Adversarial Geometry-Aware Human Motion Prediction

Liang-Yan Gui(✉), Yu-Xiong Wang, Xiaodan Liang, and José M. F. Moura

Carnegie Mellon University, Pittsburgh, USA
{lgui,yuxiongw,xiaodan1,moura}@andrew.cmu.edu

Abstract. We explore an approach to forecasting human motion in a few milliseconds given an input 3D skeleton sequence based on a recurrent encoder-decoder framework. Current approaches suffer from the problem of prediction discontinuities and may fail to predict human-like motion in longer time horizons due to error accumulation. We address these critical issues by incorporating local geometric structure constraints and regularizing predictions with plausible temporal smoothness and continuity from a global perspective. Specifically, rather than using the conventional Euclidean loss, we propose a novel *frame-wise geodesic loss* as a geometrically meaningful, more precise distance measurement. Moreover, inspired by the adversarial training mechanism, we present a new learning procedure to simultaneously validate the sequence-level plausibility of the prediction and its coherence with the input sequence by introducing *two global recurrent discriminators*. An unconditional, fidelity discriminator and a conditional, continuity discriminator are jointly trained along with the predictor in an adversarial manner. Our resulting *adversarial geometry-aware encoder-decoder (AGED)* model significantly outperforms state-of-the-art deep learning based approaches on the heavily benchmarked H3.6M dataset in both short-term and long-term predictions.

Keywords: Human motion prediction · Adversarial learning
Geodesic loss

1 Introduction

Consider the following scenario: a robot is working in our everyday lives and interacting with humans, for example shaking hands during socialization or delivering tools to a surgeon when assisting a surgery. In a seamless interaction, the robot is supposed to not only recognize but also anticipate human actions, such as accurately predicting limbs' pose and position, so that it can respond appropriately and expeditiously [17,25]. Such an ability of forecasting how a human moves or acts in the near future conditioning on a series of historical movements is typically addressed in human motion prediction [4,8,12,13,16,17,24,31]. In addition

L.-Y. Gui and Y.-X. Wang—Equal contributions.

V. Ferrari et al. (Eds.): ECCV 2018, LNCS 11208, pp. 823–842, 2018.
https://doi.org/10.1007/978-3-030-01225-0_48

to the above scenario of human-robot interaction and collaboration [28], human motion prediction also has great application potential in various tasks in computer vision and robotic vision, such as action anticipation [20,27], motion generation for computer graphics [29], and proactive decision-making in autonomous driving systems [35].

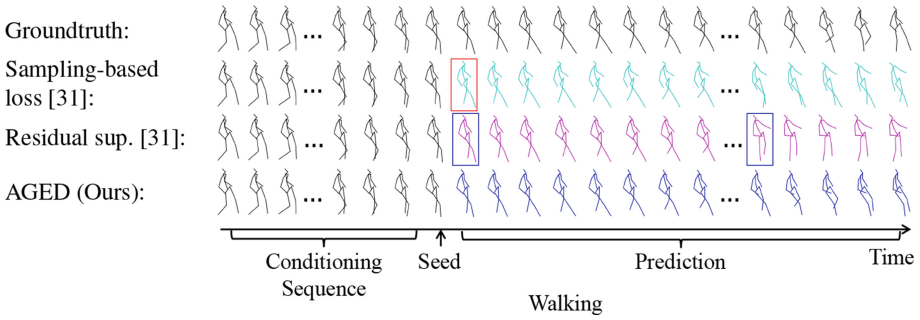

Fig. 1. Human motion prediction task. Top: the conditioning sequence and the groundtruth of the predicted sequence. Middle two: state-of-the-art prediction results (sampling-based loss and residual sup. [31]). Bottom: our prediction. The groundtruth and the input sequences are shown in black. Given the black seed motion frame in the middle, predictions are shown in color. As highlighted in the rectangles, a severe discontinuity exists between the seed motion frame and the first predicted frame for sampling-based loss (2nd row); the prediction is further away from the groundtruth than ours (3rd row, left) and error accumulates in long time horizons (3rd row, right) for residual sup. Our *single* model consistently outperforms the baselines and produces low-error, smooth, and human-like prediction. **Best viewed in color with zoom.** (Color figure online)

Modeling Motion Dynamics: Predicting human motion for diverse actions is challenging yet under-explored, because of the uncertainty of human conscious movements and the difficulty of modeling long-term motion dynamics. State-of-the-art deep learning based approaches typically formulate the task as a sequence-to-sequence problem, and solve it by using recurrent neural networks (RNNs) to capture the underlying temporal dependencies in the sequential data [31]. Despite their extensive efforts on exploring different types of encoder-decoder architectures (*e.g.*, encoder-recurrent-decoder (ERD) [12] and residual [31] architectures), they can only predict periodic actions well (*e.g.*, walking) and show unsatisfactory performance on longer-term aperiodic actions (*e.g.*, discussion), due to error accumulation and severe motion jump between the predicted and input sequences, as shown in Fig. 1. One of the main reasons is that the previous work only considers the frame-wise correctness based on a Euclidean metric at each recurrent training step, while ignoring the critical geometric structure of motion frames and the sequence-level motion fidelity and continuity from a global perspective.

Human-Like Motion Prediction: In this work, we aim to address human-like motion prediction so that the predicted sequences are more plausible and temporally coherent with past sequences. By leveraging the local frame-wise geometric structure and addressing the global sequence-level fidelity and continuity, we propose a novel model that significantly improves the performance of short-term 3D human motion prediction as well as generates realistic periodic and aperiodic long-term motion.

Geometric Structure Aware Loss Function at the Frame Level: Although the motion frames are represented as 3D rotations between joint angles, the standard Euclidean distance is commonly used as the loss function when regressing the predicted frames to the groundtruth during encoder-decoder training. The Euclidean loss fails to exploit the intrinsic geometric structure of 3D rotations, making the prediction inaccurate and even frozen to some mean pose for long-term prediction [24,32]. *Our key insight* is that the matrix representation of 3D rotations belongs to Special Orthogonal Group $SO(3)$ [43], an algebraic group with a Riemannian manifold structure. This manifold structure allows us to define a geodesic distance that is the shortest path between two rotations. We thus introduce a novel geodesic loss between the predicted motion and the groundtruth motion to replace the Euclidean loss. This geometrically more meaningful loss leads to more precise distance measurement and is computationally inexpensive.

Adversarial Training at the Sequence Level: To achieve human-like motion prediction, the model is supposed to be able to validate its entire generated sequence. Unfortunately, such a mechanism is missing in the current prediction framework. In the spirit of generative adversarial networks (GANs) [14], we introduce two global discriminators to validate the prediction while casting our predictor as a generator, and we jointly train them in an adversarial manner. To deal with sequential data, we design our discriminators as *recurrent networks*. The first *unconditional, fidelity discriminator* distinguishes the predicted sequence from the groundtruth sequence. The second *conditional, continuity discriminator* distinguishes between the long sequences that are concatenated from the input sequence and the predicted or groundtruth sequence. *Intuitively*, the fidelity discriminator aims to examine whether the generated motion sequence is human-like and plausible overall, and the continuity discriminator is responsible for checking whether the predicted motion sequence is coherent with the input sequence without a noticeable discontinuity between them.

Our contributions are three-fold. (1) We address human-like motion prediction by modeling both the frame-level geometric structure and the sequence-level fidelity and continuity. (2) We propose a novel geodesic loss and demonstrate that it is more suitable to evaluate 3D motion as the regression loss and is computationally inexpensive. (3) We introduce two complementary recurrent discriminators tailored for the motion prediction task, which are jointly trained along with the geometry-aware encoder-decoder predictor in an adversarial manner. Our full model, which we call *adversarial geometry-aware*

encoder-decoder (AGED), significantly surpasses the state-of-the-art deep learning based approaches when evaluated on the heavily benchmarked, large-scale motion capture (mocap) H3.6M dataset [22]. Our approach is also general and can be potentially incorporated into any encoder-decoder based prediction framework.

2 Related Work

Human Motion Prediction: Human motion prediction is typically addressed by state-space models. Traditional approaches focus on bilinear spatio-temporal basis models [1], hidden Markov models [7], Gaussian process latent variable models [50,53], linear dynamic models [38], and restricted Boltzmann machines [45,47–49]. More recently, driven by the advances of deep learning architectures and large-scale public datasets, various deep learning based approaches have been proposed [4,8,12,13,16,24,31], which significantly improve the prediction performance on a variety of actions.

RNNs for Motion Prediction: In addition to their success in machine translation [26], image caption [58], and time-series prediction [52,57], RNNs [44,54,55] have become the widely used framework for human motion prediction. Fragkiadaki *et al.* [12] propose a 3-layer long short-term memory (LSTM-3LR) network and an encoder-recurrent-decoder (ERD) model that use curriculum learning to jointly learn a representation of pose data and temporal dynamics. Jain *et al.* [24] introduce high-level semantics of human dynamics into RNNs by modeling human activity with a spatio-temporal graph. These two approaches design action-specific models and restrict the training process on subsets of the mocap dataset. Some recent work explores motion prediction for general action classes. Ghosh *et al.* [13] propose a DAE-LSTM model that combines an LSTM-3LR with a dropout autoencoder to model temporal and spatial structures. Martinez *et al.* [31] develop a simple residual encoder-decoder and multi-action architecture by using one-hot vectors to incorporate the action class information. The residual connection exploits first-order motion derivatives to decrease the motion jump between the predicted and input sequences, but its effect is still unsatisfactory. Moreover, error accumulation has been observed in the predicted sequence, since RNNs cannot recover from their own mistake [5]. Some work [12,24] alleviates this problem via a noise scheduling scheme [6] by adding noise to the input during training; nevertheless, this scheme makes the prediction discontinuous and makes the hyper-parameters difficult to tune. While our approach is developed in deterministic motion prediction, it can be potentially extended to probabilistic prediction [4,38,53].

Loss Functions in Prediction Tasks: The commonly used Euclidean loss (*i.e.*, the ℓ_2 loss, and to a lesser extent ℓ_1 loss) [24,31] in prediction tasks can cause the model to average between two possible futures [32] and thus result in blurred video prediction [34] or unrealistic mean motion prediction [24], increasingly worse when predicting further in the future. An image gradient difference loss is

proposed to address this issue for pixel-level video prediction [32], which is not applicable in our task. Here, by taking into the account the intrinsic geometric structure of the motion frames, we adopt a more effective geodesic metric [18,21] to measure 3D rotation errors.

GANs: GANs [2,14] have shown impressive performance in various generation tasks [10,30,32,40,41,51,56,60]. Rather than exploring different objectives in GANs, we investigate how to improve human motion prediction by leveraging the adversarial training mechanism. Our model is different from standard GANs in three ways. (1) Architecture: the discriminators in GANs are mainly convolutional or fully-connected networks [4,23,32,59]; by contrast, our generator and discriminators are both with RNN structures so as to deal with sequences. (2) Training procedure: two discriminators are used at the same time to address the fidelity and continuity challenges, respectively. (3) Loss function: we combine a geodesic (regression) loss with GAN adversarial losses to benefit from both of them. From a broader perspective, our approach can be viewed as imposing (and yet not explicitly enforcing) certain regularizations on the predicted motion, which is loosely related to the classical smoothing, filtering, and prediction techniques [11] but is more trainable and adaptable to real human motion statistics.

3 Adversarial Geometry-Aware Encoder-Decoder Model

Figure 2 illustrates the framework of our adversarial geometry-aware encoder-decoder (AGED) model for human motion prediction. The encoder and decoder constitute the *predictor*, which is trained to minimize the distance between the predicted future sequence and the groundtruth sequence. The standard Euclidean distance is commonly used as the regression loss function. However, it makes the predicted skeleton non-smooth and discontinuous for short-term prediction and frozen to some mean pose for long-term prediction. To deal with such limitations, we introduce an adversarial training process and new loss functions at the global sequence and local frame levels, respectively.

Problem Formulation: We represent human motion as sequential data. Given a motion sequence, we predict possible short-term and long-term motion in the future. That is, we aim to find a mapping $\mathcal{P}$ from an input sequence to an output sequence. The input sequence of length n is denoted as $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$, where $\mathbf{x}_i \in \mathbb{R}^K \, (i \in [1, n])$ is a mocap vector that consists of a set of 3D body joint angles with their exponential map representations [33] and K is the number of joint angles. Consistent with [12,31,48], we standardize the inputs and focus on relative rotations between joints, since they contain information of the actions. We predict the future motion sequence in the next m timesteps as the output, denoted as $\widehat{\mathbf{X}} = \{\widehat{\mathbf{x}}_{n+1}, \widehat{\mathbf{x}}_{n+2}, ..., \widehat{\mathbf{x}}_{n+m}\}$, where $\widehat{\mathbf{x}}_j \in \mathbb{R}^K \;\; (j \in [n+1, n+m])$ is the predicted mocap vector at the j-th timestep. The groundtruth of the m timesteps is given as $\mathbf{X}_{\text{gt}} = \{\mathbf{x}_{n+1}, \mathbf{x}_{n+2}, ..., \mathbf{x}_{n+m}\}$.

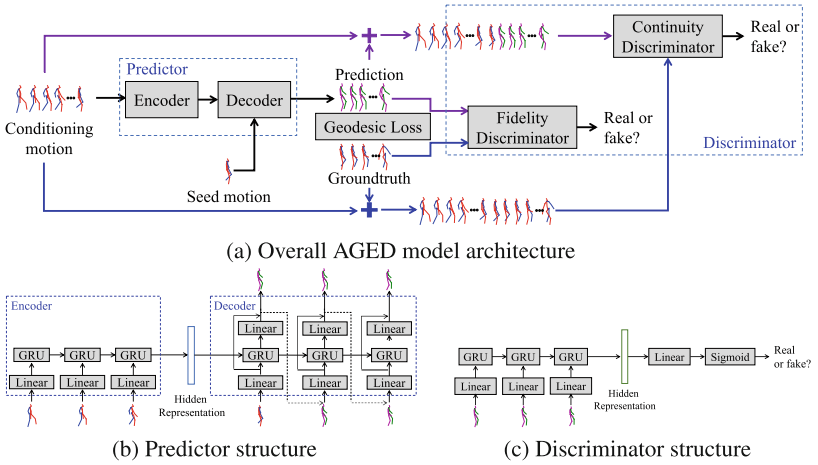

Fig. 2. An overview of our adversarial geometry-aware encoder-decoder (AGED) model. Blue-red skeletons represent the input sequence and groundtruth, and green-purple skeletons represent the prediction. An input sequence is fed into a sequence-to-sequence encoder-decoder network to produce the output sequence (b). We propose a *frame-wise geodesic loss* as a more precise distance measurement to regress the predicted sequence to the groundtruth (a). We further introduce *two global recurrent discriminators* (an unconditional, fidelity discriminator and a conditional, continuity discriminator) to validate the sequence-level plausibility of the prediction and its coherence with the input sequence (c). By jointly optimizing the geometry-aware predictor and the two discriminators in an adversarial manner, we generate the final prediction. (Color figure online)

3.1 Geometry-Aware Encoder-Decoder Predictor

Learning the predictor, *i.e.*, the mapping $\mathcal{P}$ from the input to output sequences, is cast as solving a sequence-to-sequence problem based on an encoder-decoder network architecture [31,46]. The encoder learns a hidden representation from the input sequence. The hidden representation and a seed motion frame are then fed into the decoder to produce the output sequence. Other modifications such as attention mechanisms [3] and bi-directional encoders [42] could be also incorporated into this general architecture.

We use a similar network architecture as in [31] for our predictor $\mathcal{P}$, which has achieved the state-of-the-art performance on motion prediction. The encoder and decoder consist of gated recurrent unit (GRU) [9] cells instead of LSTM [19] or other RNN variants. We use a residual connection to model the motion velocities rather than operating with absolute angles, given that the residual connection has been shown to improve prediction smoothness [31]. Each frame of the input sequence, concatenated with a one-hot vector which indicates the action class of the current input, is fed into the encoder. The decoder takes the output of itself as the next timestep input.

Geodesic Loss: At the local frame level, we introduce a geodesic loss to regress the predicted sequence to the groundtruth sequence frame by frame. Given that the motion frame is represented as 3D rotations of all joint angles, we are interested in measuring the distance between two 3D rotations. The widespread measurement is the Euclidean distance [12,24,31]. However, the crucial geometric structure of 3D rotations is ignored, leading to inaccurate prediction [24,32].

To address such an issue, we introduce a more precise distance measurement and define the new loss accordingly. For a rotation with its Euler angles $\boldsymbol{\theta} = (\alpha, \beta, \gamma)$ about rotation axis $\boldsymbol{u} = (u_1, u_2, u_3)^T$, the corresponding rotation matrix is defined as $\mathbf{R} = [\boldsymbol{\theta} \cdot \mathbf{u}]_\times$, where $\cdot$ and $\times$ denote the inner and outer products, respectively. Such 3D rotation matrices form Special Orthogonal Group SO(3) of orthogonal matrices with determinant 1 [43]. SO(3) is a Lie Group, an algebraic group with a Riemannian manifold structure. It is natural to introduce the geodesic distance to quantify the similarity between two rotations, which is the shortest path between them on the manifold. The geodesic distance in SO(3) can be defined with the angle between two rotation matrices.

Specifically, given two rotation matrices $\widehat{\mathbf{R}}$ and $\mathbf{R}$, the product $\widehat{\mathbf{R}}\mathbf{R}^T$ is the rotation matrix of the difference angle between $\widehat{\mathbf{R}}$ and $\mathbf{R}$. The angle can be calculated using the logarithm map in SO(3) [43] as

$$\log \widehat{\mathbf{R}}\mathbf{R}^T = A \frac{\arcsin\left(\|A\|_2\right)}{\|A\|_2}, \tag{1}$$

where $A = (a_1, a_2, a_3)^T$ and is computed from

$$\frac{\left(\widehat{\mathbf{R}}\mathbf{R}^T - \mathbf{R}\widehat{\mathbf{R}}^T\right)}{2} = \begin{bmatrix} 0 & -a_3 & a_2 \\ a_3 & 0 & -a_1 \\ -a_2 & a_1 & 0 \end{bmatrix}. \tag{2}$$

The geodesic distance between $\widehat{\mathbf{R}}$ and $\mathbf{R}$ is defined as

$$\mathbf{d}_G\left(\widehat{\mathbf{R}}, \mathbf{R}\right) = \left\|\log\left(\widehat{\mathbf{R}}\mathbf{R}^T\right)\right\|_2. \tag{3}$$

Based on this geodesic distance, we now define a geodesic loss $\mathcal{L}_{\text{geo}}$ between the prediction $\widehat{\mathbf{X}}$ and the groundtruth $\mathbf{X}_{\text{gt}}$. We first revert the exponential map representations $\widehat{\mathbf{x}}_j^k$, $\mathbf{x}_j^k$ of the k-th joint in the j-th frame to the Euler format $\widehat{\boldsymbol{\theta}}_j^k$, $\boldsymbol{\theta}_j^k$ [43], respectively, and calculate their corresponding rotation matrices $\widehat{\mathbf{R}}_j^k$, $\mathbf{R}_j^k$, where $k \in [1, K/3]$, $K/3$ is the number of joints (since each joint has 3D joint angles), and $j \in [n+1, n+m]$. By summing up the geodesic distances between the predicted frames and the groundtruth frames, we obtain the geodesic loss in the form of

$$\mathcal{L}_{\text{geo}}\left(\mathcal{P}\right) = \sum_{j=n+1}^{j=n+m} \sum_{k=1}^{k=K/3} \mathbf{d}_G^2\left(\widehat{\mathbf{R}}_j^k, \mathbf{R}_j^k\right). \tag{4}$$

The gradient of Eq. (4) can be computed using automatic gradient computations implemented in the software package such as PyTorch [36] given the forward function. Note that there are other distance metrics that can be defined in $SO(3)$ as well, including the one using quaternion representations [18,21]. Regarding *computing distances*, the quaternion based metric is *functionally equivalent* to our metric [18,21]. Regarding *optimization and computing gradient* as in our case, our current experimental observations indicated that the quaternion based metric led to worse results, possibly due to the need for renormalization of quaternions during optimization [15,39].

3.2 Fidelity and Continuity Discriminators

The sequence-to-sequence predictor architecture explores the temporal information of human motion and produces coarsely plausible motion. However, as shown in Fig. 1, we have observed that there exist some discontinuities between the last frames of the input sequences and the first predicted frames. For long-term prediction, the predicted motion tends to be less realistic due to error accumulation. Such phenomena were also observed in [31]. This is partially because using a frame-wise regression loss solely cannot check the fidelity of the entire predicted sequence from a global perspective. Inspired by the *adversarial training mechanism* in GANs [2,14], we address this issue by introducing *two sequence-level discriminators*.

A standard GAN framework consists of (1) a generator that captures the data distribution, and (2) a discriminator that estimates the probability of a sample being real or generated. The generator is trained to generate samples to fool the discriminator and the discriminator is trained to distinguish the generation from the real samples.

Accordingly, in our model we view the encoder-decoder predictor as a generator, and introduce two discriminators. An unconditional, fidelity discriminator $\mathcal{D}_f$ distinguishes between "short"sequences $\widehat{\mathbf{X}}$ and $\mathbf{X}_{\text{gt}}$. A conditional, continuity discriminator $\mathcal{D}_c$ distinguishes between "long" sequences $\{\mathbf{X}, \widehat{\mathbf{X}}\}$ and $\{\mathbf{X}, \mathbf{X}_{\text{gt}}\}$. Their outputs are the probabilities of their inputs to be "real" rather than "fake". Intuitively, the fidelity discriminator evaluates how smooth and human-like the predicted sequence is and the continuity discriminator checks whether the motion of the predicted sequence is coherent with the input sequence. The quality of the predictor $\mathcal{P}$ is then judged by evaluating how well $\widehat{\mathbf{X}}$ fools $\mathcal{D}_f$ and how well the concatenated sequence $\{\mathbf{X}, \widehat{\mathbf{X}}\}$ fools $\mathcal{D}_c$. More formally, following [14], we solve the minimax optimization problem:

$$\arg\min_{\mathcal{P}} \max_{\mathcal{D}_f, \mathcal{D}_c} \mathcal{L}_{\text{adv}}^{f}\left(\mathcal{P}, \mathcal{D}_f\right) + \mathcal{L}_{\text{adv}}^{c}\left(\mathcal{P}, \mathcal{D}_c\right), \tag{5}$$

where

$$\mathcal{L}_{\text{adv}}^{f}\left(\mathcal{P}, \mathcal{D}_f\right) = \mathbb{E}_{\mathbf{X}_{gt}}\left[\log\left(\mathcal{D}_f(\mathbf{X}_{\text{gt}})\right)\right] + \mathbb{E}_{\mathbf{X}}\left[\log\left(1 - \mathcal{D}_f(\mathcal{P}\left(\mathbf{X}\right))\right)\right], \tag{6}$$

$$\mathcal{L}_{\text{adv}}^{c}\left(\mathcal{P}, \mathcal{D}_c\right) = \mathbb{E}_{\{\mathbf{X}, \mathbf{X}_{gt}\}}\left[\log\left(\mathcal{D}_c(\{\mathbf{X}, \mathbf{X}_{\text{gt}}\})\right)\right] + \mathbb{E}_{\mathbf{X}}\left[\log\left(1 - \mathcal{D}_c(\{\mathbf{X}, \mathcal{P}(\mathbf{X})\})\right)\right], \tag{7}$$

and the distributions $\mathbb{E}(\cdot)$ are over the training motion sequences. Unlike the previous work [4,32,59], we design our discriminators as *recurrent networks* to deal with sequential data. Each of the discriminators consists of GRU cells to extract a hidden representation of its input sequence. A fully-connected layer with sigmoid activation is followed to output the probability that the input sequence is real.

Our entire model thus consists of a *single predictor* and *two discriminators*, extending the generator and discriminator in GANs with recurrent structures. Note that our "generator" is actually a predictor, which is the RNN encoder-decoder *without any noise inputs*. In this sense, the GAN generator maps from noise space to data space, whereas our predictor maps from past sequences to future sequences. During training, the two discriminators are learned jointly.

3.3 Joint Loss Function and Adversarial Training

We integrate the geodesic (regression) loss and the two adversarial losses, and obtain the optimal predictor by jointly optimizing the following minimax objective function:

$$\mathcal{P}^* = \arg\min_{\mathcal{P}} \max_{\mathcal{D}_f, \mathcal{D}_c} \lambda \left(\mathcal{L}_{\text{adv}}^f \left(\mathcal{P}, \mathcal{D}_f \right) + \mathcal{L}_{\text{adv}}^c \left(\mathcal{P}, \mathcal{D}_c \right) \right) + \mathcal{L}_{\text{geo}} \left(\mathcal{P} \right), \tag{8}$$

where λ is the trade-off hyper-parameter that balances the two types of losses. The predictor $\mathcal{P}$ tries to minimize the objective against the adversarial discriminators $\mathcal{D}_f$ and $\mathcal{D}_c$ that aim to maximize it.

Consistent with the recent work [23,37], our combination of a regression loss and GAN adversarial losses provides some complementary benefits. On the one hand, GAN tends to learn a better representation and tries to make prediction look real, which is difficult to achieve using standard hand-crafted metrics. On the other hand, GAN is well known to be hard to train, and easily gets stuck into local minimum (*i.e.*, not learning the distribution). By contrast, the regression loss is responsible for capturing the overall motion geometric structure and explicitly aligning the prediction with the groundtruth.

Implementation Details: We use a similar predictor architecture as in [31] for its state-of-the-art performance. The encoder and decoder consist of a single GRU cell [9] with hidden size 1,024, respectively. Consistent with [31], we found that GRUs are computationally less expensive and a single GRU cell outperforms multiple GRU cells. In addition, it is easier to train and avoids over-fitting compared with the deeper models in [12,24]. We use linear mappings between the K-dim input/output joint angles and the 1,024-dim GRU hidden state. Our two discriminators have the same architectures. For each of them, we also use a single GRU cell. Note that the frames of the sequence being evaluated are fed into the corresponding discriminator sequentially, making its number of parameters unaffected by the sequence length. Our entire model has the same inference time as the baseline model with the plain predictor [31]. The hyper-parameter λ in Eq. (8) is set as 0.6 by cross-validation. We found that the performance is generally robust with

its value ranging from 0.45 to 0.75. We use a learning rate 0.005 and a batch size 16, and we clip the gradient to a maximum ℓ_2-norm of 5. We use PyTorch [36] to train our model and run 50 epochs. It takes 35ms for forward processing and back-propagation per iteration on an NVIDIA Titan GPU.

4 Experiments

In this section, we explore the use of our adversarial geometry-aware encoder-decoder (AGED) model for human motion prediction on the heavily benchmarked motion capture (mocap) dataset [22]. Consistent with the recent work [31], we mainly focus on short-term prediction (<500 ms). We begin with descriptions of the dataset, baselines, and evaluation protocols. Through extensive evaluation, we show that our approach achieves the state-of-the-art short-term prediction performance both quantitatively and qualitatively. We then provide ablation studies, verifying that different losses and modules are complementary with each other for temporal coherent and smooth prediction. Finally, we investigate our approach in long-term prediction (>500 ms) and demonstrate its more human-like prediction results compared with baselines.

Dataset: We focus on the Human 3.6M (H3.6M) dataset [22], a large-scale publicly available dataset including 3.6 million 3D mocap data. This is an important and widely used benchmark in human motion analysis. H3.6M includes seven actors performing 15 varied activities, such as walking, smoking, engaging in a discussion, and taking pictures. We follow the standard experimental setup in [12,24,31]: we down-sample H3.6M by two, train on six subjects, and test on subject five. For short-term prediction, we are given 50 mocap frames (2 seconds in total) and forecast the future 10 frames (400 ms in total). For long-term prediction, we are given the same 50 mocap frames and forecast the future 25 frames (1 second in total) or even more (4 seconds in total).

Baselines: We compare against recent deep RNNs based approaches: (1) LSTM-3LR and ERD [12], (2) SRNN [24], (3) DAE-LSTM [13], and (4) residual sup. and sampling-based loss [31]. Following [31], we also consider a zero-velocity baseline that constantly predicts the last observed frame. As shown in [31], this is a simple but strong baseline: none of these learning based approaches quantitatively outperformed zero-velocity consistently, especially in short-term prediction scenarios.

Evaluation Protocols: We evaluate our approach under three metrics and show both quantitative and qualitative comparisons:

– (Quantitative mean angle error) For a fair comparison, we evaluate the performance using the same error measurement on subject five as in [12,24,31], which is the mean error between the predicted frames and the groundtruth frames in the angle space. Following the preprocessing in [31,48], we exclude the translation and rotation of the whole body.

Table 1. Quantitative comparisons of mean angle error between our AGED model and state-of-the-art approaches for short-term motion prediction on 4 representative activities of the H3.6M dataset. Our model variants include AGED with only the geodesic loss, AGED with two discriminators (the adversarial losses and the conventional Euclidean loss), and full AGED. Our AGED consistently outperforms the existing deep learning based approaches. While the zero-velocity baseline has slightly better performance on smoking at 80 ms prediction, ours outperforms it in all the other cases

	Walking				Eating				Smoking				Discussion			
milliseconds	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
Zero-velocity [31]	0.39	0.68	0.99	1.15	0.27	0.48	0.73	0.86	**0.26**	0.48	0.97	0.95	0.31	0.67	0.94	1.04
ERD [12]	1.30	1.56	1.84	-	1.66	1.93	2.28	-	2.34	2.74	3.73	-	2.67	2.97	3.23	-
LSTM-3LR [12]	1.18	1.50	1.67	-	1.36	1.79	2.29	-	2.05	2.34	3.10	-	2.25	2.33	2.45	-
SRNN [31]	1.08	1.34	1.60	-	1.35	1.71	2.12	-	1.90	2.30	2.90	-	1.67	2.03	2.20	-
DAE-LSTM [13]	1.00	1.11	1.39	-	1.31	1.49	1.86	-	0.92	1.03	1.15	-	1.11	1.20	1.38	-
Sampling-based loss [31]	0.92	0.98	1.02	1.20	0.98	0.99	1.18	1.31	1.38	1.39	1.56	1.65	1.78	1.80	1.83	1.90
Residual sup. [31]	0.28	0.49	0.72	0.81	0.23	0.39	0.62	0.76	0.33	0.61	1.05	1.15	0.31	0.68	1.01	1.09
AGED w/ geo (Ours)	0.28	0.42	0.66	0.73	0.22	0.35	0.61	0.74	0.30	0.55	0.98	0.99	0.30	0.63	0.97	1.06
AGED w/ adv+euc (Ours)	0.27	0.42	0.62	0.71	0.22	0.32	0.53	0.67	0.28	0.47	0.90	0.86	0.28	0.60	0.78	0.87
AGED w/ adv+geo (Ours)	**0.22**	**0.36**	**0.55**	**0.67**	**0.17**	**0.28**	**0.51**	**0.64**	0.27	**0.43**	**0.82**	**0.84**	**0.27**	**0.56**	**0.76**	**0.83**

– (Human evaluation) We also ran double-blind user studies to gauge the plausibility of the prediction as a response to the user. We randomly sample two input sequences from each of the 15 activities on H3.6M, leading to 30 input sequences. We use our model as well as sampling-based loss and residual sup. [31] (which are the top performing baselines as shown below) to generate both short-term and long-term predictions. We thus have 120 short-term motion videos and 120 long-term videos in total, including the short-term and long-term groundtruth videos. We design pairwise evaluations and 25 judges are asked to watch randomly chosen pairs of videos and then choose the one that is considered to be more realistic and reasonable.
– (Qualitative visualization) Following [12,13,24,31], we visualize some representative predictions frame by frame.

For short-term prediction in which the motion is more certain, we evaluate using all the three metrics. For long-term prediction which are more difficult to evaluate quantitatively and might not be unique [31], *we mainly focus on the user studies and visualizations*, and show some quantitative comparisons for reference.

4.1 Short-Term Motion Prediction

Prediction for less than 500 ms is typically considered as short-term prediction. Within this time range, motion is more certain and constrained by physics, and we thus focus on measuring the prediction error with respect to the groundtruth, following [12,24,31]. In these experiments, the network is trained to minimize the loss over 400 ms.

Table 2. Quantitative comparisons of mean angle error between our AGED model and top performing baselines for short-term motion prediction on the remaining 11 activities of the H3.6M dataset. Our AGED model consistently outperforms these baselines in almost all the scenarios

	Directions				Greeting				Phoning				Posing				Purchases				Sitting			
milliseconds	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
Zero-velocity [31]	0.39	0.59	0.79	0.89	**0.54**	0.89	1.30	1.49	0.64	1.21	1.65	1.83	**0.28**	**0.57**	1.13	1.37	0.62	0.88	1.19	1.27	**0.40**	1.63	**1.02**	**1.18**
Residual sup. [31]	0.26	0.47	0.72	0.84	0.75	1.17	1.74	1.83	0.23	0.43	0.69	0.82	0.36	0.71	1.22	1.48	0.51	0.97	1.07	1.16	0.41	1.05	1.49	1.63
AGED w/ geo (Ours)	0.26	0.46	0.71	0.81	0.61	0.95	1.44	1.61	0.23	0.42	0.61	0.79	0.34	0.70	1.19	1.40	0.46	0.89	1.06	1.11	0.46	0.87	1.23	1.51
AGED w/ adv+euc (Ours)	0.26	0.42	0.66	0.73	0.58	0.88	1.31	1.49	0.21	0.37	0.51	0.69	0.34	0.62	1.15	1.39	0.49	0.83	1.05	1.12	0.44	0.77	1.08	1.21
AGED w/ adv+geo (Ours)	**0.23**	**0.39**	**0.63**	**0.69**	0.56	**0.81**	**1.30**	**1.46**	**0.19**	**0.34**	**0.50**	**0.68**	0.31	0.58	**1.12**	**1.34**	**0.46**	**0.78**	**1.01**	**1.07**	0.41	**0.76**	1.05	1.19

	Sitting Down				Taking Photo				Waiting				Walking Dog				Walking Together				Average			
milliseconds	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
Zero-velocity [31]	0.39	0.74	1.07	1.19	0.25	0.51	**0.79**	**0.92**	0.34	0.67	1.22	1.47	0.60	0.98	1.36	1.50	0.33	0.66	0.94	0.99	0.40	0.71	1.07	1.21
Residual sup. [31]	0.39	0.81	1.40	1.62	0.24	0.51	0.90	1.05	0.28	0.53	1.02	1.14	0.56	0.91	1.26	1.40	0.31	0.58	0.87	0.91	0.36	0.67	1.02	1.15
AGED w/ geo (Ours)	0.38	0.77	1.18	1.41	0.24	0.52	0.92	1.01	0.31	0.64	1.08	1.12	0.51	0.87	1.21	1.33	0.29	0.51	0.72	0.75	0.32	0.62	0.96	1.07
AGED w/ adv+euc (Ours)	0.34	0.67	1.01	1.11	0.24	0.49	0.84	0.97	0.26	0.54	1.05	1.28	0.55	0.84	1.16	1.30	0.24	0.44	0.60	0.64	0.33	0.58	0.88	1.00
AGED w/ adv+geo (Ours)	**0.33**	**0.62**	**0.98**	**1.10**	**0.23**	**0.48**	0.81	0.95	**0.24**	**0.50**	**1.02**	**1.13**	**0.50**	**0.81**	**1.15**	**1.27**	**0.23**	**0.41**	**0.56**	**0.62**	**0.31**	**0.54**	**0.85**	**0.97**

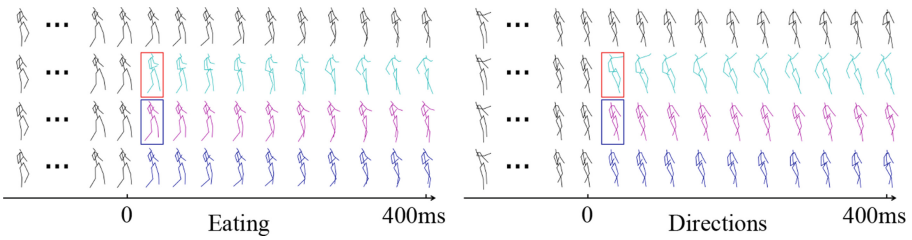

Fig. 3. Short-term motion prediction visualizations. From top to bottom: groundtruth, sampling-based loss [31], residual sup. [31], and our AGED. As highlighted in the rectangles, discontinuities exit between the inputs and the first predicted frames (2nd rows); the predictions are further away from the groundtruth than ours (3rd rows). Our AGED produces lower-error, less-jump, and smoother predictions. **Best viewed in color with zoom.** (Color figure online)

Comparisons with State-of-the-Art Deep Learning Baselines: Table 1 shows the quantitative comparisons with the full set of deep learning baselines on 4 representative activities, including walking, smoking, eating, and discussion. Table 2 compares our approach with the best performing residual sup. baseline on the remaining 11 activities. Compared with residual sup. that uses a similar predictor network but a Euclidean loss, our geodesic loss generates more precise prediction. Our discriminators further greatly boost the performance, validating that the high-level fidelity examination of the entire predicted sequence is essential for smooth and coherent motion prediction. Their combination achieves the best performance and makes our AGED model consistently outperform the existing deep learning based approaches *in all the scenarios*.

Comparisons with the Zero-Velocity Baseline: Tables 1 and 2 also summarize the comparisons with the zero-velocity approach. Although zero-velocity does not produce interesting motion, it is difficult for the existing deep learning based approaches to outperform it quantitatively in short-term prediction,

mainly on complicated actions (*e.g.*, smoking) and highly aperiodic actions (*e.g.*, sitting), which is consistent with the observations in [31]. Our AGED model shows some promising progress. (1) For *complicated motion prediction*, zero-velocity outperforms the other baselines, whereas our AGED outperforms zero-velocity, due to our adversarial discriminators. This type of action consists of small movements in upper-body, which is difficult to model as the learning based baselines only verify frame-wise predictions and ignore their temporal dependencies. By contrast, our AGED, equipped with a fidelity discriminator and a continuity discriminator, is able to check globally how smooth and human-like the entire generated sequence is, leading to significant performance improvement. (2) For *highly aperiodic motion prediction*, because these actions are very difficult to model, zero-velocity outperforms all the learning methods.

Qualitative Visualizations: Figure 3 visualizes the motion prediction results. We compare with residual sup., the best performing baseline as shown in Tables 1 and 2. Both our AGED model and residual sup. predict realistic short-term motion. One noticeable difference between them is the degree of jump (*i.e.*, discontinuity) between the last input frame and the first predicted frame. The jump in our AGED is relatively small, which is consistent with its lower prediction error and due to the introduced continuity discriminator. We also include sampling-based loss, a variant of residual sup., which shows superior qualitative visualization in long-term prediction. We observe severe discontinuities in sampling-based loss. More comparisons are shown in Fig. 1.

User Studies: Our model again outperforms the baselines by a large margin under human evaluation, as shown in Table 3. The first row summarizes the success rates of our AGED against the groundtruth and baselines. For short-term prediction, we observe that (1) our AGED has a success rate of 53.3% against the groundtruth, showing that our predictions are *on par with the groundtruth*; and (2) our AGED has a success rate of 98.6% against sampling-based loss and of 69.6% against residual sup., showing that the judges notice the jump and discontinuities between the baseline predictions and the input sequences. As a reference, the second row summarizes the success rates of the groundtruth against all the models, which have similar trends as our rates and are slightly better. These observations thus validate that users favor more realistic and plausible motion and our predictions are judged qualitatively realistic by humans.

4.2 More Ablation Studies

In addition to the comparisons between the geodesic loss and the adversarial losses in Tables 1 and 2, we conduct more thorough ablations in Table 4 to understand the impact of each loss component and their combinations. We can see that our geodesic loss is superior to the conventional Euclidean loss. This observation empirically verifies that the geodesic distance is a geometrically more meaningful and more precise measurement for 3D rotations, as also supported by the theoretical analysis in [18,21,43]. Moreover, our full model consistently outperforms its variants in short-term prediction, showing the effectiveness and complementarity of each component.

Table 3. Human voting results of short-term and long-term prediction videos. Each number represents the percentage that our prediction or the groundtruth is chosen from a pair of predictions as being more realistic and reasonable. The first row shows the percentages that our predictions are chosen against the groundtruth and baseline predictions. As a reference, the second row shows the percentages for the groundtruth. Our AGED predictions are on par with the groundtruth and significantly outperform baseline models

	Short-term				Long-term			
Model pair	Ours	Groundtruth	sampling-based loss [31]	residual sup. [31]	Ours	Groundtruth	sampling-based loss [31]	residual sup. [31]
Ours vs.	n/a	53.3%	98.6%	69.6%	n/a	48.7%	83.5%	93.1%
Groundtruth vs.	46.7%	n/a	99.7%	75.7%	51.3%	n/a	83.7%	94.9%

Table 4. Ablation analysis for short-term prediction. Some of the results are included from Table 1 for completeness. We compare our geodesic loss with the conventional Euclidean loss as the predictor regression loss and evaluate the impact of different discriminators and their combinations. Our full AGED model achieves the best performance, showing that the different components complement each other

milliseconds			Walking				Eating				Smoking				Discussion			
reg loss	fid dis	con dis	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
n/a	✓	✓	1.35	1.33	1.31	1.55	1.29	1.22	1.38	1.41	1.39	1.51	1.53	1.69	1.37	1.22	1.15	1.51
euc			0.28	0.49	0.72	0.81	0.23	0.39	0.62	0.76	0.33	0.61	1.05	1.15	0.31	0.68	1.01	1.09
euc	✓		0.27	0.43	0.66	0.74	0.23	0.35	0.58	0.71	0.28	0.52	0.94	0.90	0.42	0.62	0.87	0.93
euc		✓	0.26	0.42	0.63	0.71	0.22	0.34	0.54	0.68	0.28	0.48	0.92	0.91	0.39	0.63	0.86	0.96
euc	✓	✓	0.27	0.42	0.62	0.71	0.22	0.32	0.53	0.67	0.28	0.47	0.90	0.86	0.28	0.60	0.78	0.87
geo			0.28	0.42	0.66	0.73	0.22	0.35	0.61	0.74	0.30	0.55	0.98	0.99	0.30	0.63	0.97	1.06
geo	✓		0.24	0.39	0.62	0.71	0.22	0.32	0.56	0.68	0.27	0.46	0.89	0.87	0.33	0.59	0.80	0.91
geo		✓	0.24	0.39	0.58	0.68	0.21	0.30	0.52	0.66	0.27	0.45	0.84	0.86	0.34	0.57	0.81	0.90
geo	✓	✓	**0.22**	**0.36**	**0.55**	**0.67**	**0.17**	**0.28**	**0.51**	**0.64**	**0.27**	**0.43**	**0.82**	**0.84**	**0.27**	**0.56**	**0.76**	**0.83**

4.3 Long-Term Motion Prediction

Long-term prediction (>500 ms) is more challenging than short-term prediction due to error accumulation and the uncertainty of human motion. Given that long-term prediction is difficult to evaluate quantitatively [31], we mainly focus on the qualitative comparisons in Fig. 4 and the user studies in Table 4. For completeness, we provide representative quantitative evaluation in Table 5. Here the network is trained to minimize the loss over 1 second. While residual sup. [31] achieves the best performance among the baselines in short-term prediction, it is shown that sampling-based loss [31], a variant of residual sup. without residual connections and one-hot vector inputs, qualitatively outperforms it in long-term prediction. We show that our AGED model *consistently* outperforms these two baselines in both short-term and long-term predictions.

Table 5. Representative quantitative comparisons of mean angle error between our AGED model and state-of-the-art approaches for long-term motion prediction on 4 activities of the H3.6M dataset. Our AGED model consistently achieves the best performance

	Walking		Eating		Smoking		Discussion	
milliseconds	560	1000	560	1000	560	1000	560	1000
Zero-velocity [31]	1.35	1.32	1.04	1.38	1.02	1.69	1.41	1.96
ERD [12]	2.00	2.38	2.36	2.41	3.68	3.82	3.47	2.92
LSTM-3LR [12]	1.81	2.20	2.49	2.82	3.24	3.42	2.48	2.93
SRNN [31]	1.90	2.13	2.28	2.58	3.21	3.23	2.39	2.43
DAE-LSTM [13]	1.55	1.39	1.76	2.01	1.38	1.77	1.53	1.73
Sampling-based loss [31]	1.36	1.59	1.48	1.55	1.78	2.31	1.77	1.61
Residual sup. [31]	0.93	1.03	0.95	1.08	1.25	1.50	1.43	1.69
AGED w/ geo (Ours)	0.89	1.02	0.92	1.01	1.15	1.43	1.33	1.56
AGED w/ adv+euc (Ours)	0.87	0.99	0.87	0.96	1.16	1.38	1.31	1.39
AGED w/ adv+geo (Ours)	**0.78**	**0.91**	**0.86**	**0.93**	**1.06**	**1.21**	**1.25**	**1.30**

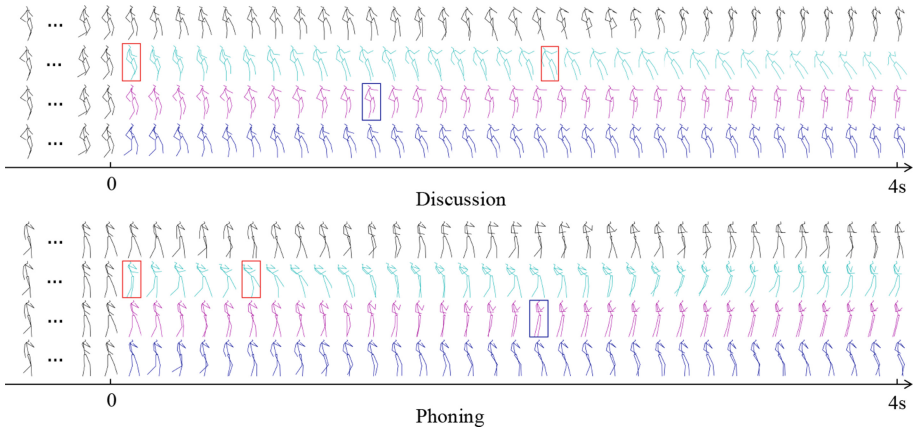

Fig. 4. Long-term motion prediction visualizations. From top to bottom for each activity: groundtruth, sampling-based loss [31], residual sup. [31], and our AGED. As highlighted in the rectangles, discontinuities exit between the inputs and the first predicted frames (2nd rows, left) and predictions drift away to unrealistic motions (2nd rows, right); predictions converge to mean poses (3rd rows). Our AGED produces more realistic, continuous, and human-like predictions. **Best viewed in color with zoom.** (Color figure online)

Qualitative Visualizations: Figure 4 shows some representative comparisons on discussion and phoning activities, which are challenging aperiodic actions. We observe that the generated predictions by residual sup. converge to mean poses and the predictions of sampling-based loss often drift away from the input sequences, making them unrealistic anymore. Our model, however, produces more plausible, continuous, and human-like prediction in long time horizons (4 seconds).

User Studies: As shown in Table 3, our model significantly improves long-term prediction based on human evaluation. Our AGED has a success rate of 48.7% against the groundtruth, showing that our predictions are still *comparable with the groundtruth*. Moreover, our AGED has success rates of 83.5% and 93.1% against sampling-based loss and residual sup., respectively, which are much larger margins of improvement compared with the corresponding rates in short-term prediction. These results demonstrate that the judges consider that our predictions are more realistic and plausible.

5 Conclusions

We present a novel adversarial geometry-aware encoder-decoder model to address the challenges in human-like motion prediction: how to make the predicted sequences temporally coherent with past sequences and more realistic. At the frame level, we propose a new geodesic loss to quantify the difference between the predicted 3D rotations and the groundtruth. We further introduce two recurrent discriminators, a fidelity discriminator and a continuity discriminator, to validate the predicted motion sequence from a global perspective. Training integrates the two objectives and is conducted in an adversarial manner. Extensive experiments on the heavily benchmarked H3.6M dataset show the effectiveness of our model for both short-term and long-term motion predictions.

Acknowledgments. We thank Richard Newcombe, Hauke Strasdat, and Steven Lovegrove for insightful discussions at Facebook Reality Labs where L.-Y. Gui was a research intern. We also thank Deva Ramanan and Hongdong Li for valuable comments.

References

1. Akhter, I., Simon, T., Khan, S., Matthews, I., Sheikh, Y.: Bilinear spatiotemporal basis models. ACM Trans. Graph. (TOG) **31**(2), 17:1–17:12 (2012)
2. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein GAN, January 2017. arXiv preprint arXiv:1701.07875
3. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: International Conference on Learning Representations (ICLR), San Diego, CA, USA, May 2015
4. Barsoum, E., Kender, J., Liu, Z.: HP-GAN: probabilistic 3D human motion prediction via GAN, November 2017. arXiv preprint arXiv:1711.09561
5. Bengio, S., Vinyals, O., Jaitly, N., Shazeer, N.: Scheduled sampling for sequence prediction with recurrent neural networks. In: Advances in Neural Information Processing Systems (NIPS), Montréal, Canada, pp. 1171–1179, December 2015
6. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: International Conference on Machine Learning (ICML), Montréal, Canada, pp. 41–48, June 2009
7. Brand, M., Hertzmann, A.: Style machines. In: ACM International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH), New Orleans, LA, USA, pp. 183–192, July 2000

8. Bütepage, J., Black, M.J., Kragic, D., Kjellström, H.: Deep representation learning for human motion prediction and classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 1591–1599, July 2017
9. Cho, K., Van Merriënboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: encoder-decoder approaches. In: Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST), Doha, Qatar, pp. 103–111, October 2014
10. Denton, E.L., Chintala, S., Fergus, R.: Deep generative image models using a Laplacian pyramid of adversarial networks. In: Advances in Neural Information Processing Systems (NIPS), Montréal, Canada, pp. 1486–1494, December 2015
11. Einicke, G.A.: Smoothing, filtering and prediction: estimating the past, present and future. InTech, February 2012
12. Fragkiadaki, K., Levine, S., Felsen, P., Malik, J.: Recurrent network models for human dynamics. In: IEEE International Conference on Computer Vision (ICCV), Las Condes, Chile, pp. 4346–4354, December 2015
13. Ghosh, P., Song, J., Aksan, E., Hilliges, O.: Learning human motion models for long-term predictions. In: International Conference on 3D Vision (3DV), Qingdao, China, pp. 458–466, October 2017
14. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems (NIPS), Montréal, Canada, pp. 2672–2680, December 2014
15. Grassia, F.S.: Practical parameterization of rotations using the exponential map. J. Graph. Tools **3**(3), 29–48 (1998)
16. Gui, L.Y., Wang, Y.X., Ramanan, D., Moura, J.M.F.: Few-shot human motion prediction via meta-learning. In: European Conference on Computer Vision (ECCV), Munich, Germany, September 2018
17. Gui, L.Y., Zhang, K., Wang, Y.X., Liang, X., Moura, J.M.F., Veloso, M.M.: Teaching robots to predict human motion. In: IEEE/RSJ International Conference on Intelligent Robots (IROS), Madrid, Spain, October 2018
18. Hartley, R., Trumpf, J., Dai, Y., Li, H.: Rotation averaging. Int. J. Comput. Vis. (IJCV) **103**(3), 267–305 (2013)
19. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
20. Huang, D.-A., Kitani, K.M.: Action-reaction: forecasting the dynamics of human interaction. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8695, pp. 489–504. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10584-0_32
21. Huynh, D.Q.: Metrics for 3D rotations: comparison and analysis. J. Math. Imaging Vis. **35**(2), 155–164 (2009)
22. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **36**(7), 1325–1339 (2014)
23. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 5967–5976, July 2017
24. Jain, A., Zamir, A.R., Savarese, S., Saxena, A.: Structural-RNN: deep learning on spatio-temporal graphs. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 5308–5317, June–July 2016

25. Jong, M.D., et al.: Towards a robust interactive and learning social robot. In: International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), Stockholm, Sweden, July 2018
26. Kiros, R., et al.: Skip-thought vectors. In: Advances in Neural Information Processing Systems (NIPS), Montréal, Canada, pp. 3294–3302, December 2015
27. Koppula, H., Saxena, A.: Learning spatio-temporal structure from RGB-D videos for human activity detection and anticipation. In: International Conference on Machine Learning (ICML), Atlanta, GA, USA, pp. 792–800, June 2013
28. Koppula, H.S., Saxena, A.: Anticipating human activities using object affordances for reactive robotic response. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **38**(1), 14–29 (2016)
29. Kovar, L., Gleicher, M., Pighin, F.: Motion graphs. In: ACM International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH), San Antonio, TX, USA, pp. 473–482, July 2002
30. Liang, X., Lee, L., Dai, W., Xing, E.P.: Dual motion GAN for future-flow embedded video prediction. In: IEEE International Conference on Computer Vision (ICCV), Venice, Italy, pp. 1762–1770, October 2017
31. Martinez, J., Black, M.J., Romero, J.: On human motion prediction using recurrent neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 4674–4683, July 2017
32. Mathieu, M., Couprie, C., LeCun, Y.: Deep multi-scale video prediction beyond mean square error. In: International Conference on Learning Representations (ICLR), San Juan, PR, USA, May 2016
33. Murray, R.M., Li, Z., Sastry, S.S., Sastry, S.S.: A Mathematical Introduction to Robotic Manipulation, 1st edn. CRC Press, Boca Raton (1994)
34. Oh, J., Guo, X., Lee, H., Lewis, R.L., Singh, S.: Action-conditional video prediction using deep networks in Atari games. In: Advances in Neural Information Processing Systems (NIPS), Montréal, Canada, pp. 2863–2871, December 2015
35. Paden, B., Čáp, M., Yong, S.Z., Yershov, D., Frazzoli, E.: A survey of motion planning and control techniques for self-driving urban vehicles. IEEE Trans. Intell. Veh. (T-IV) **1**(1), 33–55 (2016)
36. Paszke, A., et al.: Automatic differentiation in PyTorch. In: Advances in Neural Information Processing Systems (NIPS) Workshops, Long Beach, CA, USA, December 2017
37. Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., Efros, A.: Context encoders: feature learning by inpainting. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, LV, USA, pp. 2536–2544, June–July 2016
38. Pavlovic, V., Rehg, J.M., MacCormick, J.: Learning switching linear models of human motion. In: Advances in Neural Information Processing Systems (NIPS), Vancouver, Canada, pp. 981–987, December 2001
39. Pennec, X., Thirion, J.P.: A framework for uncertainty and validation of 3-D registration methods based on points and frames. Int. J. Comput. Vis. (IJCV) **25**(3), 203–229 (1997)
40. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: International Conference on Learning Representations (ICLR), San Juan, PR, USA, May 2016
41. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: International Conference on Machine Learning (ICML), New York, USA, pp. 1060–1069, June 2016

42. Ren, M., Kiros, R., Zemel, R.: Exploring models and data for image question answering. In: Advances in Neural Information Processing Systems (NIPS), Montréal, Canada, pp. 2953–2961, December 2015
43. Rossmann, W.: Lie Groups: An Introduction Through Linear Groups, vol. 5. Oxford University Press, Oxford (2002)
44. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature **323**(6088), 533–536 (1986)
45. Sutskever, I., Hinton, G.E., Taylor, G.W.: The recurrent temporal restricted Boltzmann machine. In: Advances in Neural Information Processing Systems (NIPS), Vancouver, Canada, pp. 1601–1608, December 2009
46. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems (NIPS), Montréal, Canada, pp. 3104–3112, December 2014
47. Taylor, G.W., Hinton, G.E.: Factored conditional restricted Boltzmann machines for modeling motion style. In: International Conference on Machine Learning (ICML), Montréal, Canada, pp. 1025–1032, June 2009
48. Taylor, G.W., Hinton, G.E., Roweis, S.T.: Modeling human motion using binary latent variables. In: Advances in Neural Information Processing Systems (NIPS), Vancouver, Canada, pp. 1345–1352, December 2007
49. Taylor, G.W., Sigal, L., Fleet, D.J., Hinton, G.E.: Dynamical binary latent variable models for 3D human pose tracking. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, pp. 631–638, June 2010
50. Urtasun, R., Fleet, D.J., Geiger, A., Popović, J., Darrell, T.J., Lawrence, N.D.: Topologically-constrained latent variable models. In: International Conference on Machine Learning (ICML), Helsinki, Finland, pp. 1080–1087, July 2008
51. Vondrick, C., Pirsiavash, H., Torralba, A.: Generating videos with scene dynamics. In: Advances in Neural Information Processing Systems (NIPS), Barcelona, Spain, pp. 613–621, December 2016
52. Walker, J., Marino, K., Gupta, A., Hebert, M.: The pose knows: video forecasting by generating pose futures. In: IEEE International Conference on Computer Vision (ICCV), Venice, Italy, pp. 3352–3361, October 2017
53. Wang, J.M., Fleet, D.J., Hertzmann, A.: Gaussian process dynamical models for human motion. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **30**(2), 283–298 (2008)
54. Werbos, P.J.: Backpropagation through time: what it does and how to do it. Proc. IEEE **78**(10), 1550–1560 (1990)
55. Williams, R.J., Zipser, D.: A learning algorithm for continually running fully recurrent neural networks. Neural Comput. **1**(2), 270–280 (1989)
56. Wu, J., Zhang, C., Xue, T., Freeman, B., Tenenbaum, J.: Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. In: Advances in Neural Information Processing Systems (NIPS), Barcelona, Spain, pp. 82–90, December 2016
57. Xue, T., Wu, J., Bouman, K., Freeman, B.: Visual dynamics: probabilistic future frame synthesis via cross convolutional networks. In: Advances in Neural Information Processing Systems (NIPS), Barcelona, Spain, pp. 91–99, December 2016

58. Yang, Z., Yuan, Y., Wu, Y., Cohen, W.W., Salakhutdinov, R.R.: Review networks for caption generation. In: Advances in Neural Information Processing Systems (NIPS), Barcelona, Spain, pp. 2361–2369, December 2016
59. Zhou, Y., Berg, T.L.: Learning temporal transformations from time-lapse videos. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 262–277. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_16
60. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE International Conference on Computer Vision (ICCV), Venice, Italy, pp. 2223–2232, October 2017