf{c}}_{t+1}^T = d_l(e_l(\mathbf{x}_1^t)). \tag{3}$$

With decoder d_l being auto-regressive, we are not constrained to a fixed time horizon and as such T can be as large as needed. We represent both e_l and d_l with single layer GRUs.

For training, we utilize the categorical cross-entropy as loss function:

$$\mathcal{L}_{\text{sym}} = \frac{1}{T-t} \sum_{\tau=t+1}^{T} \sum_{j=1}^{|C|} c_\tau \log(\hat{c}_{\tau j}) \tag{4}$$

where $|C|$ is the total number of discrete intention labels, c_τ denotes the reference label at time step τ, and $\hat{c}_{\tau j}$ denotes the predicted probability of the j-th class at time step τ. We will discuss in Section 3.3 how the reference labels c_τ are computed for the training set.

In order to generate plausible sequences of future intentions, we furthermore add an adversarial loss:

$$\begin{aligned}\mathcal{L}_{\text{sym}}^{\text{adv}} = \min_{d_l} \max_{D_{\text{label}}} \mathbb{E}_{\mathbf{c}}\big[\log D_{\text{label}}(\mathbf{c})\big] \\ + \mathbb{E}_{\mathbf{x}}\big[1 - \log D_{\text{label}}(d_l(\mathbf{x}))\big]\end{aligned} \tag{5}$$

where $\mathrm{D}_{\text{label}}$ is a one-hidden-layer feed forward network.

3.2. Human Motion Anticipation

In order to sample sequences of future human poses from $p(\mathbf{x}_{t+1}^T | \mathbf{x}_1^t, \hat{\mathbf{c}}_{t+1}^T)$, we utilize a conditional GAN [24] with normal distributed noise vector $z \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$ as shown in Figure 3. It is conditioned on the past human motion sequence $\mathbf{x}_1^t$ and the forecast intent $\hat{\mathbf{c}}_{t+1}^T$.

Specifically, we first encode $\mathbf{x}_1^t$ into the vector h_p using the recurrent pose encoder e_p, *i.e.*, $h_p = e_p(\mathbf{x}_1^t)$. We then concatenate h_p and z and auto-regressively generate future poses for $t < \tau \leq T$ using the pose decoder d_p:

$$(\hat{x}_\tau, h_\tau) = d_p(\hat{x}_{\tau-1} \oplus f(\hat{\mathbf{c}}_\tau^{\tau-1+\gamma}) \mid h_{\tau-1}) \tag{6}$$

where $h_t = h_p \oplus z$, $\hat{x}_t = x_t$, and $\oplus$ denotes the concatenation of two vectors. The pose encoder e_p consists of a single layer GRU while the pose decoder d_p consists of a three layer GRU.

The pose decoder d_p, however, not only depends for each frame τ on the previous generated pose $\hat{x}_{\tau-1}$ and the previous hidden state $h_{\tau-1}$, but also on $f(\hat{\mathbf{c}}_\tau^{\tau-1+\gamma})$, *i.e.*, on the intention which is forecast already γ frames ahead. If $\gamma = 1$, the decoder does not look ahead and it takes only the estimated intention labels until the current frame into account. We will show in the experiments that this results in less plausible sequences since the decoder cannot prepare the transition between two types of motions if they change, *e.g.*, between leaning down and standing as shown in Figure 1. If we allow the decoder to look ahead, the transitions are more smooth and plausible. We found that $\gamma = 10$ $(0.4s)$ is sufficient to obtain good results. Before adding the probabilities $\hat{\mathbf{c}}_\tau^{\tau-1+\gamma}$ to the decoder, we aggregate them by f, which is a temporal convolutional layer with kernel size γ.

During training, we optimize the adversarial loss

$$\begin{aligned}\mathcal{L}_{adv} = \min_{d_p} \max_{D_{\text{pose}}} \mathbb{E}_{\mathbf{x}}\big[\log D_{\text{pose}}(\mathbf{x})\big] \\ + \mathbb{E}_{\mathbf{x}|\mathbf{c}|z}\big[1 - \log D_{\text{pose}}(d_p(\mathbf{x}, \mathbf{c}, z)\big]\end{aligned} \tag{7}$$

where D_{pose} is a two-hidden-layer feed forward network. While there is usually not a high variability of the plausible human motion directly after the last observed frame but the diversity increases the longer the time horizon gets as shown in Figure 2, we additionally utilize a reconstruction loss with decreasing impact as τ increases:

$$\mathcal{L}_{\text{rec}} = \frac{1}{J \cdot (T-t)} \sum_{\tau=t+1}^{T} \sum_{j=1}^{J} \lambda(\tau) ||x_{\tau j} - \hat{x}_{\tau j}||_2 \tag{8}$$

where J is the number of joints in the pose, and $x_{\tau j}$ and $\hat{x}_{\tau j}$ denote the ground truth and model prediction of joint j at time frame τ, respectively. The weight $\lambda(\tau)$ decreases linearly over time with $\lambda(t) = 1$ and $\lambda(t+\tau_{rec}) = 0$. In our experiments, we show that $\tau_{rec} = 15$ $(0.6s)$ is sufficient.

For training the network, we use all four loss terms where the loss terms $\mathcal{L}_{\text{sym}}$ (4) and $\mathcal{L}_{\text{sym}}^{\text{adv}}$ (5) supervise the intention forecasting (green) and the loss terms $\mathcal{L}_{adv}$ (7) and $\mathcal{L}_{\text{rec}}$ (8) supervise the human motion forecasting (yellow) as shown in Figure 3.

3.3. Intention Labels

In order to obtain the intention labels $c_t \in C$ for training, we cluster the training sequences. We first cluster the poses of all training sequences using k-means and assign each frame to a cluster. Since these clusters only consider poses but not motion, we sequentially generate intention labels by detecting cycles of cluster ids in the training sequences. For all datasets, we use 8 intention labels. More details are provided in the supplementary material where we also evaluate the impact of the size of C.

4. Long-term Human Motion Quality Score

As discussed in Section 3, we need for evaluation a score that measures the plausibility of forecast human motion for longer time horizons beyond one second. Furthermore, the measure needs to measure the quality of a set of forecast sequences $\hat{\mathcal{X}}_{t+1}^T$ instead of a single sequence.

We therefore propose a novel quality measure that correlates better with human perception. The main idea is that

Long-Term					
method	walking	eating	smoking	discussion	average
[23]	0.549	0.754	1.403	1.245	0.987
[10]	**0.359**	**0.288**	0.577	1.001	<u>0.556</u>
[22]	0.841	0.909	0.824	1.733	1.077
[21]	0.590	0.821	<u>0.491</u>	1.616	0.879
[28]	0.467	<u>0.301</u>	0.751	<u>0.945</u>	0.616
Ours	<u>0.367</u>	0.621	**0.363**	**0.795**.	**0.536**

Table 1: NPSS measure from [10] for long-term motion anticipation.

a plausible sequence of poses should be close to a real sequence. For long-time horizons, however, the sequences are too long to compare them directly. Instead, we divide all sequences that have the same semantic meaning but that are not part of the training data into overlapping short motion sequences of fixed length κ. We call the short motion sequences *motion words* and we use $\kappa = 8$ for sequences with 25Hz. This results in a very large motion database $\mathcal{D}$.

When evaluating a sequence $\hat{\mathbf{x}}_{t+1}^T \in \hat{\mathcal{X}}_{t+1}^T$ for observation $\mathbf{x}_1^t$, we split the sequence into overlapping motion words as well, where we include the last $\kappa-1$ observed frames, *i.e.*, $\hat{\mathbf{x}}_{t+2-\kappa}^{t+1}, \hat{\mathbf{x}}_{t+3-\kappa}^{t+2}, \dots, \hat{\mathbf{x}}_{T+1-\kappa}^{T}$. We include the last observed frames such that the transition between observed and forecast motion is also taken into account. This is important since discontinuities between observed and forecast frames are perceived by humans as highly unrealistic. Using the motion words of all sequences of $\hat{\mathcal{X}}_{t+1}^T$, we can then compute the plausibility score by measuring the similarity of the motion words of $\hat{\mathcal{X}}_{t+1}^T$ with the motion words in $\mathcal{D}$:

$$f_{sim}\left(\hat{\mathcal{X}}_{t+1}^T\right) = \frac{1}{Z} \sum_{\hat{\mathbf{x}}_{t+1}^T \in \hat{\mathcal{X}}_{t+1}^T} \sum_{\tau=t+2-\kappa}^{T+1-\kappa} g\left(\hat{\mathbf{x}}_{\tau}^{\tau+\kappa}, \mathcal{D}\right), \quad (9)$$

where $Z = (T-t)|\hat{\mathcal{X}}_{t+1}^T|$ is the normalization factor.

For computing the plausibility of a motion word, we find the closest motion word in $\mathcal{D}$ using nearest neighbor search (NN) and compute the normalized directional motion similarity (NDMS), which is discussed in Section 4.1:

$$g\left(\hat{\mathbf{x}}_{\tau}^{\tau+\kappa}, \mathcal{D}\right) = \mathrm{NDMS}(\hat{\mathbf{x}}_{\tau}^{\tau+\kappa}, \mathrm{NN}\big(\hat{\mathbf{x}}_{\tau}^{\tau+\kappa}, \mathcal{D}\big)\big). \quad (10)$$

The function $g\left(\hat{\mathbf{x}}_{\tau}^{\tau+\kappa}, \mathcal{D}\right)$ is 1 when $\mathcal{D}$ contains the exact motion word $\hat{\mathbf{x}}_{\tau}^{\tau+\kappa}$ and it is $0 \leq g\left(\hat{\mathbf{x}}_{\tau}^{\tau+\kappa}, \mathcal{D}\right) < 1$ otherwise. Using motion words and not single poses ensures that the score evaluates motion quality and consistency and not just pose quality while the nearest neighbor approach ensures that the multi-modality of human motion is addressed. Due to the normalization factor Z, f_{sim} (9) provides a plausibility score between 0 and 1 for a set of forecast human motions.

4.1. Normalized Directional Motion Similarity

In order to compare two motion words x and y, we need to define a similarity measure. The Euclidean distance of the poses is insufficient as this favours sequences that remain close to the mean pose. Similarly, using the mean square error of the velocities favours small motion over larger motion, as we discuss in the supplementary material. Instead, we measure the similarity of the motion direction and the ratio of motion magnitudes.

Specifically, the proposed Normalized Directional Motion Similarity (NDMS) compares two motion words $\mathbf{x}, \mathbf{y}$ of length κ by

$$\mathrm{NDMS}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{t=1}^{\kappa-1} \frac{1}{J} \sum_{j=1}^{J} \Psi_t^j(\mathbf{x}, \mathbf{y})}{\kappa - 1} \quad (11)$$

$$\Psi_t^j(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\left(1 + \frac{\dot{\mathbf{x}}_{t,j}\dot{\mathbf{y}}_{t,j}^T}{||\dot{\mathbf{x}}_{t,j}|| \cdot ||\dot{\mathbf{y}}_{t,j}|| + \epsilon}\right) \cdot \frac{\min\left(||\dot{\mathbf{x}}_{t,j}||, ||\dot{\mathbf{y}}_{t,j}||\right)}{\max(||\dot{\mathbf{x}}_{t,j}||, ||\dot{\mathbf{y}}_{t,j}||) + \epsilon} \quad (12)$$

where J represents the numbers of joints of the human pose and $\dot{\mathbf{x}}_{t,j}$ is the 3D velocity of joint j at time t. The first part in (12) yields large values when the j-th joint of x and y move in the same direction while the second part yields large values when the magnitudes of the vectors are similar. To prevent division by zero, we add a small $\epsilon > 0$. This way, $\Psi_t^j(\mathbf{x}, \mathbf{y})$ produces values close to 1 when the motions of $\mathbf{x}$ and $\mathbf{y}$ are similar and values close to 0 when they are very dissimilar.

It is important to note that the proposed quality measure (10) has several advantages compared to existing measures: a) the measure penalizes discontinuities in motion; b) it penalizes unrealistic motion at a fine-grained level; c) it can be used to measure the quality of deterministic as well as stochastic approaches; d) it measures the plausibility of all forecast sequences even if they deviate from the observed future sequence; e) it correlates better than other measures with human perception.

4.2. Implementation Details

For the nearest neighbor search, we use the joint positions of *wrists*, *elbows*, *shoulders*, *hips*, *knees*, and *ankles*. For evaluation, we populate $\mathcal{D}$ with all relevant test sequences, *e.g.*, all *basketball* test sequences for evaluating *basketball*. This way, models have to produce sequences that have the same semantic meaning as the current test set (*e.g. walking*, *eating*) and not just produce common motion patterns observed in all sequences.

5. Experiments

We evaluate our method on the two standard large scale motion capture datasets: Human3.6M [14] and CMU Mocap [1]. We first analyze the quality of the forecast sequences using different measures including a user study for long-term forecasting. In the supplementary material, we

seconds:	0.4	0.8	1.2	1.6	2.0	2.4	2.8	3.2	3.6	4.0	0.4	0.8	1.2	1.6	2.0	2.4	2.8	3.2	3.6	4.0
	walking										eating									
Seq2Seq [23]	0.891	0.660	**0.673**	**0.670**	**0.647**	0.617	0.562	0.537	0.515	0.418	0.794	0.517	0.504	0.514	0.483	0.493	0.428	0.428	0.395	0.409
Trajectory [22]	0.899	0.657	0.605	0.597	0.580	0.463	0.494	0.519	0.461	0.453	0.877	0.622	**0.567**	**0.560**	**0.549**	**0.543**	0.505	**0.565**	**0.555**	0.530
History [21]	**0.929**	**0.694**	0.647	0.668	0.641	**0.657**	**0.638**	**0.646**	**0.646**	**0.630**	**0.884**	0.460	0.418	0.426	0.405	0.406	0.392	0.415	0.394	0.395
Grammer [28]	0.839	0.429	0.413	0.415	0.315	0.291	0.268	0.186	0.172	0.131	0.826	0.421	0.347	0.315	0.227	0.210	0.192	0.167	0.166	0.171
Mix&Match [4]	0.847	0.645	0.607	0.625	0.573	0.564	0.541	-	-	-	0.830	0.534	0.524	0.534	0.506	0.494	0.474	-	-	-
Ours	0.902	0.647	0.619	0.630	0.586	0.592	0.574	0.604	0.577	0.603	0.878	**0.625**	0.530	0.560	0.520	0.521	**0.531**	0.525	0.546	**0.534**
	smoking										discussion									
Seq2Seq [23]	0.685	**0.491**	0.446	**0.422**	0.426	0.432	0.420	0.380	0.365	0.367	0.833	0.503	0.484	**0.472**	0.404	0.415	0.430	0.399	0.340	0.300
Trajectory [22]	0.824	0.476	**0.468**	0.411	0.406	0.381	0.415	0.402	0.390	0.390	0.852	0.417	0.361	0.314	0.302	0.293	0.272	0.290	0.267	0.273
History [21]	**0.874**	0.455	0.389	0.404	0.404	0.394	0.388	0.399	0.397	0.367	**0.893**	0.418	0.318	0.317	0.289	0.290	0.267	0.279	0.262	0.276
Grammer [28]	0.659	0.229	0.221	0.201	0.182	0.179	0.171	0.179	0.183	0.185	0.758	0.187	0.162	0.155	0.162	0.198	0.171	0.191	0.189	0.164
Mix&Match [4]	0.646	0.420	0.429	0.419	0.408	0.404	0.411	-	-	-	0.799	0.481	0.449	0.428	0.411	0.395	0.380	-	-	-
Ours	0.800	0.474	0.456	0.412	**0.446**	**0.477**	**0.452**	**0.433**	**0.445**	**0.417**	0.838	**0.503**	**0.497**	0.466	**0.474**	**0.467**	**0.492**	**0.498**	**0.463**	**0.449**
	posing										average									
Seq2Seq [23]	0.819	0.497	0.443	0.403	0.381	0.367	0.335	0.313	0.296	0.285	0.806	0.530	**0.508**	0.478	0.445	0.438	0.414	0.383	0.361	0.339
Trajectory [22]	0.827	0.476	0.454	0.374	0.406	0.328	0.303	0.317	0.281	0.307	0.840	0.485	0.452	0.391	0.389	0.364	0.344	0.367	0.334	0.338
History [21]	**0.896**	0.419	0.343	0.317	0.246	0.224	0.218	0.211	0.204	0.198	**0.884**	0.434	0.359	0.350	0.337	0.326	0.320	0.318	0.315	0.309
Grammer [28]	0.782	0.267	0.246	0.213	0.224	0.220	0.208	0.202	0.209	0.160	0.746	0.262	0.245	0.236	0.212	0.205	0.202	0.190	0.191	0.181
Mix&Match [4]	0.777	**0.528**	0.486	0.453	0.421	0.402	0.378	-	-	-	0.770	0.500	0.480	0.465	0.444	0.430	0.419	-	-	-
Ours	0.726	0.509	**0.539**	**0.471**	**0.478**	**0.430**	**0.421**	**0.439**	**0.407**	**0.393**	0.826	**0.531**	0.507	**0.491**	**0.484**	**0.478**	**0.474**	**0.477**	**0.467**	**0.465**

Table 2: NDMS scores on Human3.6M [14] for actions *walking*, *eating*, *smoking*, *discussion* and *posing* as well as averaged over all 15 actions. For Mix-and-Match and our approaches we report the mean score over 50 samples for a given input sequence.

seconds:	0.4	0.8	1.2	1.6	2.0	2.4	2.8	3.2	3.6	4.0	0.4	0.8	1.2	1.6	2.0	2.4	2.8	3.2	3.6	4.0
	walking										eating									
VAE [37]	0.589	0.413	0.354	0.329	0.322	0.321	0.320	0.320	0.313	0.299	0.521	0.302	0.322	0.326	0.314	0.310	0.310	0.308	0.298	0.291
DLow [37]	0.586	0.422	0.376	0.345	0.346	0.344	0.331	0.325	0.312	0.296	0.519	0.288	0.315	0.318	0.320	0.320	0.320	0.320	0.305	0.292
Ours	**0.906**	**0.653**	**0.636**	**0.584**	**0.608**	**0.577**	**0.558**	**0.570**	**0.556**	**0.553**	**0.818**	**0.536**	**0.457**	**0.454**	**0.406**	**0.388**	**0.356**	**0.376**	**0.368**	**0.352**
	smoking										discussion									
VAE [37]	0.455	0.295	0.324	**0.334**	**0.330**	0.321	0.310	**0.303**	**0.295**	0.288	0.536	0.334	0.333	0.306	0.288	0.281	0.274	0.276	**0.264**	0.245
DLow [37]	0.454	0.280	0.324	0.322	0.315	0.308	0.297	0.294	0.288	0.280	0.536	0.331	**0.345**	**0.334**	0.315	0.307	**0.291**	0.270	0.257	0.241
Ours	**0.754**	**0.348**	**0.356**	0.319	0.327	**0.345**	**0.310**	0.294	0.286	**0.322**	**0.859**	**0.470**	0.340	0.331	**0.336**	**0.320**	0.283	**0.293**	0.262	**0.274**
	posing										average									
VAE [37]	0.519	0.355	0.334	0.280	0.260	0.264	0.265	0.261	0.246	0.234	0.542	0.342	0.331	0.309	0.294	0.290	0.288	0.286	0.277	0.265
DLow [37]	0.521	0.367	**0.360**	0.332	0.315	0.291	0.271	0.261	0.260	0.241	0.541	0.342	0.341	0.325	0.311	0.298	0.290	0.282	0.274	0.263
Ours	**0.724**	**0.398**	0.301	**0.353**	**0.366**	**0.320**	**0.324**	**0.312**	**0.309**	**0.297**	**0.818**	**0.444**	**0.379**	**0.380**	**0.366**	**0.349**	**0.330**	**0.328**	**0.320**	**0.320**

Table 3: NDMS scores on Human3.6 [14] using the 17 3D joint representation from DLow [37]. We report the mean score over 50 samples for a given input sequence.

	walking	eating	smoking	discussion
Seq2Seq [23]	0.750	0.353	0.312	0.188
Trajectory [22]	0.903	0.625	0.114	0.227
History [21]	0.902	0.221	0.356	0.279
Grammer [28]	0.161	0.324	0.167	0.171
Mix&Match [4]$\star$	0.875	0.617	0.445	0.523
DLow [37]	0.190	0.578	0.449	0.428
Ours	**0.938**	**0.792**	**0.633**	**0.714**

Table 4: User study for the results on Human3.6M [14]. 28 users were randomly asked to judge 4 seconds of forecast human motion. The users could only choose between *realistic* or *not realistic* where we count realistic as 1 and not realistic as 0. In the table we report the mean values and sequences valued close to 1 are deemed highly realistic. $\star$ indicates sequence length of 3.2 seconds.

Diversity (Human3.6M)					
[36]	[34]	[5]	Mix&Match [4]	DLow [37]	Ours
0.26	1.70	0.48	3.52	**4.71**	3.07
Diversity (CMU)					
0.41	**3.00**	0.43	2.63	2.90	2.40

Table 5: Average pairwise distance (APD) of recent state-of-the-art methods on Human3.6M [14] and CMU [1]. Results for DLow [37] are taken from [3].

provide additional results for short-term forecasting and an additional analysis of the quality measure.

5.1. Comparison to State-of-the-Art

Long-Term Forecasting: For evaluating long-term human motion forecasting, we first report NPSS as described in [10], utilizing the publicly available implementation. The results of the long-term time scale of $2-4$ seconds can be seen in Table 1 where our method slightly outperforms current state-of-the art methods. Grammar [28] achieves competitive results. We will, however, later show that the sequences that are generated by Grammar are less realistic than the sequences of other state-of-the-art methods. This indicates that NPSS is not a very reliable measure for the plausibility of the forecast human motion.

We therefore compare the methods using the proposed NDMS metric (see Section 4) with motion word size $\kappa = 8$ on Human3.6M. For each of the 15 actions in Human3.6M, we calculate the scores independently where we populate the database $\mathcal{D}$ with the test sequences of the given action

only - to ensure that the forecast sequences are semantically meaningful and consistent with the action. The results for up to 4 seconds are reported in Tables 2 and 3.

The results in Table 2 show that our approach outperforms stochastic and deterministic methods in terms of quality. On cyclic motion such as walking, [21] produces very strong results over long time periods. However, the motion freezes on non-periodic motion such as *discussion* and *posing*. As expected, other approaches including deterministic approaches [23, 22] perform fairly well for short sequences up to 1.2 seconds. For such short time horizon, the results are quite similar to our approach. However, for longer time horizons the benefit of forecasting the intention becomes evident and our approach outperforms the other methods by a large margin. Since DLow [37] uses a different skeleton representation than the other methods, we also report the NDMS score for the skeleton from [37] in Table 3. On average, our approach outperforms DLow and a variational autoencoder (VAE). It needs to be noted that both DLow and VAE suffer from a motion discontinuity between the observed frames and the forecast frames. The NDMS score is therefore relatively low for the shortest time horizon (0.4s).

To validate our results, we conducted a user study with 28 individuals who were given random sequences of length of 4 seconds. The users had then to rate each sequence whether it was realistic or not. Our results can be seen in Table 4. When we compare, for instance, the results for *walking* to the results reported in Tables 2 and 3, we observe a high similarity between the human perception and the NDMS metric. For a cyclic motion like *walking*, our approach performs best followed by [22] and [21]. The generative grammars [28] start showing unrealistic artifacts after around 2 seconds of walking, which is captured both by our user study as well as by our evaluation score. DLow [37] performs better than generative grammars, but by far worse than the other methods. This is due to the discontinuity at the beginning, which is more prominent for walking than for the other activities where the person often stands at the beginning, but also due to the very high diversity of the generated sequences. The forecast sequences quickly generate motions that are very unlikely to occur after a walking motion. Indeed, Table 5 shows that [4, 37] have a higher diversity among the forecast sequences, but more of the generated sequences are perceived as unrealistic as shown in Table 4. The results in Tables 2, 3, and 4 show that our approach achieves a higher forecast quality than the state-of-the-art for long time horizons, both for cyclic as well as non-cyclic motions.

As already mentioned, the generative grammars [28] achieve competitive NPSS scores, as can be seen in Table 1. This, however, is not supported by our user study and the qualitative results. This shows the weakness of NPSS, which does not occur for the proposed NDMS score. The Pearson Correlation Coefficient of NPSS to the ground truth is -0.238 while our motion similarity scores a correlation of 0.901.

5.2. Ablation Study

Impact of Loss Functions: In Figure 4, we show the impact of the loss functions. While the blue curve corresponds to the proposed method, the red curve ($\mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{rec}}$) is a special case in which we do not forecast the intention. The plots show that the NDMS scores are much lower when we do not forecast the intent. This is in particular visible for walking and eating. Without intent the model always converges to a mean motion, which, in Human3.6M, is standing and gesticulating with hands. Because of this, using no intent performs competitively only on Discussion, which is mostly made up of a person standing and gesticulating.

If we remove only one of the loss terms $\mathcal{L}_{\text{sym}}$ (orange), $\mathcal{L}_{\text{sym}}^{\text{adv}}$ (green), or $\mathcal{L}_{\text{adv}}$ (purple), the NDMS score decreases. Without $\mathcal{L}_{\text{adv}}$ (purple), our method furthermore loses the ability to generate multiple samples for a given input sequence. Removing L_{rec} (not plotted) results in unrealistic motion and poses. Without $\mathcal{L}_{\text{sym}}^{\text{adv}}$ we observed that the network sometimes predicts the same intention label for an unrealistic long time.

Impact of $\lambda(\tau)$: For the reconstruction loss L_{rec} (8), we use the weighting function $\lambda(\tau)$ which linearly decreases from 1 to 0 until τ_{rec}. In Figure 5a, we compare three settings: τ_{rec}=15, τ_{rec}=30, and no weighting at all. In the latter case, $\lambda(\tau)$=1. We observe that all settings produce similar early results but that decaying the reconstruction loss to 0 yields the best results over long time horizons. The reason for this is that the adversarial learning scheme has a greater influence with a more aggressive reconstruction decay, which allows more realistic motion over longer time horizons. However, early motion smoothness is not impeded by this. For this reason we set τ_{rec}=15.

Effect of Clustering: In Section 3.3, we describe our method to obtain frame-wise symbolic labels in an unsupervised way. For clustering, we merge cycles to avoid high frequent changes of labels. In Figure 5b, we compare our clustering approach with a naive clustering where we only apply k-means to the training data. We observe that the naive clustering results in more volatile predictions as the human motion generator tries to catch up to the fast-changing symbolic labels. When using our clustering, on the other hand, we observe more stable predictions with higher quality results.

Impact of γ: The parameter γ in (6) defines for how many frames ahead the intention is forecast. If $\gamma = 1$, the decoder d_p does not look ahead and takes only the estimated intention labels until the current frame into account. In Figure 5c, we evaluate four different values for γ: no look-ahead,

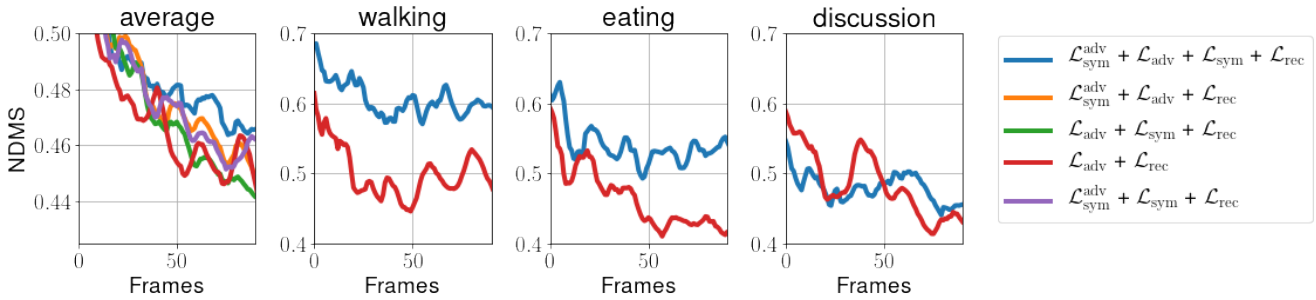
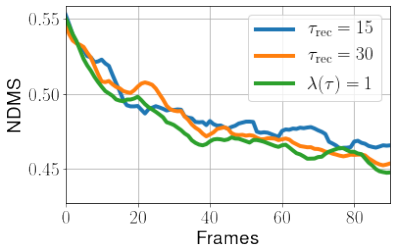
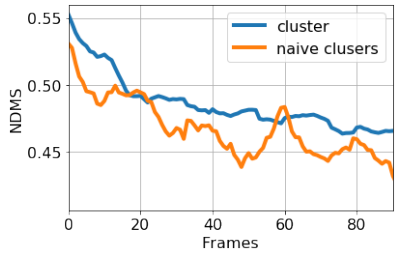


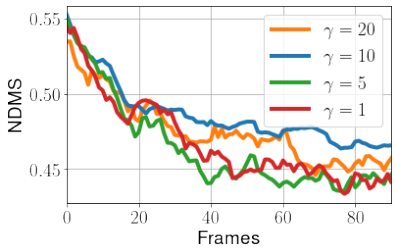
Figure 4: Impact of loss functions. The NDMS score is averaged over 50 samples per input sequence. *average*: NDMS scores averaged over all actions of Human3.6M [14]. *walking, eating, discussion*: Comparison of motion forecasting without (red) and with (blue) intention forecasting for the corresponding action.



(a) Comparing various values of τ_{rec} for the reconstruction loss. In case of $\lambda(\tau)$=1, we use always 1 as weight.



(b) Comparing naive clustering with more elaborate clustering that merges cyclic patterns into cohesive clusters.



(c) Comparing various γ look-ahead values for human motion anticipation.

Figure 5: NDMS scores averaged over all actions of Human3.6M [14] using 50 samples per input sequence.

5 frames look-ahead, 10 frames look-ahead, and 20 frames look-ahead. We observe that small values of γ substantially decrease the NDMS score. This shows that it is very important to forecast the intention ahead of time. However, when γ is too large, it also reduces the quality since only the next upcoming action is relevant for a smooth motion transition and longer look-ahead times distract the pose decoder. As a default parameter, we thus set $\gamma = 10$.

5.3. Very long motion forecasting

For state-of-the-art method evaluation we obtained sequences of up to 4 seconds. Our method, however, is capable of generating much longer sequences, even for non-cyclic motion. Figure 6 shows that our method produces consistent results over very long time horizons of 30 seconds. This is consistent with our observation that the motion, even for non-cyclic motion such as Eating, remains realistic for very long time horizons.

6. Conclusion

In this work, we presented an approach that forecasts multiple plausible sequences of human motion for a single observation[1]. In this way, the model can deal with the uncertainty of the future. In order to ensure that the forecast sequences remain plausible even for longer time horizons,

[1]Source code is available at `https://github.com/jutanke/human_motion_ndms`

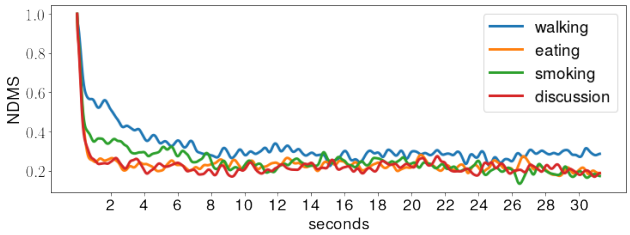


Figure 6: Average NDMS score for very long forecasting of 30 seconds on Human3.6M [14].

we proposed a novel network that not only forecasts the human motion but also the intention. By forecasting the intention ahead of time, the network generates plausible transitions between actions. Furthermore, we presented a new quality score that allows to compare methods that generate multiple sequences even for long time horizons. We demonstrated that the new similarity score correlates better with human judgement than NPSS and that the method produces superior results for long-term human motion anticipation.

Acknowledgment

The work has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) GA 1927/8-1 and GA 1927/5-2 (FOR 2535 Anticipating Human Behavior).

References

[1] CMU. Carnegie-Mellon Mocap Database.

[2] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. Structured prediction helps 3d human motion modelling. In *International Conference on Computer Vision*, 2019.

[3] Sadegh Aliakbarian, Fatemeh Sadat Saleh, Lars Petersson, Stephen Gould, and Mathieu Salzmann. Contextually plausible and diverse 3d human motion prediction. *arXiv preprint arXiv:1912.08521*, 2020.

[4] Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. A stochastic conditioning scheme for diverse human motion prediction. In *Conference on Computer Vision and Pattern Recognition*, 2020.

[5] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. In *Conference on Computer Vision and Pattern Recognition Workshops*, 2018.

[6] Judith Bütepage, Michael J Black, Danica Kragic, and Hedvig Kjellstrom. Deep representation learning for human motion prediction and classification. In *Conference on Computer Vision and Pattern Recognition*, 2017.

[7] Judith Bütepage, Hedvig Kjellström, and Danica Kragic. Anticipating many futures: Online human motion prediction and generation for human-robot interaction. In *International Conference on Robotics and Automation*, 2018.

[8] Yujun Cai, Lin Huang, Yiwei Wang, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Xu Yang, Yiheng Zhu, Xiaohui Shen, et al. Learning progressive joint propagation for human motion prediction. In *European Conference on Computer Vision*, 2020.

[9] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *International Conference on Computer Vision*, 2015.

[10] Anand Gopalakrishnan, Ankur Mali, Dan Kifer, Lee Giles, and Alexander G Ororbia. A neural temporal model for human motion prediction. In *Conference on Computer Vision and Pattern Recognition*, 2019.

[11] Liang-Yan Gui, Yu-Xiong Wang, Xiaodan Liang, and José MF Moura. Adversarial geometry-aware human motion prediction. In *European Conference on Computer Vision*, 2018.

[12] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *Transactions on Graphics*, 2016.

[13] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation, 2018.

[14] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *Pattern Analysis and Machine Intelligence*, 2014.

[15] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *Conference on Computer Vision and Pattern Recognition*, 2016.

[16] Taesoo Kwon and Jessica Hodgins. Momentum-mapped inverted pendulum models for controlling dynamic human motions. *Transactions on Graphics*, 2017.

[17] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional sequence to sequence model for human dynamics. In *Conference on Computer Vision and Pattern Recognition*, 2018.

[18] Jessica Hodgins Libin Liu. Learning basketball dribbling skills using trajectory optimization and deep reinforcement learning. *Transactions on Graphics*, August 2018.

[19] Libin Liu and Jessica Hodgins. Learning to schedule control fragments for physics-based characters using deep q-learning. *Transactions on Graphics*, 2017.

[20] Zhenguang Liu, Shuang Wu, Shuyuan Jin, Qi Liu, Shijian Lu, Roger Zimmermann, and Li Cheng. Towards natural and accurate future motion prediction of humans and animals. In *Conference on Computer Vision and Pattern Recognition*, 2019.

[21] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *European Conference on Computer Vision*, 2020.

[22] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. *International Conference on Compuver Vision*, 2019.

[23] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Conference on Computer Vision and Pattern Recognition*, 2017.

[24] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[25] Brian Paden, Michal Čáp, Sze Zheng Yong, Dmitry Yershov, and Emilio Frazzoli. A survey of motion planning and control techniques for self-driving urban vehicles. *Transactions on Intelligent Vehicles*, 2016.

[26] Dario Pavllo, Christoph Feichtenhofer, Michael Auli, and David Grangier. Modeling human motion with quaternion-based neural networks. *arXiv preprint arXiv:1901.07677*, 2019.

[27] Dario Pavllo, David Grangier, and Michael Auli. Quaternet: A quaternion-based recurrent model for human motion. In *British Machine Vision Conference*, 2018.

[28] AJ Piergiovanni, Anelia Angelova, Alexander Toshev, and Michael S Ryoo. Adversarial generative grammars for human activity prediction. *European Conference on Computer Vision*, 2020.

[29] Alejandro Hernandez Ruiz, Juergen Gall, and Francesc Moreno-Noguer. Human motion prediction via spatio-temporal inpainting. *International Conference on Computer Vision*, 2019.

[30] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, 2015.

[31] Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *Transactions on Graphics*, 2019.

[32] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, 2014.

[33] G W Taylor, G E Hinton, and S T Roweis. Modeling human motion using binary latent variables. In *Advances in Neural Information Processing Systems*, 2007.

[34] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *International Conference on Computer Vision*, 2017.

[35] Borui Wang, Ehsan Adeli, Hsu-kuang Chiu, De-An Huang, and Juan Carlos Niebles. Imitation learning for human pose prediction. In *International Conference on Computer Vision*, 2019.

[36] Xinchen Yan, Akash Rastogi, Ruben Villegas, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, Ersin Yumer, and Honglak Lee. Mt-vae: Learning motion transformations to generate multimodal human dynamics. In *European Conference on Computer Vision*, 2018.

[37] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *European Conference on Computer Vision*, 2020.

[38] Yi Zhou, Zimo Li, Shuangjiu Xiao, Chong He, Zeng Huang, and Hao Li. Auto-conditioned recurrent networks for extended complex human motion synthesis. *International Conference on Learning Representations*, 2018.