Bidirectional Transformer GAN for Long-Term Human Motion Prediction

MENGYI ZHAO, Beihang University, China
HAO TANG, Computer Vision Lab, ETH Zurich, Switzerland
PAN XIE, SHULING DAI, Beihang University, China
NICU SEBE, WEI WANG, University of Trento, Italy

The mainstream motion prediction methods usually focus on short-term prediction, and their predicted long-term motions often fall into an average pose, *i.e.* the freezing forecasting problem [27]. To mitigate this problem, we propose a novel Bidirectional Transformer-based Generative Adversarial Network (BiTGAN) for long-term human motion prediction. The bidirectional setup leads to consistent and smooth generation in both forward and backward directions. Besides, to make full use of the history motions, we split them into two parts. The first part is fed to the Transformer encoder in our BiTGAN while the second part is used as the decoder input. This strategy can alleviate the exposure problem [37]. Additionally, to better maintain both the local (*i.e.*, frame-level pose) and global (*i.e.*, video-level semantic) similarities between the predicted motion sequence and the real one, the soft dynamic time warping (Soft-DTW) loss is introduced into the generator. Finally, we utilize a dual-discriminator to distinguish the predicted sequence at both frame and sequence levels. Extensive experiments on the public Human3.6M dataset demonstrate that our proposed BiTGAN achieves state-of-the-art performance on long-term (4*s*) human motion prediction, and reduces the average error of all actions by 4%.

Additional Key Words and Phrases: Long-Term Human Motion Prediction; Bidirectional Generation; Transformer; GAN; DTW

1 INTRODUCTION

Being able to predict the future motion of a person is essential for autonomous agents, *e.g.*, assistant robots [21] and self-driving cars [29], in order to understand human behaviors during human-agent interactions, human-robot collaboration [28], and robot navigation. For instance, it is important to understand the behavior of pedestrians and make proactive decisions for autonomous vehicle systems [29].

Due to the complexity of predicting high-dimensional features jointly, it is challenging to capture the various motion patterns *e.g.*, spatial-temporal dependencies for long-term motions. Specifically, one problem in the motion prediction task is that the predicted future poses are often the static average poses with the highest probability. This problem is also known as the freezing prediction problem [27]. Moreover, it is also difficult to measure the similarity between two human motion sequences explicitly and semantically.

To address the aforementioned issues, recent motion prediction methods mainly use recurrent neural networks (RNNs) [1, 12, 18, 35], feed-forward networks [23, 33, 34], and GANs [15, 46]. However, the RNNs struggle to encode long-term historical information for high-dimensional time-series data like human motion. For instance, Martinez *et al.* [35] showed that RNNs have problems with the discontinuity of the predicted sequence at the last seen frame, as well as a prediction that converges towards the mean pose of the ground-truth data for long-term

Authors' addresses: Mengyi Zhao, Beihang University, China, zhaomengyi@buaa.edu.cn; Hao Tang, Computer Vision Lab, ETH Zurich, Switzerland, hao.tang@vision.ee.ethz.ch; Pan Xie, Shuling Dai, Beihang University, China, {panxie,sldai}@buaa.edu.cn; Nicu Sebe, Wei Wang, University of Trento, Italy, {niculae.sebe,wei.wang}@unitn.it.

1551-6857/2023/1-ART $15.00
https://doi.org/10.1145/3579359

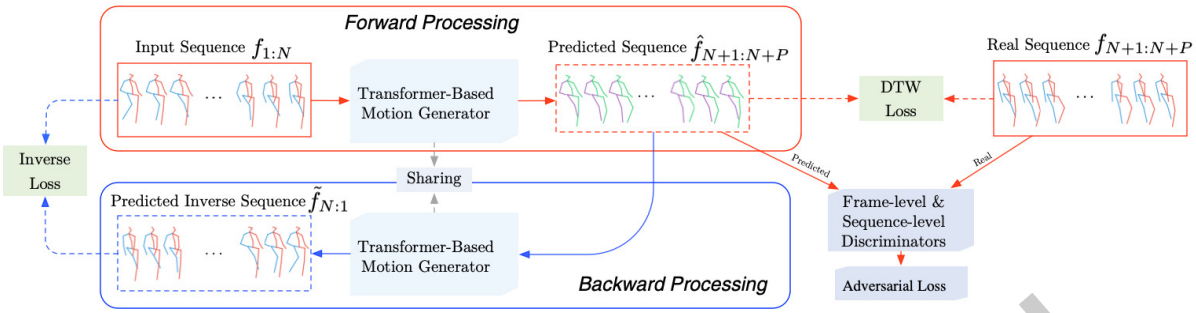

Fig. 1. Overview of the proposed BiTGAN. It consists of a Transformer-based motion generator and a dual-discriminator. The **motion generator** has a bidirectional processing loop. In the **forward** processing (red rectangle shown on the top), the input is the history motion sequence, while the output is the predicted future motion sequence. In the **backward** processing (blue rectangle shown at the bottom), the predicted motion sequence will be flipped and fed to the same motion generator as the input, and the output is the history motion sequence. An inverse loss is used to compute the difference between the predicted historical motion and the real one. By adding the adversarial loss, we make the predicted frames look realistic to the real ones in an adversarial way. The **dual-discriminator** contains a sequence-based discriminator and a frame-based one. Both try to distinguish the predicted motion sequence from the real one but at two different levels, *i.e.*, sequence level, and frame level. Besides, the Soft-DTW loss is used to better measure the similarity between the predicted motion sequence and the real one.

predictions. While the feed-forward networks can achieve more realistic predictions than RNNs, they still suffer from long-term predictions. In addition, conventional approaches compare human motion sequences based on estimating the $L2$ displacement error [35]. As shown by Martinez *et al.* [35], such measure tends to ignore the specific motion characteristics, since the same representative poses repeated over a sequence may result in a better match to a reference sequence compared to visually similar motion with different poses.

To sum up, previous works, in general, cannot capture long-range relationships, resulting in incoherent and unnatural prediction results. To address this limitation, recently, Transformer has been employed [4, 5]. Cao *et al.* [5] proposed a Transformer-based framework to exploit the scene context, while Cai *et al.* [4] exploited Transformer with the global attention mechanism to capture the long-range spatial correlations and temporal dependencies. Although these methods have achieved better performance, the freezing prediction problem still persists.

To address the aforementioned problems, in this paper, we propose a novel Bidirectional Transformer Generative Adversarial Network (BiTGAN), which can effectively exploit historical information and model the long-term relationships among frames. BiTGAN consists of a novel Transformer-based motion generator and a dual-discriminator (see Figure 1). We build our motion generator based on Transformer to model the long-range relationships between historical frames and predicted frames.

Specifically, to tackle the freezing forecasting problem in long-term prediction, we use the bidirectional generation strategy, leading to consistent forward and backward prediction results. The intuitive idea behind it is that if the generated future poses converge to the mean pose for a long time, then the backward predicted results will become worse with fewer dynamics compared to the history motion. Therefore, we add this constraint to penalize the predictions. Such bidirectional generation is inspired by the cycle consistency in the image-to-image translation task [49]. However, there are two major differences: (i) Our generator is different in content (human poses) but the same in style (running, jumping, etc.), while cycle consistency is originally used to enforce the same content (but different styles) in both input and output. (ii) We utilize forward and backward predictions to

preserve the motion consistency temporally in the video domain rather than the spatial correspondence in the image domain. In this way, our generator can predict the future motion frames from the input history motion frames in the forward processing, while in the meantime, it can predict the history motion frames from the predicted motion frames in the backward processing as shown in Figure 1.

In addition, there is a context discrepancy problem between the training and testing stages in the standard Transformer, which is known as the exposure bias problem [37]. The discrepancy refers to the fact that the decoder of the Transformer uses ground truth as input at training while using the predictions as input at testing. The distributions of the predictions and the ground truth have a discrepancy. To avoid this problem, instead of using the predicted sequence as the input of the decoder, we use part of the ground truth history sequence as the input of the decoder. In this way, we can keep the consistency between training and testing. Moreover, we use 10 frames as a mini-batch during training and testing and this setup allows batch-wise inference.

Besides, to better maintain the high-level semantic similarity between the predicted motion sequence and the real one both locally and globally, we add the soft dynamic time warping (Soft-DTW) loss to regularize our motion generator. Different from the traditional L2 loss, which is widely used in motion prediction tasks, the Soft-DTW loss is robust to shifts or dilatations across the time dimension, but it has been rarely used.

We also employ a GAN-based dual-discriminator [38], which is composed of a sequence-based discriminator and a frame-based discriminator. Both discriminators aim at distinguishing the predicted sequences from both local frame and global sequence levels. Extensive experiments on the Human3.6M dataset show that our BiTGAN achieves new state-of-the-art results on human motion prediction for both periodical and non-periodical actions and substantially improves the accuracy of the predicted long-term (4s) poses.

The contributions of this paper are summarized as follows:

- We propose a novel bidirectional Transformer GAN (BiTGAN) for long-term motion prediction. The novel bidirectional generation strategy can alleviate the freezing forecasting problem.
- For the Transformer-based generator, we design a novel data split strategy to alleviate the exposure bias problem [37].
- We introduce a new dynamic time warping (DTW) metric to evaluate the underlying similarity between two-time series and show that our method performs better.

2 RELATED WORK

Traditional Recursive Human Motion Prediction. Following the success of deep learning methods in computer vision, various deep learning models have been investigated for human motion prediction, such as RNNs and GANs.

With the rise and impressive performance of RNNs in sequence-to-sequence tasks, researchers leverage RNNs to learn temporal dependencies for motion prediction [12, 14, 18, 22, 30, 35]. Jain *et al.* [18] developed a structural-RNN incorporated with fixed spatial-temporal graphs to model human motion. However, the model is trained for each motion individually, incurring a high computational cost. Tang *et al.* [43] incorporated an attention module to summarize the recent pose history, followed by an RNN-based prediction network. Moreover, based on the observations from the statistics research on hand motion [16], Martinez *et al.* [35] proposed a residual architecture to model first-order motion derivatives, *i.e.*, velocities instead of human poses. They show that their simple method outperforms previous works.

Although these RNN-based approaches performed an interesting exploration, one can still observe unsatisfactory aspects in the predicted motion sequences. In order to fix these limitations, several works use feed-forward networks other than RNNs to model human pose [3, 23, 26, 33, 34, 40, 48]. For example, Butepage *et al.* [3] proposed a deep learning fully-connected network that investigates different strategies to encode temporal, and historical information and generalizes well to new, unseen motions. Li *et al.* [23] presented an approach

based on convolutional neural networks (CNNs) to human motion modeling. The hierarchical structure of a CNN enables it to effectively capture both spatial and temporal correlations. This method is more effective than the RNN-based ones, but the manually-selected size of the convolutional window heavily influences the temporal encoding. Mao *et al.* [34] show that encoding the short-term history in the frequency space using the discrete cosine transform (DCT) followed by a graph convolutional network (GCN) to encode spatial and temporal connections leads to better performance. They [33] further introduce a motion attention mechanism that allows capturing the motion recurrence in the long-term history. The work related to ours is [34] and [33], which also use DCT to encode motions, leverage GCNs as predictors as well as Mean Per Joint Position Error (MPJPE)[17] as the evaluation metric. However, there are three important differences between our approach and theirs. First, our method is developed to alleviate the problem of freezing forecasting for long-term (4s) prediction by the proposed bidirectional setup, while [33] and [34] are mainly designed for mid-term (1s) forecasting. Second, we design a Transformer-based Generative Adversarial Network for long-term human motion prediction. Finally, we introduce a novel dynamic time warping (DTW) metric to better measure the semantic similarity between the predicted motion sequence and the real one at video-level.

Apart from RNNs and feed-forward networks, some other models have also been proposed. For instance, to facilitate more realistic human motion prediction and alleviate the discontinuity problem, Gui *et al.* [15] incorporated adversarial training mechanisms to simultaneously validate the global plausibility and coherence of the predicted motions. Recently, Lyu *et al.* [31] reformulate the human motion problem based on stochastic differential equations and GANs. Moreover, to dynamically adjust the focus of the model, Li *et al.* [24] presented a more generic motion forecasting framework with dynamic key information selection and ranking procedures based on reinforcement learning and hybrid attention mechanism. Wang *et al.* [46] also introduced imitation learning under a reinforcement learning formulation, which is computationally effective. However, one can still observe that these methods cannot predict coherent and semantic-consistent motion sequences. More importantly, most existing human motion prediction works mentioned above only forecast human motions for maximum 1s, which is insufficient in many applications *e.g.*, human-robot interaction.

Vision Transformers for Motion Prediction. Transformer is a state-of-the-art attention-based approach in natural language processing (NLP) [7, 9, 32, 36, 47] and computer vision [10, 13, 19]. It was originally proposed for NLP [44] and has recently been successfully applied for many computer vision tasks. For human motion related tasks, there have also been several works that use Transformer-based methods. For instance, Duan *et al.* [11] use Transformer to solve the motion completion problem under different application scenarios with a unified framework. Li *et al.* [25] presented a two-stream motion Transformer generative model, which can capture long-term dependencies and generate music-conditioned flexible poses. Li *et al.* [27] designed a cross-modal Transformer-based architecture, which can generate realistic 3D dance motion given music and effectively prevent freezing or drifting problem, which is common for long-term motion generation. The advantages of the Transformer lie in the self-attention mechanism, which can capture global dependencies. Some of the recent methods in the field of motion forecasting adopt Transformer as well [4, 5]. For instance, Cao *et al.* [5] applied the standard Transformer network to predict 3D poses, but it requires an extra scene image as the input, which is different from the setup in this paper.Cai *et al.* [4] proposed to leverage the Transformer-based architecture to simultaneously capture the long-range spatial and temporal dependencies of human motion by treating the sequential joints with encoded temporal features as the input.

Our proposed Transformer-based approach is different from these methods. We focus on developing a novel bidirectional Transformer generative adversarial network (*i.e.*, BiTGAN) for long-term human motion prediction. Moreover, to tackle the freezing forecasting problem for long-term prediction, we propose a novel bidirectional generation strategy in the generator. In the backward prediction, as shown in Figure 1, we reuse the generated future frames as the input to the Transformer, and the outputs are the history motion frames and pose an extra inverse loss. Intuitively, in the backward prediction, if the input generated future frames are uniform

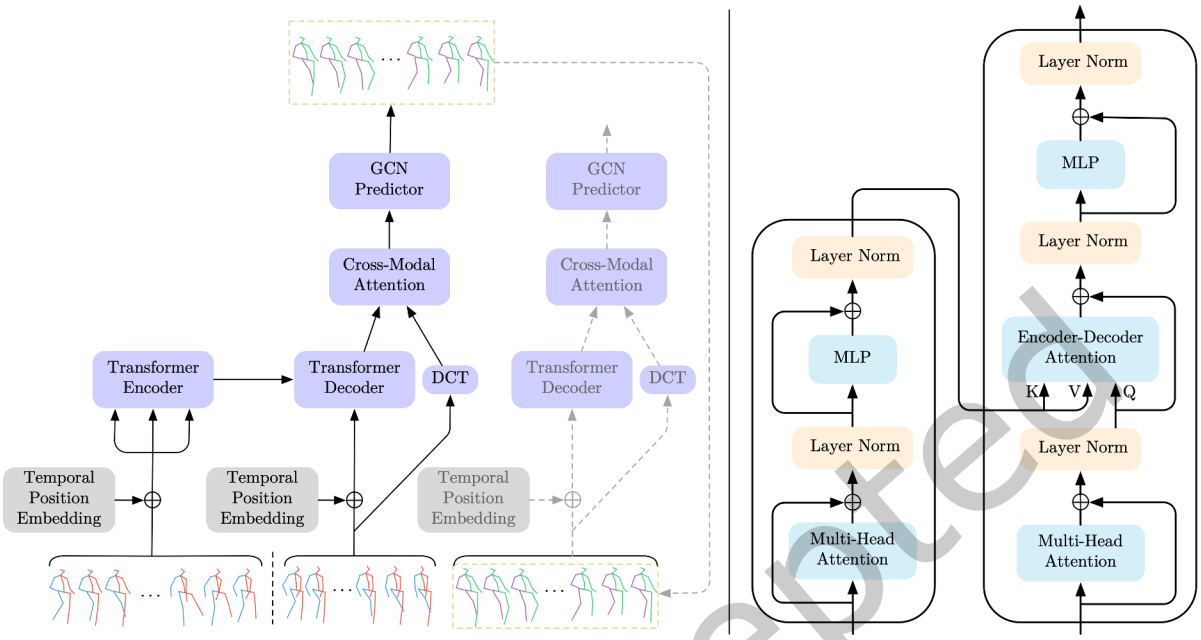

Fig. 2. Overview of our Transformer-based motion generator. It has five components, *i.e.*, Transformer encoder, Transformer decoder, DCT, cross-modal attention, and GCN predictor. The earlier history motions are encoded directly by the Transformer encoder. The recent history motions are the inputs of the Transformer decoder, and in the meanwhile, the Transformer decoder aggregates the encoded earlier history motions. The DCT further encodes the temporal information of the recent historical motions. The cross-modal attention is employed to fuse the outputs of the Transformer decoder and DCT coefficients. The GCN predictor with learnable adjacency matrices models the spatial relationship among joints. During testing, as illustrated by the black dashed line, we add the predicted frames as input and reversely forecast future poses.

freezing average pose, the predicted history poses will deviate far away from the ground truth. Thus, posing an extra inverse loss may help us to calibrate the forward prediction results and force the model to capture more human motion dynamics for a long-term span. Note that the inverse loss has a close relationship with the cycle consistency loss in [49]. The major difference is that the cycle consistency loss in [49] computes the loss between the original image and the reconstructed image, whereas the proposed inverse loss computes the loss between the original motion sequence and the backward predicted motion sequence. Our emphasis is on the sequential motion pattern consistency in the video domain rather than the correspondence of appearance structures in the image domain.

3 METHODOLOGY

In this work, our goal is to tackle the task of human motion prediction. Given the sequence motion coordinates $f_{1:N}$, of one person for the past N steps, we aim to predict the future motion for the next P steps, *i.e.*, $f_{N+1:N+P}$. In our case, $f_k \in \mathbb{R}^{3J}$ describes the 3D coordinates of human joints, where J represents the number of the human joints. Figure 1 illustrates the framework of our GAN-based bidirectional motion prediction model. The inverse loss in the backward direction deals with the long-term average prediction problem. The Transformer-based motion generator is employed for long-term motion prediction. Moreover, we observe that the Euclidean distance

can only measure the physical difference between two individual frames at the same time stamp, failing to capture the high-level long-term semantic difference. To solve this problem, we introduce a GAN-based dual-discriminator which evaluates the predictions implicitly both at the frame level and the sequence level which encodes the semantics. A soft-DTW loss is used to measure the similarity at the sequence level as it is robust to shifts or dilatations across the time dimension.

3.1 Inverse Loss in the Bidirectional Generation

Inspired by the unsupervised vision tasks, *e.g.*, domain adaption [2] and image translation [42, 49], with large enough capacity, a network can map a set of input images to any permutation of images in the target domain, and any of these learned mappings can produce an output distribution that matches the target distribution. To narrow down the possible mapping functions, even more, Zhu *et al.* [49] argue that the learned mapping functions should be cycle-consistent, meaning that for each image x from domain X, the image translation cycle should be able to return x to its original form, as follows:

$$x \rightarrow G(x) \rightarrow Q(G(x)) \approx x. \tag{1}$$

Eq. 1 formulates the cycle consistency. Similarly, as shown in Figure 1, for the previous sequence $f_{1:N} = \{f_1, \cdots, f_N\}$, our motion prediction network G should also satisfy the forward-backward motion consistency. Let $\tilde{f}_{N:1} = \left\{\tilde{f}_N, \cdots, \tilde{f}_1\right\}$ be the reversed prediction. We pose an inverse loss (*i.e.*, $L2$ loss) to measure the distance between $f_{1:N}$ and $\tilde{f}_{N:1}$. To make the two sequences match each other in the time dimension, we reverse $f_{1:N}$ to $f_{N:1}$ before computing the inverse loss:

$$\mathcal{L}_{inv} = L_2(f_{N:1}, \tilde{f}_{N:1}), \tag{2}$$

where $\tilde{f}_{N:1} = Q(G(f_{1:N}))$. Note that we share the network parameter between G and Q to reduce the model capacity.

3.2 Transformer-Based Motion Generator

Transformer-Based Prediction. The tremendous success of Transformers has been recently notable for their use to model long-range dependencies between sequential data. As illustrated in Figure 1, we propose a Transformer-based motion generator for the bidirectional prediction with the weight-sharing strategy. In particular, as shown in the left part of Figure 2, in order to modify the Transformer to fit the motion prediction task, we propose to define the latest observed frames as queries of the Transformer decoder, which is different from the most visual Transformer architectures such as [10]. Firstly, our modified Transformer generator can avoid the exposure bias [37] problem existing in the original Transformer structure. Exposure bias refers to the discrepancy of context between the training and testing stages. Specifically, it is the scenario of the model trained to predict the next object using the ground truth as context, while during inference only conditioned on the previous sequence generated by the resulting model. This discrepancy results in error accumulation among the sequence since the model has to predict based on a never seen distribution during training. While our proposed Transformer generator has the same context at training and inference with no gap between them. Secondly, for every predicting step, our modified Transformer generator allows for parallel computation not only at training but also at testing, which can significantly reduce the inference time consumption compared to the original Transformer.

The detailed structure of the Transformer encoder and decoder is depicted in the right part of Figure 2. We maintain the standard structure of the Transformer module described in [44] for simplicity. The Transformer consists of several stacked encoder and decoder blocks. Each encoder block is constructed by multi-head self-attention, layer normalization (LN), residual connections, as well as a position-wise feed-forward multi-layer perceptron (MLP) block. The MLP contains two layers with a GELU non-linearity, while each decoder block has

an extra encoder-decoder attention layer compared to the encoder. In addition, before the encoder and decoder block, a learnable temporal positional encoding is conducted.

Specifically, given the motion sequence $f_{1:N} = \{f_1, \cdots, f_N\}$, we split it to $\{f_1, ..., f_K\}$ and $\{f_{K+1}, ..., f_N\}$ as encoder input and decoder input, respectively. For each frame f_i, we add the position embedding to retain the positional information. The input of Transformer encoder $\mathbf{F}^E$ with the size of $[K, J, d]$ is the result after adding position embedding, where J is the number of human joints and d is the embedding dimension of each joint. While the input of Transformer decoder is $\mathbf{F}^D$ with the size of $[N-K, J, d]$. The Transformer structure is the same as the one in [44]. Notice that there is no mask when computing the multi-head attention in our Transformer decoder. The output of Transformer encoder and decoder is $\mathbf{O}^{enc}$ and $\mathbf{O}^{dec}$. The size of these two outputs is the same as their inputs, respectively.

Cross-Modal Attention. We utilize the cross-modal attention proposed in [41] as shown in Figure 2, whose inputs are the output of the Transformer decoder $\mathbf{O}^{dec}$ and the DCT coefficients $\mathbf{O}^{dct}$. The output of the cross-modal attention is the refined feature, which will be fed into the GCN predictor. Specifically, we directly perform a matrix multiplication between $\mathbf{O}^{dec}$ and $\mathbf{O}^{dct}$, and apply a Softmax layer to produce a correlation matrix $\mathbf{A}$,

$$\mathbf{A}_{ji} = \frac{\exp(\mathbf{O}_i^{dec}\mathbf{O}_j^{dct})}{\sum_{i=1}^{n}\exp(\mathbf{O}_i^{dec}\mathbf{O}_j^{dct})}, \tag{3}$$

where $\mathbf{A}_{ji}$ measures the impact of the i-th position of $\mathbf{O}^{dec}$ on the j-th position of the frequency code $\mathbf{O}^{dct}$. In this crossing way, the model can capture more joint influence between $\mathbf{O}^{dec}$ and $\mathbf{O}^{dct}$, producing a richer feature.

We then perform a matrix multiplication between $\mathbf{O}^{dec}$ and the transpose of $\mathbf{A}$ and reshape the result to the original size of $\mathbf{O}^{dec}$. Finally, we multiply the result by a scale parameter α and conduct an element-wise sum operation with the original $\mathbf{O}^{dec}$ to obtain the refined feature $\hat{\mathbf{O}}^{dec}$,

$$\hat{\mathbf{O}}^{dec} = \alpha\sum_{i=1}^{n}(\mathbf{A}_{ji}\mathbf{O}_i^{dec}) + \mathbf{O}^{dec}, \tag{4}$$

where α is 0 in the beginning but is gradually updated. By doing so, each frame of the refined feature $\hat{\mathbf{O}}^{dec}$ is a weighted sum of all frames of $\mathbf{O}^{dec}$ and the previous $\mathbf{O}^{dec}$. Thus, it has a global contextual view between $\hat{\mathbf{O}}^{dec}$ and $\hat{\mathbf{O}}^{dct}$, and it selectively aggregates useful contexts according to the correlation matrix $\mathbf{A}$.

At the same time, we can update $\mathbf{O}^{dct}$ using the cross-model attention model. Similar to Equation (3), we first perform a matrix multiplication between $\mathbf{O}^{dct}$ and $\mathbf{O}^{dec}$, and apply a Softmax layer to produce another correlation matrix $\mathbf{B}$. Similar to Equation (4), we then perform a matrix multiplication between $\mathbf{O}^{dct}$ and the transpose of $\mathbf{B}$ and reshape the result to the same size of $\mathbf{O}^{dct}$. After that, we multiply the result by a scale parameter β and conduct an element-wise sum operation with the original feature $\mathbf{O}^{dct}$. Finally, we concatenate both $\hat{\mathbf{O}}^{dec}$ and $\hat{\mathbf{O}}^{dct}$ in channel-wise, and feed the result to the GCN predictor to produce the motion sequence.

3.3 Adversarial Dual-Discriminator

Frame-Based Discriminator. To achieve frame-based prediction between the generated sequence and the real sequence, we use a frame-based discriminator D_F in [38] as one of the objectives of the proposed motion generator G. The adversarial loss of the frame-based discriminator (D_F) can be expressed as follows,

$$\mathcal{L}_{D_F} = \sum_{i=N+1}^{N+P}\left(\mathbb{E}_{f_i}\left[\log D_F(f_i)\right] + \mathbb{E}_{\hat{f}_i}\left[\log(1 - D_F(\hat{f}_i))\right]\right). \tag{5}$$

By doing so, we make the generated frame $\hat{f}_k$ look realistic to the real frame f_i in an adversarial way.

Sequence-Based Discriminator. Another objective of the proposed motion generator G is to achieve temporal coherence of the generated skeleton sequence. For example, when a man moves his left hand, his right hand should keep still for multiple frames. Thus, we utilize a sequence-based discriminator D_S proposed in [38] to achieve the coherence between consecutive frames of the generated sequences $\mathcal{F}_P = \hat{f}_{N+1:N+P}$ and the real one $\mathcal{F}_P^{real} = f_{N+1:N+P}$. Therefore, the adversarial loss of sequence-based discriminator D_S is defined as,

$$\mathcal{L}_{D_S} = \mathbb{E}_{\mathcal{F}_P^{real}}\left[\log D_S(\mathcal{F}_P^{real})\right] + \mathbb{E}_{\mathcal{F}_P}\left[\log(1 - D_S(\mathcal{F}_P))\right]. \tag{6}$$

Thus, the final adversarial loss of the GAN framework is the sum of both Equations (5) and (6),

$$\mathcal{L}_{adv} = \mathcal{L}_{D_F} + \mathcal{L}_{D_S}. \tag{7}$$

3.4 Optimization Objective

Mean Per Joint Position Loss. Similar to prior work [33, 34], we use the Mean Per Joint Position Error (MPJPE) proposed in [17].

$$\mathcal{L}_{\text{MPJPE}} = \frac{1}{J \times P} \sum_{k=N+1}^{N+P} \sum_{j=1}^{J} \left\| \hat{f}_{k,j} - f_{k,j} \right\|^2, \tag{8}$$

where $f_k \in \mathbb{R}^3$, $\hat{f}_k \in \mathbb{R}^3$ are the ground truth and predicted motions at future time step k respectively.

Soft-DTW Loss. To measure the overall similarity between two-time series signals, we use the Soft-DTW loss proposed in [6]. One advantage of Soft-DTW is that it is differentiable with respect to all of its arguments. It is derived from the original DTW discrepancy. Different from the Euclidean distance, DTW is robust to time shifts or dilatations.

Specifically, given two sequences $\mathbf{x} = (x_1, \ldots, x_m)$ and $\mathbf{y} = (y_1, \ldots, y_n)$, the Soft-DTW loss $\mathcal{L}_{\text{Soft-DTW}}$ is defined as:

$$\text{DTW}^{\gamma}(\mathbf{x}, \mathbf{y}) = -\gamma \log \sum_{i=1}^{m} \exp\left(-\frac{\langle \mathbf{C}, \Delta(\mathbf{x}, \mathbf{y}) \rangle}{\gamma}\right), \tag{9}$$

where $\gamma > 0$ is the smoothing parameter, C is the alignment matrix. $\Delta(\mathbf{x}, \mathbf{y}) := \left[\delta\left(x_i, y_j\right)\right]_{ij} \in \mathbb{R}^{m \times n}$ is the cost matrix, where δ is the substitution-cost function, which in most cases is the squared Euclidean distance.

Overall Optimization Objective. We use four different losses as our full optimization objective, *i.e.*, adversarial loss $\mathcal{L}_{adv}$, inverse loss $\mathcal{L}_{inv}$, Soft-DTW loss $\mathcal{L}_{\text{Soft-DTW}}$, and the mean per joint position loss $\mathcal{L}_{\text{MPJPE}}$, which can be expressed as,

$$\min_{G} \max_{D_s, D_f} \mathcal{L} = \lambda_{adv}\mathcal{L}_{adv} + \lambda_{mpjpe}\mathcal{L}_{\text{MPJPE}} + \lambda_{dtw}\mathcal{L}_{\text{Soft-DTW}} + \lambda_{inv}\mathcal{L}_{inv}, \tag{10}$$

where λ_{adv}, λ_{mpjpe}, λ_{sd} and λ_{inv} are the weights, measuring the corresponding contributions of each loss to the total loss $\mathcal{L}$. Here, we experimentally set these weights. However, we found that lower $\mathcal{L}_{inv}$ will obtain better performance since the inverse loss produced by the reconstructed input sequence is not that important compared to the losses $\mathcal{L}_{adv}$, $\mathcal{L}_{\text{MPJPE}}$ and $\mathcal{L}_{\text{Soft-DTW}}$ with the real data.

4 EXPERIMENTS

In this section, we evaluate our method on the publicly available Human3.6M dataset [17] which is one of the widest benchmark datasets in motion prediction. The dataset contains 15 activities collected from 7 different human subjects. The high-quality 3D data are recorded by using a Vicon motion capture system. For each clip in the dataset, there are 32 joints with 3D locations captured for one person. To enable the comparison of the results, we follow the data processing settings of previous works. In particular, we calculate the 3D joint coordinates using forward kinematics on a standard skeleton as in [33, 34]. As in [15, 23, 33–35, 45], we delete the global

rotation, translation as well as constant angles. As in [15, 23, 33–35] we down sample the frame rate from 50 to 25. At inference time, we also test our method on Subject 5 as in [15, 33–35] and report our results on 256 sub-sequences per action for 3D coordinates as in [33].

Evaluation Metrics. We evaluate our model in terms of two metrics. Specifically, to measure the similarity of the pose sequences directly, we evaluate our model in terms of the widely used Mean Per Joint Position Error (MPJPE) [17] in millimeter. To capture the semantic relationship between two sequences, DTW error is also reported. In particular, for the 3D position of each joint, MPJPE is computed by the average $L2$ distance between ground truth and our prediction motion sequence overall predicted time steps. While the DTW error estimates similarly of two sequences.

Implementation Details. The proposed network is implemented in PyTorch framework. Following the same settings of [33], our method is trained using the Adam optimizer [20] for 50 epochs with data batches of size 32 for Human3.6M. For Equation (10), note that we split $\mathcal{L}_{\text{MPJPE}}$ and $\mathcal{L}_{\text{Soft-DTW}}$ to two parts, which are $\mathcal{L}^{A}_{\text{MPJPE}}$ and $\mathcal{L}^{B}_{\text{MPJPE}}$, $\mathcal{L}^{A}_{\text{Soft-DTW}}$ and $\mathcal{L}^{B}_{\text{Soft-DTW}}$, respectively. Here, $\mathcal{L}^{A}_{\text{MPJPE}}$ represents the loss of earlier predicted frames, while the $\mathcal{L}^{B}_{\text{Soft-DTW}}$ represents the latter part of the predicted sequence. Accordingly, λ_{mpjpe} and λ_{sd} are separated to λ^{a}_{mpjpe} and λ^{b}_{mpjpe}, λ^{a}_{sd} and λ^{b}_{sd}, respectively.

Specifically, both λ^{a}_{mpjpe} and λ^{a}_{sd} are set to 9.8, while both λ^{b}_{mpjpe} and λ^{b}_{sd} are set to 0.1. This selection is based on the assumption that the first predicted poses will affect the later prediction due to the recursive forecasting mechanism, as shown in Figure 2. In addition, λ_{inv} is set to 0.1 because of the reason described in Section 3, and λ_{adv} is set to 1.

State-of-the-Art Comparisons. We compare our approach with one RNN-based methods, residual sup. [35] and three feed-forward models, ConvSeq2Seq [23], LTD [34], and His. Rep. [33], which constitutes the state of the art. For the prediction sequence lasting within 1000ms, we take the results from [33] directly. Otherwise, we use the results of His. Rep. [33], we utilized the pre-trained model released by the authors for Human3.6M, then predict the longer motion sequence recursively.

4.1 Quantitative Results

Following the settings of our baselines [23, 33–35, 43], we present the results for mid-term and long-term prediction. Specifically, in order to make comparisons with recent SOTAs conveniently, we defined those two time scales as $(500, 1000]ms$ and $(1000, 4000]ms$.

For Human3.6M, our model is trained with past 50 frames as input and predict future 10 frames. We further produce poses literately in a recursive way by concatenating the predictions with the history.

Results on Human3.6M. We compare our method with seven state-of-the-art methods, including Residual sup. [35], ConvSeq2Seq [23], LTD [34], His. Rep. [33], DMGNN [26], MSR-GCN [8], and STSGCN [39]. As shown in Table 1, Table 2 and Table 3, we provide the results for mid-term and long-term prediction in 3D space, respectively. Methods His. Rep. [33], MSR-GCN [8] and STSGCN [39] released their codes publicly, we used the results from their pre-trained models or re-trained the models and tested under the same protocol for fair comparison.

The results in Table 1 and Table 2 indicate that our method outperforms all the competing methods on average for both mid-term and long-term prediction. Note that we surpass LTD-50-25 [34], DMGNN [26], and STSGCN [39] almost all the time. In particular, as shown in Table 2, for the MPJPE metric, our method exceeds STSGCN [39] and His. Rep. [33] by 12.8 and 7.8 averaging for the 4s prediction respectively. Also, we outperform STSGCN [39] and His. Rep. [33] for 14 and 13 (15 totally) action classes in long-term prediction. Besides, Table 2 illustrates the effectiveness of our method especially for those acyclic motions, *e.g.*, "walking dog", "posing" and "directions", we even outperform the previous most competitive baseline MSR-GCN [8] by a large margin, which

Table 1. MPJPE error of 3D joint position on Human3.6M for mid-term prediction. The error is measured in millimeter. The two numbers after the method name "LTD" indicate the number of observed frames and that of predicted frames respectively. Best results in bold.

	Walking				Eating				Smoking				Discussion			
Milliseconds	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000
Residual sup. (CVPR'17) [35]	71.6	72.5	76.0	79.1	74.9	85.9	93.8	98.0	78.1	88.6	96.6	102.1	109.5	122.0	128.6	131.8
ConvSeq2Seq (CVPR'18) [23]	72.2	77.2	80.9	82.3	61.3	72.8	81.8	87.1	60.0	69.4	77.2	81.7	98.1	112.9	123.0	129.3
LTD-50-25 (ICCV'19) [34]	50.7	54.4	57.4	60.3	51.5	62.6	71.3	75.8	50.5	59.3	67.1	72.1	88.9	103.9	113.6	118.5
LTD-10-25 (ICCV'19) [34]	51.8	56.2	58.9	60.9	50.0	61.1	69.6	74.1	51.3	60.8	68.7	73.6	87.6	103.2	113.1	118.6
LTD-10-10 (ICCV'19) [34]	53.1	59.9	66.2	70.7	51.1	62.5	72.9	78.6	49.4	59.2	66.9	71.8	88.1	104.5	115.5	121.6
His. Rep. (ECCV'20) [33]	**47.4**	**52.1**	**55.5**	**58.1**	50.0	61.4	70.6	75.7	**47.6**	**56.6**	**64.4**	**69.5**	86.6	102.2	113.2	119.8
DMGNN (CVPR'20) [26]	73.4	-	-	95.8	58.1	-	-	86.7	50.9	-	-	72.2	**81.9**	-	-	138.3
STSGCN (ICCV'21) [39]	58.0	60.7	64.1	70.2	57.4	69.7	77.9	82.6	55.5	65.6	72.3	76.1	91.1	105.3	114.2	118.9
MSR-GCN (ICCV'21) [8]	53.3	55.4	58.1	63.7	50.8	61.4	69.7	75.4	50.5	59.5	67.1	72.1	87.0	101.9	**111.4**	116.8
Ours	49.8	55.0	58.5	60.5	**48.5**	**59.2**	**68.2**	**73.0**	48.4	57.5	65.0	70.0	85.8	**101.2**	111.6	**116.4**

	Directions				Greeting				Phoning				Posing				Purchases				Sitting			
Milliseconds	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000
Residual sup.(CVPR'17) [35]	101.1	114.5	124.5	129.1	126.1	138.8	150.3	153.9	94.0	107.7	119.1	126.4	140.3	159.8	173.2	183.2	122.1	137.2	148.0	154.0	113.7	130.5	144.4	152.6
ConvSeq2Seq (CVPR'18) [23]	86.6	99.8	109.9	115.8	116.9	130.7	142.7	147.3	77.1	92.1	105.5	114.0	122.5	148.8	171.8	187.4	111.3	129.1	143.1	151.5	82.4	98.8	112.4	120.7
LTD-50-25 (ICCV'19) [34]	74.2	88.1	99.4	**105.5**	104.8	119.7	132.1	136.8	68.8	83.6	96.8	105.1	110.2	137.8	160.8	174.8	99.2	114.9	127.1	134.9	79.2	96.2	110.3	118.7
LTD-10-25 (ICCV'19) [34]	76.1	91.0	102.8	108.8	104.3	120.9	134.6	140.2	68.7	84.0	97.2	105.1	109.9	136.8	158.3	171.7	99.4	114.9	127.9	135.9	78.5	95.7	110.0	118.8
LTD-10-10 (ICCV'19) [34]	**72.2**	**86.7**	**98.5**	105.8	103.7	120.6	134.7	140.9	67.8	83.0	96.4	105.1	107.6	136.1	159.5	175.0	98.3	115.1	130.1	139.3	76.4	93.1	106.9	115.7
His. Rep. (ECCV'20) [33]	73.9	88.2	100.1	106.5	101.9	118.4	132.7	138.8	67.4	82.9	96.5	105.0	107.6	136.8	161.4	178.2	**95.6**	**110.9**	**125.0**	**134.2**	76.4	93.1	107.0	115.9
DMGNN (CVPR'20) [26]	110.1	-	-	115.8	152.5	-	-	157.7	78.9	-	-	98.6	163.9	-	-	310.1	118.6	-	-	153.8	**60.1**	-	-	**104.9**
STSGCN (ICCV'21) [39]	79.9	95.0	103.9	109.6	106.3	119.9	130.1	**136.1**	73.1	87.9	100.6	108.3	119.7	146.3	165.4	178.4	106.8	122.1	134.1	141.0	84.7	102.4	114.8	121.4
MSR-GCN (ICCV'21) [8]	75.8	89.9	100.5	105.9	106.3	120.0	131.5	136.3	67.9	82.5	**95.8**	104.7	112.5	140.1	162.8	176.5	99.2	114.0	126.9	134.4	77.6	94.0	107.7	115.9
Ours	73.3	87.9	99.7	106.3	**101.1**	**117.8**	**131.3**	136.4	**67.3**	**82.3**	94.9	**103.2**	**107.1**	**134.6**	**156.7**	**171.0**	99.0	113.7	127.1	135.1	76.0	**92.0**	**105.4**	114.4

	Sitting down				Taking photo				Waiting				Walking dog				Walking together				Average			
Milliseconds	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000
Residual sup. (CVPR'17) [35]	138.8	159.0	176.1	187.4	110.6	128.9	143.7	153.9	105.4	117.3	128.1	135.4	128.7	141.1	155.3	164.5	80.2	87.3	92.8	98.2	106.3	119.4	130.0	136.6
ConvSeq2Seq (CVPR'18) [23]	106.5	125.1	139.8	150.3	84.4	102.4	117.7	128.1	87.3	100.3	110.7	117.7	122.4	133.8	151.1	162.4	72.0	77.7	82.9	87.4	90.7	104.7	116.7	124.2
LTD-50-25 (ICCV'19) [34]	100.2	118.2	133.1	143.8	75.3	93.5	108.4	118.8	77.2	90.6	101.1	108.3	107.8	120.3	136.3	146.4	56.0	60.3	63.1	65.7	79.6	93.6	105.2	112.4
LTD-10-25 (ICCV'19) [34]	99.5	118.5	133.6	144.1	76.8	95.3	110.3	120.2	75.1	88.7	99.5	106.9	105.8	**118.7**	**132.8**	**142.2**	58.0	63.6	67.0	69.6	79.5	94.0	105.6	112.7
LTD-10-10 (ICCV'19) [34]	96.2	115.2	130.8	142.2	72.5	90.9	105.9	116.3	73.4	88.2	99.8	107.5	109.7	122.8	139.0	150.1	55.7	61.3	66.4	69.8	78.3	93.3	106.0	114.0
His. Rep. (ECCV'20) [33]	97.0	116.1	132.1	143.6	**72.1**	**90.4**	**105.5**	**115.9**	74.5	89.0	100.3	108.2	108.2	120.6	135.9	146.9	**52.7**	**57.8**	**62.0**	**64.9**	**77.3**	91.8	104.1	112.1
DMGNN (CVPR'20) [26]	122.1	-	-	168.8	91.6	-	-	120.7	106.0	-	-	136.7	194.0	-	-	182.3	83.4	-	-	115.9	103.0	-	-	137.2
STSGCN (ICCV'21) [39]	105.2	124.8	139.2	148.4	84.2	104.6	116.6	126.3	80.8	95.7	106.4	113.6	115.4	128.1	141.6	151.5	58.9	62.3	66.9	72.5	85.1	99.4	109.9	117.0
MSR-GCN (ICCV'21) [8]	102.4	122.7	139.6	149.3	77.7	96.9	112.3	121.9	74.8	87.8	98.2	105.5	107.7	120.8	135.7	145.7	56.2	60.9	**65.0**	69.5	80.0	93.9	105.5	112.9
Ours	**96.2**	**114.5**	**129.9**	**141.3**	74.2	92.6	107.4	117.7	**72.9**	**87.3**	**97.7**	**104.9**	**105.4**	120.4	136.4	148.3	54.3	59.7	64.2	67.3	**77.3**	**91.7**	**103.6**	**111.1**

are 10 and 11 for 4s prediction respectively. The most probable reason is that the other methods tend to generate representative static poses or failed to forecast the long horizon dynamics.

Furthermore, we also provide the comparison results of the DTW metric in Table 3. Our approach surpasses all the baselines on average. Specifically, the proposed method is much better than the previous method STSGCN [39] by a margin of 2.4×10^7 averaging for 4s prediction. We also achieve the best results on 9 activities such as "eating", "smoking", "discussion", and so on. This performance gain clearly indicates that our method can produce sequences more similar to the ground truth globally.

4.2 Qualitative Results

We also provide qualitative results in Figure 3 and Figure 4 for long-term and mid-term prediction including smoking, eating, taking photo, walking dog and greeting actions. Compared with His. Rep. [33], STSGCN [39] and MSR-GCN [8], our approach can predict more dynamic and accurate future poses, which can capture both the key poses of the defined classes and the underlying dynamics for long-term prediction.

For instance, for the smoking action (see Figure 3 (a)), we can predict someone lighting a cigarette with right hand, then put this hand down, while His. Rep. [33] can only predict the movement of lighting a cigarette, and the frames in the black dashed box illustrate that the smoking poses almost stand still in more than a quarter of the forecast duration. For the greeting action (see Figure 4 (a)), we can also forecast the underlying pattern which is rising hands to greet and then laying down them, while His. Rep. [33] can only forecast the movement of rising hands and quickly produce freezing motion (as shown in the black dashed box).

Furthermore, we can learn the action-conditioned pattern for parts of the human body. For example, for the walking dog action (see Figure 3 (d)), our prediction is sitting while walking the dog, but His. Rep. [33] can not predict the dynamic of human arms and tend to converge to the mean pose for a long time.

Table 2. MPJPE error of 3D joint position on Human3.6M for long-term prediction. Best results in bold.

	Milliseconds	1200	1400	1600	1800	2000	2200	2400	2600	2800	3000	3200	3400	3600	3800	4000
Walking	His. Rep. (ECCV'20) [33]	**59.1**	61.6	**66.5**	**72.2**	**73.4**	79.8	**83.0**	84.9	84.5	93.8	94.7	96.7	100.3	109.8	106.5
	STSGCN (ICCV'21) [39]	81.6	84.3	86.8	91.7	95.4	95.9	98.2	100.2	104.2	125.9	109.5	108.8	110.4	120.9	112.3
	MSR-GCN (ICCV'21) [8]	62.8	**60.9**	68.0	76.8	78.8	**79.7**	83.6	**84.4**	**83.2**	**88.9**	**93.0**	**94.7**	**96.4**	**102.0**	**102.3**
	Ours	61.9	63.4	71.4	77.4	81.0	84.5	90.8	93.0	94.2	101.8	107.6	107.7	111.4	120.2	119.7
Eating	His. Rep. (ECCV'20) [33]	82.6	87.9	90.7	93.1	96.1	98.6	100.3	103.5	106.3	106.1	105.7	108.0	110.6	112.3	116.1
	STSGCN (ICCV'21) [39]	105.5	112.0	113.8	119.5	124.0	126.3	127.1	128.5	132.0	153.2	133.4	135.6	136.2	143.2	136.8
	MSR-GCN (ICCV'21) [8]	85.2	88.9	91.3	94.4	96.3	97.3	**98.7**	**100.2**	**102.7**	**103.0**	**104.2**	**106.7**	**108.4**	**108.3**	**111.7**
	Ours	**80.6**	**85.3**	**89.4**	**92.4**	**95.4**	**97.0**	99.8	102.0	104.8	104.0	104.2	105.8	107.6	109.3	114.0
Smoking	His. Rep. (ECCV'20) [33]	**77.2**	**84.0**	**89.4**	94.8	101.8	106.8	110.4	114.0	117.5	122.5	125.7	130.9	134.9	139.6	142.2
	STSGCN (ICCV'21) [39]	112.9	121.5	126.2	132.9	140.6	145.7	147.7	152.1	155.8	170.1	160.8	161.3	164.1	168.9	165.7
	MSR-GCN (ICCV'21) [8]	84.9	93.4	99.2	104.5	111.3	117.0	120.4	123.6	125.3	130.0	132.0	135.0	138.6	142.7	144.7
	Ours	77.7	84.4	89.5	**94.4**	**100.5**	**105.3**	109.8	**113.1**	**117.1**	**122.4**	**125.7**	**129.7**	**133.8**	**136.8**	**140.2**
Discussion	His. Rep. (ECCV'20) [33]	131.3	138.6	144.3	149.0	151.1	152.4	159.3	163.1	163.5	166.6	166.9	169.7	170.7	173.5	174.0
	STSGCN (ICCV'21) [39]	138.1	142.4	147.3	149.0	152.2	154.1	158.0	162.3	161.4	176.7	164.7	168.3	169.0	173.8	168.0
	MSR-GCN (ICCV'21) [8]	132.1	136.5	140.4	144.9	147.2	147.2	151.4	153.7	154.3	157.4	159.0	162.6	164.2	167.1	166.3
	Ours	**125.2**	**130.8**	**136.0**	**140.0**	**142.6**	**143.2**	**148.4**	**151.5**	**152.2**	**155.6**	**157.4**	**160.0**	**161.8**	**163.9**	**163.2**
Directions	His. Rep. (ECCV'20) [33]	116.3	121.4	126.7	130.0	132.9	136.1	138.0	139.1	141.8	147.4	152.8	153.8	153.8	154.8	156.7
	STSGCN (ICCV'21) [39]	140.2	145.3	147.1	151.2	153.1	155.6	156.5	157.4	157.2	183.0	161.2	160.2	158.4	168.7	156.4
	MSR-GCN (ICCV'21) [8]	126.3	132.9	138.8	143.9	146.4	149.6	151.9	152.9	154.6	158.4	162.4	161.3	159.5	158.1	157.1
	Ours	**116.1**	**121.0**	**126.4**	**129.7**	**132.1**	**134.4**	**135.6**	**136.0**	**138.2**	**142.6**	**146.7**	**147.3**	**146.6**	**145.8**	**147.1**
Greeting	His. Rep. (ECCV'20) [33]	148.2	152.4	153.1	150.2	150.6	151.8	150.7	151.3	155.5	157.8	159.0	158.4	159.9	160.1	158.6
	STSGCN (ICCV'21) [39]	158.6	157.8	161.9	160.9	163.3	163.9	162.7	164.2	166.4	172.1	167.0	166.0	163.9	163.5	164.8
	MSR-GCN (ICCV'21) [8]	151.1	154.9	154.0	151.9	150.9	150.8	148.3	149.2	153.4	154.2	155.3	155.3	154.1	**152.7**	**151.7**
	Ours	**144.3**	**147.3**	**148.0**	**144.4**	**142.4**	**144.2**	**145.1**	**144.5**	**148.3**	**149.5**	**151.7**	**150.1**	**152.0**	153.0	152.3
Phoning	His. Rep. (ECCV'20) [33]	118.9	132.4	144.2	153.4	162.3	169.6	176.0	181.4	187.6	192.5	195.4	196.6	200.1	202.2	203.4
	STSGCN (ICCV'21) [39]	142.3	150.3	157.6	163.8	167.9	170.7	173.9	178.6	182.6	196.1	190.5	190.7	191.2	193.4	191.0
	MSR-GCN (ICCV'21) [8]	120.2	131.5	141.8	150.4	157.4	162.3	167.0	171.5	176.4	180.7	184.5	186.4	188.9	190.2	191.8
	Ours	**116.8**	**128.9**	**140.4**	**148.6**	**156.5**	**161.1**	**167.0**	**171.8**	**177.0**	**181.0**	**183.7**	**184.2**	**186.7**	**188.4**	**189.3**
Posing	His. Rep. (ECCV'20) [33]	202.5	220.4	233.3	246.4	254.6	257.6	256.3	254.8	254.9	252.0	251.2	247.2	247.5	249.8	246.6
	STSGCN (ICCV'21) [39]	206.1	212.9	221.5	229.9	234.3	237.2	237.5	239.1	243.1	249.6	244.0	240.6	239.8	238.0	235.9
	MSR-GCN (ICCV'21) [8]	208.4	221.3	227.7	233.6	236.3	236.6	236.3	238.6	241.0	240.3	239.9	235.9	234.4	235.0	232.1
	Ours	**191.8**	**205.9**	**216.4**	**228.0**	**236.5**	**238.7**	**237.7**	**236.7**	**237.9**	**235.0**	**233.5**	**226.9**	**224.7**	**224.2**	**220.6**
Purchases	His. Rep. (ECCV'20) [33]	145.9	154.0	162.5	166.2	172.2	181.6	186.7	189.5	192.8	193.7	194.9	197.3	207.3	211.7	213.5
	STSGCN (ICCV'21) [39]	158.6	166.3	170.2	173.6	176.9	184.1	186.8	188.8	190.9	210.4	193.3	195.1	199.8	210.1	200.8
	MSR-GCN (ICCV'21) [8]	147.8	153.8	160.4	163.8	169.8	178.2	181.2	182.1	183.5	183.7	184.2	185.3	193.3	197.2	199.6
	Ours	**145.2**	**151.3**	**156.6**	**158.2**	**164.0**	**172.1**	**175.9**	**177.6**	**179.4**	**179.5**	**180.6**	**183.0**	**192.0**	**195.6**	**197.1**
Sitting	His. Rep. (ECCV'20) [33]	132.3	147.7	158.5	168.2	178.9	189.2	200.7	209.7	217.1	223.1	227.2	232.0	236.0	239.7	241.9
	STSGCN (ICCV'21) [39]	177.0	187.3	198.6	206.9	213.6	217.5	221.6	224.7	227.6	221.1	232.0	231.7	232.1	226.9	231.5
	MSR-GCN (ICCV'21) [8]	140.2	156.2	166.6	175.7	185.5	193.5	200.6	205.8	209.2	**211.2**	**214.1**	**216.6**	**217.9**	**219.4**	**219.0**
	Ours	**131.3**	**146.4**	**156.7**	**165.6**	**176.1**	**185.9**	**195.7**	**202.6**	**208.8**	213.2	217.7	222.1	226.1	228.2	228.8
Sitting Down	His. Rep. (ECCV'20) [33]	159.9	177.6	191.5	204.3	217.4	231.7	246.7	257.9	268.0	276.3	283.3	290.2	294.9	299.9	301.8
	STSGCN (ICCV'21) [39]	216.2	227.2	236.9	247.1	255.3	263.8	271.5	277.3	284.3	287.8	288.3	290.2	292.2	292.3	291.2
	MSR-GCN (ICCV'21) [8]	179.3	196.7	209.5	221.1	232.1	242.6	252.8	261.0	269.4	274.8	278.9	282.7	283.8	**285.0**	**284.3**
	Ours	**157.1**	**173.7**	**188.0**	**200.7**	**213.3**	**225.9**	**239.2**	**248.5**	**258.1**	**264.9**	**272.2**	**278.6**	**283.4**	286.8	289.0
Taking Photo	His. Rep. (ECCV'20) [33]	**133.1**	**148.5**	161.0	172.3	182.3	191.7	199.8	208.1	212.2	217.2	226.1	231.1	234.7	238.9	241.5
	STSGCN (ICCV'21) [39]	181.7	196.9	202.1	211.8	221.4	230.3	233.9	240.9	243.6	254.7	249.0	253.0	253.6	257.9	251.8
	MSR-GCN (ICCV'21) [8]	153.3	168.9	179.0	189.3	197.3	203.6	208.7	213.9	216.0	217.4	221.9	224.6	**225.5**	**225.9**	**226.6**
	Ours	134.8	149.8	**160.5**	**171.3**	**181.6**	**190.7**	**198.3**	**205.7**	**208.3**	**211.8**	**219.2**	**224.2**	227.2	229.1	230.8
Waiting	His. Rep. (ECCV'20) [33]	119.5	129.5	138.1	147.8	156.3	166.6	172.4	177.0	181.9	186.2	189.7	193.1	196.6	201.1	203.7
	STSGCN (ICCV'21) [39]	154.6	161.0	167.7	174.0	175.8	179.0	182.9	185.6	189.7	201.0	193.9	194.2	195.0	197.9	194.9
	MSR-GCN (ICCV'21) [8]	130.4	139.1	144.7	152.6	157.8	163.9	169.6	172.3	**175.3**	**177.5**	**180.4**	**182.8**	**186.1**	**188.7**	**187.4**
	Ours	**115.5**	**124.7**	**132.6**	**142.5**	**151.3**	**161.1**	**167.6**	**171.6**	175.7	178.6	183.0	187.4	191.4	194.1	196.3
Walking Dog	His. Rep. (ECCV'20) [33]	**158.4**	**170.2**	**180.2**	**189.8**	**195.7**	204.3	214.7	221.5	227.2	230.6	227.6	228.0	230.7	233.7	233.5
	STSGCN (ICCV'21) [39]	181.4	192.4	197.8	206.8	213.0	222.1	231.7	238.0	242.7	256.7	250.9	251.2	255.8	264.4	262.2
	MSR-GCN (ICCV'21) [8]	166.6	177.2	187.2	194.7	200.1	207.4	215.3	224.0	230.0	233.9	235.5	238.7	245.0	250.7	250.8
	Ours	161.6	174.5	184.6	193.7	199.9	**206.9**	**214.4**	**216.3**	**220.8**	**221.9**	**216.5**	**216.6**	**221.0**	**224.7**	**227.3**
Walking Together	His. Rep. (ECCV'20) [33]	**69.2**	**72.0**	**76.7**	**82.4**	**85.7**	**87.5**	**90.0**	**93.1**	**94.3**	**95.1**	**97.1**	**98.6**	**100.5**	**102.4**	**103.9**
	STSGCN (ICCV'21) [39]	113.9	120.0	122.4	129.7	135.3	139.6	141.2	144.7	146.7	163.5	147.2	149.1	150.0	162.0	156.2
	MSR-GCN (ICCV'21) [8]	72.3	74.6	78.2	85.3	89.6	92.5	95.1	95.2	96.1	99.0	101.8	104.6	105.6	110.1	111.6
	Ours	72.6	75.2	81.7	84.4	89.3	91.4	96.5	97.0	98.9	99.3	103.7	104.1	106.0	107.8	112.0
Average	His. Rep. (ECCV'20) [33]	123.6	133.2	141.1	148.0	154.1	160.4	165.7	169.9	173.7	177.4	179.8	182.1	185.2	188.6	189.6
	STSGCN (ICCV'21) [39]	151.3	158.5	163.8	169.9	174.8	179.1	182.1	185.5	188.5	201.5	192.4	193.1	194.1	198.8	194.6
	MSR-GCN (ICCV'21) [8]	130.7	139.1	145.8	152.2	157.1	161.5	165.4	168.6	171.4	174.0	176.5	178.2	180.1	182.2	182.5
	Ours	**122.2**	**130.8**	**138.5**	**144.7**	**150.8**	**156.2**	**161.4**	**164.5**	**168.0**	**170.7**	**173.6**	**175.2**	**178.1**	**180.5**	**181.8**

Table 3. DTW error ($\times 10^7$) of different actions on Human3.6M for long-term (4s) prediction. Best results in bold.

Actions	Walking	Eating	Smoking	Discussion	Directions	Greeting	Phoning	Posing	Purchases	Sitting	Sitting Down	Taking Photo	Waiting	Walking Dog	Walking Together	**Average**
His. Rep. (ECCV'20) [33]	**3.1**	6.5	6.7	13.8	12.3	14.7	15.5	27.1	17.4	17.5	28.5	17.4	14.4	21.4	**4.3**	14.7
STSGCN (ICCV'21) [39]	4.7	8.2	9.8	12.6	13.0	14.6	14.0	24.6	16.2	18.6	28.3	20.9	15.4	22.5	9.6	15.5
MSR-GCN (ICCV'21) [8]	3.9	5.9	6.7	11.9	12.5	13.7	**12.3**	23.5	15.2	**14.6**	**24.1**	16.6	**12.9**	20.3	5.1	13.3
Ours	3.6	**5.8**	**6.5**	**11.7**	**10.9**	**13.1**	13.0	**22.5**	**15.1**	15.7	25.8	**15.8**	13.3	**19.0**	4.6	**13.1**

Fig. 3. Qualitative results of long-term (4s) prediction including smoking, eating, taking photo and walking dog actions on the Human3.6M dataset. The first two frames are the latest observed frames, the others are predicted frames. The whole sequence is down-sampled to 5 frames per second. (Best viewed when zoomed in.)

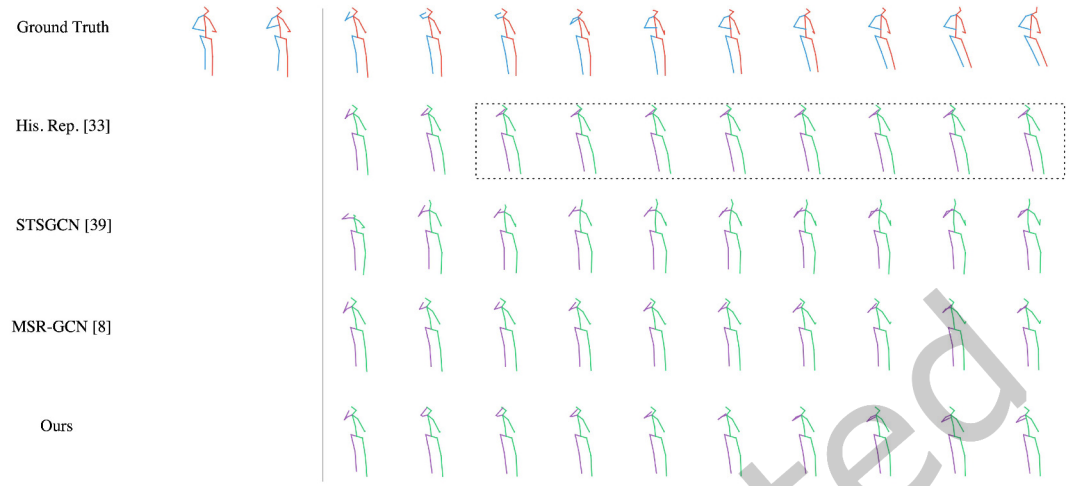

Fig. 4. Qualitative results of 2s prediction on greeting action on the Human3.6M dataset. The first two frames are the latest observed frames, the others are predicted frames. The whole sequence is down-sampled to 5 frames per second. (Best viewed when zoomed in.)

Moreover, for motions with more complexity and randomness, e.g., taking photo and walking dog, our method can generate more dynamic sequences than [33] evidently (see Figure 3), which instead produces mostly static motions. The visualization results comparison shows the consistency that using the bidirectional predictor could alleviate the average forecasting problem to a large extent.

In addition, although MSR-GCN [8] and STSGCN [39] can generate more dynamic motions than His. Rep. [33], some of them are wrong or less accurate than ours. For example, for the smoking behavior in Figure 3(b), both methods predict a person walking while smoking, but the ground truth sequence shows that the person stands there most of the predicted duration. Also, as shown in other cases in Figure 3 and Figure 4, our results are closer to the ground truth.

4.3 Ablation Studies

We conduct extensive ablation studies on the Human3.6M dataset to evaluate the effectiveness of different components of the proposed BiTGAN. As shown in Table 4, the proposed BiTGAN has five variants (*i.e.*, B1, B2, B3, B4, B5). We also compare the performance of variants by the statistics and visualizations in Figure 5 and Figure 6 respectively. In Figure 5(a), the average MPJPE errors are plotted at each future time stamp. In Figure 5(b), the MPJPE error at 4s for each action category are plotted.

- (i) **B1: Motion Transformer**. B1 only uses our designed Transformer module (as shown in Figure 2) to encode and decode the motion sequence. The output of motion Transformer is then concatenated with the DCT coefficients of the last observed sub-sequence, fed as the input of GCN to predict the future poses.
- (ii) **B2: B1 + Cross-modal Attention**. Based on B1, B2 utilizes the cross-modal attention to model the crossing relations between the outputs of Transformer decoder and DCT coefficients.
- (iii) **B3: B2 + bidirectional generation strategy**. B3 adopts the proposed bidirectional generation strategy.
- (iv) **B4: B3 + adversarial discriminators**. B4 employs the associated adversarial discriminators to encourage the outputs indistinguishable from the ground truth at both frame and sequence levels.
- (v) **B5: B4 + Soft-DTW loss**. B5 adopts the Soft-DTW loss to add more constraints to the generator.

Table 4. The ablation study of each component on Human3.6M. B6 is our BiTGAN which combines with each component

	Method	80	160	320	400	560	720	880	1000	1200	1400	1600	1800	2000	2200	2400	2600	2800	3000	3200	3400	3600	3800	4000	Ave.
	His. Rep. [33]	10.4	22.6	47.1	58.3	77.3	91.8	104.1	112.1	123.6	133.2	141.1	148.0	154.1	160.4	165.7	169.9	173.7	177.4	179.8	182.1	185.2	188.6	189.6	138.9
B1	Motion Transformer	10.7	22.9	47.3	58.8	77.7	92.3	104.4	111.2	122.5	133.0	140.6	147.3	152.5	158.2	164.5	168.7	173.1	176.1	179.5	180.4	183.2	186.3	187.9	138.0
B2	B1 + Cross-Modal Attention	10.7	23.7	47.8	59.0	78.3	92.1	104.3	112.4	122.9	132.5	140.3	147.2	150.9	157.3	163.0	167.6	172.8	175.2	178.9	179.5	182.6	185.1	186.4	137.6
B3	B2 + Bidirectional Generation	10.9	23.3	48.2	59.2	77.7	92.1	104.0	111.5	122.9	132.1	139.6	145.7	151.5	157.2	162.6	166.2	169.5	172.6	175.4	177.1	180.1	182.8	184.2	136.4
B4	B3 + Adversarial Discriminators	11.0	23.3	48.4	59.4	77.7	91.7	103.3	110.5	121.7	130.4	137.9	144.4	150.7	156.3	161.4	164.9	168.7	171.6	174.8	176.4	179.2	181.5	183.0	135.5
B5	B4 + Soft-DTW Loss	11.0	23.3	48.2	59.0	77.3	91.7	103.6	111.1	122.2	130.8	138.5	144.7	150.8	156.2	161.4	164.5	168.0	170.7	173.6	175.2	178.1	180.5	181.8	135.2

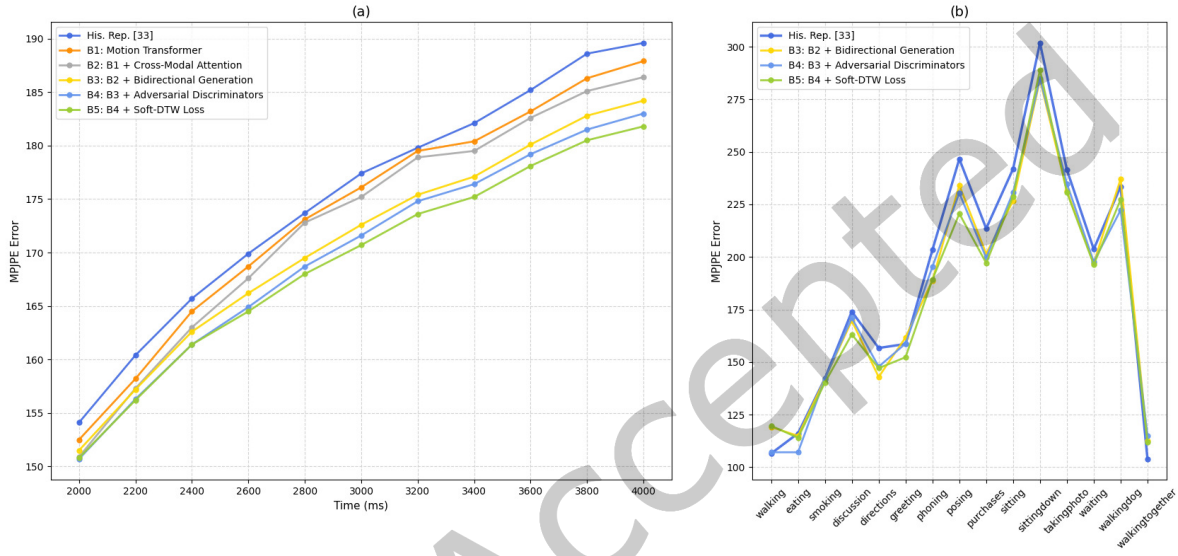

Fig. 5. (a) Comparison of average prediction error (MPJPE) over all action categories at different forecast times on the H3.6M dataset. (b) Comparison of prediction error for each action category at 4s on the H3.6M dataset.

Effect of Motion Transformer. As illustrated in Figure 5(a), by comparison B1 with His. Rep. [33], the proposed motion Transformer B1 improves performance over all the time stamps, especially for the long-term time stamp (*e.g.*, 3400ms, 3600ms, 3800ms, 4000ms). In Table 4, for 4s prediction, B1 reduces the error from 189.6 to 187.9, which shows the motion Transformer can better capture long-range dependency.

Effect of Cross-Modal Attention. In Figure 5(a), adopting the cross-modal attention to model the crossing relations between the outputs of Transformer decoder and DCT coefficients achieves a small gain over B1 all the time, which uncovers the benefits of utilizing these features effectively.

Effect of Bidirectional Generation. As shown in Figure 5(a), there is an obvious improvement from B2 to B3, highlighting the importance of the proposed bidirectional generation strategy for encoding motion patterns and human dynamics. Furthermore, as observed by the difference in performance between His. Rep. [33] and ours in Figure 6, B3 alleviates the freezing prediction problem for a long-term span, which produces poses that keep moving during the forecast period. Specifically, B3 forecasts the right hand moves forward, which is the same as the moving trend of ground truth. While many frames (in the black dashed box) are predicted by His. Rep. [33] illustrate that the predicted poses do not change in more than half of the forecast duration. Regarding the mitigation of our bidirectional structure to the average predicting problem, please refer to Section 4.2 for more comparison.

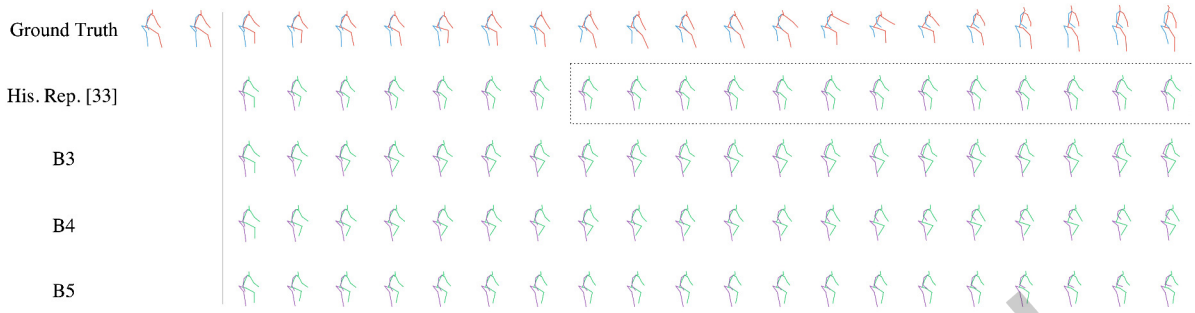

Fig. 6. Visualization of predicted poses of different methods on a sample of the H3.6M dataset for long-term prediction. From top to bottom sequences correspond to ground truth, His. Rep. [33], B3, B4, B5 (our BiTGAN) respectively. The first two frames are the latest observed frames, the others are predicted frames. The whole sequence is down-sampled to 5 frames per second. (Best viewed when zoomed in.)

Effect of Adversarial Discriminators. As illustrated Figure 5(a) and in Table 4, B4 steadily improves over B3 for long-term time stamps and achieves around 1.4 gain on the MPJPE metric at 4s. In visualization comparison (see Figure 6), B4 predicts both hands are moving, which is more closely related to the ground truth sequence in general. The boosted performance shows the advantage of adversarial discriminators (*i.e.*, frame-based discriminator and sequence-based discriminator), which leads to more realistic results by distinguishing the predicted results at frame and sequence levels.

Effect of the Soft-DTW Loss. In Table 4 and Figure 5(a), we can see that the overall performance is further facilitated by adding the Soft-DTW loss. Moreover, Figure 5(b) shows that B5 is better than B3 and B4, especially on some difficult motions *e.g.*, "posing" and "discussion". In Figure 6, B5 forecasts the movement of the right hand and legs over time, further narrowing the gap with the ground truth in the video level. These results are probably due to the fact that Soft-DTW loss enables B5 better capture the overall movement of human joints, resulting in more coherent and natural motion predictions. Note that B5 is our final model, which is significantly better than His. Rep. [33], validating the effectiveness of each component of our BiTGAN.

Effect of Loss Hyper-Parameters. We also investigate the influence of λ^a_{mpjpe}, λ^b_{mpjpe}, and λ_{inv} to the performance of our model. As shown in Table 4, we list nine different loss parameter settings (*i.e.*, L1-L9) and the corresponding results. Note that we adopt the baseline model B4 in Table 4 with the proposed motion Transformer, cross-modal attention and bidirectional generation for parameter selection since those three parts are more critical to the proposed BiTGAN. When $\lambda^a_{mpjpe} = 9.8$, $\lambda^b_{mpjpe} = 0.1$, $\lambda_{inv} = 0.1$, the prediction performance achieves the best.

Table 5. The influence of the loss hyper-parameter on Human3.6M.

	λ^a_{mpjpe}	λ^b_{mpjpe}	λ_{inv}	80	160	320	400	560	720	880	1000	1200	1400	1600	1800	2000	2200	2400	2600	2800	3000	3200	3400	3600	3800	4000	Ave.
L1	11	0.1	0.1	11.0	23.4	48.7	59.8	78.4	92.9	104.6	111.9	122.7	131.2	138.7	145.0	150.9	157.0	162.2	165.8	169.4	172.7	175.9	177.9	181.3	184.6	186.2	136.6
L2	9	0.1	0.1	11.0	23.4	48.6	59.7	78.5	93.0	105.2	112.8	124.0	133.2	140.7	146.8	152.6	158.1	163.4	167.0	170.7	174.1	177.6	179.2	181.7	184.4	186.0	137.6
L3	0.5	0.4	0.1	13.2	26.4	53.0	64.8	85.0	101.4	116.0	125.0	136.1	145.6	152.8	159.1	163.8	168.2	172.6	175.1	177.3	179.1	181.4	182.4	184.2	185.6	186.0	145.0
L4	5	4	1	13.3	26.8	53.4	65.3	85.3	101.2	115.0	123.9	134.8	144.6	151.3	157.8	162.6	167.1	170.7	172.9	175.2	177.2	179.2	180.5	182.7	184.4	185.3	143.8
L5	0.95	0.04	0.01	11.5	24.3	50.0	61.2	79.9	94.0	105.8	113.2	124.2	132.8	140.6	147.0	153.0	158.4	164.2	167.6	171.1	174.1	177.0	178.6	181.3	183.8	185.0	137.8
L6	9.5	0.4	0.1	11.5	24.2	49.8	60.9	79.6	94.0	105.7	113.1	124.5	133.3	140.9	147.3	153.1	158.7	163.8	166.9	170.0	172.5	174.9	176.9	180.0	182.8	184.5	137.3
L7	30	0.1	0.1	11.0	23.4	48.7	60.0	78.7	93.2	104.9	112.0	122.9	131.7	139.2	145.4	151.4	157.4	162.7	166.2	169.7	173.1	175.7	177.4	180.5	183.2	184.3	136.6
L8	7	2	1	12.3	25.3	51.1	62.9	82.8	98.9	113.0	122.4	134.1	144.7	152.3	159.3	164.1	168.6	172.4	174.9	176.8	178.4	180.1	180.6	181.9	183.6	184.2	143.7
L9	9.8	0.1	0.1	10.9	23.3	48.2	59.2	77.7	92.1	104.0	111.5	122.9	132.1	139.6	145.7	151.5	157.2	162.6	166.2	169.5	172.6	175.4	177.1	180.1	182.8	184.2	136.4

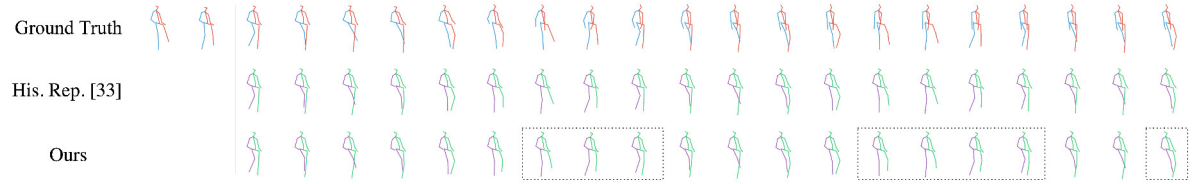

Fig. 7. Failure case (walking action) of our human motion prediction method for long-term (4s) prediction on the Human3.6M dataset. The first two frames are the latest observed frames, the others are predicted frames. The whole sequence is down-sampled to 5frames per second. (Best viewed when zoomed in.)

4.4 Limitations

Although our BiTGAN can produce dynamic poses for long-term prediction, it also has some limitations. Figure 7 illustrates a failure case of walking activity of our method. As we can see, for some frames, our prediction in leg joints is less accurate, *e.g.*, the knee bends a bit more, the steps are a bit smaller than the ground truth. The most probable reason is that, for such simple poses, our model will introduce more uncertainties, which will affect the accuracy of those actions with more certainties.

Table 6. The average MPJPE error of all actions on Human3.6M for short-term prediction. The best two results are highlighted in red and blue.

Milliseconds	Residual sup. [35]	ConvSeq2Seq [23]	LTD-50-25 [34]	LTD-10-25 [34]	LTD-10-10 [34]	His. Rep. [33]	Ours
80	25.0	16.6	12.2	12.4	11.2	10.4	11.0
160	46.2	33.3	25.4	25.2	23.4	22.6	23.3

Besides, our model has competitive performance for short-term motion prediction. We have provided the short-term prediction results in Table 6. We can observe that our approach achieves the second best results (highlighted in red) while the best is His. Rep [33]. The underlying reason might be that our method is designed for long-term prediction, which tends to produce more dynamic poses. Thus, compared with the static mean pose produced by His. Rep [33], our method leads to higher errors for short-time but lower errors for a long-range prediction.

5 CONCLUSIONS

In this paper, we propose a novel bidirectional Transformer GAN (BiTGAN) for long-term human motion prediction. Our novel bidirectional generation paradigm can effectively leverage the limited training samples as well as refrain from the freezing pose generation problem, especially for long-term prediction tasks. Besides, we split the history sequence into two parts, with the earlier part being fed to the encoder and the recent one being fed to the decoder. In this way, the Transformer generator can keep the distribution consistency between training and testing, thus alleviating the exposure problem and making the inference efficient. Moreover, we also introduce a soft-DTW loss and two discriminators to improve the capacity of maintaining the similarity between the predicted sequence and the real one implicitly and semantically. Our experimental results demonstrate the superiority of the proposed BiTGAN in predicting dynamic poses for both acyclic and periodic motions.

ACKNOWLEDGMENTS

This work is supported by the National Key Research & Development Program of China (No. 2018AAA0102902).

REFERENCES

[1] Martin Arjovsky and Léon Bottou. 2017. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862* (2017).

[2] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. 2017. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3722–3731.

[3] Judith Butepage, Michael J Black, Danica Kragic, and Hedvig Kjellstrom. 2017. Deep representation learning for human motion prediction and classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6158–6166.

[4] Yujun Cai, Lin Huang, Yiwei Wang, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Xu Yang, Yiheng Zhu, Xiaohui Shen, et al. 2020. Learning progressive joint propagation for human motion prediction. In *European Conference on Computer Vision*. Springer, 226–242.

[5] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. 2020. Long-term human motion prediction with scene context. In *European Conference on Computer Vision*. Springer, 387–404.

[6] Marco Cuturi and Mathieu Blondel. 2017. Soft-dtw: a differentiable loss function for time-series. In *International Conference on Machine Learning*. PMLR, 894–903.

[7] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860* (2019).

[8] Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. 2021. MSR-GCN: Multi-Scale Residual Graph Convolution Networks for Human Motion Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11467–11476.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[11] Yinglin Duan, Tianyang Shi, Zhengxia Zou, Yenan Lin, Zhehui Qian, Bohan Zhang, and Yi Yuan. 2021. Single-Shot Motion Completion with Transformer. *arXiv preprint arXiv:2103.00776* (2021).

[12] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. 2015. Recurrent network models for human dynamics. In *Proceedings of the IEEE International Conference on Computer Vision*. 4346–4354.

[13] Kaifeng Gao, Long Chen, Yifeng Huang, and Jun Xiao. 2021. Video relation detection via tracklet based visual transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*. 4833–4837.

[14] Anand Gopalakrishnan, Ankur Mali, Dan Kifer, Lee Giles, and Alexander G Ororbia. 2019. A neural temporal model for human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12116–12125.

[15] Liang-Yan Gui, Yu-Xiong Wang, Xiaodan Liang, and José MF Moura. 2018. Adversarial geometry-aware human motion prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 786–803.

[16] James N Ingram, Konrad P Körding, Ian S Howard, and Daniel M Wolpert. 2008. The statistics of natural hand movements. *Experimental brain research* 188, 2 (2008), 223–236.

[17] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2013. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence* 36, 7 (2013), 1325–1339.

[18] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. 2016. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the ieee conference on computer vision and pattern recognition*. 5308–5317.

[19] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. 2021. Transformers in vision: A survey. *ACM Computing Surveys (CSUR)* (2021).

[20] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[21] Hema S Koppula and Ashutosh Saxena. 2015. Anticipating human activities using object affordances for reactive robotic response. *IEEE transactions on pattern analysis and machine intelligence* 38, 1 (2015), 14–29.

[22] Jogendra Nath Kundu, Maharshi Gor, and R Venkatesh Babu. 2019. Bihmp-gan: Bidirectional 3d human motion prediction gan. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 8553–8560.

[23] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. 2018. Convolutional sequence to sequence model for human dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5226–5234.

[24] Jiachen Li, Fan Yang, Hengbo Ma, Srikanth Malla, Masayoshi Tomizuka, and Chiho Choi. 2021. Rain: Reinforced hybrid attention inference network for motion forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 16096–16106.

[25] Jiaman Li, Yihang Yin, Hang Chu, Yi Zhou, Tingwu Wang, Sanja Fidler, and Hao Li. 2020. Learning to Generate Diverse Dance Motions with Transformer. *arXiv preprint arXiv:2008.08171* (2020).

[26] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. 2020. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

214–223.
[27] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. 2021. Learn to Dance with AIST++: Music Conditioned 3D Dance Generation. *arXiv preprint arXiv:2101.08779* (2021).
[28] Hongyi Liu and Lihui Wang. 2017. Human motion prediction for human-robot collaboration. *Journal of Manufacturing Systems* 44 (2017), 287–294.
[29] Yicheng Liu, Jinghuai Zhang, Liangji Fang, Qinhong Jiang, and Bolei Zhou. 2021. Multimodal motion prediction with stacked transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7577–7586.
[30] Zhenguang Liu, Shuang Wu, Shuyuan Jin, Qi Liu, Shijian Lu, Roger Zimmermann, and Li Cheng. 2019. Towards natural and accurate future motion prediction of humans and animals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10004–10012.
[31] Kedi Lyu, Zhenguang Liu, Shuang Wu, Haipeng Chen, Xuhong Zhang, and Yuyu Yin. 2021. Learning Human Motion Prediction via Stochastic Differential Equations. In *Proceedings of the 29th ACM International Conference on Multimedia*. 4976–4984.
[32] Xin Man, Deqiang Ouyang, Xiangpeng Li, Jingkuan Song, and Jie Shao. 2022. Scenario-Aware Recurrent Transformer for Goal-Directed Video Captioning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 18, 4 (2022), 1–17.
[33] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. 2020. History repeats itself: Human motion prediction via motion attention. In *European Conference on Computer Vision*. Springer, 474–489.
[34] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. 2019. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9489–9497.
[35] Julieta Martinez, Michael J Black, and Javier Romero. 2017. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2891–2900.
[36] Nicola Messina, Giuseppe Amato, Andrea Esuli, Fabrizio Falchi, Claudio Gennaro, and Stéphane Marchand-Maillet. 2021. Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 17, 4 (2021), 1–23.
[37] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732* (2015).
[38] Xuanchi Ren, Haoran Li, Zijian Huang, and Qifeng Chen. 2020. Self-supervised dance video synthesis conditioned on music. In *Proceedings of the 28th ACM International Conference on Multimedia*. 46–54.
[39] Theodoros Sofianos, Alessio Sampieri, Luca Franco, and Fabio Galasso. 2021. Space-time-separable graph convolutional network for pose forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11209–11218.
[40] Pengxiang Su, Zhenguang Liu, Shuang Wu, Lei Zhu, Yifang Yin, and Xuanjing Shen. 2021. Motion Prediction via Joint Dependency Modeling in Phase Space. In *Proceedings of the 29th ACM International Conference on Multimedia*. 713–721.
[41] Hao Tang, Song Bai, Li Zhang, Philip HS Torr, and Nicu Sebe. 2020. Xinggan for person image generation. In *European Conference on Computer Vision*. Springer, 717–734.
[42] Hao Tang, Dan Xu, Wei Wang, Yan Yan, and Nicu Sebe. 2018. Dual generator generative adversarial networks for multi-domain image-to-image translation. In *Asian Conference on Computer Vision*. Springer, 3–21.
[43] Yongyi Tang, Lin Ma, Wei Liu, and Weishi Zheng. 2018. Long-term human motion prediction by modeling motion context and enhancing motion dynamic. *arXiv preprint arXiv:1805.02513* (2018).
[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762* (2017).
[45] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. 2018. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 601–617.
[46] Borui Wang, Ehsan Adeli, Hsu-kuang Chiu, De-An Huang, and Juan Carlos Niebles. 2019. Imitation learning for human pose prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7124–7133.
[47] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237* (2019).
[48] Bo Zhang, Rui Zhang, Niccolo Bisagno, Nicola Conci, Francesco GB De Natale, and Hongbo Liu. 2021. Where Are They Going? Predicting Human Behaviors in Crowded Scenes. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 17, 4 (2021), 1–19.
[49] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2223–2232.