Aggregated Multi-GANs for Controlled 3D Human Motion Prediction

Zhenguang Liu[1], Kedi Lyu[2*], Shuang Wu[3*], Haipeng Chen[2*], Yanbin Hao[4], Shouling Ji[5]

[1]Zhejiang Gongshang University, [2]Jilin University, [3]Nanyang Technological University,
[4]University of Science and Technology of China, [5]Zhejiang University
liuzhenguang2008@gmail.com, lvkd19@mails.jlu.edu.cn, wushuang@outlook.sg, chenhp@jlu.edu.cn,
haoyanbin@hotmail.com, sji@zju.edu.cn

Abstract

Human motion prediction from historical pose sequence is at the core of many applications in machine intelligence. However, in current state-of-the-art methods, the predicted future motion is confined within the same activity. One can neither generate predictions that differ from the current activity, nor manipulate the body parts to explore various future possibilities. Undoubtedly, this greatly limits the usefulness and applicability of motion prediction. In this paper, we propose a generalization of the human motion prediction task in which control parameters can be readily incorporated to adjust the forecasted motion. Our method is compelling in that it enables manipulable motion prediction across activity types and allows customization of the human movement in a variety of fine-grained ways. To this aim, a simple yet effective composite GAN structure, consisting of local GANs for different body parts and aggregated via a global GAN is presented. The local GANs game in lower dimensions, while the global GAN adjusts in high dimensional space to avoid mode collapse. Extensive experiments show that our method outperforms state-of-the-art. Our codes are released[1].

Introduction

An important component of our capacity to interact with the external world resides in the ability to predict the future (Martinez, Black, and Romero 2017), based on existing cues and past experience. Having a mental representation of how other humans are likely to act is crucial for us to avoid dangers and make decisions. Likewise, the ability for machines to anticipate and model human motion dynamics is very much coveted. Modeling and predicting human motion is at the heart of many important applications in the computer vision domain such as video making, human posture tracking (Taylor et al. 2010), as well as the robotics domain such as regulating a robot's response and behaviour in human-machine interactions (Koppula and Saxena 2013; Koppula, Jain, and Saxena 2016).

Human motion prediction is challenging since humans do not follow fixed laws of motion as is the case for inanimate objects. The inherent stochastic nature of human behaviour invokes complexity in the form of non-linearity and high dimensionality. Earlier approaches have typically utilized shallow latent-variable models such as hidden Markov models (Lehrmann, Gehler, and Nowozin 2014) or Gaussian processes (Wang, Fleet, and Hertzmann 2005) to represent the motion dynamics with hidden states. In recent years, many deep learning approaches were proposed and have yielded improved results. A first class of work utilised recurrent neural networks (RNNs) such as Long Short Term Memory (LSTM) (Fragkiadaki et al. 2015), Gated Recurrent Unit (GRU) (Martinez, Black, and Romero 2017; Pavllo et al. 2019) or refined RNN architectures tailored for modeling motion dynamics. Another line of work built upon the success of deep generative models including Variational Autoencoders (VAEs) (Habibie et al. 2017; Yan et al. 2018) and Generative Adversarial Networks (GANs) (Barsoum, Kender, and Liu 2018; Kundu, Gor, and Babu 2019).

A major shortcoming of prior works is that they can only predict a single output sequence for any given observed motion sequence. However, given our incomplete knowledge of the present, the future should not be represented as a single deterministic state but rather a spectrum of plausible out-

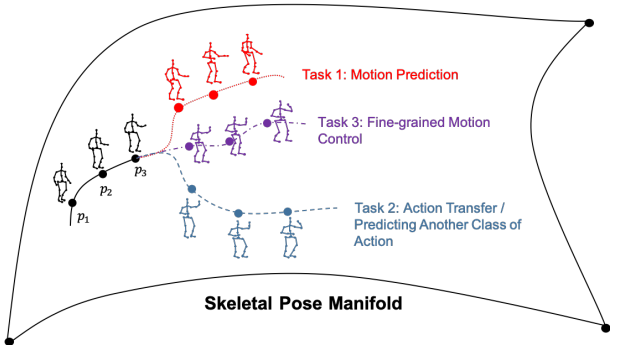

Figure 1: Problem illustration. Mathematically, an observed motion sequence traces out a trajectory on the skeletal pose manifold. The problem of human motion prediction amounts to extrapolating this trajectory in a smooth and coherent manner. We are interested in tasks: 1) predicting the most likely motion (e.g., continuation of the walking action); 2) predicting the transition to another specified action (e.g. transition from current action to eating); 3) fine-grained control over body parts, e.g., requiring the upper body performing eating action while the lower body engaging in walking.

*Corresponding Authors

[1]https://github.com/herolvkd/AM-GAN

comes. Considering only a single future outcome may be inadequate and lead to misinformed decision making. For example, in safety-crucial applications such as autonomous driving, unexpected maneuvers such as sudden overtaking must be accounted for in guiding the course of actions.

Another severe limitation of existing works is that they can only predict one single activity belonging to the same activity as the observed sequence, e.g. predicting 'walking' motion from an observed 'walking' sequence. For real applications such as video synthesis, animation, and game character movement generation, we usually want to (1) customize a future motion with fine-grained control on the body parts, e.g., requiring the whole body act in 'walking' activity but with the right arm act in 'eating', and (2) command the prediction to smoothly transfer from current activity to another activity. However, existing methods clearly fail to fulfill the above requirements. This motivates us to propose the controlled motion prediction problem. The ability to control and fine-tune the generated future motion would allow for a broader research space in human-machine interaction and would certainly be desirable for many applications.

Inspired by a divide and conquer strategy, we propose a novel Aggregated Multi-GAN (AM-GAN) framework for the task of controlled human motion prediction. The generator module in our architecture allows for conditional motion generation, which may be used to realize action transfer, *i.e.,* the predicted motion can be controlled to transfer to a specified action instead of continuing the original action. AM-GAN models the motion of central spine and four limbs with separate GANs and the final complete motion is aggregated over them. The aggregation process ensures that the overall dynamics and interaction between kinematic chains are balanced and attuned. Dividing the generative modeling of the entire human skeletal motion into sub kinematic chains reduces the complexity of high dimensional motion generation into a less complicated low dimensional one. Furthermore, such an approach also enables us to acquire fine-grained control over the generated motion in specific body parts (chains).

To summarize, our main contributions are: 1) We propose a generalized 3D human motion prediction problem that allows for adequate and fine-grained control over the predictive spectrum. 2) We design a novel Aggregated Multi-GAN (AM-GAN) framework to tackle this problem via a divide and conquer strategy. 3) AM-GAN sets the new state-of-the-art performance for quantitative short-term prediction and also generates natural and consistent long-term prediction. More importantly, AM-GAN also allows control over the prediction such as generating transitions to specified actions and fine-grained control of body part movements. The AM-GAN is, to our knowledge, the first work that can handle these tasks simultaneously (short-term prediction, long-term prediction, specified action transfer, and fine-grained motion prediction control).

Related Work

RNN based Motion Prediction Deep learning has achieved great success in various fields (Kingma et al. 2014; Butepage et al. 2017; Liu et al. 2017; Zhu et al. 2020; Xu et al. 2020). RNN is a natural candidate for modeling human motion sequence due to their suitability in handling time series. Each time step of a motion sequence is encoded as a hidden state that learns the dynamics of the motion up to the current. The final hidden state corresponding to the last observed frame is then recurrently updated to generate a future sequence. This approach has been adopted in several works (Fragkiadaki et al. 2015; Martinez, Black, and Romero 2017). A similar work (Li et al. 2018) employs convolutional layers instead of RNN units. In general, the progress within this line of work seeks to better model long-term temporal dependency and overcome the issue that the current hidden state tends to be overwhelmed by the last observed frame. (Martinez, Black, and Romero 2017) proposes to incorporate a residual connection between the RNN units while (Liu et al. 2019) proposes a hierarchical RNN architecture. The major shortcoming of this class of works is that the output is fixed to be a single deterministic future motion sequence, and would be inadequate for obtaining a representation of the spectrum of future possibilities.

GAN-based Human Motion Prediction GAN (Goodfellow et al. 2014) provides an alternative to the RNN based approach for human motion prediction. In an adversarial process, a generator network outputs possible future motion sequences conditioned on an observed input sequence and a discriminator network assesses how likely these generated sequences are real or fake. In contrast to RNN-based approaches, GAN-based approaches (Barsoum, Kender, and Liu 2018; Kundu, Gor, and Babu 2019) for human motion prediction are probabilistic in that the learned generative model can output different sequences of possible future human poses from the same input sequence by sampling an additional latent noise vector. A challenging problem with GAN approaches is mode-collapsing during training, which leads to generated outputs all lying within certain modes. (Kundu, Gor, and Babu 2019) proposed a modification of the discriminator so as to incorporate both adversarial loss as well as content loss on the predicted motion sequence.

Controllable Motion Prediction To our knowledge, few existing works assimilate control for motion prediction tasks. (Holden, Saito, and Komura 2016; Pavllo et al. 2019) incorporate control parameters in the form of path trajectories along which a predicted walking or running action is bound to traverse. Another related work (Holden et al. 2017) adapts the technique of neural style transfer for motion generation. For example, a walking motion sequence may be manifested in a jubilant fashion or a depressed manner and these different styles can be transferred onto a walking sequence without altering the essential content. However, the problem of generating transitions between different class of actions has not yet been studied in the setting of human motion prediction. Different from previous works, our approach does not focus on controlling the trajectory or expression style of a walking action. Rather, we seek to allow for predicting transitions between different actions and fine-grained control over body parts in executing a prediction.

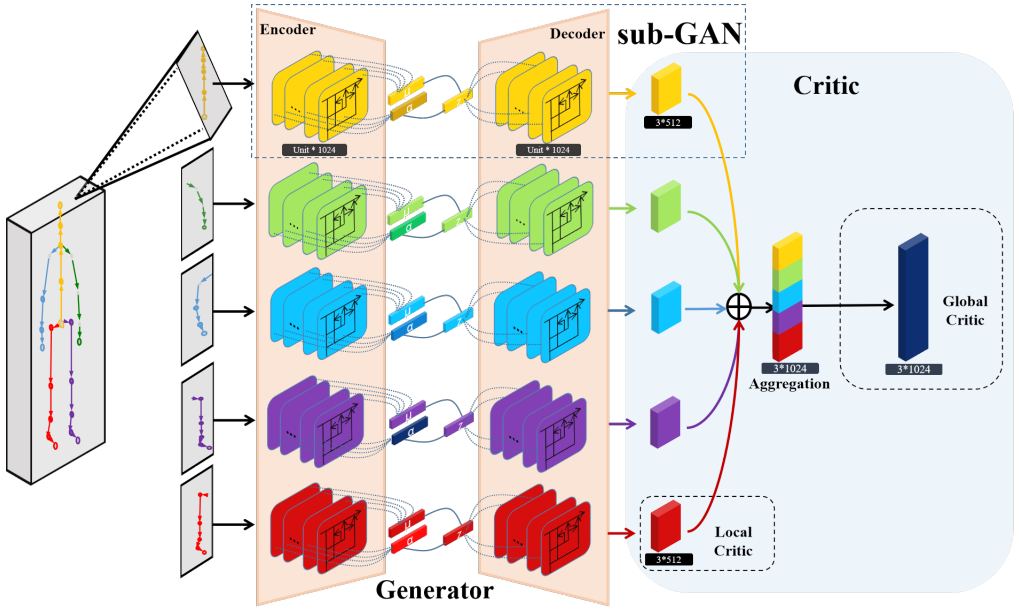

Figure 2: The proposed AM-GAN architecture. Each kinematic chain is modeled by its own sub-GAN (with matching colors). Each sub-GAN has a generator in the form of a VAE network as well as a local critic. The combined output from all sub-GANs is then aggregated and passed to a global critic. The aggregation layer and the global critic constitute the global GAN.

Our Approach

In this section, we present the details of our proposed Aggregated Multi-GAN (AM-GAN). The overview of our approach is inspired by the divide and conquer strategy. We observe that by reducing the general motion generation task entailing of the entire human skeleton to its constituent skeletal kinematic chains, the issue of mode collapsing can be effectively alleviated, leading to more accurate and diverse motion sequence forecasts.

Problem Formulation

Presented with an observed human motion sequence parameterized as $(\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_t)$ with optional control parameters, we are interested in generating a future pose sequence $(\mathbf{p}_{t+1}, \mathbf{p}_{t+2}, \cdots, \mathbf{p}_{t+T})$.

Each human skeletal pose $\mathbf{p}_i$ is composed of distinct key joints. For simplicity, we adopt 3D coordinates to characterize the locations of each joint. Mathematically, the skeletal human pose lies on a manifold, and a motion sequence traces out a trajectory on this pose manifold, which is depicted in Fig. 1. The problem of motion forecasting is to find a smooth extension to the observed trajectory, while action transfer is to find a smooth transfer from current trajectory to the target trajectory.

Method Overview and Rationale

Divide and Conquer with Multiple GANs and Aggregation Existing methods attempt to model the dynamics of entire human skeletal motion in its entirety. (Jain et al. 2016; Li et al. 2018) further seek to model spatial dependencies between joints. These attempts set the problem of motion generation on an unnecessarily high dimensional manifold. Optimizing for both the adversarial and content loss on a high dimensional manifold is harder and more likely to incur issues of non-convergence, gradient diminishing and mode collapse (Kodali et al. 2017).

Taking into account the anatomical separability of different chains and the difficulty of motion generation in high dimensional manifolds, we propose to divide the motion generation problem into that of the constituent chains and conquer the complete problem through an aggregation process that takes into account the interaction between these chains.

Corresponding to each pose $\mathbf{p}_i$, the constituent chains will be denoted as $\mathbf{c}_i^j$ where $1 \leq j \leq 5$ is the chain index (torso and four limbs). Note that we also experimented with other body divisions such as dividing each limb into two parts, but with no significant performance gain obtained.

Controlled Prediction across Action Types It is not realistic to confine human motion prediction within a single class of action, as is the case for existing methods. We propose a method that explicitly allows for predictions where one action may transit and evolve into another specified action. Therefore, instead of purely RNNs, we engage generative modeling which are better suited for this objective.

Fine-grained Control of Body Parts Previous motion prediction works are lacking in their abilities to control individual body parts in the predicted motion. We attempt to attain fine-grained control of body components and again, modeling each kinematic chain with its own specific GAN plays a major role in this. The capacity to control fine details of body parts in motion is undoubtedly a very desirable influence in fostering the versatility and usability in human motion prediction research. For example, different stylistic peculiarities or gait patterns across different individuals maybe directly integrated into the final prediction.

Architecture Overview

The overall framework of our method is illustrated in Fig. 2. There are five sub-GANs (local GANs), each taking care of a kinematic chain. The sub-GANs will now model motion generation on a lower dimensional manifold which is advantageous from a computational perspective and may be favorable in avoiding mode collapse. After a local discrimination process, the outputs from all sub-GANs are then combined via an aggregation layer. The output of the aggregation layer is then passed to the global critic for assessing the quality of the generated motion sequence in its entirety.

The conditional probability distributions learnt through the generative modeling process may then be invoked to realize action transfer in motion prediction. The sub-GANs in the AM-GAN network can further be mobilized to control the skeletal pose at a fine-grained scale.

Generator In order to effectively interpolate between latent representations for generating transitions across different action types, we have adopted a Variational Autoencoder (VAE) architecture (Kingma et al. 2014) as the generator module for each of our sub-GAN. Since the task entails sequential timeseries data, the encoder and decoder engage GRU for modeling the motion sequences.

Formally, the encoder (in sub-GAN j) maps the observed sequence $\mathbf{C}^j = \left(\mathbf{c}_1^j, \mathbf{c}_2^j, \cdots, \mathbf{c}_t^j\right)$ to a latent variable $\mathbf{z}^j$.

$$\mathbf{z}^j = \text{EN}(\mathbf{C}^j) \tag{1}$$

The decoder then recursively generates a future prediction sequence based on this latent vector $\mathbf{z}^j$.

$$\mathbf{C}'^j = \text{DE}(\mathbf{z}^j) \tag{2}$$

Aggregation Dynamically, there is interconnection and interdependence between different kinematic chains, as typified in the synchronized movements of the right arm and left leg in walking. Therefore, the individual latent motion contexts from each kinematic chain is mapped through a fully connected aggregation layer, to the final predicted pose sequence output as follows:

$$\mathbf{P}' = \text{Agg}\left(\sum_{j=1}^{5} \mathbf{C}'^j\right) \tag{3}$$

where Agg denotes the mapping function of the aggregation layer. $\mathbf{P}'$ is shorthand for the future pose sequence.

Critic The critic module consists of Local Critics for each sub-GAN D^j and a Global Critic D^G after the aggregation layer. A three-layers fully connected feedforward network is used in both types of critics. The local critics judge the generation of the five branches of the human body to ensure the accuracy of local prediction. The rationale is to first focus on the local context and reduce interference between local components or body parts at the base level. However, we cannot simply ignore the correlation and synchronization between different body parts in executing a motion. Thus we also include a global critic to judge the feasibility of the motion and to ensure the naturalness and accuracy of the entire pose.

Loss Function

Now, we zoom in on the loss function for training AM-GAN. To account for chain-level and pose-level accuracy, we consider both local loss and global loss. We further engage a stability loss at the chain-level and a consistency loss at the pose-level to ensure the authenticity and temporal smoothness of the motion.

Chain-level loss Within each sub-GAN, a generator outputs the motion prediction of a kinematic chain j. A critic network $\mathbf{D}^j$ learns to assign higher scores to real future sequences and lower scores to the generated ones. This gives the Wasserstein GAN loss for kinematic chain j as:

$$\mathcal{L}_{\text{Chain}^j} = \mathop{\mathbb{E}}_{\mathbf{C}^j} \mathbf{D}^j[C^j_{t+1:t+T}] - \mathop{\mathbb{E}}_{\mathbf{C}'^j} \mathbf{D}^j[C'^j_{t+1:t+T}] \tag{4}$$

where $\mathbf{C}^j$ and $\mathbf{C}'^j$ respectively denote the real and generated future sequences for chain j.

On top of the Wasserstein GAN loss, we further introduce a stability loss given by the variance over $\mathcal{L}_{\text{Chain}^j}$.

$$\mathcal{L}_{\text{Stability}} = \frac{1}{5}\sum_{j=1}^{5}(\mathcal{L}_{\text{Chain}^j} - \overline{\mathcal{L}_{\text{Chain}^j}})^2. \tag{5}$$

The local or chain level loss for chain j is given by a weighted sum of the Wasserstein GAN loss for the chain as well as the stability loss,

$$\mathcal{L}^j_{\text{Local}} = \alpha\mathcal{L}_{\text{Chain}^j} + \beta\mathcal{L}_{\text{Stability}} + \gamma\mathcal{L}^j_{\text{GroundTruth}} \tag{6}$$

where α, β, and γ are weight hyperparameters. $\mathcal{L}^j_{\text{GroundTruth}}$ is the difference between ground truth of chain j and its prediction.

Pose-level loss The aggregation layer aggregates and balances the five kinematic chains to give a generated pose sequence $\mathbf{P}'_{t+1:t+T}$. A global critic network $\mathbf{D}^G$ learns the authenticity of this pose, incurring a pose-level Wasserstein GAN loss as:

$$\mathcal{L}_{\text{Pose}} = \mathop{\mathbb{E}}_{\mathbf{P}} \mathbf{D}^G[P_{t+1:t+T}] - \mathop{\mathbb{E}}_{\mathbf{P}'} \mathbf{D}^G[P'_{t+1:t+T}] \tag{7}$$

where $\mathbf{P}$ denotes the real future pose sequence.

To ensure temporal smoothness, a consistency loss is incorporated as

$$\mathcal{L}_{\text{Consistency}} = \sum_{i=1}^{T} |\mathbf{P}'_{t+i} - \mathbf{P}'_{t+i-1}|^2. \tag{8}$$

The global or pose-level loss for the global GAN is given by a weighted sum of the pose Wasserstein GAN loss and the consistency loss:

$$\mathcal{L}_{\text{Global}} = \lambda\mathcal{L}_{\text{Pose}} + \mu\mathcal{L}_{\text{Consistency}} + \eta\mathcal{L}_{\text{GroundTruth}} \tag{9}$$

$\mathcal{L}_{\text{GroundTruth}}$ is the difference between the ground truth pose sequence $\mathbf{P}$ and its prediction $\mathbf{P}'$.

Action Transfer The training for action transfer is challenging since there are no public MoCap datasets involving transitions between different actions. We can only train in an unsupervised way. As illustrated in Fig. 3, we perform interpolation in the latent space to fill in the discontinuity

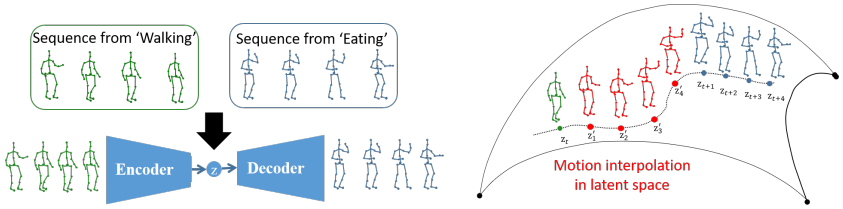

Figure 3: We perform motion interpolation in the latent space (i.e., z in VAE) and insert these motion frames (red) to fill up the discontinuity between two action sequences. Different from other two tasks, here the AM-GAN network is trained to predict an immediate future pose rather than a sequence.

between two sequences. Specifically, let $\mathbf{P} = (\mathbf{P}_1, \cdots, \mathbf{P}_t)$ be a sequence of action A_1 and $\mathbf{P}' = (\mathbf{P}'_{m+1}, \cdots, \mathbf{P}'_{m+T})$ a sequence of action A_2. We are to transfer from $\mathbf{P}$ to $\mathbf{P}'$. We may first train the VAE generator using large-scale ground truth sequences of all actions, making it able to predict the immediate future pose (**not** future sequence). Then, feed $\mathbf{P}$ and $(\mathbf{P}'_{m-t+1}, \cdots, \mathbf{P}'_m)$ respectively to the trained VAE network, obtaining their latent vectors $\mathbf{z}_1$ and $\mathbf{z}_2$. We perform interpolation between $\mathbf{z}_1$ and $\mathbf{z}_2$ to fill in the discontinuity between $\mathbf{P}_t$ to $\mathbf{P}'_{m+1}$. Given that $\mathbf{z}_2$ contains motion contexts towards generating prediction $\mathbf{P}'_{m+1}$, the decoded interpolations will gradually transfer from $\mathbf{P}_t$ to $\mathbf{P}'_{m+1}$.

Fine-grained pose control Breaking down the motion generation to chain level representations allows us to naturally incorporate fine-grained manipulation of each chain. Specifically, assuming that we are to generate the future sequence of an observed 'walking' sequence $\mathbf{P}$ and wants to control the right arm to act like 'direction'. We can train the proposed network for each action independently. Then, we simply feed $\mathbf{P}$ into the trained 'walking' network to generate future 'walking' action for the five chains. Similarly, we feed $\mathbf{P}$ into the trained 'direction' network to generate future 'direction' action for the right arm chain. We replace the right arm chain generated by the 'walking' network with that of the 'direction' network, and feed the new five chains to the aggregation layer to generate a natural final pose sequence.

Experiments

Dataset and preprocessing

In order to verify the ability of our model, we selected the largest and widely used data set: H3.6m (Ionescu et al. 2014) that contains 3.6 million human images with 3D poses (comprising 25 distinct skeletal joints) obtained via the Vicon mocap system. There are 15 different action classes performed by 7 subjects. Following the protocol of previous methods (Fragkiadaki et al. 2015; Jain et al. 2016), subject 5 is used for testing while the data for the other 6 subjects are used for training. The observed sequence has 25 frames and the predicted output sequence is also set to 25 frames.

Implementation Details

We use Tensorflow to implement our method. For each of our sub-GAN, we have adopted a Variational Auto-encoder (VAE) architecture (Kingma et al. 2014) as the generator module. The encoder and decoder of VAE consist of a GRU with 1,024 hidden units. The training of the GANs follows that of Wasserstein GANs (Gulrajani et al. 2017) but with our customized loss function outlined in the Loss Function section. The aggregation module is a full connected layer with 1,024 units. The local critic is a three-layer fully connected feedforward network with 512 units, while the global critic possesses 1,024 units. The Adam Optimizer (Kingma and Ba 2014) is employed with initial learning 5e-5 and the batch size is set to 16.

Quantitative Evaluation on H3.6m dataset

We benchmark our method against previous methods, including ERD (Fragkiadaki et al. 2015), SRNN (Jain et al. 2016), RRNN (Martinez, Black, and Romero 2017), LSTM-AE (Tu et al. 2018), HP-GAN (Barsoum, Kender, and Liu 2018), CEM (Li et al. 2018), BiHMP-GAN (Kundu, Gor, and Babu 2019), SKel-Net (Guo and Choi 2019), HMR (Liu et al. 2019) and QuaterNet (Pavllo et al. 2019). Following previous work, the mean angle error (MAE) is adopted as the evaluation metric. We report the results in Table 1 over four actions 'Walking', 'Eating', 'Smoking', 'Discussion'.

As shown in Table 1, our method generally achieves superior performance in quantitative accuracy over the state-of-the-arts. One case in point is that the actions have different levels of complexity. For example, walking is a more regular action and demonstrates a high degree of periodicity whereas discussion is more complex and basically aperiodic. As would be expected, prediction accuracy falls with complexity of the action, which is observed throughout the empirical results. One noteworthy remark is that our method demonstrates significant improvements for the more complex action types such as discussion. This can be attributed to the fact that by modeling body parts and skeletal kinematic chains separately, our AM-GAN approach is less affected by lack of regularity and periodicity in the overall motion pattern. Furthermore, comparing against the other two GAN frameworks HP-GAN and BiHMP-GAN, the superior performance of our AM-GAN for both short-term and long-term prediction is also an indicator to the effectiveness of the divide and conquer approach in modeling local body parts and merging them at the final stage.

For the sake of completeness, we benchmark our method over the remaining 11 H3.6m actions against the two most recent and state-of-the-art works HMR and QuaterNet. The comparison results are shown in Table 2. It can be seen that our method achieves new state-of-the-art performance quite

Methods \Time (ms)	Walking					Eating				
	80	160	320	400	1000	80	160	320	400	1000
ERD (Fragkiadaki et al. 2015)	0.93	1.18	1.59	1.78	2.24	1.27	1.45	1.66	1.80	2.02
SRNN (Jain et al. 2016)	0.81	0.94	1.16	1.30	1.80	0.97	1.14	1.35	1.46	2.11
RRNN (Martinez, Black, and Romero 2017)	0.33	0.56	0.78	0.85	1.14	0.26	0.43	0.66	0.81	1.34
LSTM-AE (Tu et al. 2018)	1.00	1.11	1.39	——	1.39	1.31	1.49	1.86	——	2.01
CEM (Li et al. 2018)	0.33	0.54	0.68	0.73	0.92	0.22	0.36	0.58	0.71	1.24
HP-GAN (Barsoum, Kender, and Liu 2018)	0.95	1.17	1.69	1.79	2.47	1.28	1.47	1.70	1.82	2.51
SKelNet (Guo and Choi 2019)	0.34	0.52	0.69	0.70	0.90	0.23	0.39	0.50	0.71	1.23
BiHMP-GAN (Kundu, Gor, and Babu 2019)	0.33	0.52	0.63	0.67	0.85	**0.20**	0.33	0.54	0.70	1.20
HMR (Liu et al. 2019)	0.34	0.52	0.67	0.69	0.90	0.22	**0.31**	0.51	0.69	1.21
QuaterNet (Pavllo et al. 2019)	**0.23**	0.37	0.59	0.66	0.86	**0.20**	0.33	0.54	0.68	1.17
AM-GAN (Ours)	**0.23**	**0.51**	**0.62**	**0.66**	**0.84**	**0.20**	**0.31**	**0.49**	**0.66**	**1.15**
Methods \Time (ms)	Smoking					Discussion				
	80	160	320	400	1000	80	160	320	400	1000
ERD (Fragkiadaki et al. 2015)	1.66	1.95	2.35	2.42	3.14	2.27	2.47	2.68	2.76	3.11
SRNN (Jain et al. 2016)	1.45	1.68	1.94	2.08	2.57	1.22	1.49	1.83	1.93	2.19
RRNN (Martinez, Black, and Romero 2017)	0.35	0.64	10.3	1.15	1.83	0.37	0.77	1.06	1.10	1.79
LSTM-AE (Tu et al. 2018)	0.92	1.03	1.15	——	1.77	1.11	1.20	1.38	——	1.73
CEM (Li et al. 2018)	0.26	0.49	0.96	0.92	1.62	0.32	0.67	0.94	1.01	1.86
HP-GAN (Barsoum, Kender, and Liu 2018)	1.71	1.89	2.33	2.42	3.20	2.29	2.61	2.79	2.88	3.67
SKelNet (Guo and Choi 2019)	0.26	0.47	0.91	0.89	1.56	0.29	0.62	0.88	1.00	1.68
BiHMP-GAN (Kundu, Gor, and Babu 2019)	0.26	0.50	0.91	**0.86**	1.11	0.33	0.65	0.91	0.95	1.77
HMR (Liu et al. 2019)	0.26	0.47	0.90	0.91	1.49	0.29	**0.55**	0.83	0.94	1.61
QuaterNet (Pavllo et al. 2019)	0.26	0.49	0.92	0.89	1.67	**0.28**	0.62	0.87	0.95	1.89
AM-GAN (Ours)	**0.25**	**0.46**	**0.88**	0.88	**1.10**	**0.28**	**0.55**	**0.81**	**0.92**	**1.58**

Table 1: Performance evaluation (in MAE) of comparison methods over 4 different action types on the H3.6m dataset

	Directions					Greeting					Phoning					Posing				
Time (ms)	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
HMR	0.38	0.58	0.77	0.90	1.43	0.52	0.85	1.25	1.40	1.73	0.56	1.09	1.50	1.61	1.80	0.24	0.53	1.12	1.42	2.50
QuaterNet	**0.36**	**0.57**	0.78	0.90	1.42	0.52	**0.81**	1.25	1.38	1.73	0.58	1.12	1.49	1.60	1.82	0.26	0.56	1.19	1.41	2.56
AM-GAN (Ours)	**0.36**	**0.57**	**0.75**	**0.89**	**1.41**	**0.51**	0.86	**1.24**	**1.36**	**1.70**	**0.54**	**1.05**	**1.48**	**1.58**	**1.79**	**0.22**	**0.51**	**1.10**	**1.41**	**2.48**
	Purchases					Sitting					Sittingdown					Takingphoto				
Time (ms)	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
HMR	0.58	**0.83**	1.19	1.25	1.99	0.38	0.61	1.01	1.15	1.66	0.39	0.75	1.12	1.28	1.91	**0.21**	0.44	0.85	1.01	1.42
QuaterNet	0.59	**0.83**	1.21	1.24	2.37	0.37	0.58	0.99	**1.11**	1.62	0.37	0.73	1.09	1.23	1.94	0.24	0.48	0.88	1.02	1.43
AM-GAN (Ours)	**0.55**	0.85	**1.18**	**1.23**	**1.95**	**0.35**	**0.60**	**0.98**	1.13	**1.60**	**0.36**	**0.72**	**1.08**	**1.20**	**1.85**	0.23	**0.41**	**0.83**	**0.99**	**1.40**
	Waiting					Walkingdog					Walkingtogether					Average				
Time (ms)	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
HMR	0.27	0.59	1.02	1.33	2.21	0.55	0.87	1.20	1.36	1.84	0.26	0.50	0.69	0.71	1.21	0.39	0.69	1.06	1.46	1.79
QuaterNet	0.29	0.58	1.03	1.30	2.38	**0.53**	0.89	1.19	1.34	1.97	0.24	**0.45**	**0.65**	**0.68**	1.31	0.38	0.69	1.05	1.47	1.82
AM-GAN (Ours)	**0.23**	**0.56**	**1.00**	**1.29**	**2.15**	**0.53**	**0.85**	**1.18**	**1.33**	**1.80**	**0.22**	**0.45**	0.71	0.73	**1.19**	**0.37**	**0.67**	**1.04**	**1.43**	**1.75**

Table 2: Performance evaluation (in MAE) over 11 remaining action types on the H3.6m dataset

consistently for both short term and long term predictions. As shown in Fig. 4, we also visualize the results of our method against current state-of-the-art HMR, observing that our method delivers more natural and smooth sequences.

Action transfer

We now consider scenarios where we control the specific action performed in the future. Some generated sequences depicting transitioning back and forth between different actions are showcased in Fig. 5, where the middle six frames are generated transition frames. Empirical results show that our method is able to smoothly and authentically transfer from one action to another.

Human pose control

We can assimilate high level controls over specific body parts with AM-GAN. The local and global level generation can be seamlessly combined to create natural and consistent motion sequences. As an example, we illustrate in the left figure of Fig. 6 on how we may control the movements of the right leg in the predicted sequence. Specifically, given a walking sequence, we generate its future 'walking' sequence

Time (ms)	80	160	320	400	1000
	Walking				
Remove Local Critic	0.40	0.63	0.73	0.85	1.01
Remove Global Critic	0.35	0.61	0.72	0.84	1.92
Remove sub-GANs	0.90	1.10	1.61	1.79	2.01
AM-GAN (Complete)	0.23	0.51	0.62	0.66	0.84
	Discussion				
Remove Local Critic	0.40	0.71	0.89	1.20	2.89
Remove Global Critic	0.29	0.62	0.87	1.00	1.19
Remove sub-GANs	2.20	2.51	2.68	2.71	3.20
AM-GAN (Complete)	0.28	0.55	0.81	0.92	1.58

Table 3: Ablation studies on different components of our AM-GAN evaluated over 'Walking' and 'Discussion'.

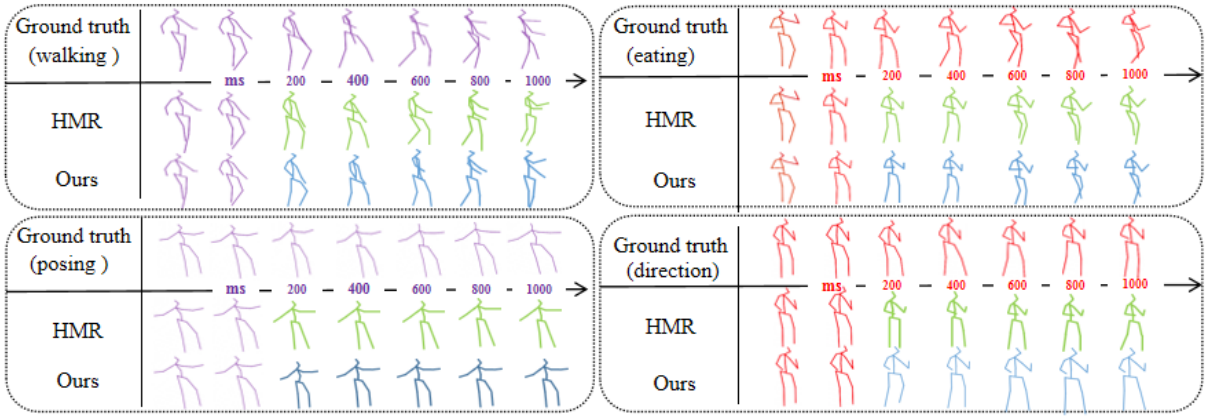

Figure 4: Some visual results.

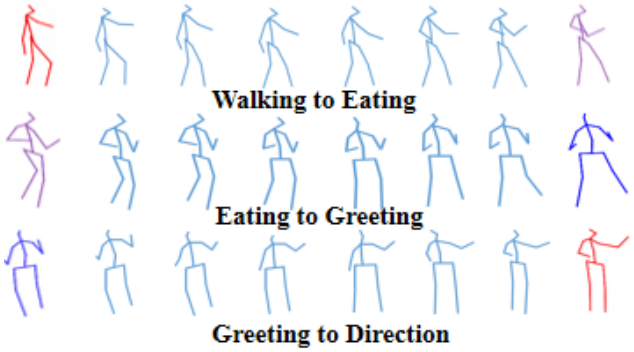

Figure 5: Action transfer examples.

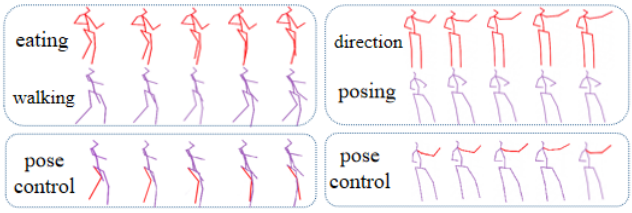

Figure 6: Fine-grained control over different chains in the predicted motion. In the left figure, we predict a 'walking' action (purple) with the right leg following the action of 'eating' (red). The result is shown on the bottom row with matching colors. In the right figure, we predict a 'posing' action with the left hand following the 'direction' action.

but enforce the right leg act like 'eating' action prediction. Similarly, in the right figure of Fig.6, we show another example where we predict the future 'posing' action sequence but command the left hand to follow a 'direction' action prediction.

Ablation Studies

We quantitatively analyze the contribution of various design components, including removing the 5 sub-GANs, removing the local critic and global critic. The empirical results are reported in Table 3. One crucial finding is that incorporating sub-GANs and a divide and conquer strategy indeed improves over the baseline with a single GAN for modeling the motion generation. Modeling the constituent chains instead of the entire pose directly reduces complexity, thus improving the short-term and long-term prediction capabilities. Furthermore, it may be observed that the improvement over the baseline is more significant in the case of aperiodic actions such as 'Discussion'. This is likely due to the fact that zooming in on individual body parts is more effective for modeling complex and irregular motion, where different body parts usually engage differently in the motion. Another observation is that the local critic enhances the short-term prediction whereas the global critic elevates long-term prediction performance.

Conclusion

In this paper, we propose a novel human motion prediction problem to incorporate control. In order to reduce the complexity of optimizing over the entire human pose manifold and to achieve fine-grained control, we employed a divide and conquer strategy and modeled each kinematic chain separately with our AM-GAN. On top of improving the state-of-the-art for both short and long-term predictions, AM-GAN also allows for controlling of the prediction spectrum. This includes generating transitions to a desired action or fine-grained control over specific body parts. We are positive that controllable motion generation would open many doors in adjacent fields.

Acknowledgements

This work was partly supported by the National Key Research and Development Program of China under No. 2020AAA0140004. The Natural Science Foundation of Zhejiang Province, China No. LQ19F020001. The National Natural Science Foundation of China No. 61902348 and No. 61772466. The Zhejiang Provincial Natural Science Foundation for Distinguished Young Scholars No. LR19F020003, and the Fundamental Research Funds for the Central Universities (Zhejiang University NGICS Platform).

References

Barsoum, E.; Kender, J.; and Liu, Z. 2018. HP-GAN: Probabilistic 3D human motion prediction via GAN. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 1418–1427.

Butepage, J.; Black, M. J.; Kragic, D.; and Kjellstrom, H. 2017. Deep representation learning for human motion prediction and classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6158–6166.

Fragkiadaki, K.; Levine, S.; Felsen, P.; and Malik, J. 2015. Recurrent Network Models for Human Dynamics. In *ICCV*.

Gao, J.; Shi, Z.; Wang, G.; Li, J.; Yuan, Y.; Ge, S.; and Zhou, X. 2020. Accurate Temporal Action Proposal Generation with Relation-Aware Pyramid Network. In *AAAI*, 10810–10817.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. C. 2017. Improved training of wasserstein gans. In *Advances in neural information processing systems*, 5767–5777.

Guo, X.; and Choi, J. 2019. Human motion prediction via learning local structure representations and temporal dependencies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 2580–2587.

Habibie, I.; Holden, D.; Schwarz, J.; Yearsley, J.; and Komura, T. 2017. A Recurrent Variational Autoencoder for Human Motion Synthesis. In *BMVC*.

Holden, D.; Habibie, I.; Kusajima, I.; and Komura, T. 2017. Fast neural style transfer for motion data. *IEEE computer graphics and applications* 37(4): 42–49.

Holden, D.; Saito, J.; and Komura, T. 2016. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)* 35(4): 138.

Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2014. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE TPAMI* 36(7): 1325–39.

Jain, A.; Zamir, A. R.; Savarese, S.; and Saxena, A. 2016. Structural-RNN: Deep Learning on Spatio-Temporal Graphs. In *CVPR*, 5308–5317.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Kingma, D. P.; Mohamed, S.; Rezende, D. J.; and Welling, M. 2014. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, 3581–3589.

Kodali, N.; Abernethy, J.; Hays, J.; and Kira, Z. 2017. On convergence and stability of gans. *arXiv preprint arXiv:1705.07215* .

Koppula, H. S.; Jain, A.; and Saxena, A. 2016. Anticipatory planning for human-robot teams. In *Experimental robotics*, 453–470. Springer.

Koppula, H. S.; and Saxena, A. 2013. Anticipating human activities for reactive robotic response. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2071–2071. IEEE.

Kundu, J. N.; Gor, M.; and Babu, R. V. 2019. Bihmp-gan: Bidirectional 3d human motion prediction gan. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 8553–8560.

Lehrmann, A. M.; Gehler, P. V.; and Nowozin, S. 2014. Efficient Nonlinear Markov Models for Human Motion. In *CVPR*, 1314–1321.

Li, C.; Zhang, Z.; Sun Lee, W.; and Hee Lee, G. 2018. Convolutional sequence to sequence model for human dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5226–5234.

Liu, Z.; Cheng, L.; Liu, A.; Zhang, L.; He, X.; and Zimmermann, R. 2017. Multiview and Multimodal Pervasive Indoor Localization. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, 109–117.

Liu, Z.; Wu, S.; Jin, S.; Liu, Q.; Lu, S.; Zimmermann, R.; and Cheng, L. 2019. Towards natural and accurate future motion prediction of humans and animals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10004–10012.

Martinez, J.; Black, M.; and Romero, J. 2017. On human motion prediction using recurrent neural networks. In *CVPR*.

Pavllo, D.; Feichtenhofer, C.; Auli, M.; and Grangier, D. 2019. Modeling human motion with quaternion-based neural networks. *International Journal of Computer Vision* 1–18.

Taylor, G. W.; Sigal, L.; Fleet, D. J.; and Hinton, G. E. 2010. Dynamical binary latent variable models for 3d human pose tracking. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 631–638. IEEE.

Tekin, B.; Rozantsev, A.; Lepetit, V.; and Fua, P. 2016. Direct prediction of 3d body poses from motion compensated sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 991–1000.

Tu, J.; Liu, H.; Meng, F.; Liu, M.; and Ding, R. 2018. Spatial-temporal data augmentation based on LSTM autoencoder network for skeleton-based human action recognition. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, 3478–3482. IEEE.

Wang, J. M.; Fleet, D. J.; and Hertzmann, A. 2005. Gaussian Process Dynamical Models. In *NIPS*, 1441–1448.

Xu, N.; Zhang, H.; Liu, A.; Nie, W.; Su, Y.; Nie, J.; and Zhang, Y. 2020. Multi-Level Policy and Reward-Based Deep Reinforcement Learning Framework for Image Captioning. *IEEE Trans. Multim.* 22(5): 1372–1383.

Yan, X.; Rastogi, A.; Villegas, R.; Sunkavalli, K.; Shechtman, E.; Hadap, S.; Yumer, E.; and Lee, H. 2018. Mt-vae: Learning motion transformations to generate multimodal human dynamics. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 265–281.

Yuan, Y.; and Kitani, K. 2019. Diverse Trajectory Forecasting with Determinantal Point Processes. *arXiv preprint arXiv:1907.04967* .

Zhu, L.; Lu, X.; Cheng, Z.; Li, J.; and Zhang, H. 2020. Deep Collaborative Multi-View Hashing for Large-Scale Image Search. *IEEE Trans. Image Process.* 29: 4643–4655. doi: 10.1109/TIP.2020.2974065. URL https://doi.org/10.1109/TIP.2020.2974065.

Zhuo, T.; Cheng, Z.; Zhang, P.; Wong, Y.; and Kankanhalli, M. S. 2020. Unsupervised Online Video Object Segmentation With Motion Property Understanding. *IEEE Trans. Image Process.* 29: 237–249.