# GlocalNet: Class-aware Long-term Human Motion Synthesis

Neeraj Battan *, Yudhik Agrawal *, Sai Soorya Rao, Aman Goel, and Avinash Sharma
International Institute of Information Technology, Hyderabad
{neeraj.battan, yudhik.agrawal}@research.iiit.ac.in,
{sai.soorya, aman.goel}@students.iiit.ac.in, asharma@iiit.ac.in

## Abstract

*Synthesis of long-term human motion skeleton sequences is essential to aid human-centric video generation [8] with potential applications in Augmented Reality, 3D character animations, pedestrian trajectory prediction, etc. Long-term human motion synthesis is a challenging task due to multiple factors like, long-term temporal dependencies among poses, cyclic repetition across poses, bi-directional and multi-scale dependencies among poses, variable speed of actions, and a large as well as partially overlapping space of temporal pose variations across multiple class/types of human activities. This paper aims to address these challenges to synthesize a long-term ($> 6000$ ms) human motion trajectory across a large variety of human activity classes ($> 50$). We propose a two-stage activity generation method to achieve this goal, where the first stage deals with learning the long-term global pose dependencies in activity sequences by learning to synthesize a sparse motion trajectory while the second stage addresses the generation of dense motion trajectories taking the output of the first stage. We demonstrate the superiority of the proposed method over SOTA methods using various quantitative evaluation metrics on publicly available datasets.*

## 1. Introduction

Skeleton sequences are traditionally used for human activity/action representation & analysis [26]. Recently, human motion synthesis [3, 5, 6, 9, 21, 23] is gaining ground as it is widely used to aid human-centric video generation [8] with potential applications in Augmented Reality, 3D character animations, pedestrian trajectory prediction, etc.

Human motion synthesis is a challenging task due to multiple factors like long-term temporal dependencies among poses, cyclic repetition across poses, bi-directional and multi-scale dependencies among poses, variable speed of actions, and a large as well as partially overlapping space of temporal pose variations across multiple class/types of human activities.

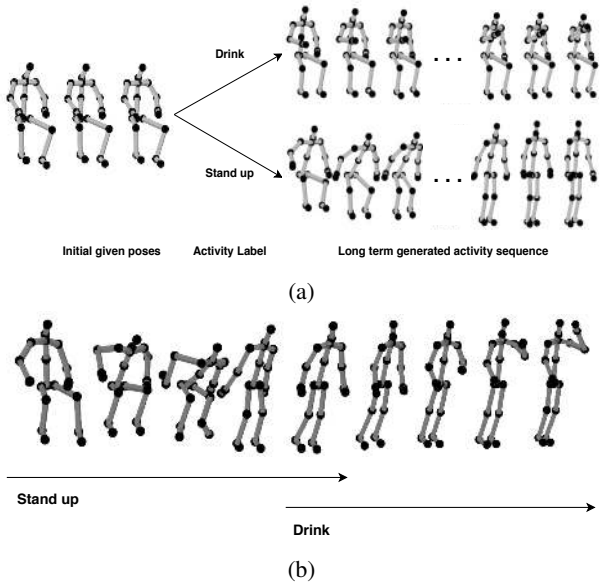


Figure 1: Motivation: a) Using the same set of sparse initial poses, our method can generate differents type of activities based on the input class label. The figure depicts two such activities - Drinking and Standing up that were synthesized from the same set of initial poses. b) Our method is also capable of transitioning across actions. The figure demonstrates the transition from Standing Up to Drinking activity.

Existing methods for human motion synthesis [3, 5, 9, 10, 14, 21] primarily uses auto-regressive models such as LSTM [13], GRU [2] and Seq2Seq [27] which aim to predict a temporally short-duration motion trajectories (of near future) given a set of few initial poses (or sometime referred as frames). However, these models do not generalize well while generating long-duration motion trajectories across multiple activity classes due to following inherent limitations. First, typically these auto-regressive mod-

*Indicates equal contribution

els are fed with temporally redundant poses and thus their Markovian dependency assumption fails to exploit the long-duration dependencies among poses. Second, the model only learns the temporally forward dependency on short-term sequences (again with temporally redundant poses) and hence miss to exploit the temporally backward long-term dependencies in poses. Third, the majority of these methods do not attempt the conditional generation across a large class of activities. This is probably because there could be a significant amount of partial overlap of short-term pose trajectories across multiple activity classes. Thus, modeling the long-term pose dependency is critical for learning a generalized model.

Recently, graph convolution networks (GCN), that are traditionally used in an action recognition task, are employed to synthesize human motion sequence. GCN based methods [32, 33] model intra-frame (joint level spatial graph) and inter-frame (frame level temporal graph) relations as one spatio-temporal graph for every sequence and perform graph convolution. However, these methods also have multiple limitations that are discussed in detail in Section 2.

This paper aims to overcome the limitations of existing methods and synthesize a long-term human motion trajectory across a large variety of human activity classes ($> 50$). We propose a two-stage activity generation method to achieve this goal, where the first stage deals with learning the long-term global pose dependencies in activity sequences by learning to synthesize a sparse motion trajectory while the second stage addresses the generation of dense motion trajectories taking the output of the first stage.

We demonstrate the superiority of the proposed method over SOTA methods using various quantitative evaluation metrics on publicly available datasets [15, 25, 1], where our method generalizes well even on $60$ activity classes. As shown in Figure 1a, our method is capable of generating the different types of activities based on input class labels and in Figure 1b we demonstrate the transition between Standing Up and Drinking activity. Following are the key contributions of our work:

- We propose a novel two-stage deep learning method to synthesize long-term ($> 6000$ ms) dense human motion trajectories.
- Our method is capable of generating class-aware motion trajectories. The proposed GloGen embed the sparse activity sequences into a lower dimensional discriminative subspace enabling generalization to a large number of activity classes.
- Proposed method can generate a new motion trajectory as a temporal sequence of multiple activity types.
- Proposed method can control the pace of generated activities, thereby enabling the generation of variable speed motion trajectories of the same activity type.
- To the best of our knowledge, our method first time demonstrates the generalization ability of any long-term ($> 6000$ ms) motion trajectory synthesis method over $60$ activity classes.

## 2. Related Work

Traditional methods [19, 24, 17, 7] used graph-based modeling of poses for motion trajectory synthesis. Majority of the recent deep learning methods aimed at short or medium-term motion synthesis and that limited to a single or small set of activity classes. [14] used foot and ground contact information to synthesize locomotion tasks over a given trajectory using a convolutional autoencoder. However, the proposed approach is limited to the locomotion task only and cannot synthesize any other type of action. In [32], the authors proposed a method to generate human motion using a graph convolution network.

RNN based approaches have performed well for action recognition, as shown in [20]. Several researchers followed a similar direction to solve the task of human motion synthesis and proposed approaches based on RNNs. Kundu et al. [18] proposed a method for the task of human motion synthesis using an LSTM autoencoder setup. The proposed network encodes and then decodes back a given motion but is not capable of generating any novel human motion. In [10], the authors proposed an approach to generate human motion using the LSTM autoencoder setup. In [12] authors proposed a variational autoencoder setup to generate human motion. In [23] the network is trained on multiple actions, but they didn't provide any way to control the type of output motion trajectory.

There has been a significant increase in applications and performance of generative models with the arrival of GAN [11]. Generative adversarial networks were originally proposed to generate images and later on for videos. Recent methods attempted to synthesize better human motion by incorporating GANs with RNNs in Seq2Seq autoencoders. In [16] Kiasari et al. proposed a method to generate human motion using labels starting poses and a random vector to synthesize human motion, but they did not provide any quantitative results in the paper, and qualitative analysis is also unsatisfactory. In [3], the authors proposed an approach to generate human motion using GAN.

Recent GCN-based method [33] models a sequence as a spatio-temporal graph and perform class conditioned graph convolution. However, their fixed size graph modeling limits their scalability to generate long-term sequences. More importantly, the size of the frame sequence that can be considered for learning the temporal dependencies across frames/poses is shown to be relatively small. Additionally,

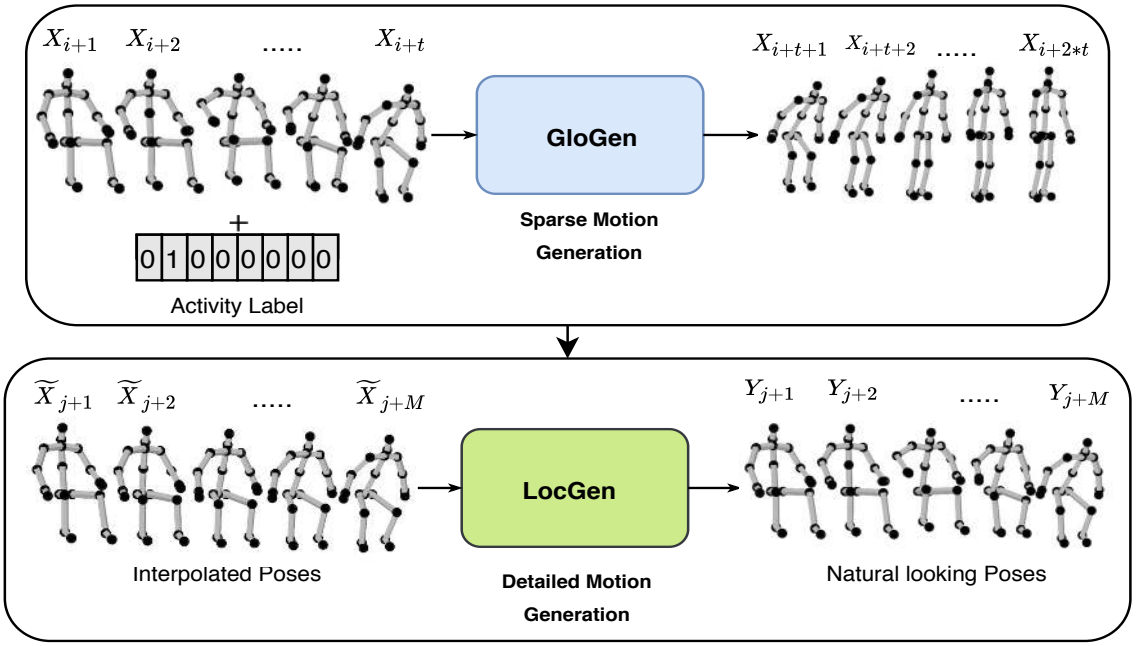


Figure 2: Overview of our two-stage framework, GlocalNet. In the first stage, GloGen generates the sparse motion trajectory of an activity, followed by the second stage, LocGen, that predicts the dense poses from the generated sparse motion.

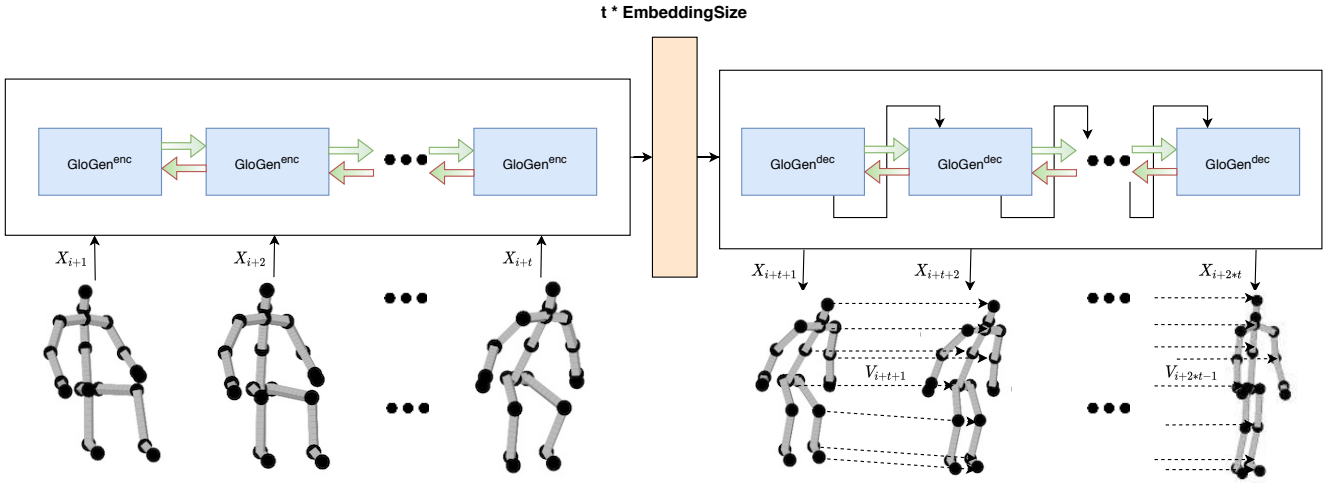


Figure 3: Architecture of GloGen network used as sparse motion trajectory generator.

since their method takes random noise as input, it lacks control using the initial state of the activity and hence is not capable of transitioning between two actions as done by our method in Figure 1b. Similarly, one can not synthesize a long duration motion sequence by repeatedly invoking their fixed length GCN generator. Another similar work in [32] proposed to synthesize very long-term sequences but fails to model class conditioning in their generative model, which is an essential aspect of motion synthesis.

## 3. Our Method: GlocalNet

Our novel two-stage human motion synthesis method attempts to address the key challenges associated with the task of long-term human motion trajectory synthesis across a large number of activity classes. More precisely, we aim to learn the long-term temporal dependencies among

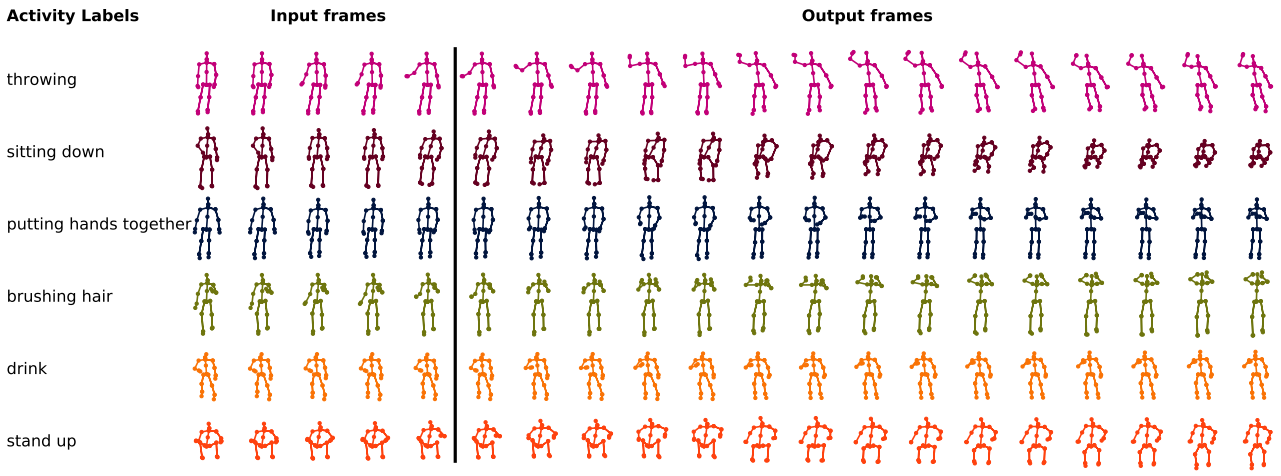


Figure 4: Output of GloGen using different activity labels and initial poses.

poses, cyclic repetition across poses, bi-directional, and multi-scale dependencies among poses. Additionally, our method attempts to incorporate class priors in the generation process to learn a discriminatory embedding space for motion trajectories, thereby addressing the generalisability aspect over a large class of human activities.

**Two Stage Motion Synthesis**

The key limitation of the existing temporal auto-regressive models like Seq2Seq is the Markovian dependency assumption, where a new set of poses is assumed to be depending upon just a few preceding poses. This impairs their capability to capture the long-term dependence among poses that are far apart and thus led to an accumulation of the prediction error (e.g., mean joint error) while attempting iterative prediction of long-term motion trajectories. We propose to overcome this limitation by splitting the process into two stages, where the first stage is employed to capture the global dependence among poses by learning temporal models on sparsely sampled poses instead of original dense motion trajectories. Thus, the second stage can subsequently deal with the generation of more detailed motion trajectories starting from sparse motion trajectories synthesized by the first stage. This also enables the additional capability to control the frame rate of the synthesized motion trajectories.

The other key drawback of the Markovian model is its incompetence to exploit the temporally backward dependencies in poses. Thus, we propose to employ the bi-directional LSTMs in the first stage to overcome this limitation. Finally, existing methods fail to generalize the motion synthesis for a large class of activity types, probably because of significant overlap among motion trajectories across multiple classes. We propose to overcome this limitation by employing a conditional generator (with class prior) in the first stage itself (while generating sparse global motion trajectories).

Such decoupling enables the first stage to learn the class-specific long-term (bi-directional) pose dependence while the second stage primarily focuses on the generation of class agnostic fine-grained dense motion trajectories given the sparse output trajectories from the first stage. Figure 2 outlines the overview of our proposed two-stage method.

### 3.1. First Stage: GloGen

The first stage is implemented as auto-regressive Seq2Seq network equipped with bi-directional LSTMs called *G*loGen, shown in Figure 3. The GloGen encoder takes as input a sequence of a sparse set of $t$ initial poses $\{X_1, X_2...X_t\}$ that are uniformly sampled from input motion trajectory during training. Here each pose $X_i$ depicts a fixed dimensional vectorial representation of the human pose. These poses are then concatenated with the action class priors encoded as one-hot vectors and fed to the encoder. Unlike traditional Seq2Seq models, we feed all the output states of the encoder i.e., $\{H_1, H_2...H_t\}$ as input to the GloGen decoder instead of just the last state. The rationale behind this choice is that all hidden states jointly capture the sparse input poses' global embedding. Finally, the decoder output is considered as the set of predicted $t$ number of poses. These predicted poses are used as input to synthesize the next set of $t$ iteratively to generate the sparse global motion.

$$H_{i+1}, H_{i+2}...H_{i+t} = GloGenEncoder(X_{i+1}, X_{i+2}...X_{i+t}) \quad (1)$$

$$X_{i+t+1}, X_{i+t+2}...X_{i+2t} = GloGenDecoder(H_{i+1}, H_{i+2}...H_{i+t}) \quad (2)$$

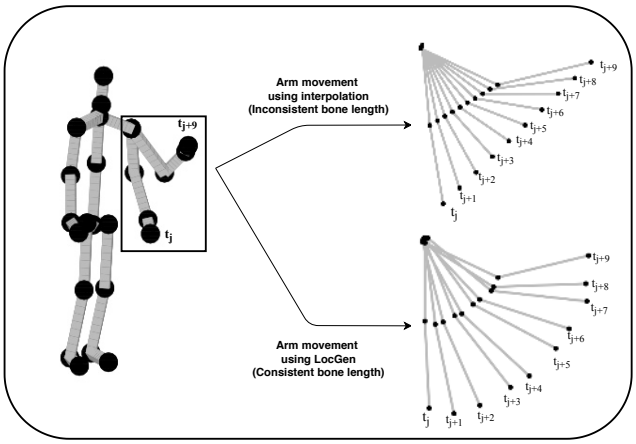


Figure 5: Comparison of linear interpolation v/s LocGen based generation of dense motion trajectories.

### 3.2. Second Stage: LocGen

Once we predict the sparse motion trajectories from GloGen, we need to process them further to obtain dense motion trajectories as the predicted pose will be far apart in pose space and hence would lack the temporal smoothness behavior. One option to obtain a dense set of poses from sparse-poses is to perform simple interpolation based upsampling in Euclidean representation of poses. However, from Figure 5, we can infer that simple interpolation is not a good option as it leads to unnatural motion trajectories. This is because the intermediate poses provided by the interpolation typically yield straight lines due to which fix bone length constraint is violated frequently, and the motion does not seem natural. Interpolation in Euler angle space is an alternate option that do not violate bone-length constraint. However, such representation of skeleton has issue that even small error in joint angles near root of kinematic tree results in large error in the joint locations for other dependent joints, while doing interpolation. Thus, we stick to Euclidean $[x, y, z]$ representation of joints in this work but other representations can also be considered.

| Method | LocGen ↓ | Interpolation ↓ |
|---|---|---|
| Vae Seq2Seq | 0.222 | 0.230 |
| Seq2Seq [23] | 0.214 | 0.223 |
| att. Seq2Seq [28] | 0.336 | 0.352 |
| acLSTM [21] | 0.328 | 0.355 |
| Our Method | **0.172** | **0.177** |

Table 1: Comparison of Our Method GlocalNet (GloGen + LocGen) in terms of Euclidean Distance per frame on NTU RGB+D(3D) dataset. LocGen majorly contributes to the qualitative results rather than quantitative.

We propose to obtain dense motion trajectories using another auto-regressive network named *L*ocGen, shown in Figure 2. Input to LocGen encoder is a set of (Euclidean) interpolated poses. The encoder first embeds the human pose into a higher dimension and then fed the hidden states to the decoder (similar to GloGen), generating more natural motion trajectories. LocGen has the same architecture as GloGen except that instead of sparse motion poses, LocGen takes interpolated dense motion trajectories as input, and there is no class prior concatenated with input poses. Thus, LocGen learns to transform interpolated trajectories into natural looking temporally smooth motion trajectories.

In order to generate interpolated poses between given two sparse-poses generated by GloGen, we use the following formulation. Let $M$ be the number of interpolated poses that need to be synthesized between two given sparse-poses $X_i$ and $X_{i+1}$. Let $\widetilde{X}_j$ be the $j$-th interpolated pose for $1 \leq j \leq M$, then we can compute $\widetilde{X}_j$ as:

$$\widetilde{X}_j = \alpha_j * X_i + (1 - \alpha_j) * X_{i+1} \tag{3}$$

where $\alpha_j = j/M$.

$\{\widetilde{X}_{j+1}, \widetilde{X}_{j+2} \ ... \ \widetilde{X}_{j+M}\}$ are given as input to the LocGen which first embeds them into the higher dimension and then use the embeddings to generate natural looking poses$\{Y_{j+1}, Y_{j+2} \ ... \ Y_{j+M}\}$.

$$Y_{j+1}, Y_{j+2}...Y_{j+M} = LocGen(\widetilde{X}_{j+1}, \widetilde{X}_{j+2}...\widetilde{X}_{j+M}) \tag{4}$$

## 4. Experiments & Results

Every model is trained individually from scratch using same setting in Table 1. All of the trained models, code, and data shall be made publicly available, along with a working demo. Please refer to our supplementary material for an extended set of video results.

### 4.1. Datasets

**Human 3.6M [15]:** Following the same pre-processing procedure as in [30], we down-sampled 50 Hz video frames to 16 Hz to obtain better representative and larger variation 2D human motions. The skeletons consist of 15 major body joints, which are represented in 2D. We consider ten distinctive classes of actions in our experiments, that includes sitting down, walking, direction, discussion, sitting, phoning, eating, posing, greeting, and smoking.

**NTU RGB+D(3D) [25]:** This dataset contains around 56,000 samples on 60 classes performed by 40 subjects and recorded with 3 different cameras. Hence, it provides

| Models | cross-view | | cross-subject | |
|---|---|---|---|---|
| | MMD$_{avg}\downarrow$ | MMD$_{seq}\downarrow$ | MMD$_{avg}\downarrow$ | MMD$_{seq}\downarrow$ |
| SkeletonVAE [12] | 1.079 | 1.205 | 0.992 | 1.136 |
| SkeletonGAN [6] | 0.999 | 1.311 | 0.698 | 0.788 |
| c-SkeletonGAN [30] | 0.371 | 0.398 | 0.338 | 0.402 |
| SA-GCN [33] | 0.316 | 0.335 | 0.285 | 0.299 |
| Our Method ($L_J$) | 0.213 | 0.218 | 0.201 | 0.212 |
| Our Method ($L_{MF}$) | 0.646 | 0.647 | 0.601 | 0.625 |
| **Our Method ($L_J + L_{MF}$)** | **0.195** | **0.197** | **0.177** | **0.187** |

Table 2: Comparison of Our Method (GloGen) in terms of MMD on NTU RGB+D(2D).

a good benchmark to test 3D human motion synthesis. We have used the available Cross-Subject split provided by the dataset for our experiments. We resort to standard pre-processing steps adopted by existing methods [18].

**NTU RGB+D(2D) [25]:** To compare with previous works [33], we follow the same setting to obtain 2D coordinates of 25 body joints and consider the same ten classes to run experiments. We use the available Cross-View and Cross-Subject splits.

**CMU Dataset [1]:** The dataset is given as sequences of the 3D skeleton with 57 joints. We evaluate our method on three distinct classes from the CMU motion capture database, namely, martial arts, Indian dance, and walking similar to [21].

### 4.2. Implementation Details

**Network Training:** We use Nvidia's GTX 1080Ti, with 11GB of VRAM to train our models. For training GLoGen, the output dimension of our Encoder is 200. We are using 1 layered Bi-LSTM as our Encoder as well as Decoder. Dropout regularization with a 0.25 discard probability, was used for the layers. We use the AdamW optimizer [22] with an initial learning rate of 0.002, to get optimal performance

| Models | MMD$_{avg}\downarrow$ | MMD$_{seq}\downarrow$ |
|---|---|---|
| E2E [31] | 0.991 | 0.805 |
| EPVA [31] | 0.996 | 0.806 |
| adv-EPVA [31] | 0.977 | 0.792 |
| SkeletonVAE [12] | 0.452 | 0.467 |
| SkeletonGAN [6] | 0.419 | 0.436 |
| c-SkeletonGAN [30] | 0.195 | 0.218 |
| SA-GCN [33] | 0.146 | 0.134 |
| Our Method | **0.103** | **0.102** |

Table 3: Comparison of Our Method (GloGen) in terms of MMD on Human 3.6M.

on our setup. We use MSE loss to calculate our objective function. Similar to [33], we set the predicted action sequence length for Human 3.6M and NTU RGB+D(2D) datasets to be 50 and input sequence length to be 10. We set the batch size for training to be 100, for testing to be 1000. For datasets CMU and NTU RGB+D(3D), a batch size of 64 is used. For training on NTU RGB+D(3D) with all 60 classes, we use input action sequence length to be 5 and predicted sequence length of sparse poses to be 15 for GloGen and then using LocGen, we generate 4 new poses for every pair of adjacent sparse-poses.

**Loss Function:** Loss function is calculated on joint locations and motion flow. We use the following loss function to train out network $L_J$ and $L_{MF}$.

$$L = (\lambda_1 * L_J) + (\lambda_2 * L_{MF}) \tag{5}$$

The joint loss $L_J$ in Equation 6 gives the vertex-wise Euclidean distance between the predicted joints $X_i$ and ground truth joints $\hat{X_{i+1}}$.

$$L_J = \sum_{i=1}^{t} ||X[i] - \hat{X[i]}||_2 \tag{6}$$

In order to enforce smoothness in temporal sequence, we minimize the motion flow loss $L_{MF}$ defined in Equation 7, which gives the Euclidean distance between the predicted motion flow $V_i$ and ground truth motion flow $\hat{V_{i+1}}$.

$$L_{MF} = \sum_{i=1}^{t-1} ||V[i] - \hat{V[i]}||_2 \tag{7}$$

Where, motion flow for the i$^{\text{th}}$ frame $\hat{V_{i+1}}$. is the difference between joint locations $\hat{X_{i+1}}$ and $\hat{X}_i$.

$$\hat{V}_i = \hat{X_{i+1}} - \hat{X}_i \tag{8}$$

### 4.3. Evaluation Metrics

**Maximum Mean Discrepancy:** The metric is based on a two-sample test to measure the discrepancy of two distributions based on their samples. The metric has been used in

| Method | 80ms | 160ms | 240ms | 320ms | 400ms | 480ms | 560ms | 640ms |
|---|---|---|---|---|---|---|---|---|
| **Walking** | | | | | | | | |
| acLSTM [21] | 1.05 | 1.77 | 2.20 | 2.46 | 2.66 | 2.79 | 2.99 | 3.24 |
| Scheduled Sampling [4] | 0.42 | 0.56 | 0.71 | 0.83 | 0.93 | 0.99 | 1.02 | 1.05 |
| Seq2Seq [23] | **0.09** | **0.13** | **0.24** | **0.42** | 0.74 | 1.22 | 1.85 | 2.79 |
| Our Method | 0.36 | 0.47 | 0.52 | 0.60 | **0.62** | **0.65** | **0.71** | **0.82** |
| **Indian Dance** | | | | | | | | |
| acLSTM [21] | 0.685 | 0.99 | 1.22 | 1.53 | 1.89 | 2.08 | 2.27 | 2.55 |
| Scheduled Sampling [4] | 1.54 | 2.24 | 2.49 | 2.52 | 2.65 | 2.90 | 2.94 | 3.12 |
| Seq2Seq [23] | **0.49** | 0.79 | 1.48 | 2.95 | 5.41 | 8.88 | 13.29 | 18.73 |
| Our Method | 0.50 | **0.56** | **0.64** | **0.68** | **0.69** | **0.69** | **0.79** | **0.84** |
| **Martial Arts** | | | | | | | | |
| acLSTM [21] | 0.52 | 0.74 | 0.95 | 1.14 | 1.35 | 1.56 | 1.73 | 1.88 |
| Scheduled Sampling [4] | 0.63 | 0.86 | 0.91 | 0.98 | 1.07 | 1.12 | 1.20 | 1.28 |
| Seq2Seq [23] | **0.28** | **0.43** | 0.87 | 1.57 | 2.53 | 3.89 | 5.83 | 8.62 |
| Our Method | 0.40 | **0.43** | **0.47** | **0.52** | **0.55** | **0.59** | **0.67** | **0.71** |

Table 4: Comparison of Our Method (GloGen) in terms of Euclidean Distance per frame on CMU dataset.

[29, 30, 33] for measuring the quality of action sequences by evaluating the similarity between generated actions and the ground truth. Similar to [30], for calculating MMD on motion dynamics which are in the form of sequential data points, the average MMD over each frame is denoted by $\text{MMD}_{avg}$ and MMD over whole sequences are denoted by $\text{MMD}_{seq}$.

**Euclidean distance:** This metric used in [21] calculates error as the euclidean distance from the ground truth for the corresponding frame.

### 4.4. Results

**Long-term Dense Motion Synthesis:** We use GlocalNet to generate long-term dense motion sequences. Table 1 shows the results on NTU RGB+D(3D) for dense motion trajectory synthesis and compare it with existing methods. All the methods were trained from scratch using the same data pre-processing [18] and have the same input(Class Label & Initial Poses). These quantitative results show the superior performance of the GlocalNet. Additionally, we report detailed results including long term motion ($> 6000$ ms) and class-wise performance in the supplementary material. We can clearly infer that our proposed solution outperforms all the existing methods. Figure 4 depicts the synthesized sparse motion trajectories obtained using the GlocalNet on NTU RGB+D(3D) dataset for six different activity classes. As we can see from the figure, the network is able to learn the global long-term temporal dependence in poses successfully across multiple classes and thus generate significantly different motion trajectories for similar initial input poses.

**Comparison with Short-term Motion Synthesis Models:** To compare with existing short-term motion synthesis models on different datasets, we use the first stage of our network(GloGen). For fair comparison, we follow the same settings as followed in these methods. Table 2 contains the quantitative results on NTU RGB+D(2D) and our method outperforms others with a good margin. Table 3 shows the results on Human 3.6M for GloGen, which outputs sparse-motion trajectory and compare with SOTA methods. These quantitative results suggest the superior performance of the GloGen over the MMD metric. Additionally, as shown in Table 4 for CMU Dataset, we report superior performance of our method over the existing ones on Euclidean per frame metric. As reported in the table, our method shows consistent performance even for longer sequences across different actions.

**Ablation Study on Loss Functions:** In order to show the importance of the proposed $L_J$ and $L_{MF}$ loss separately, we also trained our network using the individual loss components and reported the results in Table 2. As it is clearly visible, $L_{MF}$ alone is not sufficient; in combination with $L_J$ it helps improve the performance of our method. In terms of qualitative results, we observed jitters in the generated sequence (without having $L_{MF}$). Thus, $L_{MF}$ enables the network to learn generating smoother transition in skeleton sequences.

**Synthesis for Sequence of Activities:** Our network can also be used to generate a multi-activity motion trajectory by temporally varying the activity prior. To achieve this, we first synthesize the motion trajectories using the approach described in Section 3. Then we treat the final $t$ poses of

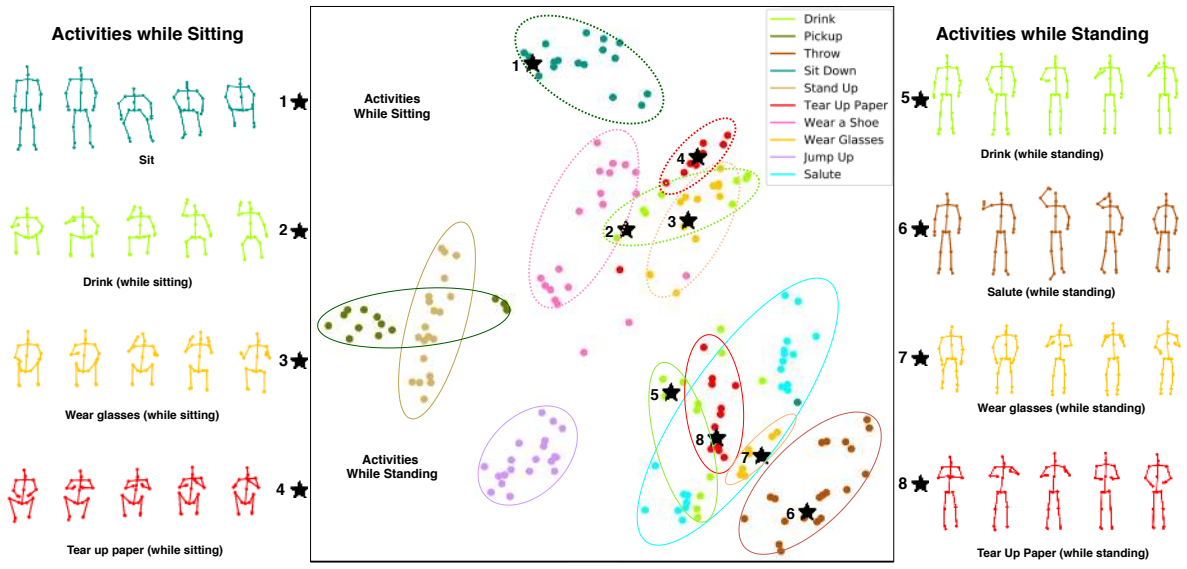


Figure 6: The t-SNE plot of GloGen embedding subspace along with the plot of selected motion trajectories where multiple samples for different classes are represented as color-coded 3D points.

the generated trajectory as the initial $t$ poses for generating the next set of $t$ poses belonging to new action class by providing the one-hot vector for the new class prior. This process is repeated to generate a new sequence with potentially multiple activity classes, in a single synthesized sequence of arbitrary length.

Figure 1b shows an example of a sparse motion trajectory where we generate poses for Stand Up activity and then use its last set of poses to generate Drink activity. Here, we can clearly visualize a smooth transition of poses across the two classes of activities.

## 5. Discussion

A major limitation of the Seq2Seq models class is that the last encoder hidden state becomes the bottleneck of the network as all the information at the input side passes through it to reach the decoder. To deal with this problem, attention architecture was proposed [28], where all the encoder hidden states are given to the decoder along with affinity scores that tell the importance of every input state corresponding to every output state. Such attention enabled Seq2Seq networks to achieve SOTA performance for the task of machine translation. However, generating motion is a different task from machine translation as we aim to predict the future poses looking at the previous ones, while modeling the long-term global dependency in far away poses. Therefore, in our method, instead of giving only the last state, we share the outputs of all states from the encoder to decoder LSTM units and predict the future poses.

**GloGen Embedding Subspace:** In order to visualize the behavior of feature embeddings, we concatenate the pose embeddings of GloGen-encoder over a sequence and project it as a point into 2D space using t-SNE. Figure 6 shows the t-SNE plot of embedding subspace along with the skeleton representation of selected motion trajectories where multiple samples for different classes are represented as color-coded 2D points. We can clearly infer from this figure that proposed GloGen projects these sequences into a discriminative subspace that enables it to handle the synthesis of different classes better. Interestingly, we can also see that some sequences from a few activities are scattered across two clusters as they can be performed while both sitting or standing, e.g., Wear glasses and Drink. Nevertheless, apart from a few outlier points due to the noisy samples present in the NTU RGB+D(3D) dataset, this plot clearly indicates the subspace's class discriminative nature.

## 6. Conclusion

In this paper, we propose a novel two-stage method for synthesizing long-term human-motion trajectories across a large variety of activity types. The proposed method can also generate new motion trajectories as a combination of multiple activity types as well as allows us to control the pace of generated activities. We demonstrate the superiority of the proposed method over SOTA methods using various quantitative evaluation metrics on publicly available datasets.

## References


[1] *CMU Graphics Lab Motion Capture Database*. http://mocap.cs.cmu.edu/.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[3] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. In *Computer Vision and Pattern Recognition Workshops*, pages 1418–1427, 2018.

[4] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Neural Information Processing Systems*, pages 1171–1179, 2015.

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

[6] Haoye Cai, Chunyan Bai, Yu-Wing Tai, and Chi-Keung Tang. Deep video generation, prediction and completion of human action sequences. In *European Conference on Computer Vision*, pages 366–382, 2018.

[7] Dan Casas, Margara Tejera, Jean-Yves Guillemaut, and Adrian Hilton. 4d parametric motion graphs for interactive animation. In *ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, pages 103–110, 2012.

[8] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5933–5942, 2019.

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

[10] Partha Ghosh, Jie Song, Emre Aksan, and Otmar Hilliges. Learning human motion models for long-term predictions. In *International Conference on 3D Vision (3DV)*, pages 458–466. IEEE, 2017.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Neural Information Processing Systems*, pages 2672–2680, 2014.

[12] Ikhsanul Habibie, Daniel Holden, Jonathan Schwarz, Joe Yearsley, and Taku Komura. A recurrent variational autoencoder for human motion synthesis. In *British Machine Vision Conference*, 2017.

[13] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[14] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics*, 35(4):1–11, 2016.

[15] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2013.

[16] Mohammad Ahangar Kiasari, Dennis Singh Moirangthem, and Minho Lee. Human action generation with generative adversarial networks. *arXiv preprint arXiv:1805.10416*, 2018.

[17] Lucas Kovar, Michael Gleicher, and Frédéric Pighin. Motion graphs. In *ACM SIGGRAPH 2008 classes*, pages 1–10. 2008.

[18] Jogendra Nath Kundu, Maharshi Gor, Phani Krishna Uppala, and Venkatesh Babu Radhakrishnan. Unsupervised feature learning of human actions as trajectories in pose embedding manifold. In *Winter Conference on Applications of Computer Vision*, pages 1459–1467. IEEE, 2019.

[19] Chan-Su Lee and Ahmed Elgammal. Human motion synthesis by motion manifold learning and motion primitive segmentation. In *International Conference on Articulated Motion and Deformable Objects*, pages 464–473. Springer, 2006.

[20] Shuai Li, Wanqing Li, Chris Cook, Ce Zhu, and Yanbo Gao. Independently recurrent neural network (indrnn): Building a longer and deeper rnn. In *Computer Vision and Pattern Recognition*, pages 5457–5466, 2018.

[21] Zimo Li, Yi Zhou, Shuangjiu Xiao, Chong He, Zeng Huang, and Hao Li. Auto-conditioned recurrent networks for extended complex human motion synthesis. *arXiv preprint arXiv:1707.05363*, 2017.

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[23] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2891–2900, 2017.

[24] Jesus Martinez-del Rincon, Michal Lewandowski, Jean-Christophe Nebel, and Dimitrios Makris. Generalized laplacian eigenmaps for modeling and tracking human motions. *IEEE transactions on cybernetics*, 44(9):1646–1660, 2013.

[25] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Computer Vision and Pattern Recognition*, pages 1010–1019, 2016.

[26] Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, and Tieniu Tan. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In *Computer Vision and Pattern Recognition*, pages 1227–1236, 2019.

[27] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Neural information Processing Systems*, pages 3104–3112, 2014.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, pages 5998–6008, 2017.

[29] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *International Conference on Computer Vision*, pages 3332–3341, 2017.

[30] Zhenyi Wang, Ping Yu, Yang Zhao, Ruiyi Zhang, Yufan Zhou, Junsong Yuan, and Changyou Chen. Learning diverse stochastic human-action generators by learning smooth latent transitions. *arXiv preprint arXiv:1912.10150*, 2019.

[31] Nevan Wichers, Ruben Villegas, Dumitru Erhan, and Honglak Lee. Hierarchical long-term video prediction without supervision. *arXiv preprint arXiv:1806.04768*, 2018.

[32] Sijie Yan, Zhizhong Li, Yuanjun Xiong, Huahan Yan, and Dahua Lin. Convolutional sequence generation for skeleton-based action synthesis. In *International Conference on Computer Vision*, pages 4394–4402, 2019.

[33] Ping Yu, Yang Zhao, Chunyuan Li, and Changyou Chen. Structure-aware human-action generation. In *European Conference on Computer Vision*, 2020.