Existence Is Chaos: Enhancing 3D Human Motion Prediction with Uncertainty Consideration

Zhihao Wang[1,2], Yulin Zhou[1,2], Ningyu Zhang[1], Xiaosong Yang[3], Jun Xiao[1], Zhao Wang[2*]
[1]Zhejiang University
[2]Ningbo Innovation Center, Zhejiang University
[3]National Centre for Computer Animation, Bournemouth University
zhao_wang@zju.edu.cn, zhihao_wang@zju.edu.cn

Abstract

Human motion prediction is consisting in forecasting future body poses from historically observed sequences. It is a long-standing challenge due to motion's complex dynamics and uncertainty. Existing methods focus on building up complicated neural networks to model the motion dynamics. The predicted results are required to be strictly similar to the training samples with L_2 loss in current training pipeline. However, little attention has been paid to the uncertainty property which is crucial to the prediction task. We argue that the recorded motion in training data could be an observation of possible future, rather than a predetermined result. In addition, existing works calculate the predicted error on each future frame equally during training, while recent work indicated that different frames could play different roles. In this work, a novel computationally efficient encoder-decoder model with uncertainty consideration is proposed, which could learn proper characteristics for future frames by a dynamic function. Experimental results on benchmark datasets demonstrate that our uncertainty consideration approach has obvious advantages both in quantity and quality. Moreover, the proposed method could produce motion sequences with much better quality that avoids the intractable shaking artefacts. We believe our work could provide a novel perspective to consider the uncertainty quality for the general motion prediction task and encourage the studies in this field. The code will be available in https://github.com/Motionpre/Adaptive-Salient-Loss-SAGGB.

Introduction

Humans have the ability to predict how an action could be extrapolated in the future. This enables humans to react timely while interacting with external world. In the field of artificial intelligence, how to enable machines to anticipate human behaviour is a paramount challenge. Whether the challenge is handled well affects many real-life applications such as autonomous driving (Paden et al. 2016; Djuric et al. 2020) and human-robotics interaction (Koppula and Saxena 2013). The task of human motion prediction can be described as giving a series of human pose sequences in the past and predicting future human pose sequences. Anticipating the future movement of the 3D human skeleton is very challenging due to the complex spatial-temporal modality and the great uncertainty of the future.

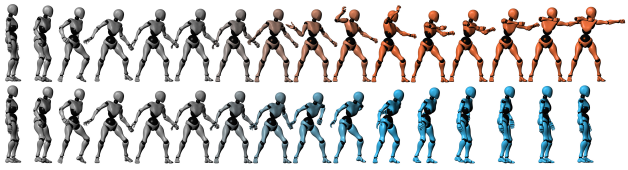


Figure 1: The uncertainty characteristic of human motion. For a certain motion clip, the recorded future motion should be an observation of possible future, rather than a predetermined result. The future motion could vary even with similar historical motion.

Current popular training pipeline of existing approaches is shown in Fig. 2. A given piece of training data sample would be divided into past pose sequences and future pose sequences $\{X_{obs}, X_{pre}\}$. The module would generate predicted future sequence $\hat{X}_{pre}$ while the parameters would be updated with the gradient of $loss(X_{pre}, \hat{X}_{pre})$. The prediction results are required to be strictly similar to the training examples. However, the uncertainty property of motion is overlooked. It would meet difficulties while dealing with the scenario that different motions may begin with the similar human motion poses. Thus, it is critical to take the uncertainty into consideration in human motion prediction.

The uncertainty of motion prediction mainly refers to its challenging variation, e.g. a difficult pose or a distant frame, especially for non-periodic behaviors. Most existing approaches treat each future frame equally. However, the uncertainty consideration of human motion is actually not equal in each future frame. In our real-life experience, the short-term future of a given action is much easier to be predicted , but the possible ground truth for movements in the longer period could be diverse. Such diverse could be varied due to the type of actions but generally become larger over time. In other word, the recorded future sequences in a motion clip should be an observation of historical sequence's possible future, rather than a predetermined result. For example, humans could predict a walking action will lift their feet off ground in the next millisecond, however, humans can walk to the left side or walk to the right side a few min-

*Corresponding Author.

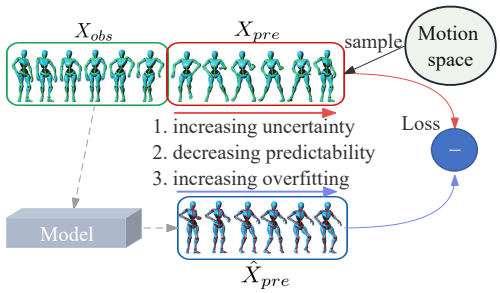


Figure 2: Training pipeline of existing prediction approaches. A given piece of training data sample would be divided into past pose sequences X_{obs} and future pose sequences X_{pre}. The predicted result $\hat{X}_{pre}$ is required to be strictly similar to the training data X_{pre}, where the uncertainty of future motion has been ignored.

utes later. The inherent relationships between the previous frames and longer future frames will generally go weaker over time.

Motivated by such insight, we aim to explicitly model the uncertainty of human motion for achieving high precision and robustness prediction performance. We challenge the widely used average approaches that do not differentiate the weight of each frame. The proposed model is encouraged to focus more on learning the credible frames accurately. Two significant pieces of evidence from existing empirical results support our idea: (1) Long-term error accumulation has been recognized as one of the biggest issues which bring performance degradation in motion prediction problems. As most models will predict the next frame conditioned on the previously predicted sequence, a small error in the initial frame will be amplified greatly due to the butterfly effects. (2) The latest work (Ma et al. 2022) proves that a different initial pose could bring sharp performance gains than an individual method. These validate that a more accurate prediction of the early frames will matter a lot in the final results.

Moreover, the long-term prediction from observation would usually have low confident, and this could vary due to the motion type. In light of these findings, a self-attention graph generate block (SAGGB) is designed to leverage the certainty connection information from the diversity of actions and the complexity of behaviors . Additionally, an Adaptive-Salient Loss is presented to make effective use of the uncertainty property to produce realism long-term prediction results. We consider properly assigning the weight to each frame, and present an active function to dynamically learn the weights for frames.

To evaluate our idea, extensive experiments are carried out on H3.6M, 3DPW and CMU Mocap datasets to study the impact of learning early frames for the final performance. The results demonstrate that our method achieve competitive performance in both short-term and long-term motion prediction tasks. Besides, our prediction results are more smooth and natural, achieve high quality without the intractable shaking effects.

In summary, the contributions of this paper are the followings:

1. The role of the uncertainty property in human motion prediction tasks has been studied, where its importance, and subsequently elaborated on its mechanism and principles are revealed. We hope our work will encourage more studies to rethink the value of uncertainty factors in the motion prediction problems.
2. A novel motion prediction work involved uncertainty consideration is proposed. A dynamic function is designed to learn the assigned weights for future frames. Extensive experimental results on the benchmark datasets validate our method outperform competitors in most short-term prediction jobs and achieve competitive performance in long-term prediction jobs. The proposed method could bring more favourable gains than existing methods in both quantity and quality.
3. We delve deeper into the uncertainty assumption and carried out extensive experiments to pursue the problem of how could the different assigned weights could affect the training and learning of the motion prediction models. Fruitful insights are given out by our ablation studies.

Related Work

Human Motion Prediction

Recently, deep learning networks became the mainstream of motion prediction. Since motion prediction was often seen as a sequence-to-sequence task, RNN-based methods were naturally used for this task. For example, LSTM-3LR (Jain et al. 2016) and ERD (Fragkiadaki et al. 2015) introduced LSTM into the task. ERD added encoder and decoder before and after LSTM to achieve better results. LSTM-3LR replaced bone information with speed information. However, RNNs suffer from error accumulation and were unable to effectively model spatial information. Then feed-forward neural networks were applied to the field. They tried to use convolution kernels to extract information in space and time. TE (Butepage et al. 2017) and QuaterNet (Pavllo, Grangier, and Auli 2018) used CNNs to capture temporal information. convSeq2Seq (Li et al. 2018) and CHA (Li et al. 2019) designed hierarchical CNNs to model spatial and temporal information simultaneously. The TrajectoryCNN (Liu et al. 2020) proposed the features of trajectory space that were more easily handled by CNN. The siMLPe (Guo et al. 2023) proposed to use simple multi-layer MLPs to extract temporal and spatial information and achieves excellent results. Some probabilistic prediction works (Yuan and Kitani 2020; Mao, Liu, and Salzmann 2022) predicted multiple possible motion sequences. However, due to the different evaluation metrics, our method does not compare with these works.

Graph Convolutional Networks (GCNs)

Recently, GCNs achieved the state-of-art results in motion prediction task. LTD (Mao et al. 2019) saw pose graph as a fully-connected graph and modeled temporal information by DCT representations which were followed by lots of later work. JDM (Su et al. 2021) and MPT (Liu et al. 2021) proposed GCNs that took velocity information as input.

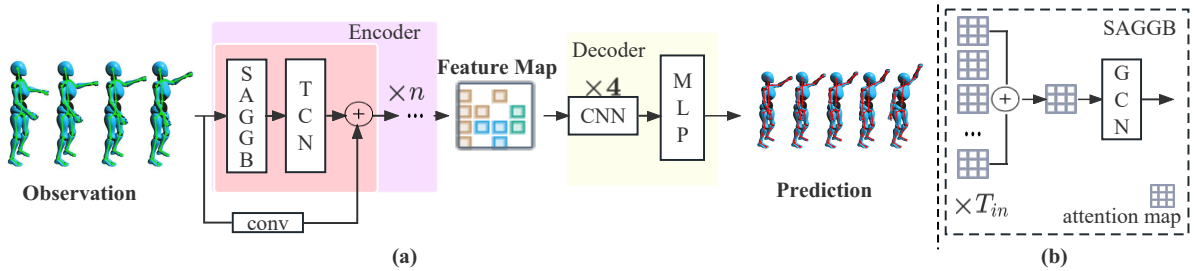


Figure 3: Overview of proposed model for human motion prediction with SAGGB. In the encoder, SAGGB leverage self attention mechanism to generate sample related graph to extract spatial information. In the decoder, we use lightweight CNNs and MLP to predict.

In the graph representation, DMGNN (Li et al. 2020) and MSR-GCN (Dang et al. 2021) used muti-scale graph and extracted information from single scales and cross scales. STSGCN (Sofianos et al. 2021) and GAGCN (Zhong et al. 2022) built space-time separable GCN to extract information in graph. SPGSN (Li et al. 2022) did scattering decomposition of the graph and GCNs were performed on all graphs. In the learning strategy, PGBIG (Ma et al. 2022) performed muti-stage prediction whose output corresponded to a smooth sequence in every stage. UA-HMP (Ding and Yin 2021) integrated probabilistic prediction into deterministic methods. AuxFormer (Xu et al. 2023) introduced a model learning framework with auxiliary task which was recovering corrupted coordinates depending on the rest coordinates. DMAG (Gao et al. 2023) used frequency decomposition and feature aggregation respectively to encode the information. Meanwhile, Transformer was also attempted for human motion prediction (Mao, Liu, and Salzmann 2020; Cai et al. 2020; Aksan et al. 2021). Self-attention mechanism of Transformer had strong modeling ability on sequence data and was able to model the dependencies of joints. In our model, we synthesize the modeling ability of self-attention mechanism and the effectiveness of graph neural networks.

Methodology

Overview

Firstly, some notations and variables in this paper would be enumerate. We denote $X_{obs} = [x_1, x_2, \cdot, x_{T_{in}}]$ as the past pose sequences and $X_{pre} = [x_{T_{in}+1}, x_{T_{in}+2}, \cdot, x_{T_{in}+T_{out}}]$ as the future pose sequences, where $x_i \in R^{N\times d}$ with N joints and d-dimension space (d is 3) is represented as human pose at time i. The goal of motion prediction is to learn a mapping function that maps X_{obs} to X_{pre}.

The overview of proposed model is showed in Fig. 3. An encoder-decoder structure is adopted to conduct end-to-end prediction. First, an encoder with proposed Self-attention Graph Blocks (SAGGBs) and Temporal Convolutional Network modules (TCNs) would encode the input into high-dimensional feature spaces. Then a lightweight decoder makes predictions for future pose sequence estimation.

The encoder contains a series of residual blocks. Each block consists of SAGGB and TCN which respectively extract the spatial and temporal information. After feature extraction through all blocks, the encoder would output a feature map of $R^{T_{in}\times V\times C_f}$, where $C_f = 128$ is used in this work. The decoder is a lightweight design that includes 4 CNNs and a MLP. The CNN module uses time dimension as feature channel to extract information. The first CNN projects the input from $R^{T_{in}\times V\times C_f}$ to $R^{T_{out}\times V\times C_f}$. The MLP projects the C_f to 3.

Self-attention Graph Generate Block

For GCNs, either natural connection or full connection applies the same graph weights among all samples, where the the diversity of actions and the complexity of behaviors are ignored. For instance, for action walking, the dependence between feet should be more of a concern but for action eating the dependence is not so important. Such dependency reflects the variation of uncertainty with motion diversity. Hence, We propose a Self-attention Graph Generate Block (SAGGB) module to generate data-driven graph to model such information. Let's recall $x_i \in R^{n\times d}$ as the human pose at time i. We apply self-attention mechanism to generate graph A_i for every pose

$$A_i = softmax(\frac{\sigma(x_i)\phi(x_i)^{\mathrm{T}}}{\sqrt{d_k}}) \tag{1}$$

where $\sigma(.)$ and $\phi(.)$ are mapping function to generate query and key for every joint; d_k is the dimension of query and key. For each pose, SAGGBs generate attention map about the joints as the graph to this pose. As can be noticed, SAGGBs generate different graphs with different weights for different poses.

However, the same pose in different pose sequences could have different means. Consequently, it's more sensible to generate special graph for every pose sequence rather than pose. We denote the graph A_{sample} for a given sample as below

$$A_{sample} = \sum_{i=1}^{T_{in}} A_i \tag{2}$$

By constructing a new graph for each sequence, the model is able to extract implicit higher-dimensional information from different input poses.Based on the special graph for

each sample, the output X^{l+1} of SAGGB block with given input X^l can be defined as:

$$X^{l+1} = \sigma(A_{sample} X^l W^l) \tag{3}$$

Loss Function

To train the model for human motion prediction, L_2 loss is usually used in existing approaches:

$$\ell = \frac{1}{T_{out} N} \sum_{t=1}^{T_{out}} \sum_{n=1}^{N} \|\hat{x}_t^n - x_t^n\|_2 \tag{4}$$

As mentioned before, such averaging form treat all timestamps equally, where problem arises in two aspects. First, the prediction of the later frames could be difficult due to the uncertainty of the action semantics. However, the loss with averaging form gives more weight to the later frames with greater uncertainty and less structured knowledge, which makes the model fit more noise and uncertainty. Second, human motion is a continuous process, thus the distribution of one frame is heavily dependent on the state of the previous frame. This dependency is passed frame by frame. Therefore, the prediction of the first frame in the prediction sequence is crucial. The first frame in prediction determines the initial position of the model from the deterministic space to the uncertain space. Meanwhile, the dependency relationship shows that the loss of the first frame prediction largely determines the expression of the entire model.

Based on the above two issues, we propose an Adaptive-Salient Loss in this work.

Adaptive Loss Although motion uncertainty varies in different actions, it mainly increases over time. To overcome this imbalance problem, inspired by multi-task model (Kendall, Gal, and Cipolla 2018), we propose an Adaptive Loss for human motion prediction. First, for the output of the model $\hat{X}_{pre}$, we define a probabilistic model as:

$$p\left(x_t|\hat{x}_t\right) = \mathcal{N}\left(\hat{x}_t, \sigma\right) \tag{5}$$

$$p\left([x_1, x_2, \ldots, x_{T_{out}}]\,|[\hat{x_1}, \hat{x_2}, \ldots, \hat{x}_{T_{out}}]\right) = \prod_{t=1}^{T_{out}} p(x_t|\hat{x_t}) \tag{6}$$

where σ denotes the uncertainty. To optimize it, we maximise the log likelihood.

$$\log p\left(x_t|\hat{x_t}\right) \propto -\frac{1}{2\sigma^2} \|x_t - \hat{x_t}\|_2 - \log \sigma \tag{7}$$

Consequently, we regard the prediction for each frame as a part of task and combine the tasks to get better predictions on the whole.

$$\log p\left(X_{pre}|\hat{X}_{pre}\right) \propto -\sum_{t=1}^{T_{out}} \left(\frac{1}{2\sigma_t^2} \|x_t - \hat{x_t}\|_2 + \log \sigma_t\right) \tag{8}$$

Accordingly, we can define the proposed Adaptive Loss as Eq. (9). Adaptive Loss reflects the increasing of uncertainty over time. At prediction sequences, σ is the increasing uncertainty over the time. During training, σ adjusts the weights of different frames hence model gives variant attention to frames.

$$L_{adaptive} = \sum_{t=1}^{T_{out}} \frac{1}{2\sigma_t^2} \|x_t - \hat{x_t}\|_2 + \sum_{t=1}^{T_{out}} \log \sigma_t \tag{9}$$

Salient Loss In Adaptive Loss, we build the probabilistic model based on the assumption that the prediction for each frame is independent. However, human motion is a continuous process and the distribution of one frame is heavily dependent on the state of the previous frame. We formulate the perception as follows:

$$x_{t+1} = f(x_t, \sigma_{t+1}, \upsilon_t) \tag{10}$$

where υ_t is temporal and motion information before timestamp t. σ_{t+1} represent uncertainty spaces in $t+1$.

Eq. (10) is intuitive that expresses the continuity and uncertainty of human motion. The continuity is also essential while making predictions. Therefore, as the initial state of the model output prediction sequence, the last input frame is crucial to the entire prediction process, which can be expressed as:

$$\hat{X}_{pre} = F(x_{T_{in}}, \sigma_{T_{in}+1:T_{in}+T_{out}}, \upsilon_{T_{in}:T_{in}+T_{out}-1}) \tag{11}$$

The last input frame is the initial state of the prediction output in the deterministic space. Eq. (11) expresses that the prediction process needs to explicitly utilize the last input frame. Many previous methods have taken advantage of this. In LTD (Mao et al. 2019) and SPGSN (Li et al. 2022), the last input frame is repeated T_{out} times after input as the input of the model, and the offset relative to the last frame is output. PGBIG (Ma et al. 2022) repeats the last frame T_{out} times as an initial estimate for the predicted sequence.

Meanwhile, the first predicted frame is the initial state of the prediction output in the uncertainty space. Hence, we can continue the progressive deduction of prediction and reformulated Eq. (11) as:

$$\hat{X}_{pre} = F(\hat{x}_{T_{in}+1}, \sigma_{T_{in}+2:T_{in}+T_{out}}, \upsilon_{T_{in}+1:T_{in}+T_{out}-1}) \tag{12}$$

Previous methods explicitly utilize the initial state of the output in the deterministic observed data as shown in Eq. (11). The initial state of the model in the prediction with uncertainty is ignored. In order to tackle this issue, we propose a Salient Loss as follows:

$$L_{salient} = \omega T_{out} \|x_1 - \hat{x}_1\|_2 + \sum_{t=1}^{T_{out}} \|x_t - \hat{x}_t\|_2 \tag{13}$$

where ω is salient factor.

Adaptive-Salient Loss Adaptive Loss indicates that early stage prediction usually has higher confident. During training, we treat σ as a set of trainable parameters and σ is learned together with the model parameters. The $\log \sigma$ is the regularization term.

action	walking				eating				smoking				discussion			
millisecond	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
DMGNN	17.3	30.7	54.6	65.2	11.0	21.4	36.2	43.9	9.0	17.6	32.1	40.3	17.3	34.8	61.0	69.8
LTD	12.3	23.0	39.8	46.1	8.4	16.9	33.2	40.7	7.9	16.2	31.9	38.9	12.5	27.4	58.5	71.7
SPGSN	10.1	19.4	34.8	41.5	7.1	14.9	30.5	37.9	6.7	13.8	28.0	34.6	10.4	23.8	53.6	67.1
PGBIG	10.2	19.8	34.5	40.3	7.0	15.1	30.6	38.1	6.6	14.1	28.2	34.7	10.0	23.8	53.6	66.7
Ours	**8.8**	**17.7**	**33.1**	**39.6**	**6.2**	**14.1**	**29.8**	**37.3**	**5.7**	**12.8**	**26.9**	**33.7**	**8.7**	**21.8**	**51.4**	**64.9**
action	directions				greeting				phoning				posing			
millisecond	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
DMGNN	13.1	24.6	64.7	81.9	23.3	50.3	107.3	132.1	12.5	25.8	48.1	58.3	15.3	29.3	71.5	96.7
LTD	9.0	19.9	43.4	53.7	18.7	38.7	77.7	93.4	10.2	21.0	42.5	52.3	13.7	29.9	66.6	84.1
SPGSN	7.4	17.2	39.8	50.3	14.6	32.6	70.6	86.4	8.7	18.3	38.7	48.5	10.7	25.3	59.9	76.5
PGBIG	7.2	17.6	40.9	51.5	15.2	34.1	71.6	87.1	8.3	18.3	38.7	48.4	10.7	25.7	60.0	76.6
Ours	**6.2**	**16.0**	**39.0**	**50.0**	**12.5**	**30.4**	**68.6**	**85.4**	**7.4**	**17.1**	**37.8**	**47.9**	**9.1**	**23.3**	**57.4**	**74.6**
action	purchases				sitting				sittingdown				takingphoto			
millisecond	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
DMGNN	21.4	38.7	75.7	92.7	11.9	25.1	44.6	50.2	15.0	32.9	77.1	93.0	13.6	29.0	46.0	58.8
LTD	15.6	32.8	65.7	79.3	10.6	21.9	46.3	57.9	16.1	31.1	61.5	75.5	9.9	20.9	45.0	56.6
SPGSN	12.8	28.6	61.0	74.4	9.3	19.4	**42.3**	**53.6**	14.2	27.7	56.8	70.7	8.8	18.9	41.5	**52.7**
PGBIG	12.5	28.7	60.1	**73.3**	8.8	19.2	42.4	53.8	13.9	27.9	57.4	71.5	8.4	18.9	42.0	53.3
Ours	**10.9**	**26.8**	**59.8**	73.9	**8.1**	**18.4**	**42.3**	54.1	**12.8**	**26.3**	**55.9**	**70.3**	**7.8**	**17.9**	**41.3**	52.9
action	waiting				walkingdog				walkingtogether				average			
millisecond	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
DMGNN	12.2	24.2	59.6	77.5	47.1	93.3	160.1	171.2	14.3	26.7	50.1	63.2	17.0	33.6	65.9	79.7
LTD	11.4	24.0	50.1	61.5	23.4	46.2	83.5	96.0	10.5	21.0	38.5	45.2	12.7	26.1	52.3	63.5
SPGSN	9.2	19.8	43.1	54.1	—	—	—	—	8.9	18.2	33.8	40.9	10.4	22.3	47.1	58.3
PGBIG	8.9	20.1	43.6	54.3	18.8	39.3	73.7	86.4	8.7	18.6	34.4	41.0	10.3	22.7	47.4	58.5
Ours	**7.6**	**17.9**	**41.1**	**52.3**	**16.0**	**36.0**	**72.0**	**85.2**	**7.7**	**16.8**	**32.3**	**39.4**	**9.1**	**20.9**	**45.9**	**57.4**

Table 1: Comparisons of short-term prediction for all actions and the average on H3.6M. The best results are shown in bold. Our method outperforms all baselines in most short-term prediction cases.

Salient Loss could emphasize first frame's importance as the initial state of the prediction sequence with uncertainty. During training, we set ω as a fixed value to emphasize the importance of the initial pose.

Our final loss function is a weighted combination of Adaptive Loss and Salient Loss:

$$L = \lambda L_{Adaptive} + (1 - \lambda) L_{Salient} \tag{14}$$

where λ is weight factor.

Experiments

Datasets

Human3.6M dataset (H3.6M) is a widely used motion prediction dataset consists of 15 actions performed by 7 actors(S1,S5,S6,S7,S8,S9 and S11). The human body is represented as 32 joints. Follow (Li et al. 2022; Ma et al. 2022), global information including global rotation and movement is removed. Meanwhile, all samples are down-sampled to 25 frames per second and we test our model on S5.

CMU Motion Capture dataset (CMU Mocap) is another dataset widely used for human motion prediction. Like previous methods, we use 8 actions and choose 25 joints for each pose. Other processing is similar to H3.6M.

3D Pose in the Wild dataset (3DPW) includes both indoor and outdoor actions captured at 30Hz. Each pose has 26 joints and we use 23 of them. We conduct experiments according to the official segmented training set, validation set and test set.

Evaluation Criterion and Baselines

Follow (Ma et al. 2022), Mean Per Joint Position Error (MPJPE) is employed as criterion which calculates the average L_2 distance at each timestamp.

The proposed method is compared with DMGNN (Li et al. 2020), LTD (Mao et al. 2019), SPGSN (Li et al. 2022) and PGBIG (Ma et al. 2022). All methods are tested under the same conditions. Meanwhile, the proposed method is also compared with STSGCN (Sofianos et al. 2021) and GAGCN (Zhong et al. 2022) in H3.6M under their evaluation criteria. Inspired by (Du et al. 2023), we further employ Jitter metric to evaluate the quality of predicated motion sequences.

$$\text{Jitter} = \frac{\mu^3}{T-3} \sum_{t=0}^{T-3} (\Delta x_{t+3} - 3 * \Delta x_{t+2} + 3 * \Delta x_{t+1} - \Delta x_t) \tag{15}$$

action	walking		eating		smoking		discussion		directions		greeting		phoning		posing	
millisecond	560	1000	560	1000	560	1000	560	1000	560	1000	560	1000	560	1000	560	1000
DMGNN	73.4	95.8	58.1	86.7	50.9	72.2	81.9	138.3	110.1	115.8	152.5	157.7	78.9	**98.6**	163.9	310.1
LTD	54.1	59.8	53.4	77.8	50.7	72.6	91.6	121.5	71	101.8	115.4	148.8	69.2	103.1	114.5	173
SPGSN	**46.9**	**53.6**	**49.8**	**73.4**	46.7	68.6	—	—	70.1	100.5	—	—	66.7	102.5	—	—
PGBIG	48.1	56.4	51.1	76	46.5	69.5	**87.1**	118.2	**69.3**	**100.4**	110.2	143.5	65.9	102.7	**106.1**	**164.8**
Ours	49.0	56.3	51.1	75.1	**46.1**	**68.2**	**87.1**	**117.2**	70.6	101.8	**110.0**	**141.7**	**65.7**	101.9	109.4	165.8
action	purchases		sitting		sittingdown		takingphoto		waiting		walkingdog		walkingtogether		avearge	
millisecond	560	1000	560	1000	560	1000	560	1000	560	1000	560	1000	560	1000	560	1000
DMGNN	118.6	153.8	**60.1**	**104.9**	122.1	168.8	91.6	120.7	106	136.7	194	182.3	83.4	115.9	103	137.2
LTD	102	143.5	78.3	119.7	100	150.2	77.4	119.8	79.4	108.1	111.9	148.9	55	65.6	81.6	114.3
SPGSN	—	—	75.0	116.2	—	—	75.6	**118.2**	73.5	103.6	—	—	—	—	77.4	**109.6**
PGBIG	**95.3**	**133.3**	74.4	116.1	**96.7**	**147.8**	**74.3**	118.6	**72.2**	**103.4**	104.7	139.8	51.9	64.3	**76.9**	110.3
Ours	96.9	137.5	74.7	116.6	96.9	150.2	77.3	120.8	73.3	104.1	**103.8**	**137.3**	**51.5**	**61.7**	77.5	110.4

Table 2: Comparisons of long-term prediction for all actions and the average on H3.6M. Our method achieves competitive results with the best models. However, our model only has approximately 10% parameters as the best model.

millisecond	80	160	320	400	560	1000
STSGCN	10.1	17.1	33.1	38.3	50.8	75.6
GAGCN	10.1	16.9	32.5	38.5	50.0	72.9
Ours	**6.3**	**11.9**	**24.1**	**30.2**	**42.4**	**66.1**

Table 3: Comparison of prediction for the average on H3.6M under the evaluation criteria of STSGCN.

where Δx_t is the Euclidean distance between prediction pose and future pose at frame t, T is the total frames and μ is the frame per second.

Comparison with the State-of-the-art Methods

To show the performance of our model, we list the quantitative results for both short-term prediction(400ms) and long-term prediction(1000ms) on H3.6M, CMU Mocap and 3DPW.

H3.6M Tab. 1 shows the short-term comparisons results. Our model outperforms state-of-art methods in most cases of all actions, which shows good stability compared to other methods by taking into account uncertainty and better encoding spatial information.

The long-term comparisons is shown in Tab. 2. Our model preferentially guarantees the optimization of the early frame. It also performs competitively in the long-term prediction due to less fitting noise and uncertainty. Meanwhile, although current state-of-art methods methods could provide impressive long-term prediction MPJPE score, the motion quality of predicted results remain in doubt. This part would be detail discussed in the following paragraph and an example video is illustrated in the supplementary material.

CMU Mocap & 3DPW Tab. 4 and Tab. 5 show the comparison on CMU Mocap and 3DPW with both short-term and long-term prediction. Only average error is listed due to space limitations. We also achieve best in most cases, which demonstrates the effectiveness of our method.

millisecond	80	160	320	400	560	1000
DMGNN	13.6	24.1	47.0	58.8	77.4	112.6
LTD	9.3	17.1	33.0	40.9	55.8	86.2
SPGSN	8.3	14.8	28.6	37.0	–	**77.8**
PGBIG	7.6	14.3	29.0	36.6	50.9	80.1
Ours	**7.2**	**13.5**	**27.9**	**36.4**	**48.6**	82.1

Table 4: Comparison of prediction for the average on CMU

millisecond	100	200	400	600	800	1000
DMGNN	17.8	37.1	70.4	94.1	109.7	123.9
LTD	16.3	35.6	67.5	90.4	106.8	117.8
SPGSN	15.4	32.9	64.5	91.6	104.0	111.1
PGBIG*	13.1	**29.2**	61.0	89.6	102.6	109.4
Ours	**12.4**	**29.2**	**59.1**	**87.7**	**99.9**	**107.7**

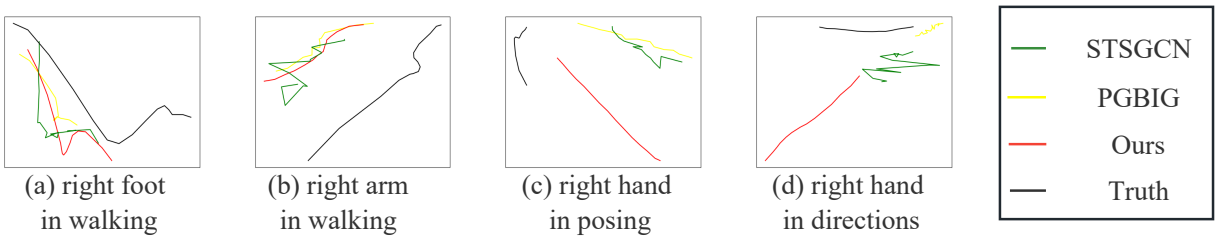
* This is a correction value that reproduced with original checkpoints

Table 5: Comparison of prediction for the average on 3DPW

Computational Complexity Analysis The computational complexity and time consuming results are shown in Tab. 6, where the model checkpoint provided by original public repository is used. Our model is smaller than previous methods as we conduct a lightweight architecture. Although the calculation cost of our model is increased due to the self-attention, the actual inference time of our model is still in advantage.

Model	LTD	SPGSN	PGBIG	Ours
Parameters	2.55M	5.67M	1.74M	0.55M
FLOPs	133.7M	549.3M	55.8M	143.5M
Inference Time	5.18ms	63.07ms	16.22ms	8.09ms

Table 6: Comparison of Computational Complexity



(a) right foot in walking (b) right arm in walking (c) right hand in posing (d) right hand in directions

Figure 4: Comparison of Qualitative Evaluation Results. Joint trajectory is recorded from 400ms to 1000ms. The shaking problems have occurred in comparison method's prediction results, which could be caused by over-fitting.

	0-1000ms	400-1000ms	800-1000ms
STSGCN	671.76	518.99	649.05
PGBIG	161.20	195.91	227.81
Ours	111.82	114.70	119.24

Table 7: Comparison of Quality of Predicted Motion in Jitter Metric (m/s^3)

	80	160	320	400	560	1000
GCN	9.7	22.2	47.9	59.7	78.4	111.8
SAGGB-3	9.3	22.0	48.2	60.1	79.5	112.6
SAGGB-6	**9.1**	**20.9**	**45.9**	**57.4**	**77.5**	**110.4**
SAGGB-8	9.1	20.9	46.4	58.3	77.7	110.9

Table 8: Ablation on SAGGB module for network architecture, evaluated on long-term prediction on H3.6M

Quality Analysis of Predicted Motion

Statistical analysis on prediction quality is shown in Tab. 7 using Jitter metric (Du et al. 2023). To be specific, the sequences predicted by our method performed better in realism and stable with the increase of prediction frames. Visualization of Joint Trajectory is shown in Fig. 4. The trajectory of a single joint point projected on the x-y plane of a prediction sequence is shown. More turns appear at the end of the PGIBIG's and STSGCN's predictions to mimic the real joint motion. Comparing to baselines, ours predicted results are more reasonable and smooth. Further details in prediction result is shown in the supplementary video.

Ablation Study

The ablation study results for hyper parameters is shown in Fig. 5. These results confirm the validity of our proposed loss and also support our hypothesis. The salient factor ω and weight factor λ is determined as 10 and 0.3, respectively. In addition, we also investigate the validity of our model structure. We have conducted the ablation study of proposed SAGGB and investigate the block number for encoder, where the result is shown on Tab. 8.

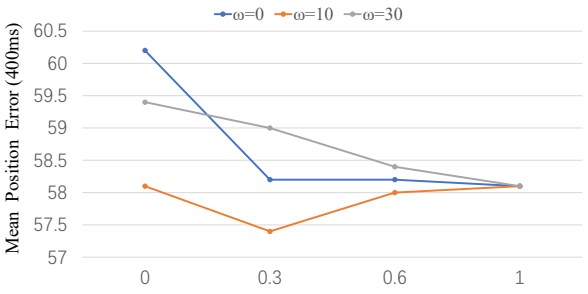


Figure 5: Ablation on value of λ and ω on H3.6M

Discussions

Our method achieves high quality and realism motion prediction. However, there are still two limitations: 1. The popular L_2 Loss, e.g. MPJPE may be not sufficient enough to fully express the quality of the prediction and be an indicator for a prediction model. Especially long-term prediction results with similar MPJPE could perform quite different on motion stable and realism. Further evaluation indicators could be conducted to improve model training and prediction quality evaluation. 2. Current benchmark datasets are originally created for either pose estimation or action recognition purpose. Hence, some key characteristics of motion prediction are not considered well, e.g. various possible action modes from various subjects. In the future, how to construct a dataset that is more suitable for human action prediction is worthy of future research.

Conclusion

To sum up, we have proposed a novel motion prediction framework with uncertainty consideration, which challenges the assumption that training data should be fully trusted equally. Extensive studies have been carried out to evaluate our insights on the heavily benchmark Human3.6M, 3DPW and CMU datasets. Our method could achieve favourable gains. More importantly, our method could tackle overfitting problem that avoids weird artifacts and generates more realistic motion sequences. Extensive ablation studies are carried out to present fruitful insights into the field. We believe this new perspective of uncertainty will inspire other researchers and facilitate the prediction related work both in academic and industries in the future.

Acknowledgments

This research has been supported by National Key Research and Development Project of China (Grant No. 2021ZD0110702), Natural Key Research and Development Project of Zhejiang Province (Grant No. 2023C01043), Ningbo Natural Science Foundation (Grant No. 2023Z236) , Information Technology Center and State Key Lab of CAD&CG, Zhejiang University.

References

Aksan, E.; Kaufmann, M.; Cao, P.; and Hilliges, O. 2021. A spatio-temporal transformer for 3d human motion prediction. In *2021 International Conference on 3D Vision (3DV)*, 565–574. IEEE.

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

Cai, Y.; Huang, L.; Wang, Y.; Cham, T.-J.; Cai, J.; Yuan, J.; Liu, J.; Yang, X.; Zhu, Y.; Shen, X.; et al. 2020. Learning progressive joint propagation for human motion prediction. In *European Conference on Computer Vision*, 226–242. Springer.

Dang, L.; Nie, Y.; Long, C.; Zhang, Q.; and Li, G. 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.

Ding, P.; and Yin, J. 2021. Uncertainty-aware Human Motion Prediction. *arXiv preprint arXiv:2107.03575*.

Djuric, N.; Radosavljevic, V.; Cui, H.; Nguyen, T.; Chou, F.-C.; Lin, T.-H.; Singh, N.; and Schneider, J. 2020. Uncertainty-aware short-term motion prediction of traffic actors for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2095–2104.

Du, Y.; Kips, R.; Pumarola, A.; Starke, S.; Thabet, A.; and Sanakoyeu, A. 2023. Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 481–490.

Fragkiadaki, K.; Levine, S.; Felsen, P.; and Malik, J. 2015. Recurrent network models for human dynamics. In *Proceedings of the IEEE international conference on computer vision*, 4346–4354.

Gao, X.; Du, S.; Wu, Y.; and Yang, Y. 2023. Decompose More and Aggregate Better: Two Closer Looks at Frequency Representation Learning for Human Motion Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6451–6460.

Guo, W.; Du, Y.; Shen, X.; Lepetit, V.; Alameda-Pineda, X.; and Moreno-Noguer, F. 2023. Back to mlp: A simple baseline for human motion prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 4809–4819.

Jain, A.; Zamir, A. R.; Savarese, S.; and Saxena, A. 2016. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the ieee conference on computer vision and pattern recognition*, 5308–5317.

Kendall, A.; Gal, Y.; and Cipolla, R. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7482–7491.

Koppula, H. S.; and Saxena, A. 2013. Anticipating human activities for reactive robotic response. In *IROS*, 2071. Tokyo.

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

Li, M.; Chen, S.; Zhang, Z.; Xie, L.; Tian, Q.; and Zhang, Y. 2022. Skeleton-Parted Graph Scattering Networks for 3D Human Motion Prediction. *arXiv preprint arXiv:2208.00368*.

Li, M.; Chen, S.; Zhao, Y.; Zhang, Y.; Wang, Y.; and Tian, Q. 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.

Li, Y.; Wang, Z.; Yang, X.; Wang, M.; Poiana, S. I.; Chaudhry, E.; and Zhang, J. 2019. Efficient convolutional hierarchical autoencoder for human motion prediction. *The Visual Computer*, 35(6): 1143–1156.

Liu, X.; Yin, J.; Liu, J.; Ding, P.; Liu, J.; and Liu, H. 2020. Trajectorycnn: a new spatio-temporal feature learning network for human motion prediction. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(6): 2133–2146.

Liu, Z.; Su, P.; Wu, S.; Shen, X.; Chen, H.; Hao, Y.; and Wang, M. 2021. Motion prediction using trajectory cues. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13299–13308.

Ma, T.; Nie, Y.; Long, C.; Zhang, Q.; and Li, G. 2022. Progressively Generating Better Initial Guesses Towards Next Stages for High-Quality Human Motion Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6437–6446.

Mao, W.; Liu, M.; and Salzmann, M. 2020. History repeats itself: Human motion prediction via motion attention. In *European Conference on Computer Vision*, 474–489. Springer.

Mao, W.; Liu, M.; and Salzmann, M. 2022. Weakly-supervised action transition learning for stochastic human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8151–8160.

Mao, W.; Liu, M.; Salzmann, M.; and Li, H. 2019. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9489–9497.

Paden, B.; Čáp, M.; Yong, S. Z.; Yershov, D.; and Frazzoli, E. 2016. A survey of motion planning and control techniques for self-driving urban vehicles. *IEEE Transactions on intelligent vehicles*, 1(1): 33–55.

Pavllo, D.; Grangier, D.; and Auli, M. 2018. Quaternet: A quaternion-based recurrent model for human motion. *arXiv preprint arXiv:1805.06485*.

Sofianos, T.; Sampieri, A.; Franco, L.; and Galasso, F. 2021. Space-time-separable graph convolutional network for pose forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11209–11218.

Su, P.; Liu, Z.; Wu, S.; Zhu, L.; Yin, Y.; and Shen, X. 2021. Motion prediction via joint dependency modeling in phase space. In *Proceedings of the 29th ACM International Conference on Multimedia*, 713–721.

Xu, C.; Tan, R. T.; Tan, Y.; Chen, S.; Wang, X.; and Wang, Y. 2023. Auxiliary tasks benefit 3d skeleton-based human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9509–9520.

Yuan, Y.; and Kitani, K. 2020. Dlow: Diversifying latent flows for diverse human motion prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, 346–364. Springer.

Zhong, C.; Hu, L.; Zhang, Z.; Ye, Y.; and Xia, S. 2022. Spatio-Temporal Gating-Adjacency GCN for Human Motion Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6447–6456.