Generating Smooth Pose Sequences for Diverse Human Motion Prediction

Wei Mao[1], Miaomiao Liu[1], Mathieu Salzmann[2,3]

[1]Australian National University; [2]CVLab, EPFL; [3]ClearSpace, Switzerland

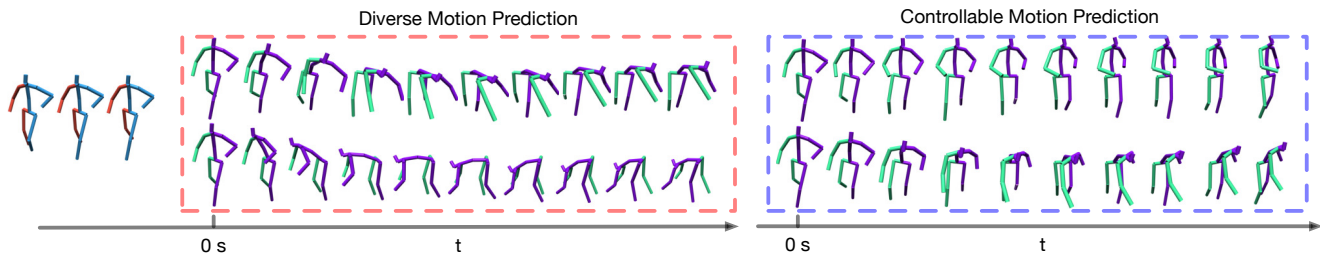{wei.mao, miaomiao.liu}@anu.edu.au, mathieu.salzmann@epfl.ch

Figure 1: **Diverse and controllable motion prediction.** Given a past human motion, shown with blue and red skeletons, our model can predict diverse future motions, as shown in the red box. It also allows us to produce future motions with the same lower-body motion but diverse upper-body ones, i.e., to perform controllable motion prediction, as shown in the blue box.

Abstract

Recent progress in stochastic motion prediction, i.e., predicting multiple possible future human motions given a single past pose sequence, has led to producing truly diverse future motions and even providing control over the motion of some body parts. However, to achieve this, the state-of-the-art method requires learning several mappings for diversity and a dedicated model for controllable motion prediction. In this paper, we introduce a unified deep generative network for both diverse and controllable motion prediction. To this end, we leverage the intuition that realistic human motions consist of smooth sequences of valid poses, and that, given limited data, learning a pose prior is much more tractable than a motion one. We therefore design a generator that predicts the motion of different body parts sequentially, and introduce a normalizing flow based pose prior, together with a joint angle loss, to achieve motion realism. Our experiments on two standard benchmark datasets, Human3.6M and HumanEva-I, demonstrate that our approach outperforms the state-of-the-art baselines in terms of both sample diversity and accuracy. The code is available at https://github.com/wei-mao-2019/gsps

1. Introduction

Predicting future human motions from historical pose sequences has broad applications in autonomous driving [41], animation creation in the game industry [50] and human robot interaction [32]. Most existing work focuses on deterministic prediction, namely, predicting only the most likely future sequence [16, 40, 34, 38, 37]. However, future human motion is naturally diverse, especially over a long-term horizon (> 1s).

Most of the few attempts to produce diverse future motion predictions exploit variational autoencoders (VAEs) to model the multi-modal data distribution [51, 54, 4]. These VAEs-based models are trained to maximize the motion likelihood. As a consequence, and as discussed in [55], because training data cannot cover all possible diverse motions, test-time sampling tends to concentrate on the major data distribution modes, ignoring the minor ones, and thus limiting the diversity of the output. To address this Yuan *et al*. [55] proposed to learn multiple mapping functions, which produce multiple predictions that are explicitly encouraged to be diverse. While this framework indeed yields high diversity, it requires training several mappings in parallel, and separates the training of such mappings from that of the VAE it employs for prediction. Furthermore, while the approach was shown to be applicable to controllable motion prediction, doing so requires training a dedicated model and does not guarantee the controlled part of the motion, e.g., the lower body, to be truly fixed to the same motion in different predictions as the remaining body parts vary.

In this paper, we introduce an end-to-end trainable approach for diverse motion prediction that does not require learning several mappings to achieve diversity. Our framework yields fully controllable motion prediction; one can strictly fix the motion of one portion of the human body and generate diverse predictions for the other portion only.

To this end, we rely on the observation that diverse future motions are composed of valid human poses organized in smooth sequences. Therefore, instead of learn-
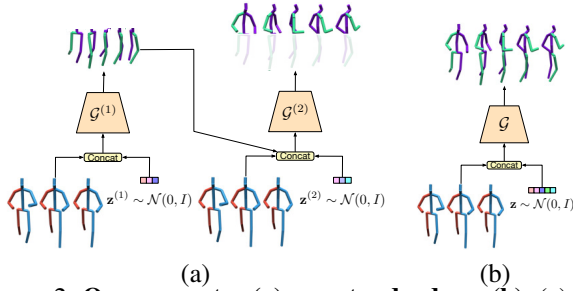
Figure 2: **Our generator (a) vs a standard one (b).** (a) We predict the motions of different body parts sequentially. (b) Existing methods directly produce the whole motion.

ing a *motion* prior/distribution, for which sufficiently diverse training data is hard to obtain, we propose to learn a *pose* prior and enforce a hard constraint on the predicted poses to form smooth sequences and to satisfy human kinematic constraints. Specifically, we model our pose prior as a normalizing flow [45, 15], which allows us to compute the data log-likelihood exactly, and further promote diversity by maximizing the distance between pairs of samples during training.

To achieve controllable motion predictions, as illustrated in Fig. 2 (a), we generate the future motions of the different body parts of interest in a sequential manner. Our design allows us to produce diverse future motions that share the same partial body motion, e.g., the same leg motion but diverse upper-body motions. This is achieved by fixing the latent codes for some body parts while varying those of the other body parts. In contrast to [55], our approach allows us to train a single model that achieves both non-controllable and controllable motion prediction.

Our contributions can be summarized as follows: (i) We develop a unified framework achieving both diverse and part-based controllable human motion prediction, using a pre-ordered part sequence; (ii) We propose a *pose* prior and a *joint angle* constraint to regularize the training of our generator and encourage it to produce smooth pose sequences. Such strategy overcomes the difficulty of learning the distribution of diverse *motions* as other VAE-based methods do.

Our experiments on standard human motion prediction benchmarks demonstrate that our approach outperforms the state-of-the-art methods in terms of sample diversity and accuracy.

2. Related Work

Human motion prediction. Early attempts at human motion prediction [9, 46, 49, 53] relied on non-deep learning approaches, such as Hidden Markov Model [9] and Gaussian Process latent variable models [53]. Despite their success on modeling simple periodic motions, more complicated ones are typically better handled via deep neural net-

works. Deep learning based methods can be roughly categorized into deterministic approaches and stochastic ones.

Deterministic models focus on predicting the most likely human future motion sequence given historical observations [16, 27, 10, 40, 44, 19, 34, 3, 38, 52, 18, 37, 11, 39]. Motivated by the success of RNNs for sequence modeling [48, 31], many such methods employ recurrent architectures [16, 27, 40, 44, 19, 18]. Nevertheless, feed-forward models have more recently been shown to effectively leverage the human kinematic structure and long motion history [34, 38, 3, 10]. In any event, while deterministic human motion prediction models have achieved promising results, especially for short-term prediction (< 0.5s), they struggle with predictions for long-term horizons (> 1s). This is because human motion is an inherently stochastic process, where one observed motion can lead to multiple possible future ones.

Addressing this has been the focus of stochastic motion prediction methods. Existing ones are mainly based on deep generative models [51, 35, 6, 22, 33, 54, 4, 55], such as variational autoencoders (VAEs) [29] and generative adversarial networks (GANs) [17]. In the context of VAE-based ones, Yan *et al.* [54] proposed to jointly learn a feature embedding for motion reconstruction and a feature transformation to model the motion mode transitions; Aliakbarian *et al.* [4] introduced a perturbation strategy for the random variable to prevent the generator from ignoring it. In both cases, once the generator is trained, possible future motions are obtained by feeding it randomly-sampled latent codes. However, as argued in [55], such likelihood-based sampling strategy concentrates on the major mode(s) of the data distribution while ignoring the minor ones. To address this, Yuan *et al.* [55] introduced a learnable sampling strategy equipped with a prior explicitly encouraging the diversity of the future predictions obtained from a *pre-trained* generator. Despite their promising performance, their future predictions are constrained by the use of a *pre-trained* generator. As an alternative to VAE-based models, GAN-based methods [35, 6, 22, 33] train the generator jointly with a discriminator. While, in principle, one could also employ a diversity-promoting prior in these methods, in practice, the resulting additional constraints further complicate the inherently-difficult training process [5]. As such, existing GAN-based methods tend to produce limited diversity. Here, instead of using a discriminator to regularize the generation process, we employ a normalizing-flow-based *pose* prior, accounting for the fact that training data can more easily cover the diversity of poses than that of motions, and encourage the resulting poses to be valid and form smooth sequences.

3D human pose prior. In the 3D human pose estimation literature, many works [8, 43, 57] have attempted to learn 3D pose priors to avoid invalid human poses. Such priors
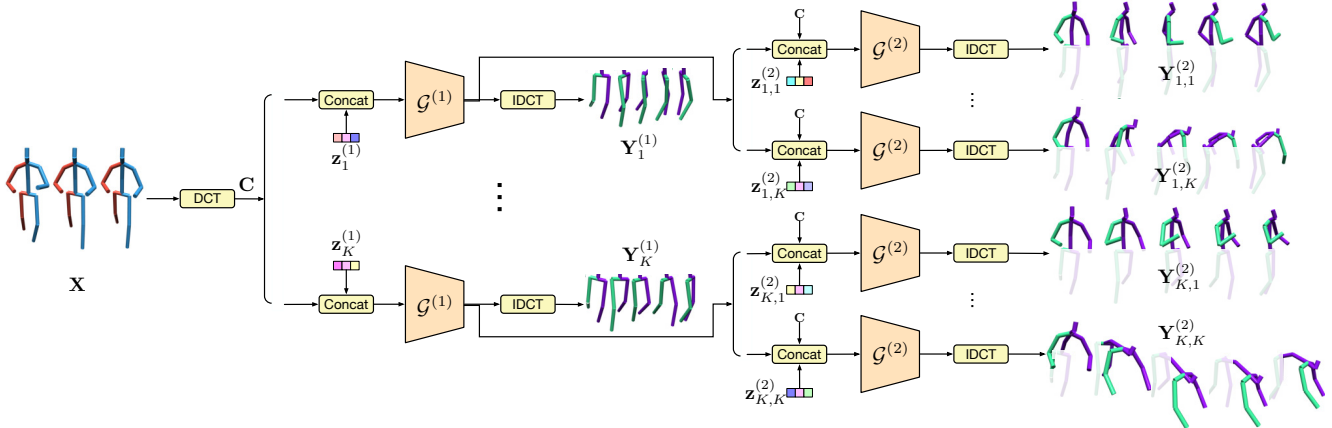
Figure 3: **Overview of our approach.** Given a past motion $\mathbf{X}$, we first sample K latent codes $\{\mathbf{z}_i^{(1)}\}_{i=1}^{K}$ and decode them to future lower-body motions $\{\mathbf{Y}_i^{(1)}\}_{i=1}^{K}$. For each such future motion $\mathbf{Y}_i^{(1)}$, we again sample K latent codes $\{\mathbf{z}_i^{(2)}\}_{i=1}^{K}$ to generate future upper-body motions $\{\mathbf{Y}_{i,j}^{(2)}\}_{j=1}^{K}$.

include Gaussian mixture models [8] and VAEs [43]. However, as discussed in [57], these priors only approximate the pose log-likelihood and may lead to instability [57]. To evaluate the exact log-likelihood, Zanfir *et al*. [57] thus proposed a normalizing flow based pose prior. Normalizing flows (NFs) [45, 15] have recently become popular for density estimation precisely because they allow one to compute the exact log-likelihood. Here, we use a pre-trained NF model to evaluate the log-likelihood of the generated future poses, and by maximizing the log-likelihood, encourage the generator to produce realistic poses.

Joint angle constraint. Joint angle limits have been explored for the task of human pose estimation [21, 12, 1, 13]. In particular, Akhter *et al*. [1] introduced a pose-dependent joint limit function learned from a motion capture dataset. However, their function is non-differentiable, which makes it ill-suited for deep neural networks. Our angle loss is similar to that of [13]. However, unlike Dabral *et al*. [13] who only consider angle constraints on arms and legs, consisting of manually-defined valid angle ranges, we encompass all angles between different body parts/joints and exploit the training data to determine the valid angle ranges.

Controllable motion prediction. To the best of our knowledge, Dlow [55] constitutes the only attempt at controllable motion prediction, e.g, fixed lower-body motion but diverse upper-body motions. This was achieved by training a dedicated model, different from the uncontrollable one, yet sill not guaranteeing absolute control of the body parts that should undergo the same motion. By contrast, we develop a unified model that can achieve both controllable and uncontrolled diverse motion prediction, while guaranteeing the controlled parts to truly follow a fixed motion.

Controllable motion prediction/generation has also been studied in computer graphics, specifically for virtual character control [24, 23, 36]. These works focus on generating human motions for a specific goal, such as following a

given path or performing pre-defined actions. In particular, [36] relies on a motion VAE to capture motion dynamics and searches for motions that achieve the desired task via sampling policies, while accounting for neither motion diversity nor the detailed body movements. By contrast, our work aims to predict diverse future motions and control the detailed motion of body parts.

3. Our Approach

Let us now introduce our approach to diverse and controllable motion prediction. We represent a human pose $\mathbf{x} \in \mathbb{R}^D$ as the concatenation of 3D joint coordinates in a single frame. Given a past motion sequence $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_H]^T$ of H frames, we aim to predict a set of poses $\mathbf{Y} = [\mathbf{x}_{H+1}, \mathbf{x}_{H+2}, \cdots, \mathbf{x}_{H+T}]^T$ representing a possible future motion. To this end, we rely on a deep generative model, which we design to yield diversity and allow for controllable motion prediction, as discussed below.

3.1. Diverse Motion Prediction

In this section, we first briefly review the use of deep generative models in the context of human motion prediction and then introduce our solution for diverse prediction.

Deep Generative Models. Let $p(\mathbf{Y}|\mathbf{X})$ denote the data distribution of future motions $\mathbf{Y} \in \mathcal{Y}$, where $\mathcal{Y}$ is the set of all possible future motions, conditioned on past motions $\mathbf{X} \in \mathcal{X}$, with $\mathcal{X}$ encompassing all possible motion history. By introducing a latent variable $\mathbf{z} \in \mathcal{Z}$, the data distribution can be reparameterized as $p(\mathbf{Y}|\mathbf{X}) = \int p(\mathbf{Y}|\mathbf{X}, \mathbf{z})p(\mathbf{z})d\mathbf{z}$, where $p(\mathbf{z})$ is a Gaussian distribution. Generating a future motion $\mathbf{Y}$ can then be achieved by sampling a latent variable $\mathbf{z}$ and mapping it to $\mathbf{Y}$ using a deterministic *generator* function $\mathcal{G} : \mathcal{Z} \times \mathcal{X} \to \mathcal{Y}$. Formally, this is expressed as

$$\mathbf{z} \sim p(\mathbf{z}) \,, \tag{1}$$
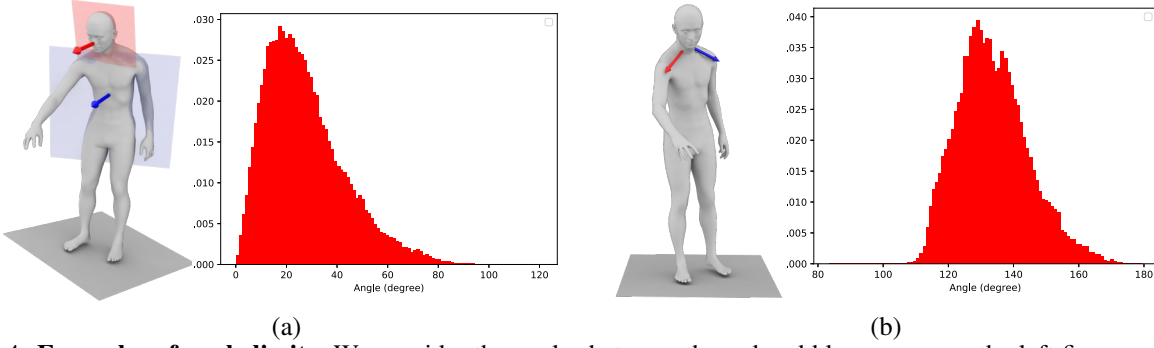$$\mathbf{Y} = \mathcal{G}(\mathbf{z}, \mathbf{X}) \,, \tag{2}$$

Figure 4: **Examples of angle limits.** We consider the angles between the red and blue arrows on the left figures, and show their distributions in the right plots. We define the valid scope based on the minimal and maximal angles. (a) Angle between the head and body; (b) Angle between the left and right shoulders.

where the generator $\mathcal{G}$ is commonly implemented as a deep neural network. To train such a generator, VAE-based methods [54, 4] typically rely on maximizing the evidence lower bound (ELBO) of the motion likelihood, i.e., minimizing the data reconstruction error and the KL divergence. However, this does not encourage diversity across different randomly-sampled vectors $\mathbf{z}$; instead it focuses on maximizing the likelihood of the training motions, and test-time diversity is thus limited by that of the training motions.

To overcome this, we propose to explicitly generate K future motions $\{\hat{\mathbf{Y}}_j\}_{j=1}^K$ for each sample during training, and explicitly promote their diversity. Because we aim for diversity, we cannot encourage all generated motions to match the ground truth. Therefore, we redefine the reconstruction error so that at least one of them is close to the ground truth. This yields the loss

$$\mathcal{L}_r = \min_j \|\hat{\mathbf{Y}}_j - \mathbf{Y}\|^2 \,, \qquad (3)$$

where $j \in \{1, 2, \cdots, K\}$ indexes over the generated motions for one sample.

This loss, however, imposes constraints on only one of the K generated motions. To better constrain the others, we leverage the intuition that, while the dataset contains only one ground-truth future motion for each past motion, several sequences have similar past motions. For each past motion, we therefore search for the training samples with similar past motions, based on a distance threshold, and take their future motion as pseudo ground truth. Let $\{\mathbf{Y}_p\}_{p=1}^P$ denote the resulting pseudo ground truths. Then, we define the multi-modal reconstruction error

$$\mathcal{L}_{mm} = \frac{1}{P} \sum_{p=1}^P \min_j \|\hat{\mathbf{Y}}_j - \mathbf{Y}_p\|^2 \,, \qquad (4)$$

where $j \in \{1, 2, \cdots, K\}$. It encourages our generator to cover each pseudo ground truth with at least one sampled motion.

To further explicitly encourage diversity across the generated motions, following [55], we use the diversity-promoting loss

$$\mathcal{L}_d = \frac{2}{K(K-1)} \sum_{j=1}^K \sum_{k=j+1}^K e^{-\frac{\|\hat{\mathbf{Y}}_j - \hat{\mathbf{Y}}_k\|_1}{\alpha}} \,, \qquad (5)$$

where α is a normalizing factor.

One drawback of this diversity loss, however, is that it may lead the model to produce unrealistic and physically-invalid motions, particularly if $P < K$ above, which leaves some motions unregularized. The most straightforward way to overcome this would be to learn a motion prior. This, however, would require an impractically large amount of training data. Instead, we therefore leverage the observation that natural motions are composed of valid human poses and thus introduce the pose prior and angle losses discussed below.

Pose prior. We model our pose prior using a normalizing flow [45, 15], which is an invertible transformation that aims to transfer an unknown data distribution to a distribution with a tractable density function, e.g., a Gaussian distribution. In other words, we model the 3D human pose distribution $p(\mathbf{x})$ by learning a bijective and differentiable function $f(\cdot)$, which maps a pose sample $\mathbf{x} \sim p(\mathbf{x})$ to a latent representation $\mathbf{h} = f(\mathbf{x})$ following a standard Gaussian distribution, i.e., $\mathbf{h} \sim \mathcal{N}(0, \mathbf{I})$.

Following the normalizing flow literature [45, 15], this allows us to compute the likelihood of a pose $\mathbf{x}$ as

$$p(\mathbf{x}) = g(\mathbf{h}) \left| \det\left(\frac{\partial f}{\partial \mathbf{x}}\right) \right| \,, \qquad (6)$$

where $g(\mathbf{h}) = \mathcal{N}(\mathbf{h}|0, \mathbf{I})$ and $\det(\frac{\partial f}{\partial \mathbf{x}})$ is the determinant of the Jacobian matrix of $f(\cdot)$.

In practice, the function f is modeled via a deep network. Considering model size and inference efficiency, we choose a simple network with only 3 fully-connected layers, which is in stark contrast with the much larger architectures in the recent normalizing flow literature [15, 28, 42, 25]. To ensure that $f(\cdot)$ is invertible, we compute the weights of each layer via a QR decomposition and use monotonic activation functions. More details about our normalizing flow architecture are provided in the supplementary material.

Given a dataset $\mathcal{D}$ of valid human poses, we learn the function f by maximizing the log-likelihood of the samples

in $\mathcal{D}$. This can be written as

$$f^* = \arg\max_f \sum_{\mathbf{x} \in \mathcal{D}} \log p(\mathbf{x}) \tag{7}$$

$$= \arg\max_f \sum_{\mathbf{x} \in \mathcal{D}} \log g(\mathbf{h}) + \log \left| \det \left(\frac{\partial f}{\partial \mathbf{x}} \right) \right| . \tag{8}$$

Given the trained function f, we then define a loss function to encourage our generator to produce valid poses. Specifically, we write it as the negative log-likelihood of a generated human pose $\hat{\mathbf{x}}$, which yields

$$\mathcal{L}_{nf} = -\log p(\hat{\mathbf{x}}) \tag{9}$$

$$= -\log g(\hat{\mathbf{h}}) - \log \left| \det \left(\frac{\partial f^*}{\partial \hat{\mathbf{x}}} \right) \right| , \tag{10}$$

where $\hat{\mathbf{h}} = f(\hat{\mathbf{x}})$ and $g(\hat{\mathbf{h}}) = \mathcal{N}(\hat{\mathbf{h}}|0, \mathbf{I})$.

Joint angle loss. In addition to our learnt pose prior, we leverage the fact that human movements are constrained by the physiological structure of the human body, e.g., one cannot turn their head fully backwards. In our context, this means that the angles between some body parts are limited to a certain range. Here, instead of manually encoding these different ranges, we discover them by analysing a valid human pose dataset $\mathcal{D}$.

To this end, as shown in Fig. 4, we first compute unit length vectors, which represent either the orientations of body parts or the directions of limbs. A body part orientation is the normal of the plane defined by 3 joints. For example, the torso plane, shown in blue in Fig. 4 (a), is defined by the left and right shoulders and the pelvis. A limb direction is defined by 2 joints of the limb. We then compute the angle between those vectors for every pose in $\mathcal{D}$ and define the valid range based on the minimal and maximal such angle. We list all angles and their valid ranges in the supplementary material.

Let $\{a_j\}_{j=1}^L$ denote the L pre-defined angles, and l_{a_j} and u_{a_j} the lower bound and upper bound of angle a_j, respectively. Given a human pose $\hat{\mathbf{x}}$, we write our joint angle loss for a_j as

$$\mathcal{L}_{a_j} = \begin{cases} (a_j(\hat{\mathbf{x}}) - l_{a_j})^2, & \text{if } a_j < l_{a_j} \\ (a_j(\hat{\mathbf{x}}) - u_{a_j})^2, & \text{if } a_j > u_{a_j} \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

where $a_j(\hat{\mathbf{x}})$ is the angle value calculated from pose $\hat{\mathbf{x}}$. We then combine the different angles in our final angle loss $\mathcal{L}_a = \sum_{j=1}^L \mathcal{L}_{a_j}$.

Predicting smooth trajectory. In the stochastic motion prediction literature, human movements are commonly represented by a sequence of 3D joint coordinates. With such a representation, encouraging the generated poses to be valid, as discussed above, does not ensure that the resulting sequence will look natural. To enforce temporal smoothness, we therefore adopt a trajectory representation based on the

Discrete Cosine Transform (DCT), as suggested in [38]. In particular, using a reduced number of low-frequency DCT components guarantees the resulting trajectory to be smooth as shown in Fig. 5.
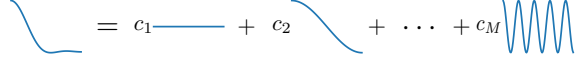

Figure 5: **A smooth trajectory** can be compactly represented by a linear combination of predefined DCT bases [2].

Specifically, given the past motion sequence $\mathbf{X}$, we first replicate the last pose T times to generate a temporal sequence $\tilde{\mathbf{X}}$ of length $H + T$, where T is the length of the future sequence to predict. We then compute the DCT coefficients of this sequence $\mathbf{C}$ as

$$\mathbf{C} = \tilde{\mathbf{X}}\mathbf{T} , \tag{12}$$

where $\mathbf{T} \in \mathbb{R}^{(H+T) \times M}$ and each column of $\mathbf{T}$ represents a predefined DCT basis; $\tilde{\mathbf{X}} \in \mathbb{R}^{D \times (H+T)}$ and each row of $\tilde{\mathbf{X}}$ is the trajectory of a joint coordinate; $\mathbf{C} \in \mathbb{R}^{D \times M}$ and each row of $\mathbf{C}$ represents the first $M \leq H + T$ DCT coefficients for a trajectory.

We then make our generator predict the DCT coefficients of the future motion $\hat{\mathbf{C}}$ given those of the past motion $\mathbf{C}$. Given the predicted coefficients, we recover the future motion via inverse DCT as

$$\hat{\mathbf{Y}} = \hat{\mathbf{C}}\mathbf{T}^T , \tag{13}$$

where $\hat{\mathbf{C}} \in \mathbb{R}^{D \times M}$ and $\hat{\mathbf{Y}} \in \mathbb{R}^{D \times (H+T)}$. As in [38], our generator also outputs the past motion to encourage the transition from past poses to future ones to be smooth. See the supplementary material for more detail.

3.2. Controllable Motion Prediction

The diverse motion prediction framework discussed above does not provide any control over the generated motions. For controllable motion prediction, our goal is to predict future sequences that share the same motion for parts of the body while being diverse for the other parts, e.g., same leg motion but diverse upper-body motion. To this end, instead of directly modeling the joint data distribution as discussed above, we propose to model a product of sequential conditional distributions.

Specifically, let a human motion be split into N different body part motions, that is, $\mathbf{Y} = [\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, \cdots, \mathbf{Y}^{(N)}]$, where $\mathbf{Y}^{(i)} \in \mathbb{R}^{T \times D_i}$ defines the motion of the ith body part, e.g., the left leg. Then, the future body motion distribution $p(\mathbf{Y}|\mathbf{X})$ can be expressed as

$$p(\mathbf{Y}|\mathbf{X}) = p(\mathbf{Y}^{(1)}|\mathbf{X})p(\mathbf{Y}^{(2)}|\mathbf{X}, \mathbf{Y}^{(1)}) \cdots p(\mathbf{Y}^{(N)}|\mathbf{X}, \{\mathbf{Y}^{(i)}\}_{i=1}^{N-1}) . \tag{14}$$

Each of the N conditional distributions describes the motion of a particular body part given the motion of the previous body parts. Similar to the standard deep generative model discussed above, we model each conditional distribution as

$$\mathbf{z}^{(i)} \sim p^{(i)}(\mathbf{z}) , \tag{15}$$

Method	Human3.6M [26]					HumanEva-I [47]				
	APD ↑	ADE ↓	FDE ↓	MMADE ↓	MMFDE ↓	APD ↑	ADE ↓	FDE ↓	MMADE ↓	MMFDE ↓
ERD [16]	0	0.722	0.969	0.776	0.995	0	0.382	0.461	0.521	0.595
acLSTM [58]	0	0.789	1.126	0.849	1.139	0	0.429	0.541	0.530	0.608
Pose-Knows [51]	6.723	0.461	0.560	0.522	0.569	2.308	0.269	0.296	0.384	0.375
MT-VAE [54]	0.403	0.457	0.595	0.716	0.883	0.021	0.345	0.403	0.518	0.577
HP-GAN [6]	7.214	0.858	0.867	0.847	0.858	1.139	0.772	0.749	0.776	0.769
BoM [7]	6.265	0.448	0.533	0.514	0.544	2.846	0.271	0.279	0.373	0.351
GMVAE [14]	6.769	0.461	0.555	0.524	0.566	2.443	0.305	0.345	0.408	0.410
DeLiGAN [20]	6.509	0.483	0.534	0.520	0.545	2.177	0.306	0.322	0.385	0.371
DSF [56]	9.330	0.493	0.592	0.550	0.599	4.538	0.273	0.290	0.364	0.340
DLow [55]	11.741	0.425	0.518	0.495	0.531	4.855	0.251	0.268	0.362	0.339
Ours	**14.757**	**0.389**	**0.496**	**0.476**	**0.525**	**5.825**	**0.233**	**0.244**	**0.343**	**0.331**

Table 1: **Quantitative results** on Human3.6M and HumanEva-I. Our model consistently outperforms others on all metrics.

$$\mathbf{Y}^{(i)} = \mathcal{G}^{(i)}(\mathbf{z}^{(i)}, \mathbf{X}, \{\mathbf{Y}^{(j)}\}_{j=1}^{i-1}), \qquad (16)$$

where $i \in \{1, 2, \cdots, N\}$ and $p^{(i)}(\mathbf{z})$ is a standard Gaussian distribution $\mathcal{N}(0, 1)$. In other words, to compute a future motion, we sample different random variables $\{\mathbf{z}^{(i)}\}_{i=1}^{N}$ and decode them sequentially to obtain the future motion of each body part. By fixing some random variables $\{\mathbf{z}^{(i)}\}_{i=1}^{J}$ while varying the others $\{\mathbf{z}^{(i)}\}_{i=J+1}^{N}$, our model can generate multiple times the same motion for body parts $\{\mathbf{Y}^{(i)}\}_{i=1}^{J}$ while producing diverse motions for the other body parts $\{\mathbf{Y}^{(i)}\}_{i=J+1}^{N}$.

Note that our generator design lets us achieve both diverse and controllable motion prediction at the same time. During training, for each past motion of the i-th body part, we generate K motions for the $(i+1)$-th body part leading to K^N future motions given one past motion. For example, as illustrated in Fig. 3, we sample K different leg motions $\{\mathbf{Y}_j^{(1)}\}_{j=1}^{K}$ and for each leg motion $\mathbf{Y}_j^{(1)}$, we generate K different upper-body motions $\{\mathbf{Y}_{j,k}^{(2)}\}_{k=1}^{K}$. In this context, we then re-write our diversity-promoting loss as a per-body-part loss, leading to

$$\mathcal{L}_{d_i} = \frac{2}{K(K-1)} \sum_{j=1}^{K} \sum_{k=j+1}^{K} e^{-\frac{\|\hat{\mathbf{Y}}_{\cdot,j}^{(i)} - \hat{\mathbf{Y}}_{\cdot,k}^{(i)}\|_1}{\alpha^{(i)}}}, \qquad (17)$$

where $i \in \{1, 2, \cdots, N\}$ is the index of the body part, and $\mathbf{Y}_{\cdot,k}^{(i)}$ is the k-th future motion for the i-th body part[1]. Altogether, our final training loss is then expressed as

$$\mathcal{L} = \lambda_{nf}\mathcal{L}_{nf} + \lambda_a\mathcal{L}_a + \sum_{i=1}^{N} \lambda_{d_i}\mathcal{L}_{d_i} + \lambda_r\mathcal{L}_r + \lambda_{mm}\mathcal{L}_{mm}, \quad (18)$$

where N is the number of body parts.

In practice, inspired by [38], we define each generator $\mathcal{G}^{(i)}$ as a Graph Convolutional Network (GCN) [30] with several graph convolution layers. Given a feature map $\mathbf{F} \in \mathbb{R}^{D \times |\mathbf{F}|}$, each such layer computes a transformed feature map $\mathbf{F}' \in \mathbf{R}^{D \times |\mathbf{F}'|}$ as $\mathbf{F}' = \tanh(\mathbf{AFW})$, where

---

[1]Note that we eliminate the sample index of previous body parts for simplicity. For example, the k-th future motion of the 2nd body part should be represented as $\mathbf{Y}_{j,k}^{(2)}$, where $j \in \{1, 2, \cdots, K\}$ is the sample index of the 1st body part.

$\mathbf{A} \in \mathbb{R}^{D \times D}$ represents a fully connected graph with learnable connectivity and $\mathbf{W} \in \mathbb{R}^{|\mathbf{F}| \times |\mathbf{F}'|}$ is a matrix of trainable weights. The details of our network architecture are provided in the supplementary material.

4. Experiments

4.1. Dataset

Following [55], we evaluate our method on 2 motion capture datasets, Human3.6M [26] and HumanEva-I [47], and adopt the same training and testing settings as in [55] on both datasets.

Human3.6M consists of 7 subjects performing 15 actions. We use 5 subjects (S1, S5, S6, S7, S8) for training and the remaining 2 subjects (S9, S11) for testing. We use the original frame rate (50 Hz) and a 17-joint skeleton. We remove the global translation. Our model is trained to observe 25 past frames (0.5s) and predict 100 future frames (2s).

HumanEva-I consists of 3 subjects performing 5 actions, depicted by videos captured at 60 Hz. A person is represented by a 15-joint skeleton. We adopt the official train/test split [47] and also remove the global translation. The model predicts 60 future frames (1s) given 15 past frames (0.25s).

4.2. Metrics, Baselines & Implementation

Metrics. We follow the same evaluation protocol as in [55] to measure diversity and accuracy. (1) To measure the prediction diversity, we use the Average Pairwise Distance (APD) defined as $\frac{2}{K(K-1)} \sum_{i=1}^{K} \sum_{j=i+1}^{K} \|\hat{\mathbf{Y}}_i - \hat{\mathbf{Y}}_j\|_2$. (2) To measure the reconstruction accuracy over *the whole sequence*, we use the Average Displacement Error (ADE) computed as $\frac{1}{T} \min_i \|\hat{\mathbf{Y}}_i - \mathbf{Y}\|_2$. (3) To measure the reconstruction accuracy of *the last future pose*, we use the Final Displacement Error (FDE) defined as $\min_i \|\hat{\mathbf{Y}}_i[T] - \mathbf{Y}[T]\|_2$. We further report (4) the multi-modal version of ADE (MMADE), similar to $\mathcal{L}_{mm}$, and (5) the multi-modal version of FDE (MMFDE).

Baselines. We compare our method with 3 types of baselines. (1) Deterministic motion prediction methods, includ-
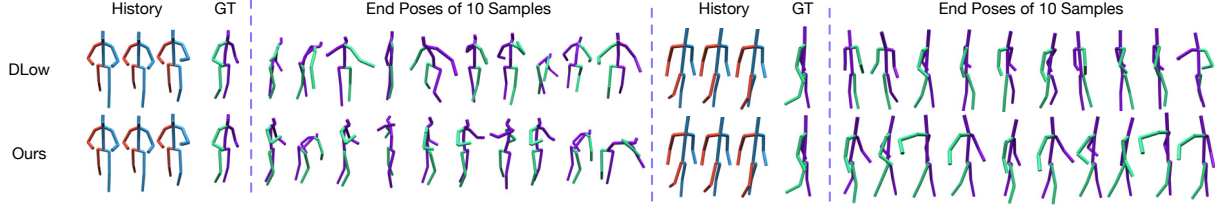
Figure 6: **Results of diverse motion prediction.** We show the past motion with red and blue skeletons, and future poses with green and purple ones.
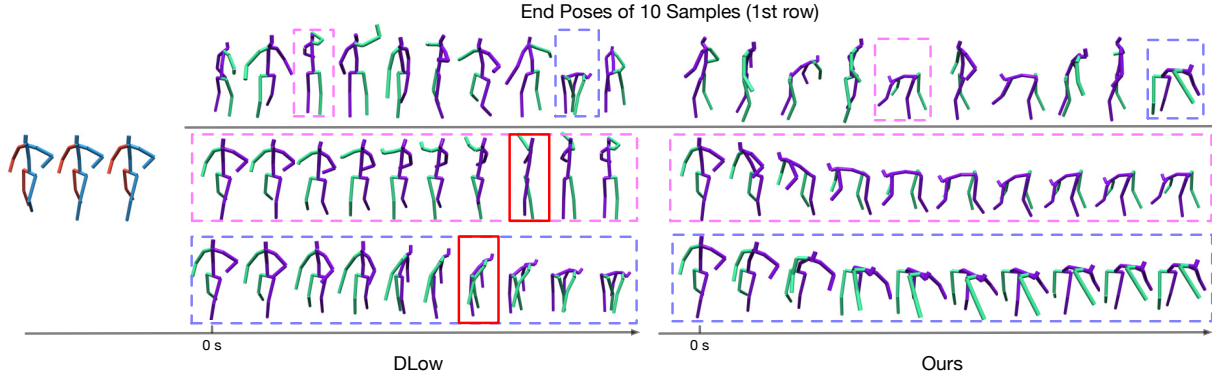

Figure 7: **Visualization of poses on Human3.6M.** In the first row, we show the end poses of 10 samples. Below, we then show different frames corresponding to the two samples highlighted by the magenta and blue dashed boxes. As highlighted by the red boxes, DLow [55] yields invalid poses. This is because it lacks a pose level and kinematics prior.
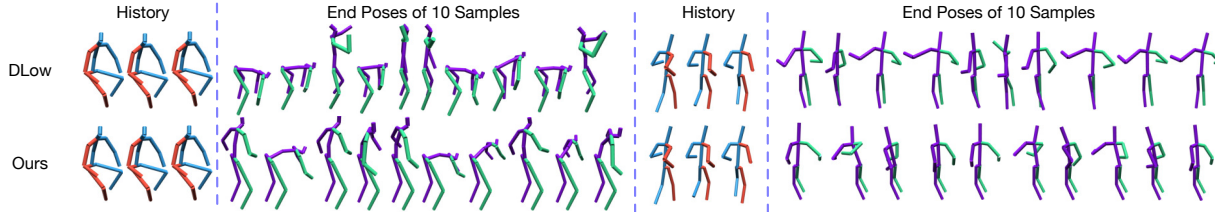

Figure 8: **Controllable motion prediction.** Our approach can generate future motions with the same lower-body motion but diverse upper-body ones.

APD	Human3.6M [26]		HumanEva-I [47]	
	Lower ↓	Upper ↑	Lower ↓	Upper ↑
BoM [7] w/ Eq. 14	0	4.408	0	1.319
DLow [55]-Control	1.071	12.741	0.937	4.671
DLow [55] w/ RS	0.780	7.280	0.571	1.821
Ours	**0**	**13.150**	**0**	**5.096**

Table 2: APD for controllable motion prediction. From first to last row, we show results of BoM [7] with our conditional formulation (Eq. 14), the controllable version of DLow [55], the controllable version of DLow [55] with rejection sampling (RS) and our model respectively. We outperform DLow [55] on both lower-body and upper-body. Note that DLow [55] uses a different model here, whereas our results are obtained using the same model as in Table 1.

ing **ERD** [16] and **acLSTM** [58]; (2) Stochastic motion prediction methods without diversity-promoting technique, including CVAE based methods, **Pose-Knows** [51] and **MT-VAE** [54], as well as a CGAN based method, **HP-GAN** [6]; (3) Diverse motion prediction methods, including **BoM** [7],

GMVAE [14], **DeLiGAN** [20], **DSF** [56], and **DLow** [55]. The results of all baselines are directly reported from[55].

Implementation. We set the hidden size of the generator $\mathcal{G}^{(i)}$ to 256 and the dimension of the random variable $\mathbf{z}^{(i)}$ to 64. To compare with the controllable version of DLow [55], we divide a human pose into 2 parts: lower and upper body ($N = 2$). We also provide qualitative results for $N > 2$ in our supplementary material. The number of samples K is set to 10. Thus, during training, we predict 10 future lower body motions, and for each motion, we generate 10 upper body motions. For Human3.6M, the model is trained using a batch size of 16 for 500 epoch with 5000 training examples per epoch. The weights of the different loss terms $(\lambda_{nf}, \lambda_a, \lambda_{d_1}, \lambda_{d_2}, \lambda_r, \lambda_{mm})$ and the normalizing factors (α_1, α_2) are set to $(0.01, 100, 8, 25, 2, 1)$ and $(100, 300)$, respectively. For HumanEva-I, the model is trained using a batch size of 16 for 500 epoch with 2000 training examples per epoch. The weights of different loss terms $(\lambda_{nf}, \lambda_a, \lambda_{d_1}, \lambda_{d_2}, \lambda_r, \lambda_{mm})$ and the normal-

$\mathcal{L}_{nf}$	$\mathcal{L}_a$	$\mathcal{L}_d$	$\mathcal{L}_r$	$\mathcal{L}_{mm}$	Human3.6M [26]						HumanEva-I [47]					
					APD ↑	ADE ↓	FDE ↓	MMADE ↓	MMFDE ↓	NLL ↓	APD ↑	ADE ↓	FDE ↓	MMADE ↓	MMFDE ↓	NLL ↓
	✓	✓	✓	✓	15.257	0.389	0.497	0.477	0.527	251.190	**7.574**	0.245	0.293	0.416	0.430	363.838
✓		✓	✓	✓	19.608	0.397	0.534	0.506	0.575	97.774	6.569	0.235	0.276	0.406	0.410	119.673
✓	✓		✓	✓	6.318	**0.370**	**0.488**	0.478	0.529	**64.731**	2.048	**0.214**	0.259	0.384	0.398	**75.375**
✓	✓	✓		✓	**20.030**	0.479	0.562	0.513	0.569	89.726	6.778	0.567	0.625	0.606	0.633	109.777
✓	✓	✓	✓		18.079	0.394	0.538	0.520	0.587	91.977	6.474	0.234	0.283	0.415	0.421	113.630
✓	✓	✓	✓	✓	14.757	0.389	0.496	**0.476**	**0.525**	74.872	5.826	0.233	**0.244**	**0.343**	**0.331**	103.306

Table 3: **Ablation Study** on Human3.6M and HumanEva-I.

izing factors (α_1, α_2) are set to $(0.01, 100, 5, 10, 2, 1)$ and $(15, 50)$, respectively. Additional implementation details are provided in the supplementary material.

4.3. Results

Diverse Motion Prediction. In Table 1, we compare our results with the baselines on Human3.6M and HumanEva-I. For all stochastic motion prediction baselines, the results are computed over 50 future motions per test sequence.

Our method consistently outperforms all baselines on all metrics. In general, stochastic motion prediction methods outperform deterministic ones in accuracy (ADE, FDE, MMADE, MMFDE). The reason is that, for multi-modal datasets, deterministic prediction models tend to predict an average mode, which leads to higher errors. For stochastic motion prediction, there is a trade-off between sample diversity (APD) and accuracy. One can achieve high diversity while sacrificing some performance in accuracy, e.g., DSF [56], or vice versa, e.g., BoM [7].

Let us now focus on DLow [55], which constitutes the state of the art. Although, its learnable sampling strategy balances diversity and accuracy well, leading to better results than the other baselines, our method outperforms it on all metrics for both datasets. Note that, on Human3.6M, our method achieves 8% lower reconstruction error (ADE) with 25% higher sample diversity (APD). The qualitative comparisons in Fig. 6 further evidence that our predictions are closer to the GT and more diverse. We further provide a detailed comparison in Fig. 7, which shows that DLow [55] still produces some invalid poses, highlighted with red boxes. This can be caused by its lack of a pose-level prior.

Controllable Motion Prediction. We also compare our method with DLow for controllable motion prediction in Table 2. Here, the prediction model aims to predict future motions with the *same* lower-body motion but *diverse* upper-body motions. Our method gives a *full* control of the lower-body with *higher diversity* on upper-body. By contrast, DLow [55] cannot guarantee the lower-body motion for different samples to be exactly the same. Although rejection sampling[2] helps to achieve a better control of the lower-body motion (third row), the diversity of upper body motion also drops. In Fig. 8, we compare our results with

those of DLow [55], which further supports our conclusions. Moreover, DLow [55] requires a different model for controllable motion prediction, whereas our method yields a unified model able to achieve diverse and controllable motion prediction jointly. We further adapt our conditional formulation (Eq. 14) to one of the most recent baselines (BoM [7]). The results in Table 2 confirm that it also applies to other generators.

4.4. Ablation Study

In Table 3, we evaluate the influence of our different loss terms. In general, there is a trade-off between the diversity loss $\mathcal{L}_d$ and the other losses. Without the diversity loss, the model yields the best ADE at the cost of diversity. By contrast, removing any other loss term leads to higher diversity but sacrifices performance on the corresponding accuracy metrics. Note that, in Table 3, we also report the negative log-likelihood (NLL) of the poses obtained from our pose prior to demonstrate their quality. Although, without the pose prior loss $\mathcal{L}_{nf}$ (first row), we can achieve higher diversity and almost the same accuracy, such diversity gain comes with a dramatic decrease in pose quality (NNL). In other words, while some samples are accurate, many others are not realistic. For qualitative comparison, please refer to the supplementary material.

5. Conclusion

In this paper, we have introduced an end-to-end trainable approach for both diverse and controllable motion prediction. To overcome the likelihood sampling problem that reduces sample diversity, we have developed a normalizing flow based pose prior together with a joint angle loss to encourage producing realistic poses, while enforcing temporal smoothness. To achieve controllable motion prediction, we have designed our generator to decode the motion of different body parts sequentially. Our experiments have demonstrated the effectiveness of our approach. Our current model assumes a predefined sequence of body parts, thus not allowing one to control an arbitrary part at test time. We will focus on addressing this in the future.

---

[2]For each test sequence, we sampled 1000 future motions and chose the 50 with lower-body motion closest to the target one.

References

[1] Ijaz Akhter and Michael J Black. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *CVPR*, pages 1446–1455, 2015. 3

[2] Ijaz Akhter, Yaser Sheikh, Sohaib Khan, and Takeo Kanade. Nonrigid structure from motion in trajectory space. In *Advances in neural information processing systems*, pages 41–48, 2009. 5

[3] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. Structured prediction helps 3d human motion modelling. In *ICCV*, pages 7144–7153, 2019. 2

[4] Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. A stochastic conditioning scheme for diverse human motion prediction. In *CVPR*, pages 5223–5232, 2020. 1, 2, 4

[5] Martín Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. In *ICLR*, 2017. 2

[6] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. In *CVPRW*, pages 1418–1427, 2018. 2, 6, 7

[7] Apratim Bhattacharyya, Bernt Schiele, and Mario Fritz. Accurate and diverse sampling of sequences based on a "best of many" sample objective. In *CVPR*, pages 8485–8493, 2018. 6, 7, 8

[8] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, pages 561–578. Springer, 2016. 2, 3

[9] Matthew Brand and Aaron Hertzmann. Style machines. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 183–192. ACM Press/Addison-Wesley Publishing Co., 2000. 2

[10] Judith Butepage, Michael J. Black, Danica Kragic, and Hedvig Kjellstrom. Deep representation learning for human motion prediction and classification. In *CVPR*, July 2017. 2

[11] 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 *ECCV*, pages 226–242. Springer, 2020. 2

[12] Jixu Chen, Siqi Nie, and Qiang Ji. Data-free prior model for upper body pose estimation and tracking. *IEEE Transactions on Image Processing*, 22(12):4627–4639, 2013. 3

[13] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Abhishek Sharma, and Arjun Jain. Learning 3d human pose from structure and motion. In *ECCV*, pages 668–683, 2018. 3

[14] Nat Dilokthanakul, Pedro AM Mediano, Marta Garnelo, Matthew CH Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016. 6, 7

[15] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *ICLR*, 2017. 2, 3, 4

[16] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *ICCV*, pages 4346–4354, 2015. 1, 2, 6, 7

[17] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014. 2

[18] Anand Gopalakrishnan, Ankur Mali, Dan Kifer, Lee Giles, and Alexander G Ororbia. A neural temporal model for human motion prediction. In *CVPR*, pages 12116–12125, 2019. 2

[19] Liang-Yan Gui, Yu-Xiong Wang, Xiaodan Liang, and José MF Moura. Adversarial geometry-aware human motion prediction. In *ECCV*, pages 786–803, 2018. 2

[20] Swaminathan Gurumurthy, Ravi Kiran Sarvadevabhatla, and R Venkatesh Babu. Deligan: Generative adversarial networks for diverse and limited data. In *CVPR*, pages 166–174, 2017. 6, 7

[21] Lorna Herda, Raquel Urtasun, and Pascal Fua. Hierarchical implicit surface joint limits for human body tracking. *Computer Vision and Image Understanding*, 99(2):189–209, 2005. 3

[22] Alejandro Hernandez, Jurgen Gall, and Francesc Moreno-Noguer. Human motion prediction via spatio-temporal inpainting. In *ICCV*, pages 7134–7143, 2019. 2

[23] Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *TOG*, 36(4):1–13, 2017. 3

[24] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *TOG*, 35(4):1–11, 2016. 3

[25] Chin-Wei Huang, David Krueger, Alexandre Lacoste, and Aaron Courville. Neural autoregressive flows. In *International Conference on Machine Learning*, pages 2078–2087. PMLR, 2018. 4

[26] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 36(7):1325–1339, jul 2014. 6, 7, 8

[27] Ashesh Jain, Amir Roshan Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatiotemporal graphs. In *CVPR*, pages 5308–5317, 2016. 2

[28] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *NIPS*, volume 29. Curran Associates, Inc., 2016. 4

[29] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[30] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017. 6

[31] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *NIPS*, pages 3294–3302, 2015. 2

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

[33] Jogendra Nath Kundu, Maharshi Gor, and R Venkatesh Babu. Bihmp-gan: Bidirectional 3d human motion prediction gan. In *AAAI*, volume 33, pages 8553–8560, 2019. 2

[34] Xiu Li, Hongdong Li, Hanbyul Joo, Yebin Liu, and Yaser Sheikh. Structure from recurrent motion: From rigidity to recurrency. In *CVPR*, June 2018. 1, 2

[35] Xiao Lin and Mohamed R Amer. Human motion modeling using dvgans. *arXiv preprint arXiv:1804.10652*, 2018. 2

[36] Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel Van De Panne. Character controllers using motion vaes. *TOG*, 39(4):40–1, 2020. 3

[37] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *ECCV*, pages 474–489. Springer, 2020. 1, 2

[38] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *ICCV*, pages 9489–9497, 2019. 1, 2, 5, 6

[39] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Multi-level motion attention for human motion prediction. *IJCV*, pages 1–23, 2021. 2

[40] Julieta Martinez, Michael J. Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *CVPR*, July 2017. 1, 2

[41] 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. *IEEE Transactions on intelligent vehicles*, 1(1):33–55, 2016. 1

[42] George Papamakarios, Iain Murray, and Theo Pavlakou. Masked autoregressive flow for density estimation. In *NIPS*, 2017. 4

[43] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019. 2, 3

[44] Dario Pavllo, David Grangier, and Michael Auli. Quaternet: A quaternion-based recurrent model for human motion. In *BMVC*, 2018. 2

[45] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *ICML*, pages 1530–1538. PMLR, 2015. 2, 3, 4

[46] Hedvig Sidenbladh, Michael J Black, and Leonid Sigal. Implicit probabilistic models of human motion for synthesis and tracking. In *ECCV*, pages 784–800. Springer, 2002. 2

[47] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1-2):4, 2010. 6, 7, 8

[48] Ilya Sutskever, James Martens, and Geoffrey E Hinton. Generating text with recurrent neural networks. In *ICML*, pages 1017–1024, 2011. 2

[49] Graham W Taylor, Geoffrey E Hinton, and Sam T Roweis. Modeling human motion using binary latent variables. In *NeurIPS*, pages 1345–1352, 2007. 2

[50] Herwin Van Welbergen, Ben JH Van Basten, Arjan Egges, Zs M Ruttkay, and Mark H Overmars. Real time animation of virtual humans: a trade-off between naturalness and control. In *Computer Graphics Forum*, volume 29, pages 2530–2554. Wiley Online Library, 2010. 1

[51] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *ICCV*, pages 3332–3341, 2017. 1, 2, 6, 7

[52] Borui Wang, Ehsan Adeli, Hsu-kuang Chiu, De-An Huang, and Juan Carlos Niebles. Imitation learning for human pose prediction. In *ICCV*, pages 7124–7133, 2019. 2

[53] Jack M Wang, David J Fleet, and Aaron Hertzmann. Gaussian process dynamical models for human motion. *TPAMI*, 30(2):283–298, 2008. 2

[54] 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 *ECCV*, pages 265–281, 2018. 1, 2, 4, 6, 7

[55] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *ECCV*, pages 346–364. Springer, 2020. 1, 2, 3, 4, 6, 7, 8

[56] Ye Yuan and Kris M Kitani. Diverse trajectory forecasting with determinantal point processes. In *ICLR*, 2019. 6, 7, 8

[57] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. In *ECCV*, pages 465–481. Springer, 2020. 2, 3

[58] Yi Zhou, Zimo Li, Shuangjiu Xiao, Chong He, Zeng Huang, and Hao Li. Auto-conditioned recurrent networks for extended complex human motion synthesis. In *ICLR*, 2018. 6, 7