

MUGL: Large Scale Multi Person Conditional Action Generation with Locomotion

Shubh Maheshwari
maheshwarishubh98@gmail.com
CVIT, IIIT Hyderabad
Hyderabad 500032, INDIA

Debtanu Gupta*
debtanu.gupta@research.iiit.ac.in
CVIT, IIIT Hyderabad
Hyderabad 500032, INDIA

Ravi Kiran Sarvadevabhatla
ravi.kiran@iiit.ac.in
CVIT, IIIT Hyderabad
Hyderabad 500032, INDIA

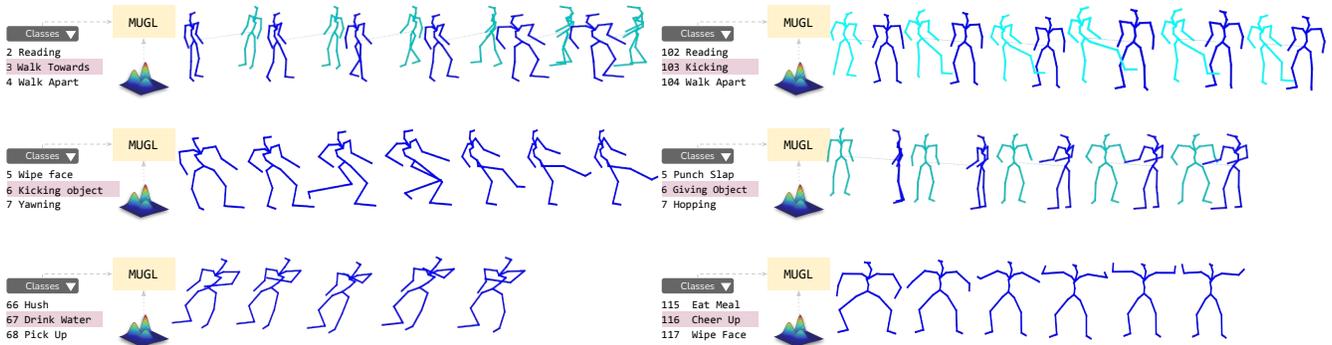


Figure 1: Examples of single and multi-person human pose action sequences generated by our model, MUGL. The user can select from among a large number of action classes (120). The changing length of dotted line connecting kinematic tree root joints of individuals in two-person actions indicates change in their relative global positions. Note the distinction from single person actions where individual frames correspond to timesteps, i.e. global displacements are absent. Also note the variable length of sequences generated. Additional results in Fig. 3.

ABSTRACT

We introduce MUGL, a novel deep neural model for large-scale, diverse generation of single and multi-person pose-based action sequences with locomotion. Our controllable approach enables variable-length generations customizable by action category, across more than 100 categories. To enable intra/inter-category diversity, we model the latent generative space using a Conditional Gaussian Mixture Variational Autoencoder. To enable realistic generation of actions involving locomotion, we decouple local pose and global trajectory components of the action sequence. We incorporate duration-aware feature representations to enable variable-length sequence generation. We use a hybrid pose sequence representation with 3D pose sequences sourced from videos and 3D Kinect-based sequences of NTU-RGBD-120. To enable principled comparison of generation quality, we employ suitably modified strong baselines during evaluation. Although smaller and simpler compared to baselines, MUGL provides better quality generations, paving the way for practical and controllable large-scale human action generation.

1 INTRODUCTION

The ability to synthesize novel and diverse human motion at scale while satisfying structural and kinematic body constraints has significant applications in animation, robotics and human-object interaction. Over the years, various approaches have been proposed

including physics based simulation [31], key-framing [23] and data-driven methods [34, 35]. Apart from pixel-based video data, the availability of reliable motion capture systems [9, 20, 39] has enabled pose-based human action synthesis [19, 25, 30]. Unlike videos, pose-based action representations and data have the significant advantage of decoupling the action-centric aspects from distracting or privacy-violating aspects (e.g. entity appearance, background).

However, pose-based human action synthesis is very challenging because of spatiotemporal correlation arising from body joint movement and the inter/intra category variety among actions. The challenge is further amplified by the presence of multiple actors. Due to the resulting complexity and lack of large-scale datasets [9, 20, 33], many approaches focus on synthesis of small action sets of fixed duration and simple, single-person actions [19, 25, 30].

In this backdrop, the availability of the large-scale NTU-RGBD dataset [27] is a promising development. The dataset contains 120 diverse single, multi-person activities performed indoors by a large subject pool and captured from multiple viewpoints using Kinect RGBD cameras. However, a fundamental bottleneck exists at the raw data stage itself. For a number of instances and action classes (e.g. ‘handshake’, ‘pushing’), the Kinect-based 3D joint data is temporally incoherent and noisy. Naïvely using this data for training would cause generative models to learn sub-optimal representations and result in unrealistic, poor-quality generations. Given recent advances, utilizing 3D human pose estimated directly from RGB videos is a promising alternative which addresses some issues mentioned above. Therefore, we source 3D human poses from RGB

*Equal contribution

videos across all the 120 action classes of the NTU-RGBD dataset. In doing so, we obtain a large-scale, good quality hybrid pose sequence representation which inherits advantages of the original viz. access to global trajectory (from Kinect-based sequences), large number of action classes, diversity in viewpoints and variety in action dynamics. In addition, the controlled indoor setting enables a principled comparison of generative model quality.

Despite the availability of large-scale datasets such as NTU-RGBD-120, existing approaches have not shown results beyond a small number of categories and single-person actions [13, 38]. The absence of locomotion with respect to global frame makes these methods unproductive for actions such as ‘walking’ or ‘running’. Also, actions involving locomotion and multi-person interaction (e.g. ‘kicking’) cannot be modelled properly. In addition, existing approaches generate fixed-length sequences causing certain generated actions (e.g. ‘throw’, ‘salute’, ‘take off headphone’) to appear unnatural.

To address these shortcomings, we propose MUGL, a novel generative model which enables large-scale diverse generation of single and multi-person human actions of *variable duration, with locomotion*. Notably, we accomplish controllable generation with a single, unified model. Our contributions are summarized below:

- We propose MUGL, a novel efficient deep network for large-scale controllable generation of multi-person human action sequences with variability in action duration and with locomotion (Sec. 4).
- We introduce a hybrid pose sequence representation with 3D pose sequences sourced from videos and 3D Kinect-based sequences of NTU-RGBD-120 (Sec. 5.1).
- MUGL outperforms strong baselines and generates visibly more realistic sequences for **all 120 categories** of NTU-120 (Sec. 5.4).

Additional details can be found in our project page skeleton.iit.ac.in/mugl.

2 RELATED WORK

Action prediction: A small initial action sequence is used to condition the generation of the full version in action prediction task. One set of approaches employ a sequence-to-sequence paradigm to predict joints [6, 11] or joint velocities [29]. Other approaches use adversarial generative models [5, 24] and conditioned autoregressive models [19] to synthesize human leg motion. Gao et al. [14] also use an autoregressive setup involving disjoint part grouping of skeleton joints. However, they train separate models for each action category. To produce sequences, Battan et al. [6] use a two stage approach involving sparse and dense motion prediction. In general, however, end-to-end action synthesis is more challenging due to the absence of input priming used for action prediction.

Traditional Action Generation: Some early works model motion sequences as a directed graph of pose subsequences sampled from motion capture corpus [3, 4, 23]. These methods are storage-intensive and generalization is limited to producing memorized sequences. Another set of approaches use Restricted Boltzmann Machines (RBM) [17] for probabilistic human motion generation [34, 35]. Though possessing the advantage of learning a probabilistic distribution, these methods involve conditioning on

proper choice of initial state which inhibits scaling to complex, large number of activities.

Modern Action Generation: Peng et al. [31] use reinforcement learning to create realistic motion clips imitating a broad range including locomotion acrobatics. Peng et al. [32] further extend the earlier work using deep pose estimation to incorporate publicly available RGB videos. However, these works mainly focus on using a single motion clip for action sequence generation. Yan et al. [37] propose to generate the action sequence using a graph convolution based GAN model with generation conditioned on a latent vector sampled from a Gaussian process. Yu et al. [38] propose a novel self-attention based GCN method with category conditioning of a GAN network. However, the focus is on generation of 2-D skeletons and only 10 action categories are considered. Guo et al. [13] propose a conditional VAE and also use RGB-based skeletons. However, their generations are confined to a small number of single-person action categories (13).

In a slightly different paradigm, some approaches use natural language action descriptions to condition the generation process. Chaitanya et al. [2]’s approach involves learning a joint embedding space for both action and language representation. However, the model generates a single sample for a given action description, i.e. stochasticity is absent. Xiao et al. [26] adopt a similar approach but within a GAN-based setting which enables stochastic sequence generation. Hyemin et al. [1] propose an attention-based sequence to sequence generator confined to upper body parts.

Action sequence representation: Many methods use joint 3-D positions to represent human skeleton pose [1, 2, 14, 37]. But this representation has a disadvantage of non-constrained bone length and motion beyond normal articulation range. To utilize the advantage of constrained bone length, some methods represent pose using joint rotations expressed as quaternions [30]. The discontinuous nature of these representations, however, results in sub-optimal pose embeddings. To overcome this, we employ a continuous 6-D representation for joint rotation (Sec. 3.1).

Action sequence embedding: Since human actions can be viewed as a sequence of human poses, many approaches [2, 14] use Recurrent Neural Networks(RNN) to obtain action sequence embeddings. However, RNN-based methods often fail to model temporal structures at multiple scales and in generating long motion sequences. A recent group of approaches view the action as a spatiotemporal graph and apply graph convolutions to represent the same [37, 38]. However, these methods are computationally expensive, even for a small number of action classes. We use Convolutional Neural Networks(CNN) in our approach which are computationally efficient for representing the spatiotemporal structure of the action sequence.

Locomotion: Estimation of locomotion is crucial for tasks such as multi-person interaction, object interaction, path planning. Cao et al. [7] use 2D pose histories and a scene image to plan a path towards reaching each goal. Habibie et al. [15] propose a VAE conditioned on a user-supplied control signal (desired trajectory). Similarly Holden et. al. [18] use a CNN autoencoder to create a low dimension representation and user-provided instructions to edit and synthesize long range motion. These approaches do not allow generation controllable by user at action category level. Phase functional neural networks [19] use cyclic functions to compute

weights of the network during inference for high quality, real time character control. These methods are generally confined to only locomotion classes such as walking, running and terrain traversal and are unable to scale to large number of categories, including non-locomotion action classes.

Overall, unlike our work, existing methods do not cover a large number of action classes, do not produce variable-length sequences and do not generate multi-person activity sequences with relative global displacements.

3 PRELIMINARIES

Problem formulation: An action sequence \mathcal{X} is associated with a class label $c \in C$ and can involve up to P individuals for up to T timesteps. Formally, we denote the sequence as $\mathcal{X} = \{[X^{(1)}, X^{(2)}, \dots, X^{(P)}]_t\}$, $1 \leq p \leq P$ and $1 \leq t \leq T$. $[X^{(i)}]_t$ represents the J -joint pose configuration of i -th person at time step t within the global coordinate frame. Note that the number of people involved in an action (P) and number of time steps can vary across action classes. Our goal is a model which stochastically generates a variable length action sequence \mathcal{X} conditioned on class label c .

In our setting, the number of action classes $|C|$ in NTU dataset is 120, maximum number of people in actions P is 2, number of joints J is 24 and timesteps T is 64.

3.1 Action Sequence Representation

Apart from actions performed in-place (without locomotion), our dataset includes actions involving locomotion and relative global motion involving multiple people. To accommodate these kinds of actions, we consider each action sequence \mathcal{X} to have a decoupled global component and a local component as described below.

Global component (\mathcal{X}_g): This contains global trajectory sequence of the root node associated with per-timestep kinematic pose tree on a per-person basis. Instead of maintaining independent per-person global trajectories, we model relative displacements of individuals with respect to the first person (see ‘Global component’ in Fig. 2). Let the first person’s root node global trajectory be denoted as $G^{(1)} = [g_1, g_2, \dots]$ where $g_i \in \mathbb{R}^3$. Let the relative displacement sequence for the j -th person’s root node ($1 < j \leq P$) be $D^{(j)} = [d_1, d_2, \dots]$ where $d_i \in \mathbb{R}^3$. Thus, the global trajectory for j -th person’s root node is given by $G^{(j)} = G^{(1)} + D^{(j)}$. Note that $G^{(1)}$ and $D^{(j)}$, $1 < j \leq P$ together comprise the global component \mathcal{X}_g .

Local component (\mathcal{X}_l): This comprises of per-person kinematic pose tree sequences whose per-timestep kinematic tree root nodes are all considered to be at the global origin (see Fig. 2). Representing the pose tree with local 3-D position based joint representations seems a straightforward choice. However, in practice, this generates unnatural-looking motion arising from non-constrained bone length and joint movement beyond normal articulation range. To overcome these issues, we use a fixed reference pose and forward kinematics to model the pose tree’s relative joint displacements as a rotation matrix. Using the procedure of Zhou et.al. [40] (purple shaded box in Fig. 2), we adopt a continuous 6-D representation for the rotation matrix. This choice enables us to bypass restrictive post-processing required to maintain orthogonality of the usual 3×3 joint rotation matrix. Consequently, we model the local pose sequences comprising the action as $\mathcal{X}_l = \{[X_l^{(1)}, X_l^{(2)}, \dots, X_l^{(P)}]_t\}$

where $[X_l^{(i)}]_t \in \mathbb{R}^{J \times 6}$, i.e. 6-D rotation representation of J joints, $1 \leq i \leq P$. For sequences with a single person, the reference sequence is duplicated P times for consistent processing.

4 OUR APPROACH (MUGL)

MUGL comprises of three encoder modules to represent the local component of action sequence, the corresponding global component and the sequence length. Corresponding to these modules, three decoder counterpart modules exist (see Fig. 2). We first provide a brief overview of our latent generative framework. We subsequently describe the previously mentioned modules (Sec. 4.2, 4.4) and the associated latent representation (Sec. 4.3).

4.1 Conditional Gaussian Mixture VAE

In a traditional Variational Auto Encoder (VAE)[21] generative model, we maximize the so-called variational lower bound (ELBO) i.e. $\mathcal{L}(\theta, \phi; \mathcal{X}) = \mathbb{E}_{q_\phi(z|\mathcal{X})}[\log p_\theta(\mathcal{X}|z)] - D_{KL}(q_\phi(z|\mathcal{X})||p(z))$, where $q_\phi(z|\mathcal{X})$ encodes variational approximation of the latent surrogate z ’s distribution conditioned on input data \mathcal{X} while $p_\theta(\mathcal{X}|z)$ approximates the latent-conditioned data likelihood. D_{KL} stands for KL-divergence. A popular choice for latent prior distribution $p(z)$ is the unimodal standard Gaussian, i.e. $p(z) = \mathcal{N}(0, I)$ because it enables a closed-form solution for computing KL-divergence. However, the unimodality can become a capacity bottleneck especially when data is expected to contain a large number of clusters due to attribute-based (e.g. category, action dynamics) similarity among samples. Therefore, we adopt an extension of VAE known as Gaussian Mixture VAE (GMVAE) [10] wherein the latent prior is a Mixture of Gaussian. Furthermore, we use a conditioning strategy to incorporate additional action attribute information as part of the inference and generation process. We dub the resulting model C-GMVAE. In this framework, the encoder models $q_\phi(y, z|\mathcal{X}, \mathbf{a})$ while the decoder models $p_\theta(\mathcal{X}|y, z, \mathbf{a})$ where \mathbf{a} represents conditioning information (action class c and viewpoint v). y denotes the weight distribution over Gaussian components of the mixture. The ELBO in this case can be written as:

$$\mathcal{L}(\theta, \phi, \mathbf{a}; \mathcal{X}) = \mathbb{E}_{q_\phi(y, z|\mathcal{X}, \mathbf{a})}[\log p_\theta(\mathcal{X}|y, z, \mathbf{a}) - D_{KL}(q_\phi(y, z|\mathcal{X}, \mathbf{a})||p(y, z|\mathbf{a}))] \quad (1)$$

4.2 Encoder Modules

Recall that we represent action sequence \mathcal{X} in terms of global component \mathcal{X}_g and local component \mathcal{X}_l (Sec. 3.1).

Local Pose Encoder: The local component $\mathcal{X}_l \in \mathbb{R}^{T \times 6 \times (J \times P)}$ is processed in two phases. The first phase (Spatial Pose Encoder) contains a single 2D residual CNN block and learns features for each timestep in \mathcal{X}_l . The second phase (Temporal Encoder) applies multiple 2D residual CNN convolutions to downsample the timesteps T . Finally, the output from the Temporal Encoder is flattened to produce the spatio-temporal local features f_l (see Fig. 2).

Global Trajectory Encoder: The global component $\mathcal{X}_g \in \mathbb{R}^{T \times 3 \times P}$ is downsampled via a series of 1D convolutions across the timestep dimension. The result is transformed by a linear layer to obtain a flattened feature representation f_g . Note that the dimension P in \mathcal{X}_g arises from global trajectory for first person and $(P - 1)$ relative displacement sequences for the other $(P - 1)$ people involved in the action.

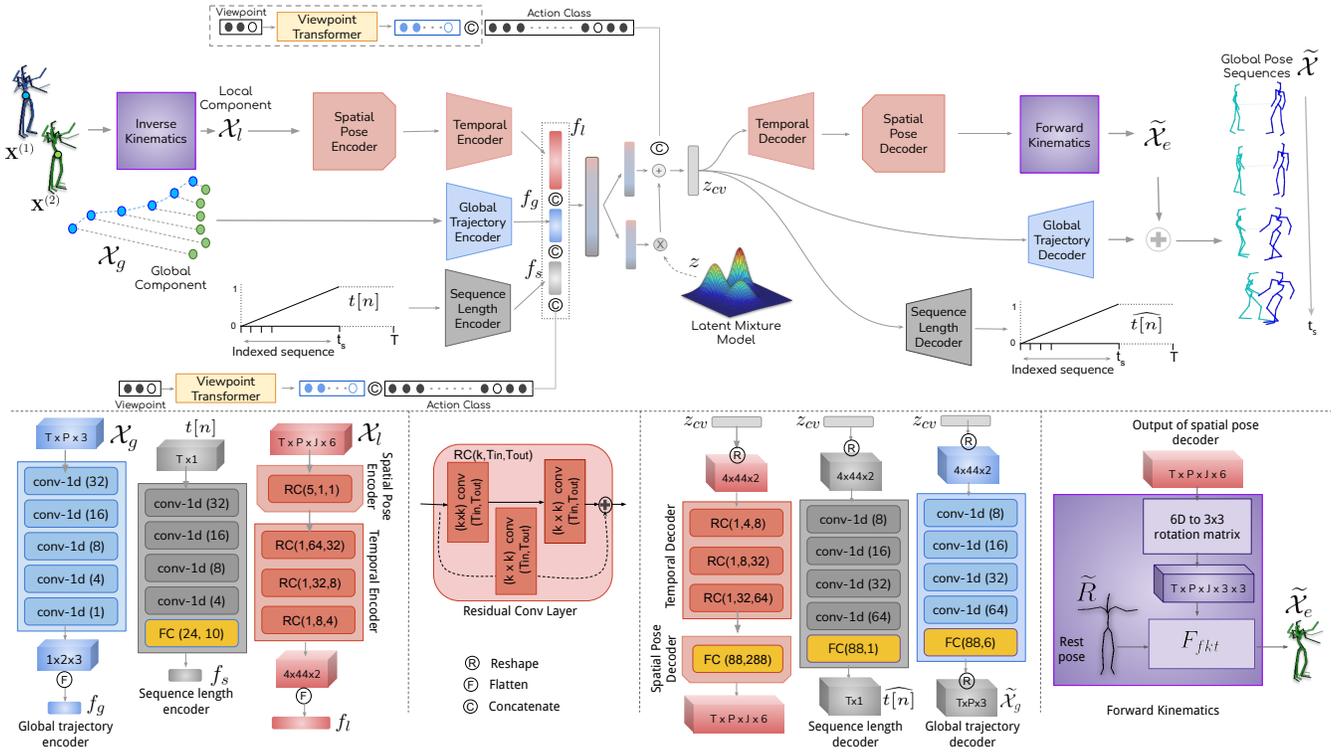


Figure 2: Architecture of MUGL. The green and blue circular markers at torso joint on $X^{(1)}$, $X^{(2)}$ indicate the shared origin of action’s local component. The dotted boundary around transformed viewpoint indicates that a fixed viewpoint is used during inference. During training, the input sequence’s viewpoint is utilized. Refer to Sec. 4 for details.

Sequence Length Encoder: Let t_s denote the length of a particular training sequence. To encode the sequence length, we first define an indexed sequence $t[n]$ of length T as follows:

$$t[n] = \begin{cases} \frac{n}{t_s-1} & \text{if } 0 \leq n < t_s. \\ 1 & \text{if } t_s \leq n < T. \end{cases}$$

where T is the maximum possible sequence length. Note that $t[n]$ is a normalized non-decreasing sequence of length T whose values lie in $[0, 1]$ (see Fig. 2). The Sequence Length Encoder transforms $t[n]$ to f_s via 1-D convolutions.

Encoder Representation: The local pose sequence representation f_l , global trajectory sequence representation f_g and sequence length representation f_s are then concatenated. During training, this concatenated representation is conditionally modulated with action class-label c (See Fig 2). In addition, we also condition on a transformed version of viewpoint v to enrich the encoded representation. The result is mapped to the parametric representations of the C-GMVAE’s variational approximation distribution $q_\phi(y, z|\mathcal{X}, \mathbf{a} = (c, v))$ (Equation 1).

4.3 Latent Representation

We first sample a latent vector z from the mixture of K Gaussian components (Sec. 4.1). Similar to the conditioning applied on the encoder representation, the latent vector z is conditioned on class label and transformed viewpoint, by concatenation (see Fig. 2). The resulting is transformed by a linear layer and reshaped to obtain z_{cv} .

Note that the training sequence’s viewpoint is used for conditioning only during training. During generation, the viewpoint is set to a fixed default value.

4.4 Decoders

Sequence Decoder: Complementary to encoding phase (Sec. 4.2), the Local Spatial and Temporal Decoders perform spatiotemporal upsampling. Starting from the conditioned latent representation z_{cv} , the local pose components of generated action sequence with 6-D rotation representations for each joint are obtained. Instead of transpose convolutions in the local Spatial Decoder (mirroring their Spatial Encoder counterparts), we use a multi-layer perceptron. Empirically, we found this to generate better quality sequences.

Forward Kinematics: Similar to [28] this module (shown as purple shaded box in Fig. 2) computes 3-D positions of each joint from the local representation output by Temporal Decoder. First, it converts the 6-D rotation representation into 3×3 rotation matrix corresponding to every joint, i.e. $\tilde{\mathcal{X}}_t = \{[X_t^{(1)}, X_t^{(2)}, \dots, X_t^{(p)}]_t\}$ where $X_t^{(i)} \in \mathbb{R}^{J \times 3 \times 3}$. The forward kinematic function then takes $\tilde{\mathcal{X}}_t$, vertices of rest pose $\tilde{\mathcal{R}}$ (see Fig. 2) and the kinematic tree as input. It applies transformation F_{fkt} which outputs 3-D positions of joints $\tilde{\mathcal{X}}_e = \{[X_e^{(1)}, X_e^{(2)}, \dots, X_e^{(p)}]_t\}$, where pose instance $X_e^{(i)} \in \mathbb{R}^{J \times 3}$. The transformation F_{fkt} is given as:

Model	Class conditioning	Support for multi-person actions	Variable sequence length generation	Input	# parameters	Inference time (sec)	Training time (hrs)
VAE-LSTM[15]	✗	✗	✗	3-D joints	1468K	0.03	45.14
CS-GCN[37]	✗	✗	✗	3-D joints	7744K	0.60	10.74
SA-GCN[38]	✓(10)	✗	✗	2-D joints	14502K	3.31	71.90
action2motion[13]	✓(13)	✗	✗	Lie Space	503K	3.31	40.27
MUGL (ours)	✓(120)	✓	✓	6-D[40]	922K	0.02	18.92

Table 1: A comparative summary of models used for evaluation.

# Classes	# Viewpoint	Frame rate	Sequence length			Hyperparameters				
			Max	Min	Avg	λ_{rot}	λ_{global}	λ_{len}	Batch size	θ_s
120	3	8.25	256	5	69.72	10	1	2	100	0.97

Table 2: Salient dataset attributes and hyperparameter choices.

$$F_{fkt}(X_l^{(i(c))}, \tilde{\mathcal{R}}) = \begin{cases} [0, 0, 0] & \text{if root joint.} \\ X_l^{(i(c))} \cdot (\tilde{\mathcal{R}}^c - \tilde{\mathcal{R}}^p) & \\ + F_{fkt}(X_l^{(i(p))}, \tilde{\mathcal{R}}) & \text{otherwise.} \end{cases}$$

where $X_l^{(i(c))}$ and $X_l^{(i(p))}$ indicate 3×3 rotation matrix of child and parent joints respectively in the kinematic tree of pose instance $X_l^{(i)}$. $\tilde{\mathcal{R}}^c$ and $\tilde{\mathcal{R}}^p$ indicate 3D joint position of child and parent joints respectively in the kinematic tree of rest pose $\tilde{\mathcal{R}}$. Finally, the 3-D joint positions of pose instance $X_e^{(i)} \in \mathbb{R}^{J \times 3}$ is given as, $X_e^{(i)} = [f_{fkt}(X_l^{(i(1))}, \tilde{\mathcal{R}}), f_{fkt}(X_l^{(i(2))}, \tilde{\mathcal{R}}), \dots, f_{fkt}(X_l^{(i(J))}, \tilde{\mathcal{R}})]$.

This process ensures consistency in the bone lengths throughout the sequence. Note that $\tilde{\mathcal{X}}_e$ represents local pose structure sequence, i.e. joint positions considering root joint of pose kinematic tree (torso) to be at the global origin for each time step. Also, the transformation from 6D space to 3×3 rotation matrix space is continuous and the forward kinematic function is differentiable with respect to 3-D rotation. This enables incorporation of $\tilde{\mathcal{X}}_j$ into the overall optimization procedure (Sec. 4.5).

Finally, note that the ‘Inverse Kinematics’ module in the encoder portion of MUGL performs the opposite of the process described above, i.e. it maps the 3-D joint based local pose sequence to the 6-D rotation based counterpart.

Global Trajectory Decoder: The conditioned latent sample z_{cv} is gradually upsampled via a series of 1-D CNN layers to a T timestep sequence. The resulting sequence is transformed by a linear layer to generate the global trajectory $\tilde{\mathcal{X}}_g \in \mathbb{R}^{T \times P \times 3}$ of the reference (first) person and relative displacements of the remaining $(P - 1)$ people involved in the activity (Sec. 4.2).

The local pose sequences $\tilde{\mathcal{X}}_e$ are appropriately offset using the information from global trajectory decoder $\tilde{\mathcal{X}}_g$ to obtain the final generated sequence $\tilde{\mathcal{X}}$ (see Fig. 2). Since our procedure uniformly generates sequences for P individuals, we consider only the first person’s sequence for single person action classes.

Sequence length Decoder: The latent representation z_{cv} is transformed via a linear layer and the resulting output is processed via 1-D convolutions to obtain a $\mathbb{R}^{T \times 1}$ sequence. This sequence is

transformed by a non-negative activation (ReLU). Subsequently, the cumulative sum sequence of the activation transformed sequence is obtained. This ensures that the resulting sequence is non-decreasing. The sequence elements are then transformed via a sigmoid activation to obtain $\widehat{t}[n]$, the decoder analogue of the indexed sequence (Eqn. 4.2). The location of first element in $\widehat{t}[n]$ that is greater than or equal to a fixed threshold θ_s is considered the length of the sequence \widehat{t}_s . More precisely, $\widehat{t}_s = 1 + \underset{j}{\operatorname{argmin}} [t[j] \geq \theta_s], 0 \leq j < T$.

4.5 Optimization

Optimizing the overall VAE-based framework requires a tradeoff between reconstruction loss for training data and distribution approximation loss for latent space (Equation 1). The Sequence Decoder generates joint rotation representation in 6-D space which is integrated with Forward Kinematics module to obtain 3-D joint coordinates. During optimization, the (MSE) loss for 6-D rotation representation space \mathcal{L}_{6D} and 3-D space \mathcal{L}_{3D} together comprise the local motion reconstruction loss $\mathcal{L}_{local}^{rec} = \lambda_{rot} \mathcal{L}_{6D} + \mathcal{L}_{3D}$ where λ_{rot} is a tradeoff hyperparameter. We combine $\mathcal{L}_{local}^{rec}$ with global trajectory (MSE) loss $\mathcal{L}_{global}^{rec}$, sequence length (MSE) loss \mathcal{L}_{len}^{rec} and the KL-divergence loss of GMVAE, \mathcal{L}_{KL} . The effective final loss function is defined as:

$$\mathcal{L} = (\mathcal{L}_{local}^{rec} + \lambda_{global} \mathcal{L}_{global}^{rec}) + \lambda_{len} \mathcal{L}_{len}^{rec} + \lambda_{KL} \mathcal{L}_{KL} \quad (2)$$

where λ_{rot} , λ_{global} , λ_{len} and λ_{KL} are hyperparameters. Note that the local and global losses are computed only for the original, non-padded extent of the training sequences.

4.6 Implementation

We train MUGL architecture using Adam optimizer with initial learning rate 0.015, integrated with a learning scheduler which decays the learning by 0.5 with step size 10 and train the model for 200 epochs. The value of hyper-parameters in the loss function (Equation 2) are empirically set and the hyper-parameter for KL-divergence λ_{KL} is determined by a cyclic annealing schedule [12].

Although the original frame rate of NTU-RGBD is 33.33 fps, such high resolution is not required for training. Therefore, we subsample the sequences by a factor of 4 (See Table 2).

During inference, we sample from the mixture of $K = |C|$ Gaussian components (i.e. K is same as number of action classes). The latent vector is conditioned on one-hot representations of desired action class. The resulting vector is fed in parallel to local and global components of the Sequence Decoder. The local component

generates an action sequence in the form of 6D rotation representation. The representation is transformed to its 3×3 counterpart and together with the reference rest pose, forward kinematics is applied to obtain the local pose. This is combined with the output from Global Trajectory Decoder module to obtain the initial action sequence. The effective sequence length obtained from Sequence Length Decoder (Sec. 4.4) is used to trim the initial action sequence to the final result. We perform class distribution based oversampling for minibatches to compensate for the relatively small number of action classes involving leg movement in our dataset. To match the ground truth data’s frame rate we increase MUGL’s output sequence resolution in the temporal dimension from 64 timesteps to 256 using bicubic interpolation.

We conduct all experiments on a machine with an Intel Xeon E5-2640 v4 and Nvidia GeForce GTX 1080 Ti 11GB GPUs with Ubuntu 16.04 OS. Our code uses Python-3.7 and PyTorch-0.4 library. See Table 2 for hyperparameter settings.

5 EXPERIMENTS

5.1 Hybrid Pose Sequence Representation

The original NTU dataset [27] contains Kinect-based 3D skeleton data with temporally incoherent, noisy and uncurated joint data. Naively using this data for training causes models to generate poor quality action sequences. Therefore, we use VIBE [22], a state of the art RGB-based 3D pose estimator to obtain skeleton sequences from video sequences of NTU-RGBD-120 dataset [27]. Unlike Kinect, VIBE utilizes the complete video to estimate 3D pose with minimal variance in bone lengths across the sequence. The obtained pose estimates are much more reliable due to additional context-based filtering. Since VIBE only provides local pose (See.3.1), we use corresponding samples from the original NTU-120 Kinect sequence to obtain the global trajectory of each subject. For multi-person activities, associating the per-person global trajectories to the VIBE-based pose counterparts is done based on similarity in each subject’s orientation.

We ultimately produce hybrid pose sequences spanning the 120 action classes and 114,480 samples. We preprocess sequences (padding, cropping) to ensure a uniform length of 256 timesteps. For training and evaluation, we follow the cross-setup protocol defined for NTU-RGBD-120 [27]. In this protocol, action sequences from half of the total camera setups are used for training and the sequences captured from the other half are used for evaluation.

5.2 Baselines

We compare MUGL with four baseline generative models, namely SA-GCN [38], action2motion [13], CS-GCN [37] and VAE-LSTM [15] (see Table 1). We use available implementations for the baselines except for CS-GCN [37], which we implement from scratch and extend by adding action class conditioning. We perform a similar implementation-based extension for VAE-LSTM [15].

5.3 Generation Quality Measures

To measure the quality of generated sequences, we use two variants of Maximum mean discrepancy (MMD) [36].

MMD-A: Maximum Mean Discrepancy captures similarity between generated and test set (ground-truth) sample distributions [6, 36, 38].

		NTU-VIBE-94 (Single person classes only)	
Model	MMD-A ↓	MMD-S ↓	
SA-GCN[38]	0.68 \pm 0.12	0.43 \pm 0.02	
action2motion[13]	0.57 \pm 0.11	0.52 \pm 0.03	
CS-GCN[37]	1.09 \pm 0.17	0.56 \pm 0.01	
VAE-LSTM[15]	1.11 \pm 0.17	0.54 \pm 0.01	
MUGL	0.34\pm0.12	0.17\pm0.01	
		NTU-VIBE-120	
MUGL (Multi person)	0.45\pm0.15	0.36\pm0.03	

Table 3: Model comparison in terms of generative quality scores on NTU-VIBE dataset. ↓ indicates that the motions are better when metric is lower. For fair comparison, all the baselines are trained on 94 single-person classes. Single person and multi-person variant of MUGL are not directly comparable but exhibit similar trends.

For our setting (MMD-A), the base similarity is measured on a per-timestep basis for sequence pairs g, e sampled from generated set G and test set E . Let $g_t \in \mathbb{R}^{J \times 3}$ and $e_t \in \mathbb{R}^{J \times 3}$ represent the t -th timestep poses of the sampled pair and having same action class. The base similarity (MMD-A) is computed as $\mathcal{K}(g_t, g_t) + \mathcal{K}(e_t, e_t) - 2\mathcal{K}(g_t, e_t)$ where \mathcal{K} is a similarity kernel. In particular, we employ the RBF kernel [8].

MMD-S: Unlike MMD-A, MMD-S is computed on the whole sequence. Let, g, e be sequences chosen from generated set G and test set E , where $g, e \in \mathbb{R}^{T \times J \times 3}$. We flatten g, e to get a vector representation of the whole sequence. MMD-S is computed as $\mathcal{K}(g, g) + \mathcal{K}(e, e) - 2\mathcal{K}(g, e)$.

Issues with feature-based generative quality measures: In existing works, measures such as Fréchet Inception Distance (FID) [16], Diversity score (DS) [13] and Multimodality score (MS) [13] are commonly reported. These measures rely on a pre-trained classifier’s feature representation of the input (real or generated). Most skeleton action classifiers employ a preprocessing step in which root joints of pose sequence are translated to origin and the sequence is transformed so that the person shoulder faces x-axis. Doing so eliminates variability arising from camera viewpoint and enables good classification performance. However, as Fig. 4 shows, the preprocessing severely distorts the original action’s dynamics. Removing preprocessing would seem a possible solution. However, doing so causes a significant drop in classifier performance (to less than 20%), making the resulting feature representations unreliable.

Therefore, generative quality scores (FID, MS, DS) obtained from feature representations of such preprocessed action sequences cannot be considered reliable for evaluating action sequences.

5.4 Results

To compute performance scores for a given model, we uniformly generate 300 samples per action class. Since baseline models are confined to single-person generation and have no provision for multi-person setting, we use a single-person variant of the dataset with 94 action classes for training and evaluation of all the models. The quantitative results can be viewed in Table 3. MUGL’s scores are visibly better than the baselines. Empirically, MMD scores (Table. 3)

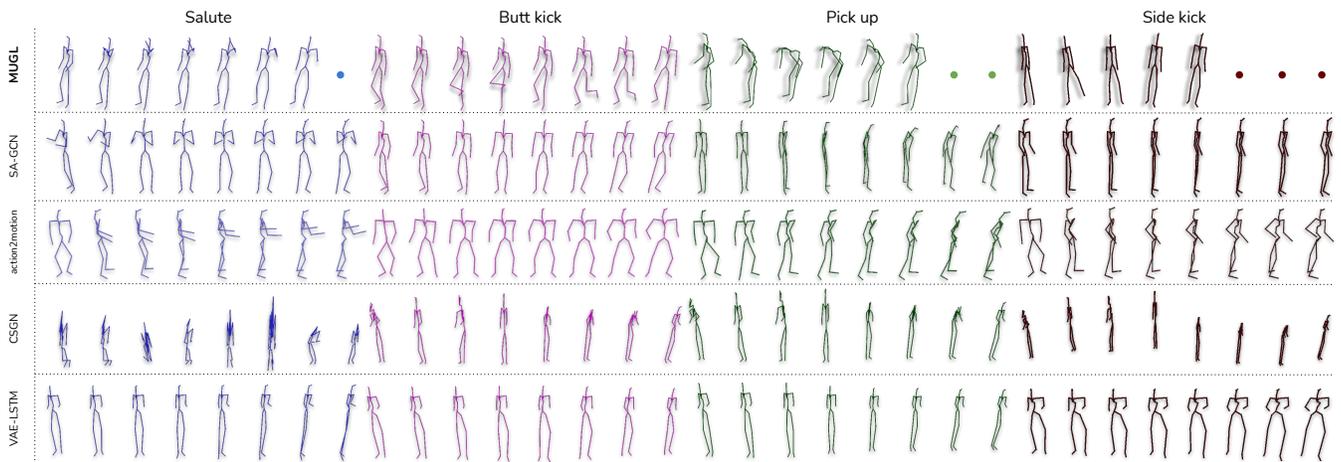


Figure 3: Comparing conditionally generated single-person action sequences across models. Also, note the variable sequence length of MUGL’s examples. (ref. Sec. 5.4).

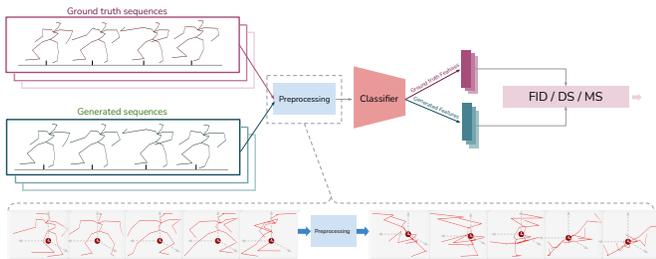


Figure 4: The typical pipeline for computing feature representation based generative quality scores is shown. The sequence at bottom right shows the effect of preprocessing. For e.g., even though rotation (about vertical axis) might be a signature of original action, we see that preprocessing can distort and eliminate such signature components. Quality score (e.g. FID) based on feature representations of such sequences fail to capture the key action dynamics. We empirically observed that these scores correlate poorly with visual quality of category-conditioned action generations.

correlate well with observed visual quality. The advantage of our approach can also be appreciated from a comparison of compute attributes for various models given in Table 1. We compute the quality measures for two-person action generation as well (see bottom row of Table 3). Since the model configurations (one-person and two-person) are different, the scores are not directly comparable. However, it can be seen that two-person generation scores are similar in range to the one-person variant.

Fig. 3 provides a qualitative illustration with generated samples from some of the action categories. The results, including Fig. 1, demonstrate the efficacy of our approach in generating diverse action-conditioned variable-length action sequences. This is especially apparent for actions involving significant limb movement.

To explore the variability of performance due to design choices (action representation, architectural components and optimization), we computed scores for ablative variants of MUGL trained on all 120 NTU-120 classes. From the results in Table 4, we observe the following: (i) performance is lower with vanilla VAE at sequence level (MMD-S) (ii) MLP in Spatial Decoder noticeably improves performance compared to the usual 2D Transpose Convolution (iii) class and viewpoint conditioning matters (iv) performance is lower with raw 3D joint representation (v) generating fixed length sequences (by removing variable sequence length encoder-decoder) negatively affects performance.

Overall, the results demonstrate the suitability of our design choices. The benefit of our decoupled local and global sequence modelling, the choice of CNNs for computational efficiency, importance of variable sequence estimation, the increased latent distribution capacity due to Gaussian Mixture model are all evident.

6 CONCLUSION

We have introduced MUGL, a novel deep neural model for variable-length pose-based action generation with locomotion. Our controllable method enables diverse generation of single and multi-person actions at scale. MUGL outperforms strong baselines qualitatively and quantitatively in terms of generated sequences. Our generated sequences can be used for rendering, e.g. using skeleton pose as α and a desired shape profile as β parameter configurations in SMPL system [28]. Going forward, we intend to explore MUGL’s performance on other human pose action datasets.

ACKNOWLEDGEMENTS

We acknowledge Qualcomm Innovation Fellowship’s support to Debanu Gupta and Google Cloud’s GPU credits for academic research.

REFERENCES

- [1] Hyemin Ahn, Timothy Ha, et al. 2018. Text2Action: Generative Adversarial Synthesis from Language to Action. In *ICRA*. 5915–5920.

Ablation Type	Pipeline Component	Ablation Details	MMD-A ↓	MMD-S ↓
Architectural	VAE (Sec. 4.1)	Reduce VAE latent dimensions (0.5x)	0.47	0.47
		Increase VAE latent dimensions (2x)	0.48	0.55
		Reduce GMM components ($K = 60$)	0.41	0.39
		Increase GMM components ($K = 240$)	0.47	0.39
		Vanilla VAE: Unimodal Gaussian ($K = 1$)	0.45	0.41
	Encoder & Decoder (Sec. 4.2,4.4)	MLP in Spatial Encoder	0.46	0.39
		Transpose 2D conv in Spatial Decoder	0.65	0.64
		No class, viewpoint conditioning in Sequence Encoder 3-D joint representation as input, output	0.48 0.46	0.41 0.45
	Viewpoint conditioning (Sec. 4.3)	Using one-hot vector/remove viewpoint transformer	0.49	0.41
No viewpoint conditioning		0.45	0.39	
No viewpoint conditioning in latent		0.79	0.75	
	Variable sequence length	Remove variable sequence length encoder-decoder	1.77	0.79
Optimization	Sequence Decoder (Sec. 4.4)	No 3D Loss	0.74	0.37
		No Rotation Loss	0.56	0.39
MUGL (multi-person)			0.45	0.36

Table 4: Performance scores for MUGL ablative variants.

- [2] C. Ahuja and L. Morency. 2019. Language2Pose: Natural Language Grounded Pose Forecasting. In *3DV*.
- [3] Okan Arıkan. 2003. Motion Synthesis from Annotations. *ACM SIGGRAPH* 22, 3 (2003), 402–408.
- [4] Okan Arıkan and D. A. Forsyth. 2002. Interactive Motion Generation from Examples. In *CGIT*. 483–490.
- [5] Emad Barsoum, John Kender, et al. 2017. HP-GAN: Probabilistic 3D human motion prediction via GAN. *ArXiv* (2017).
- [6] Neeraj Bhattan, Yudhik Agrawal, Sai Soorya Rao, Aman Goel, and Avinash Sharma. 2021. GlocalNet: Class-Aware Long-Term Human Motion Synthesis. In *WACV*. 879–888.
- [7] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qizhi Cai, Minh Vo, and Jitendra Malik. 2020. Long-term human motion prediction with scene context.
- [8] Yin-Wen Chang, Cho-Jui Hsieh, Kai-Wei Chang, Michael Ringgaard, and Chih-Jen Lin. 2010. Training and Testing Low-degree Polynomial Data Mappings via Linear SVM. *Journal of Machine Learning Research* 11, 48 (2010), 1471–1490. <http://jmlr.org/papers/v11/chang10a.html>
- [9] CMUMocap. 2003. CMU Graphics Lab Motion Capture Database Converted to FBX. (2003).
- [10] Nat Dilokthanakul, Pedro AM Mediano, et al. 2016. Deep Unsupervised Clustering with Gaussian Mixture Variational Autoencoders. *arXiv* (2016).
- [11] Katerina Fragkiadaki, Sergey Levine, et al. 2015. Recurrent network models for human dynamics. In *ICCV*. 4346–4354.
- [12] Hao Fu, Chunyuan Li, et al. 2019. Cyclical Annealing Schedule: A Simple Approach to Mitigating KL Vanishing. *arXiv* (2019).
- [13] Chuan Guo, Xinxin Zuo, et al. 2020. Action2Motion: Conditioned Generation of 3D Human Motions. *ACMMM* (2020).
- [14] Xiao Guo and Jongmoo Choi. 2019. Human motion prediction via learning local structure representations and temporal dependencies. In *AAAI*, Vol. 33. 2580–2587.
- [15] Ikhsanul Habibie, Daniel Holden, et al. 2017. A Recurrent Variational Autoencoder for Human Motion Synthesis. In *BMVC*.
- [16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Nash Equilibrium. *arXiv* (2017).
- [17] G.E. Hinton and R.R. Salakhutdinov. 2006. Reducing the Dimensionality of Data with Neural Networks. *Science* 313 (08 2006), 504–7.
- [18] Daniel Holden et al. 2016. A Deep Learning Framework for Character Motion Synthesis and Editing. *SIGGRAPH* 35, 4, Article 138 (2016).
- [19] Daniel Holden, Taku Komura, and Jun Saito. 2017. Phase-functioned neural networks for character control. *ACM ToG* 36 (07 2017), 1–13.
- [20] Catalin Ionescu, Dragos Papava, et al. 2014. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE TPAMI* 36, 7 (2014), 1325–1339.
- [21] Diederik P. Kingma and Max Welling. 2019. An Introduction to Variational Autoencoders. *arXiv* (2019).
- [22] Muhammed Kocabas, Nikos Athanasiou, et al. 2020. VIBE: Video Inference for Human Body Pose and Shape Estimation. In *CVPR*.
- [23] Lucas Kovar. 2002. Motion Graphs. *ACM SIGGRAPH* 21, 3 (2002), 473–482.
- [24] Jogendra Nath Kundu, Maharshi Gor, et al. 2019. BiHMP-GAN: Bidirectional 3D Human Motion Prediction GAN. *AAAI* (2019).
- [25] Zimo Li, Yi Zhou, et al. 2018. Auto-conditioned recurrent networks for extended complex human motion synthesis. In *ICLR*.
- [26] Xiao Lin and Mohamed R. Amer. 2018. Human Motion Modeling using DVGANs. (2018).
- [27] Jun Liu, Amir Shahroudy, et al. 2020. NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. *TPAMI* 42, 10 (2020), 2684–2701.
- [28] Matthew Loper, Naureen Mahmood, et al. 2015. SMPL: A Skinned Multi-Person Linear Model. *SIGGRAPH Asia* 34, 6 (2015), 248:1–248:16.
- [29] Julieta Martinez, Michael J Black, et al. 2017. On human motion prediction using recurrent neural networks. In *CVPR*. 2891–2900.
- [30] Dario Pavlo, David Grangier, et al. 2018. QuaterNet: A Quaternion-based Recurrent Model for Human Motion. In *BMVC*.
- [31] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. 2018. DeepMimic: Example-guided Deep Reinforcement Learning of Physics-based Character Skills. *ACM SIGGRAPH* 37, 4, Article 143 (2018), 143:1–143:14 pages.
- [32] Xue Bin Peng, Angjoo Kanazawa, Jitendra Malik, Pieter Abbeel, and Sergey Levine. 2018. SFV: Reinforcement Learning of Physical Skills from Videos. *ACM SIGGRAPH*, Article 178 (2018).
- [33] Matthias Plappert, Christian Mandery, and Tamim Asfour. 2016. The KIT Motion-Language Dataset. *Big Data* 4, 4 (2016), 236–252.
- [34] I. Sutskever and G. Hinton. 2007. Learning multilevel distributed representations for high-dimensional sequences. In *AISTATS*, Vol. 2.
- [35] Ilya Sutskever, Geoffrey E Hinton, et al. 2008. The Recurrent Temporal Restricted Boltzmann Machine. In *NIPS*. 1601–1608.
- [36] Ilya O Tolstikhin, Bharath K Sriperumbudur, and Bernhard Schölkopf. 2016. Minimax estimation of maximum mean discrepancy with radial kernels. *NIPS* (2016), 1938–1946.
- [37] Sijie Yan, Zhizhong Li, et al. 2019. Convolutional Sequence Generation for Skeleton-Based Action Synthesis. (2019).
- [38] Ping Yu, Yang Zhao, et al. 2020. Structure-aware human-action generation. In *ECCV*. 18–34.
- [39] Zhengyou Zhang. 2012. Microsoft Kinect Sensor and Its Effect. *IEEE MultiMedia* 19 (2012), 4–12.
- [40] Yi Zhou, Connelly Barnes, et al. 2019. On the continuity of rotation representations in neural networks. In *CVPR*. 5745–5753.