Neurocomputing 511 (2022) 34–42

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

# Spatial–temporal modeling for prediction of stylized human motion

Chongyang Zhong [a,b], Lei Hu [a,b], Shihong Xia [a,b,*]

[a] Institute of Computing Technology, Chinese Academy of Sciences, China
[b] University of Chinese Academy of Sciences, China

ARTICLE INFO

*Article history:*
Available online 29 August 2022

*Keywords:*
stylized motion
transformer
human motion prediction
spatial–temporal modeling
constant variance GMM

ABSTRACT

Human motion prediction refers to forecasting human motion in the future given a past motion sequence, which has significant applications in human tracking, automatic motion generation, autonomous driving, human-robotics interaction, etc. Previous works usually used RNN-based methods, focusing on modeling the temporal dynamics of human motion, which have made great effort on content motions. However, it is unclear for their performance on stylized motion, which is with more expressive emotions and states of the human motion. Different styles within the same motion type have similar motion patterns but also subtle variances. This makes it difficult to be predicted. The main idea of this paper is to learn the spatial characteristic of stylized motion and combine it with the temporal dynamics to achieve accurate prediction. We adopt a transformer-based style encoder to learn the motion representation in the pose space and then maps it to the latent space modeled by the constant variance Gaussian mixture model; meanwhile, we use the hierarchical multi-scale RNN as a temporal encoder to capture the temporal dynamics of human motion; finally, we feed the spatial and temporal features into the prediction decoder to predict the next frame. Our experiments on the Human 3.6 M and Stylized MotionDatasets demonstrate that our model has comparable prediction performance with the state-of-the-art motion prediction works on Human 3.6 M and outperforms previous works on stylized human motion prediction.



## 1. Introduction

Human motion prediction helps us know the human movement trend in the future and make responses to it accordingly, which has very significant applications in automatic motion generation, autonomous driving, human-robotics interaction, target tracking, and motion planning, etc.

Motion prediction works use traditional statistical modeling methods at the very beginning to make short predictions of simple motion types, such as hidden Markov model [3] and Gaussian process hidden variable model [25]. The deep neural networks have activated a lot of works in the field of motion prediction. Among these works, the Recurrent Neural Network(RNN)-based methods dominate over other methods due to its well-known characteristics of processing temporal series data. The first RNN-based motion prediction work is ERD (Encoder-Recurrent-Decoder)[6], which achieves relatively accurate motion prediction in the short-term. Unfortunately, it is unable to predict long-term human motion because of the accumulation of errors. Besides, it can't handle multi-type motions either. Following their research, other works give some new solutions to increase the time range of motion prediction and improve prediction accuracy. Furthermore, they achieve a breakthrough from single-action prediction to multi-action prediction [13,18,29,9,4]. These motion prediction works focus on learning the temporal dynamics of the motion sequence, which makes them effective for locomotion and multi-type motion with large motion type variances.

With the increasing application requirements in human-robotics interaction, VR/AR, and games, human motions are required to be more expressive to reflect people's emotions and states. Factually, when people are moving, they have some semantic information such as emotions (anger, depressed) and states (childlike, old) instead of swinging hands and feet mechanically, which is called motion style [27,26,21]. Unlike different motion types, motion styles are mostly determined by the subtle variances under the same motion type. And we observe that traditional motion prediction works mentioned above cannot be easily converted into stylized motion prediction due to the lack of detailed spatial modeling. We think there are two main challenges in stylized motion prediction: 1. Style feature is semantic information, which is difficult to extract. And the data distributions of different stylized motions are both overlapped and different (we will discuss

* Corresponding author at: Institute of Computing Technology, Chinese Academy of Sciences, China.
*E-mail addresses:* zhongchongyang@ict.ac.cn (C. Zhong), hulei19z@ict.ac.cn (L. Hu), xsh@ict.ac.cn (S. Xia).

https://doi.org/10.1016/j.neucom.2022.08.075
0925-2312/

this in 5.2.1), leading to ambiguities in prediction; 2. Stylized motions are structural time series whose temporal and spatial characteristics are highly coupled and need to be modeled simultaneously.

In this paper, we propose an auto-regressive network to deal with stylized human motion prediction. Considering the characteristics of stylized motion, we simultaneously model the spatial–temporal characteristics to learn effective low-dimensional representations. For spatial modeling, we first adopt transformer as the style encoder to learn a latent spatial representation. Transformer has been proven to be capable of learning the correlation between different elements of the input vector, which is particularly suitable to learn the interdependence of the relevant joints. Secondly, we adopt the Gaussian mixture model to learn the data distribution of stylized motion in latent space. For the accuracy of prediction, we use a variant called Constant-Variance VAE, which is proven to be able to increase the stability of VAE network by fixing the variances of the Gaussian model as a constant. For temporal modeling, we draw on the idea of hierarchical multi-scale RNN to capture the complex temporal dynamics of human motion. Finally, we concatenate the features from spatial modeling and temporal modeling as the input of a prediction decoder to predict the pose of the next frame. And the prediction results obtained in this time step become the input for the next step to achieve auto-regressive motion prediction. Since the motion style is a kind of semantic information and is difficult to be expressed by a single frame, we intuitively use N sequential frames as the input of the network. In addition, we feed the spatial features into an additional reconstruction decoder to enhance the learning ability of the style encoder.

The main contributions of our work can be summarized as follows:

- 1.We propose a transformer-based style encoder to extract the style features of stylized motion, which is combined with a latent space built by constant-variance Gaussian mixture model to model the subtle differences of different motion styles.
- 2.We propose an auto-regressive network structure to simultaneously model the spatial and temporal characteristics of stylized motion, which first achieved accurate predictions of stylized motion.
- 3.We carry out extensive experiments on Human 3.6 M [12] and Stylized Motion Datasets [27] both quantitatively and qualitatively to demonstrate that the prediction performance of our method is comparable with the state-of-the-art works on Human 3.6 M and outperforms previous works on Stylized Motion Datasets.

## 2. Related Works

Motion prediction is usually divided into deterministic prediction and probabilistic prediction. When we aim to accurately forecast the trend of human motion in the future, deterministic prediction is commonly used to get more accurate prediction results, where RNN-based network structures are frequently used. When users ask for as many plausible future motions as possible, probabilistic prediction is a better choice to obtain rich and reasonable prediction results, which usually uses the network structure based on VAE or GAN. We will briefly review related works from the following aspects.

### 2.1. Deterministic prediction

The purpose of deterministic prediction is to accurately predict the human sequences with a discriminant model. As we all know, RNN is superior to other networks in processing temporal series data [22,15], which makes it become the mainstream method of deterministic prediction. RNN is firstly used in human motion modeling as a component of the classic structure named ERD (Encoder-Recurrent-Decoder) in [6], which makes short-term (560 ms) motion prediction for single-action. For long-term motion prediction, SRNN (Structural RNN) [13] uses graph models to encode the structural characteristics of human motion and successfully extends the prediction time range to 1000 ms. However, RNN-based networks have two obvious drawbacks: 1. There is an obvious jump between the last frame of the input motion and the first frame of prediction, which is called first frame discontinuity; 2. The prediction errors will accumulate over time, especially on the testing set.

The first frame discontinuity problem is solved to some extent in [18] by directly connecting the input to the output of the network, which is known as residual network. In this way, they model the transition of velocity instead of poses to alleviate the discontinuity, which is convenient and effective. On the other hand, a series of works have been proposed to solve the problem of error accumulation. Some works aim to force the network to directly face the unseen input data with errors during the training process, which is significant to enhance the robustness of the network. Auto-conditioned RNN [29] achieves this by feeding the output of the network itself and Ground Truth alternately into the network at every certain number of time steps. Furthermore, sampling-based loss [18] uses completely the output of the network itself instead of Ground truth as the input of the next frame to enhance the error processing ability. Another effective way is to modify the architecture by adding an auto-encoder network with dropout operation between RNN cells to avoid network overfitting [9]. Besides this, capturing the temporal dynamics of different time scales is proved to be helpful to model the temporal dependence of human motion in [4]. They built a hierarchical multi-scale RNN to learn the motion dynamics of different time intervals and mixed them to get more accurate prediction results.

The above RNN-based methods focus on capturing temporal dynamics of human motion while ignoring the subtle modeling of pose space. Since stylized motions have similar motion patterns, which make the variances more subtle to be modeled, these methods will produce ambiguity when predicting stylized motion. Therefore, our model not only uses RNN to capture the temporal dynamics but also uses a transformer-based encoder to extract the style features. The combination of spatial and temporal characteristics eliminates the ambiguity in the prediction and enables better modeling of stylized motion.

### 2.2. Probabilistic prediction

Probabilistic prediction usually relies on generative network structures, among which VAE [14] and GAN [10] are commonly used. This type of method builds a probability model on the existing motion data to predict a variety of results through random sampling. Using the structure of conditional-VAE [20], Pose-VAE [24] extracts motion features from videos to generate new motions. Due to the simple assumption of its latent space distribution, it is not capable to generate a rich variety of motions. To solve this problem, a neural network Q is used to learn K mapping functions and then maps the latent space to K different subspaces [28]. Through sampling from the different subspaces, they obtain diverse human motions. With the great success of Generative Adversarial Network (GAN) in the computer vision community, GAN is introduced to human motion prediction to optimize the quality of the prediction. HP-GAN [1] adds a random variable with a standard normal distribution into RNN to generate various motion prediction results. Based on this work, BiHMP-GAN [16] gives content information to the random variables, which not only

enables random motion prediction but also makes a deterministic prediction to a certain extent by controlling random variables. [21] propose to use meta-learning to generate diverse stylized human motion. They mainly focus on motion generation and do not carry on in-depth research on stylized motion prediction.

Probabilistic prediction methods use probability distribution to model latent space, which can partly model the data distribution of stylized motion. However, such a scheme increases the diversity of prediction, leading to the reduction of accuracy. To achieve more accurate stylized motion prediction, we combine the idea of Constant-Variance VAE [7] and Gaussian mixture model to model the subtle variances of different motion styles, which strikes a balance between diversity and accuracy.

## 3. Problem Formulation

Let us first formulate the problem to be solved. Traditional motion prediction, that is, given a past motion sequence $X_{1:n} = \{x_1, x_2, \ldots, x_n\}$, where $x_i$ represents a single pose of motion sequences, to predict a motion sequence of future t time steps $X_{n+1:n+t}$. This problem can be modeled with a conditional probability:

$$p(x_{n+1}, x_{n+2}, \ldots, x_{n+t} | x_1, x_2, \ldots, x_n) \tag{1}$$

The difference between stylized motion and ordinary motion is that stylized motion contains both inherent content information and style information. Style information expresses people's emotions and states, such as happiness, angry, depressed, old, sexy, etc. It is a more refined modeling of human motion that could make human motion more rich and expressive. Similar to motion recognition, motion style is semantic information based on the whole motion sequence and thus is difficult to be described by a single frame. So we define stylized motion as $x_{1:i} = \{c_{1:i}, S\}$, where S is the style information of the entire motion sequence, and $c_{1:i}$ is the content information of each frame, which is specifically expressed as:

$$c_i = \{p_x, p_y, p_z, v_x, v_y, v_z, \theta_1, \theta_2, \ldots, \theta_m\} \tag{2}$$

where $p_x, p_y, p_z$ represent the $x, y, z$ global joint position of the root joint respectively, and $v_x, v_y, v_z$ represent the $x, y, z$ global joint velocity of the root joint respectively, $\theta_i$ represents the root joint angle and local joint angles of other joints represented by an exponential map, and the total dimension is 81. We believe that the position and velocity information of the root joint contains the temporal dynamics of the motion, while the local joint angles of each joint contain the spatial characteristics of the pose. To obtain more accurate prediction results, we use N motion frames to predict the N + 1th frame instead of directly predicting the motion sequence of future t time steps, and then iteratively achieve sequence prediction. Therefore, we formalize the stylized motion prediction problem as the following formula:

$$p(c_{N+1} | c_1, c_2, \ldots, c_N, S) \tag{3}$$

## 4. Our Method

### 4.1. Overview

The overview of our network structure is shown in Fig. 1. Firstly, we use N frames of motion frames as the condition to model the stylized motion in the spatial dimension and feed the local joint angles into the spatial transformer to extract the features in the pose space. Secondly, the dimensionality-reduced *stylefeature* is obtained by using the constant variance Gaussian mixture model to model the latent space of stylized motion. At the same time, we feed the position and velocity of the root joint in the condition frames into hierarchical multi-scale LSTM to capture the temporal dynamics of human motion, which is named *temporalfeature*. Finally, we combine style and temporal features and feed them into the prediction decoder to get the prediction result of the N + 1th frame. In addition, we also feed the *stylefeature* into the reconstruction decoder to reconstruct the motion of the Nth frame to strengthen the feature extraction ability of the style encoder. And a residual connection is added between input and output. We will introduce our research methods in detail in the next chapter.

### 4.2. Solutions

#### 4.2.1. Spatial modeling of stylized motion

Transformer [23] has achieved fruitful results in the field of Natural Language Processing, where the self-attention mechanism can model the interconnection of each word in a sentence, which is very similar to the relationship between each joint of the human body. Therefore, we use the transformer as a style encoder to extract style features in pose space.

We first map the local joint angles of $c_i$ to a variable in a D-dimensional space $E_i = \{e_1, e_2, \ldots, e_m\}$ with a fully connected layer, and then use 3 different weight matrices $W_Q, W_K, W_V$ to compute $Q, K, V$ respectively:

$$\begin{aligned} Q &= E_i W_Q \\ K &= E_i W_K \\ V &= E_i W_V \end{aligned} \tag{4}$$

Next, we use $H$ different linear transformation matrices to project $Q, K, V$ to different subspaces respectively, and use the projected values to calculate multi-head attentions:

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{5}$$

where:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{D}}\right)V \tag{6}$$

Finally, we concatenate multi-head attentions and use $W^O$ to compute the *spatialfeature*:

$$spatialfeature = Concat(head_1, head_2, \ldots, head_H)W^O \tag{7}$$

The single motion frame is not enough to describe the style of the motion sequence, hence the features extracted from a single motion frame have limited expressive ability. As a consequence, we feed continuous N motion frames into the style encoder, do the above operations for each frame, and then concatenate the features of the N frames together to get *spatialfeature*.

After getting the *spatialfeature*, we map it to a low-dimensional latent space $p(S)$ through a fully connected layer. The vanilla VAE [14] uses the standard normal distribution to model latent space, which makes it difficult to model stylized motion with both crossover and discrepant data distribution. As a result of this, we consider using the Gaussian mixture model instead. However, during the experiment, we find that the effect of motion prediction is not satisfactory when using the Gaussian mixture model. Since we want to achieve accurate prediction, and the variances caused by the randomness of the Gaussian mixture model will be harmful to the accuracy, it is better to find a model less stochastic. That's why we learn from the idea of Constant-Variance VAE [7] and fix the variance of each Gaussian model in the Gaussian mixture model to a constant $\sigma$ [8]. The style encoder only needs to learn the mean $\mu_i$ and the weight $\omega_i$ of the M Gaussian models in the

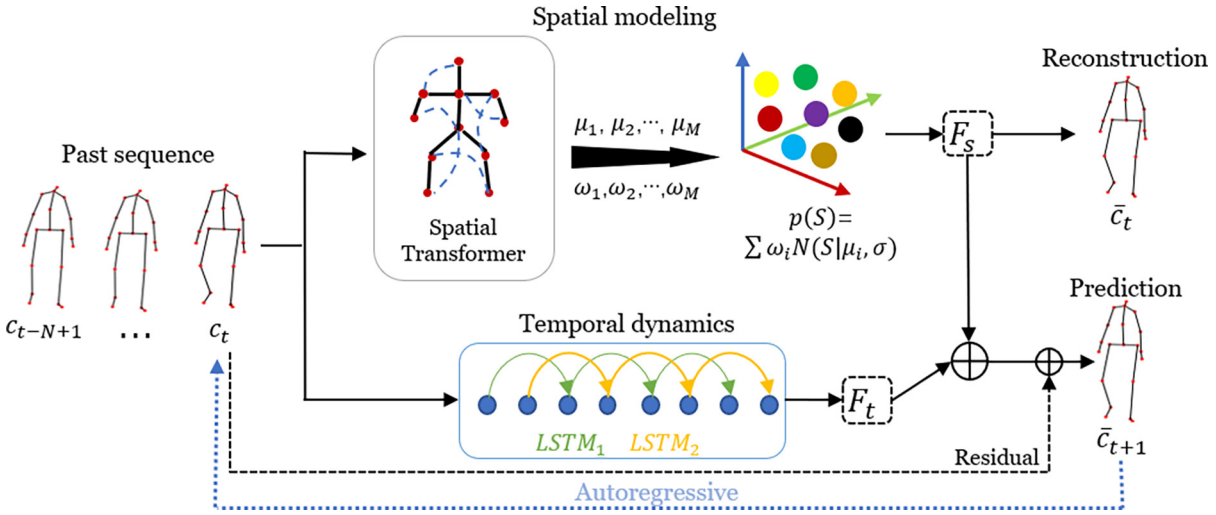


**Fig. 1.** The overview of our stylized motion prediction network.

Gaussian mixture model. The latent space can be modeled using the following formula:

$$p(S) = \sum_i \omega_i N(S|\mu_i, \sigma) \tag{8}$$

We additionally added a reconstruction decoder composed of a 3-layer fully connected layer (288–512-81) and feed the *stylefeature* $F_s$ sampled from the latent space into the reconstruction decoder to enhance the performance of *stylefeature*.

#### 4.2.2. Capture temporal dynamics

RNN-related networks are suitable for capturing the temporal dynamics of human motion because of their well-known ability to handle temporal series data. However, most of the current works feed motion pose into the network frame by frame, which is not capable to model the temporal dynamics well because the transition between adjacent frames is extremely small. Therefore, we use 2 LSTM networks to form a multi-scale RNN network [5] and feed the position and velocity of the root joint of the $m, m+2, m+4, \ldots$frame into $LSTM_1$ and the $m+1, m+3, m+5, \ldots$th frame to $LSTM_2$. Then combine the outputs of the two LSTM networks and feed them into a fully connected layer to obtain the multi-scale *temporalfeature*, which is denoted as $F_t$. The purpose of this operation is to extract better temporal features by modeling the dynamics of human motion at different time scales. After getting $F_t$ and $F_s$, we concatenate them together and feed them into the prediction decoder consisting of 3 fully connected layers (256–512-81), and output the prediction result of the $N+1$th frame.

#### 4.2.3. Training

In our network structure, the RNN cell uses LSTM with a hidden state dimension of 1024. The dimensions of $F_s, F_t, D, Q, K$ and $V$ are **64**, **32**, **32**, **32**, **32**, **32**, respectively. $H, M, N$, and $\sigma$ are **8**, **8**, **5**, and **0.5** respectively. During training, we input the $1:N$ frames, finally output the prediction result of the $N+1$th frame and the reconstruction result of the $N$th frame. During the test, we abandon the reconstruction encoder and only output the prediction results of the $N+1$th frame, and then splice it after the $N$th frame. On the next turn, the $2:N+1$ frames are used as the input to realize auto-regressive prediction. The loss functions are as follows:

$$L = \omega_1 L_{KL} + \omega_2 L_{pre} + \omega_3 L_{recon} + \omega_4 L_{smooth} \tag{9}$$

where

$$\begin{aligned} L_{KL} &= \|\sum_i \mu_i \omega_i\|^2 \\ L_{recon} &= \|X_N - \widetilde{X}_N\|^2 \\ L_{pre} &= \|X_{N+1} - \widetilde{X}_{N+1}\|^2 \\ L_{smooth} &= \|(X_{N+1} - X_N) - \left(\widetilde{X}_{N+1} - \widetilde{X}_N\right)\|^2 \end{aligned} \tag{10}$$

where $L_{KL}$ is the KL divergence. Because we set the variance of the latent space to a constant, there is no variance term in the KL divergence. $L_{pre}$ and $L_{smooth}$ are the loss of the prediction encoder, which are to ensure the accuracy of the prediction result of the next frame and the predicted motion as smooth as possible respectively. $L_{recon}$ is the loss of the reconstruction encoder to ensure the accuracy of the reconstruction result. The $\omega_1, \omega_2, \omega_3, \omega_4$ used in the experiment are 0.1, 1, 0.2, and 0.05 respectively. Between the input and output of the network, we use the residual structure proposed in [18]. At the same time, we also adopt the training strategy named scheduled sampling [2] to enhance the generalization ability of the network.

## 5. Experiments

### 5.1. Datasets

The datasets used in our experiments include Human 3.6 M [12] and Stylized Motion Datasets [27]. We will introduce these two datasets as follows:

#### 5.1.1. Human 3.6 M

Human 3.6 M is a large human motion dataset that is mostly used in the field of motion prediction. They release the data of 15 motion categories from 7 subjects, including walking, running, smoking, discussion, etc. The frame rate is 50 Hz, and each character has 32 joint points. We use the 3D angle data to make the prediction, including the position of the root joint and the joint angles of all joints.

#### 5.1.2. Stylized Motion Datasets

Stylized Motion Datasets is a stylized motion dataset created by Xia [27] including 8 motion styles: angry, old, depressed, sexy, childlike, strutting, proud, neutral, and 5 motion categories: walking, running, jumping, punching, kicking. The frame rate is 120 Hz, and each character includes 25 joint points. We used the walking sequences of the dataset.

### 5.2. Results

#### 5.2.1. Stylized Motion Datasets analysis

We use the NPSS proposed in [11] to analyze the difference between Stylized Motion Datasets and Human 3.6 M. NPSS can measure the similarity of two sets of motion sequences in the data distribution degree. We randomly selected two sets of motion from Human 3.6 M: the single type of motion and multiple types of motion, each of which containing 100 sequences with a length of 60 frames. Meanwhile, from the walking sequences of the Stylized Motion Datasets, 100 sequences with the same length are also randomly selected. We divide the obtained 3 sets of movements into two groups randomly and calculate their NPSS. After repeating the above steps 10 times, we get the means and variances shown in Table 1. The smaller the mean, the more similar the extracted motion patterns are, and the variance means how different the distribution of the selected sequences are. The results in Table 1 show that the single type of motions in Human 3.6 M has similar motion patterns and data distributions because of its low mean and variance. On the contrary, multiple motion types have obvious differences in motion patterns and data distributions. The different styles of the Stylized Motion Datasets are similar in terms of motion patterns, but the data distribution is more complicated than the single type motion in Human 3.6 M. This shows that our analysis and understanding of the Stylized Motion Datasets is reasonable, and it also explains to some extent that why the existing works cannot achieve good performance on Stylized Motion Datasets.

#### 5.2.2. Baseline and implementation details

We compare the previous RNN-based works including ERD [6], SRNN [13], Seq2Seq [18], TP-RNN [4] on Human 3.6 M and Stylized Motion Datasets, as well as the zero-velocity baseline proposed in [18]. For the fairness of the experiment, we convert the pose of Stylized Motion Datasets to the skeleton of the Human 3.6 M through motion retargeting [19] and also downsampling the frequency of the two data sets to 25 Hz. At the same time, we adopt the code that these works publicly release on the Internet. For Human 3.6 M, we use their pre-trained model. For Stylized Motion Datasets, we use the experimental settings mentioned in their papers. We use Nvidia 2080Ti to train our network for a total 150 epochs. The learning rate is initialized to 0.005, and the decay rate for each epoch is 0.98. The P of scheduled sampling is initialized to 1, and each epoch is reduced by 0.025. The batch size is set to 32, and the optimizer is Adam. According to the experimental settings of previous works, our past sequence length is set to 10 frames, and the predicted length is 40 frames.

#### 5.2.3. Evaluations

(1) **Comparison on Human 3.6 M:** Following previous works, we make long-term and short-term predictions respectively on Human 3.6 M. In the experiment, we use the data of subject 5 for testing and the others for training, randomly extract 8 past sequences for prediction, and then calculate the Euler distance between the prediction results and Ground Truth as the prediction error. Among them, TP-RNN and our method are action-agnostic, and the others are for specific types. The experimental results are shown in Table 2 and Table 3.
Experimental results show that our work is comparable to the current state-of-the-art method TP-RNN in terms of short-term and long-term prediction. From 80 ms to 1000 ms, our prediction results are close to or better than TP-RNN, which shows that our prediction method is also effective on classic datasets. It is worth noting that some of our prediction results are not as good as TP-RNN on Human 3.6 M. According to our observation, this is reasonable because our model mainly solves spatial modeling and the RNN structure used in temporal modeling is much simpler than TP-RNN, in the meanwhile, the sequences of the same motion type in Human 3.6 M have similar patterns and relatively less spatial variances, which makes our style encoder cannot fully play its role. However, with the help of our style encoder, a simple RNN network structure can also achieve comparable prediction performance to the state-of-the-art works on classic datasets.
(2) **Comparison on stylized walking:** For Stylized Motion Datasets, we make short and long term predictions on the Walking motion of all 7 styles, including angry, old, sexy, childlike, strutting, proud, and depressed. We divide the data into training set, validation set, and test at 5:1:1, and the rest of the experimental settings are consistent with the settings of Human 3.6 M. Because our method and TP-RNN are both action-agnostic and have better performances than other methods on Human 3.6 M, we only compare our method with TP-RNN here. For the ablation study, we also compare the prediction results after

**Table 1**
The NPSS measurement of Human 3.6 M and Style Walking. ST means single type and MT means multiple types.

| | $\mu$ | $\sigma$ |
|---|---|---|
| H36M(ST) | 0.9090 | 0.0439 |
| H36M(MT) | 2.3146 | 0.1885 |
| Style Walking | 1.1957 | 0.1014 |

**Table 3**
MAE Comparison of the long-term prediction of the 4 main motion types on Human 3.6 M. The best results are shown in bold.

| | Walking | | Eating | | Smoking | | Discussion | |
|---|---|---|---|---|---|---|---|---|
| milliseconds | 560 | 1000 | 560 | 1000 | 560 | 1000 | 560 | 1000 |
| ERD | 2.04 | 2.41 | 2.35 | 2.44 | 3.71 | 3.80 | 2.88 | 2.92 |
| SRNN | 1.88 | 2.13 | 2.28 | 2.55 | 3.30 | 3.25 | 2.40 | 2.45 |
| Seq2seq | 0.88 | 0.95 | 0.96 | 1.35 | 1.24 | 1.85 | 1.42 | 1.78 |
| Zero-velocity | 1.35 | 1.32 | 1.04 | 1.38 | 1.02 | 1.69 | 1.41 | 1.96 |
| TP-RNN | **0.75** | **0.77** | **0.85** | 1.16 | 1.01 | **1.66** | **1.39** | **1.78** |
| Ours | 0.78 | 0.82 | 0.88 | **1.15** | **0.99** | 1.71 | 1.42 | 1.82 |

**Table 2**
MAE Comparison of the short-term prediction of the 4 main motion types on Human 3.6 M. The best results are shown in bold.

| | Walking | | | | Eating | | | | Smoking | | | | Discussion | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| milliseconds | 80 | 160 | 320 | 400 | 80 | 160 | 320 | 400 | 80 | 160 | 320 | 400 | 80 | 160 | 320 | 400 |
| ERD | 0.93 | 1.18 | 1.59 | 1.78 | 1.27 | 1.45 | 1.66 | 1.80 | 1.66 | 1.95 | 2.35 | 2.42 | 2.27 | 2.47 | 2.68 | 2.76 |
| SRNN | 0.81 | 0.94 | 1.16 | 1.30 | 0.97 | 1.14 | 1.35 | 1.46 | 1.45 | 1.68 | 1.94 | 2.08 | 1.22 | 1.49 | 1.83 | 1.93 |
| Seq2seq | 0.28 | 0.49 | 0.72 | 0.81 | 0.23 | 0.39 | 0.62 | 0.76 | 0.33 | 0.61 | 1.05 | 1.15 | 0.31 | 0.68 | 1.01 | 1.09 |
| Zero-velocity | 0.39 | 0.68 | 0.99 | 1.15 | 0.27 | 0.48 | 0.73 | 0.86 | 0.26 | 0.48 | 0.97 | 0.95 | 0.31 | 0.67 | **0.94** | 1.04 |
| TP-RNN | **0.25** | **0.41** | **0.58** | **0.65** | **0.20** | 0.33 | **0.53** | **0.67** | **0.26** | 0.47 | **0.88** | **0.90** | **0.30** | 0.66 | 0.96 | **1.04** |
| Ours | 0.28 | **0.41** | 0.61 | 0.66 | 0.22 | **0.32** | 0.57 | 0.71 | 0.28 | **0.46** | 0.91 | 0.93 | 0.33 | **0.64** | 1.01 | 1.08 |

removing the style encoder. The experimental results are shown in Table 4.

From the results in Table 4, we can discover that our method performs better than TP-RNN on Stylized Motion Datasets in almost all time steps, which shows the effectiveness of our method for stylized motion prediction. There are three experimental results worth noting:

(a) In the 160–560 ms interval, our method performs better than TP-RNN in most styles while the performance of TP-RNN in the same time period is superior to ours on Human3.6 M. The reason for this phenomenon is that TP-RNN is a deterministic prediction method that focuses on capturing the temporal dynamics. Resulting in their better prediction effects on datasets containing single motion type and multiple motion types with large variances like Human 3.6 M. On the other hand, our method works on the subtle modeling of the spatial structure to achieve better prediction of stylized motion with both overlapping and different data distribution.

(b) The prediction result of our method at the 80 ms is worse than TP-RNN in most styles. It is reasonable about this because that TP-RNN uses the velocity as motion representation of a single frame, which can effectively enhance the smoothness of temporal modeling, making their initial prediction effect better than our method. However, the velocity mainly expresses the trend of motion over time, which is not as helpful as position and angle for spatial characteristics modeling. We try to add joint velocity into our motion representation during our experiment, and it turns out unsatisfactory.

(c) Another time step worth paying attention to is 1000 ms, where our method has a much better prediction effect than TP-RNN. These results prove that our method can effectively avoid ambiguity in the prediction, which leads to better performances in long-term prediction. This is also the reason why we focus on modeling spatial characteristics using the style encoder.

(3) **Qualitative evaluation**: We visualize some of the prediction results of Stylized Motion Datasets for qualitative evaluation. As with quantitative evaluation, we compared TP-RNN, our method, and the result of our method after removing the style encoder. For a more intuitive perspective on the difference between the predictions and Ground Truth, we draw them in the same coordinate system. Corresponding to the quantitative evaluation, we show the results of 80 ms, 160 ms, 320 ms, 400 ms, 560 ms, and 1000 ms. From Fig. 3, we can see that the prediction results of TP-RNN and ours (w/o SE) become worse after 560 ms, and the motion predicted by our method is the most consistent with Ground Truth, which fully demonstrates the effectiveness of style encoder + Constant variance GMM.

(4) **Ablation study**: At first, we would like to discuss the choice of N in the experiment. Recall that motion style is a kind of semantic information expressed by sequences instead of a single frame. Therefore, we use N continuous frames as the input of our network. We try various options of N(from 2 to 6) in the experiment to find out which is the most effective. For comparison, we compute the mean prediction error of all 7 motion styles with the same setting in 5.2.3. The results are shown in Fig. 2. According to Fig. 2, when N is relatively small, the input sequence has insufficient style information for extracting the style feature. With N gradually increasing, the style information contained in the input sequence is gradually enriched, thus the prediction performance becomes better. However, too many conditional frames increase the chances of overfitting, which will reduce the prediction effect of the model on the testing set. After experimental verification, the prediction performance is best when N = 5, so we finally chose this setting.

To verify the effect of the style component we proposed, we do an ablation study both quantitatively and qualitatively. First of all, we remove the whole style encoder to find out whether the accuracy of the prediction is affected. The results are shown in the fourth row of Table 4. After removing the style encoder, the prediction

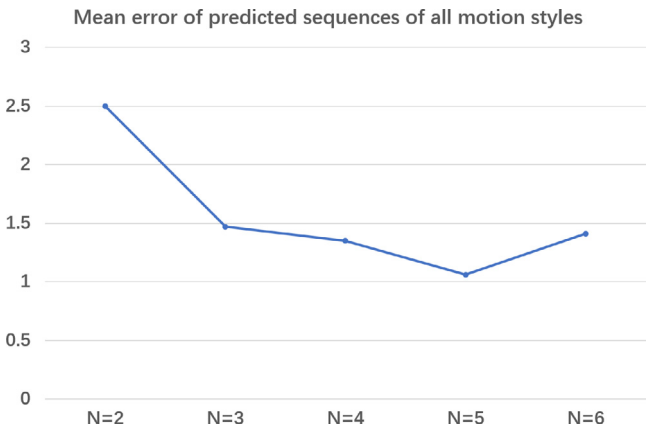


**Fig. 2.** Ablation study for options of N.

**Table 4**
MAE Comparison of 7 motion styles on Stylized Motion Datasets. SE stands for style encoder.The best results are shown in bold.

| | Angry | | | | | | Old | | | | | | Sexy | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| milliseconds | 80 | 160 | 320 | 400 | 560 | 1000 | 80 | 160 | 320 | 400 | 560 | 1000 | 80 | 160 | 320 | 400 | 560 | 1000 |
| TP-RNN | **0.21** | 0.52 | 0.89 | 1.74 | 1.99 | 2.74 | 0.33 | **0.54** | 0.98 | **1.42** | 1.87 | 2.45 | **0.24** | 0.44 | 0.77 | **1.31** | 1.74 | 1.88 |
| Ours(w/o SE) | 0.62 | 0.74 | 0.97 | 1.87 | 1.89 | 3.14 | 0.53 | 0.78 | 1.00 | 1.78 | 1.99 | 2.93 | 0.45 | 0.70 | 0.90 | 1.45 | 1.97 | 2.41 |
| Ours | 0.32 | **0.45** | **0.84** | **1.64** | **1.71** | **1.99** | **0.31** | 0.55 | **0.87** | 1.67 | **1.80** | **1.93** | 0.28 | **0.44** | **0.68** | 1.44 | **1.68** | **1.71** |
| | Childlike | | | | | | Strutting | | | | | | Proud | | | | | |
| milliseconds | 80 | 160 | 320 | 400 | 560 | 1000 | 80 | 160 | 320 | 400 | 560 | 1000 | 80 | 160 | 320 | 400 | 560 | 1000 |
| TP-RNN | **0.28** | 0.64 | 1.15 | 1.77 | 1.87 | 2.36 | **0.24** | 0.56 | 0.87 | 1.53 | 1.97 | 2.42 | **0.32** | 0.69 | 0.84 | 1.33 | 1.90 | 2.38 |
| Ours(w/o SE) | 0.59 | 0.89 | 1.67 | 1.99 | 2.19 | 3.22 | 0.36 | 0.87 | 1.01 | 1.97 | 2.11 | 2.91 | 0.78 | 0.85 | 1.11 | 1.55 | 2.33 | 3.02 |
| Ours | 0.29 | **0.45** | **0.89** | **1.35** | **1.50** | **1.80** | 0.27 | **0.45** | **0.78** | **1.47** | **1.69** | **1.97** | 0.34 | **0.56** | **0.70** | **1.15** | **1.38** | **1.66** |
| | Depressed | | | | | | Average of all 7 | | | | | | | | | | | |
| milliseconds | 80 | 160 | 320 | 400 | 560 | 1000 | 80 | 160 | 320 | 400 | 560 | 1000 | | | | | | |
| TP-RNN | **0.25** | 0.59 | 0.79 | **1.02** | 1.54 | 2.33 | **0.27** | 0.57 | 0.90 | 1.45 | 1.84 | 2.38 | | | | | | |
| Ours(w/o SE) | 0.66 | 0.85 | 1.18 | 1.33 | 1.86 | 2.91 | 0.57 | 0.81 | 1.12 | 1.71 | 0.25 | 2.93 | | | | | | |
| Ours | 0.33 | **0.46** | **0.59** | 1.11 | **1.40** | **1.82** | 0.31 | **0.48** | **0.76** | **1.40** | **1.59** | **1.83** | | | | | | |

accuracy is significantly reduced, which further illustrates the effectiveness of our method. To further illustrate the significance of our style encoder, we do a qualitative ablation study by using an ordinary 3-layer fully connected Encoder + standard Gaussian distribution instead of our style encoder + Constant Variance GMM to model the latent space and visualize the latent space learned by the two methods. The result is shown in the Fig. 4.

Through Fig. 4, it can be confirmed that the data distribution of Stylized Motion Datasets is actually as we analyzed, that is, there are overlaps and differences in the data distribution. When we compare Fig. 4(a) and (b) carefully, we can find that the latent space data distribution learned by the fully connected Encoder + standard Gaussian distribution is relatively scattered. The motions of the same style are not gathered together, and it is impossible to distinguish between normal speed walking and fast speed walking. On the contrary, the style encoder + constant variance GMM model the latent space more concentratedly. We can not only distinguish the styles more accurately but can also distinguish between normal speed walking and fast speed walking to a certain extent.

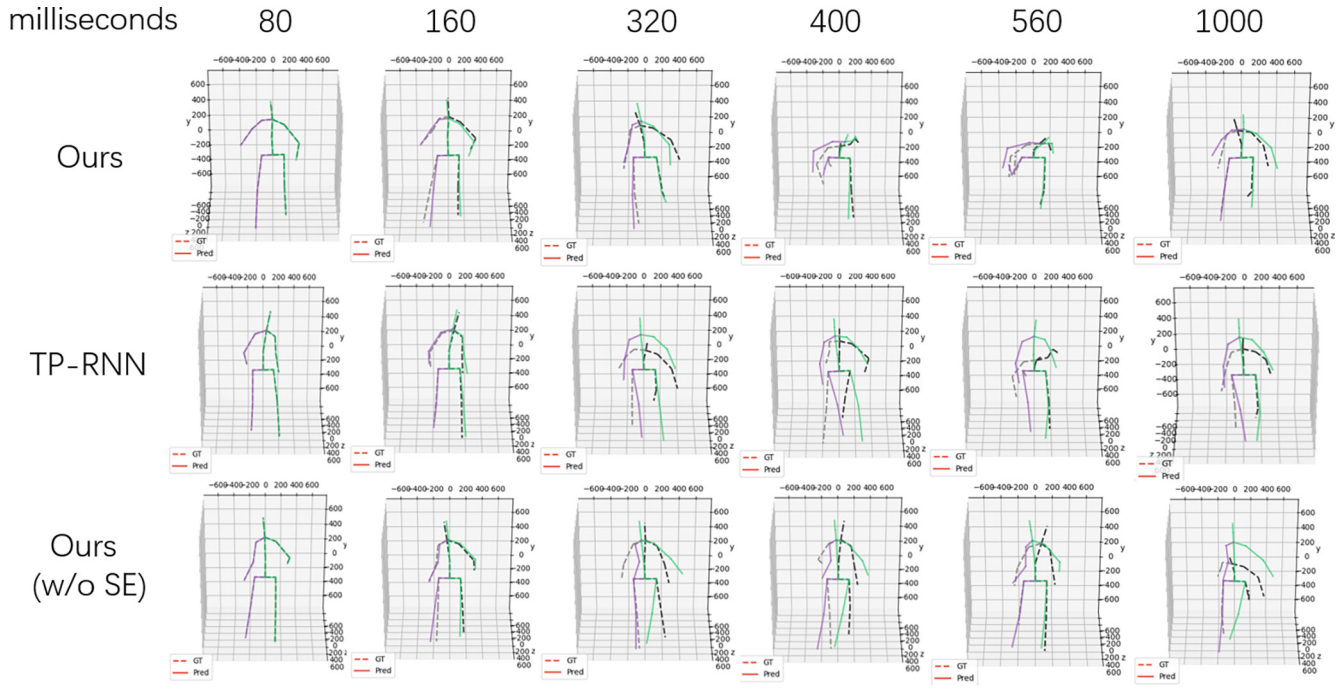


**Fig. 3.** Visual comparison of Angry Walking prediction results. The dotted line in the figure is Ground Truth and the solid line is the prediction.

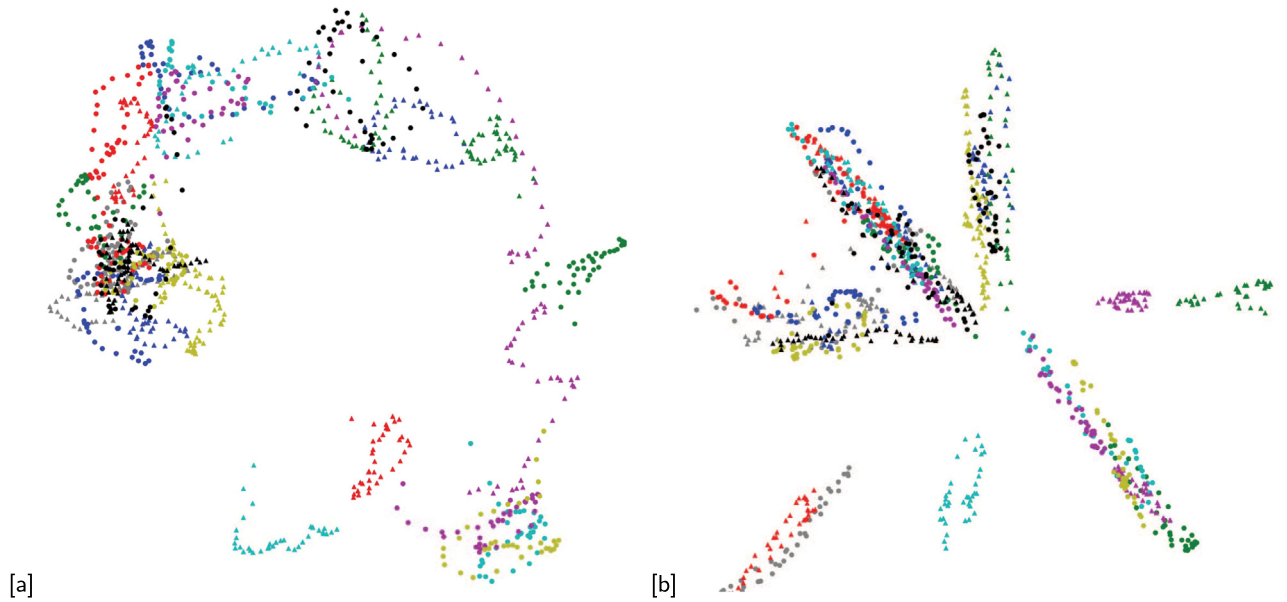


**Fig. 4.** We use T-SNE [17] to reduce the dimension of the latent space to 2 dimensions and visualize them. The different colors in the figure indicate different motion styles, the triangles indicate normal speed walking, and the circles indicate fast speed walking. (a) is the result of fully connected encoder + standard Gaussian distribution, (b) is the result of style encoder + Consta.nt variance GMM.

## 6. Conclusion

In this paper, we firstly analyzed the similarities and differences between stylized motion and ordinary motions using NPSS. Based on this observation, we proposed a method that can accurately predict stylized motion and demonstrated the effectiveness of our method through a series of experiments. Our model used transformer as the style encoder to extract stylized features. At the same time, we use a Gaussian mixture model with constant variance to model the data distribution of stylized motion in latent space. Combining the temporal dynamic captured by hierarchical multi-scale LSTM with style features extracted by style encoder, we provide a solution to the problem of the lack of detailed spatial structure modeling of motion in previous works, which led to ambiguity in prediction. Qualitative and quantitative experiments show that the predictive effect of our work on the Stylized Motion Datasets is state-of-the-art.

However, our method also has some limitations. First of all, the residual network structure does not completely solve the problem of first frame discontinuity. In the future, we hope to cope with this problem by estimating the initial hidden state of RNN. Secondly, the errors of our prediction results can increase significantly with large time step t due to the inherent randomness of human motion as well as the accumulation of errors in the prediction process. Future works may start from these two aspects for more accurate motion prediction.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This work was supported by National Key R&D Program "Science and Technology Winter Olympics" Key Special Project No.2020YFF0304701, the National Natural Science Foundation of China No.61173055.

## References

[1] E. Barsoum, J. Kender, Z. Liu, Hp-gan: Probabilistic 3d human motion prediction via gan, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2018, pp. 1418–1427.
[2] S. Bengio, O. Vinyals, N. Jaitly, N. Shazeer, Scheduled sampling for sequence prediction with recurrent neural networks, in: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1, MIT Press, Cambridge, MA, USA, 2015, pp. 1171–1179.
[3] M. Brand, A. Hertzmann, Style machines, in: Proceedings of the 27th annual conference on Computer graphics and interactive techniques, 2000, pp. 183–192.
[4] Chiu, H.k., Adeli, E., Wang, B., Huang, D.A., Niebles, J.C., 2019. Action-agnostic human pose forecasting, in: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE. pp. 1423–1432.
[5] J. Chung, S. Ahn, Y. Bengio, Hierarchical multiscale recurrent neural networks, in: International Conference on Learning Representations, 2017.
[6] K. Fragkiadaki, S. Levine, P. Felsen, J. Malik, Recurrent network models for human dynamics, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4346–4354.
[7] P. Ghosh, A. Losalka, M.J. Black, Resisting adversarial attacks using gaussian mixture variational autoencoders, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2019, pp. 541–548.
[8] Ghosh, P., Sajjadi, M.S.M., Vergari, A., Black, M., Scholkopf, B., 2020. From variational to deterministic autoencoders, in: International Conference on Learning Representations. URL:https://openreview.net/forum?id=S1g7tpEYDS.
[9] P. Ghosh, J. Song, E. Aksan, O. Hilliges, Learning human motion models for long-term predictions, in: 2017 International Conference on 3D Vision (3DV), IEEE, 2017, pp. 458–466.
[10] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y., 2014. Generative adversarial nets, in: NIPS, pp. 2672–2680. URL:http://papers.nips.cc/paper/5423-generative-adversarial-nets.
[11] A. Gopalakrishnan, A. Mali, D. Kifer, L. Giles, A.G. Ororbia, A neural temporal model for human motion prediction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 12116–12125.
[12] C. Ionescu, D. Papava, V. Olaru, C. Sminchisescu, Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments, IEEE transactions on pattern analysis and machine intelligence 36 (2013) 1325–1339.
[13] A. Jain, A.R. Zamir, S. Savarese, A. Saxena, Structural-rnn: Deep learning on spatio-temporal graphs, in: Proceedings of the ieee conference on computer vision and pattern recognition, 2016, pp. 5308–5317.
[14] Kingma, D.P., Welling, M., 2014. Auto-Encoding Variational Bayes, in: 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14–16, 2014, Conference Track Proceedings. arXiv:http://arxiv.org/abs/1312.6114v10.
[15] Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R.S., Urtasun, R., Torralba, A., Fidler, S., 2015. Skip-thought vectors, in: NIPS, pp. 3294–3302. URL:http://papers.nips.cc/paper/5950-skip-thought-vectors.
[16] J.N. Kundu, M. Gor, R.V. Babu, Bihmp-gan: Bidirectional 3d human motion prediction gan, in: Proceedings of the AAAI conference on artificial intelligence, 2019, pp. 8553–8560.
[17] L. Van der Maaten, G. Hinton, Visualizing data using t-sne, Journal of machine learning research 9 (2008).
[18] J. Martinez, M.J. Black, J. Romero, On human motion prediction using recurrent neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2891–2900.
[19] J.S. Monzani, P. Baerlocher, R. Boulic, D. Thalmann, Using an intermediate skeleton and inverse kinematics for motion retargeting, John Wiley & Sons Ltd (2010) 11–19.
[20] K. Sohn, H. Lee, X. Yan, Learning structured output representation using deep conditional generative models, Advances in neural information processing systems 28 (2015) 3483–3491.
[21] Sun, L., Tomizuka, M., Zhan, W., 2021. Multi-style human motion prediction and generation via meta-learning.
[22] Sutskever, I., Martens, J., Hinton, G.E., 2011. Generating text with recurrent neural networks, in: ICML.
[23] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, u., Polosukhin, I., 2017. Attention is all you need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA. p. 6000–6010.
[24] J. Walker, K. Marino, A. Gupta, M. Hebert, The pose knows: Video forecasting by generating pose futures, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 3332–3341.
[25] J.M. Wang, D.J. Fleet, A. Hertzmann, Gaussian process dynamical models for human motion, IEEE transactions on pattern analysis and machine intelligence 30 (2007) 283–298.
[26] Y.H. Wen, Z. Yang, H. Fu, L. Gao, Y. Sun, Y.J. Liu, Autoregressive stylized motion synthesis with generative flow, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13612–13621.
[27] S. Xia, C. Wang, J. Chai, J. Hodgins, Realtime style transfer for unlabeled heterogeneous human motion, ACM Transactions on Graphics (TOG) 34 (2015) 1–10.
[28] Y. Yuan, K. Kitani, Dlow: Diversifying latent flows for diverse human motion prediction, European Conference on Computer Vision, Springer. (2020) 346–364.
[29] Zhou, Y., Li, Z., Xiao, S., He, C., Huang, Z., Li, H., 2018. Auto-conditioned recurrent networks for extended complex human motion synthesis, in: International Conference on Learning Representations. URL:https://openreview.net/forum?id=r11Q2SlRW.

**Chongyang Zhong** received a BSc degree in automation from Tsinghua University (THU), China, in 2017. He is currently working toward a Ph.D. degree in computer science at the University of Chinese Academy of Science, supervised by Prof. Shihong Xia.

**Lei Hu** received a BSc degree in Applied Mathematics from Southwest Jiaotong University(SWJTU), China, in 2019. He is currently working toward a Ph.D. degree in computer science at the University of Chinese Academy of Science, supervised by Prof. Shihong Xia.

**Shihong Xia** received a Ph.D. degree in computer science from the University of Chinese Academy of Sciences. He is currently a professor of the Institute of Computing Technology, Chinese Academy of Sciences (ICT, CAS), and the director of the human motion laboratory. His primary research is in the area of computer graphics, virtual reality and artificial intelligence.