DSAG: A Scalable Deep Framework for Action-Conditioned Multi-Actor Full Body Motion Synthesis

Debtanu Gupta
debtanu.gupta@research.iiit.ac.in

Shubh Maheshwari
maheshwarishubh98@gmail.com

Sai Shashank Kalakonda
sai.shashank@research.iiit.ac.in

Manasvi Vaidyula
manasvi@students.iiit.ac.in

Ravi Kiran Sarvadevabhatla
ravi.kiran@iiit.ac.in
Centre for Visual Information Technology
IIIT Hyderabad, Hyderabad, INDIA 500032

Abstract

We introduce DSAG, a controllable deep neural framework for action-conditioned generation of full body multi-actor variable duration actions. To compensate for incompletely detailed finger joints in existing large-scale datasets, we introduce full body dataset variants with detailed finger joints. To overcome shortcomings in existing generative approaches, we introduce dedicated representations for encoding finger joints. We also introduce novel spatiotemporal transformation blocks with multi-head self attention and specialized temporal processing. The design choices enable generations for a large range in body joint counts (24 - 52)*, frame rates* (13 - 50)*, global body movement (in-place, locomotion) and action categories* (12 - 120)*, across multiple datasets (NTU-120, HumanAct12, UESTC, Human3.6M). Our experimental results demonstrate DSAG's significant improvements over state-of-the-art, its suitability for action-conditioned generation at scale.*

1. Introduction

A number of interesting approaches have been proposed in recent times for controllable synthesis of pose-based human motion. However, most have focused on generating fixed duration or single-person actions, often with a small number ($\leqslant$ 60) of action categories [9, 14, 18, 19, 25]. These approaches are generally trained using datasets sourced via 3D motion capture setups. However, Kinect sourced 3D action sequences of large-scale datasets [15, 11] often contain poorly estimated joints and exhibit temporal incoherence. This affects the quality of action sequences generated while scaling these models to these datasets. Obtaining 3D pose sequences from RGB videos has emerged as an alternative paradigm which addresses some of the issues mentioned above [6, 19]. Building upon this trend and seeking to overcome issues mentined above, our work makes multiple contributions.

- To address the lack of fine-grained joints in the large-scale NTU-RGBD dataset, we create and employ full body pose sequences with detailed finger joint representations (Sec. 4.1).
- We introduce several crucial improvements to state of the art approach for large-scale controllable action generation. Our framework DSAG contains 1) self-attention modules which improve quality for subtle motion actions and low frame rate datasets 2) specialized temporal processing modules which tackle high within-class action variance in training data 3) dedicated processing for finger joints at global and local level for improved realism (Sec. 3).
- DSAG noticeably outperforms existing approaches across datasets with varying frame rate, global body movement and action durations (Sec. 5).

Additional details can be found at skeleton.iiit.ac.in/dsag

2. Related Work

Human Motion Synthesis: Among recent deep-learning based generative approaches, Ping et al. [25] use a stochastically conditioned LSTM along with a novel GCN mod-

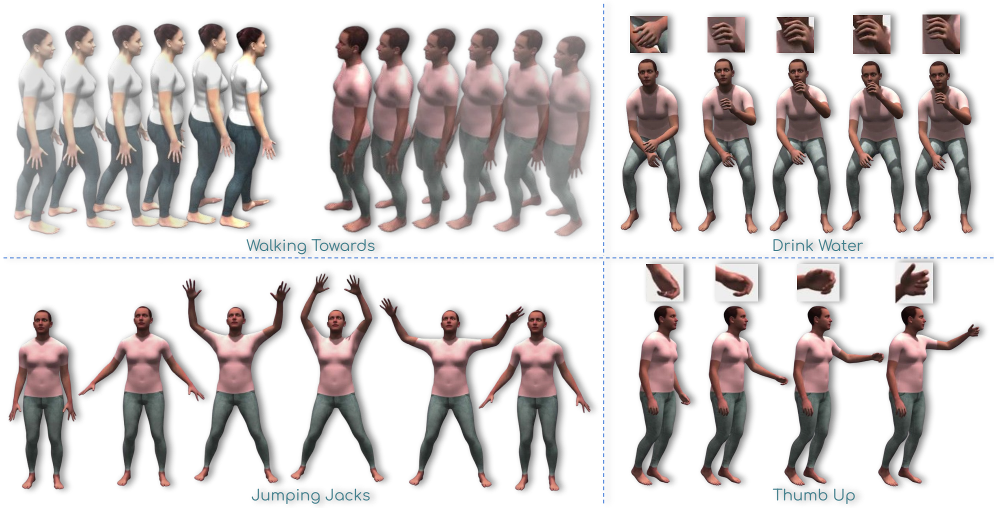

Figure 1: Renderings of action sequences generated by DSAG. Our model generates in-place and locomotory single/multi-actor variable duration *full body* sequences. It also scales across datasets containing a large range in frame-rate, joints and actions. The dotted square shows magnified detail of fingers.

ule to generate 2D skeleton sequences. Action2motion [5] uses a conditional VAE to generate 3D sequences. The ACTOR [19] framework uses a transformer based VAE to output pose parameters of the popular SMPL [16] human mesh model and generate actions. As with our approach, both action2motion and ACTOR use a rotation-based pose representation which provides stability due to the fixed bone length. However, training ACTOR is computationally expensive due to the dense nature of the mesh being optimized. Also, the aforementioned methods generate single actor sequences for a small number of categories whereas our approach can generate multi-person actions for large number of categories

Kinetic-GAN [3] builds upon CS-GCN [24] and employs a GAN-based latent mapping network to generate single actor sequences from a large number of action classes (94) of NTU-RGBD [15]. Unlike computationally expensive GCN based methods, MUGL [6] is a lightweight CNN-only architecture for large-scale, variable duration and multi-person action generation. Despite its success, MUGL outputs poor quality generations for full body joint representations and fails to generalize across datasets. By incorporating multi-headed self attention in our novel spatiotemporal block (Sec. 3.3.1), DSAG is better equipped at capturing variable frame rates and intra-class diversity of different datasets.

Full body pose sequences (with finger joints): Previous methods [19, 5, 6] do not exhibit realistic finger movement due to the insufficiently detailed fingers in raw pose representations they employ [13, 26, 16]. The introduction of full body expressive parametric model representations such as SMPL-X [17], ADAM [12] enable detailed finger joints. However, these approaches are computationally expensive. ExPose [2] provides a more efficient alternative, making it our method of choice for obtaining full body pose sequences. BABEL [20] contains finger joints. However the long tailed distribution of samples with respect to action classes makes it unsuitable for our problem. Existing approaches report results on coarse body pose models which lack detailed finger joints and fail to generate sequences of adequate quality on full body pose sequences. By decoupling body and hand components, our approach is more suited to generate good quality full body pose sequences.

3. Our approach (DSAG)

Problem formulation: We denote an action sequence associated with a class label $c \in \mathcal{C}$ as $\mathcal{X} = \{[X^{(1)}, X^{(2)}, \dots X^{(p)}]_t\}, 1 \leqslant p \leqslant P$ and $1 \leqslant t \leqslant T$. Here, P is the maximum possible actors and T is the maximum possible temporal action duration. $[X^{(i)}]_t$ is the J-joint pose configuration of i-th person at time step t. Note that the number of actors (P) and time steps (t) can vary within and across action classes. Our objective is to design a model which stochastically generates a variable duration action sequence $\mathcal{X}$ conditioned on class label c (see Fig. 2).

3.1. Action Sequence Representation

Instead of directly encoding the 3D pose sequence, we decouple its local and global components [6]. This enables dedicated representations for pose and locomotion dynamics of the action. To enhance representation quality especially for actions involving subtle finger movements, we further decouple the pose tree joints into hand-level joints (fingers) and rest of the body. In addition, we encode the number of action timesteps to enable variable-duration ac-

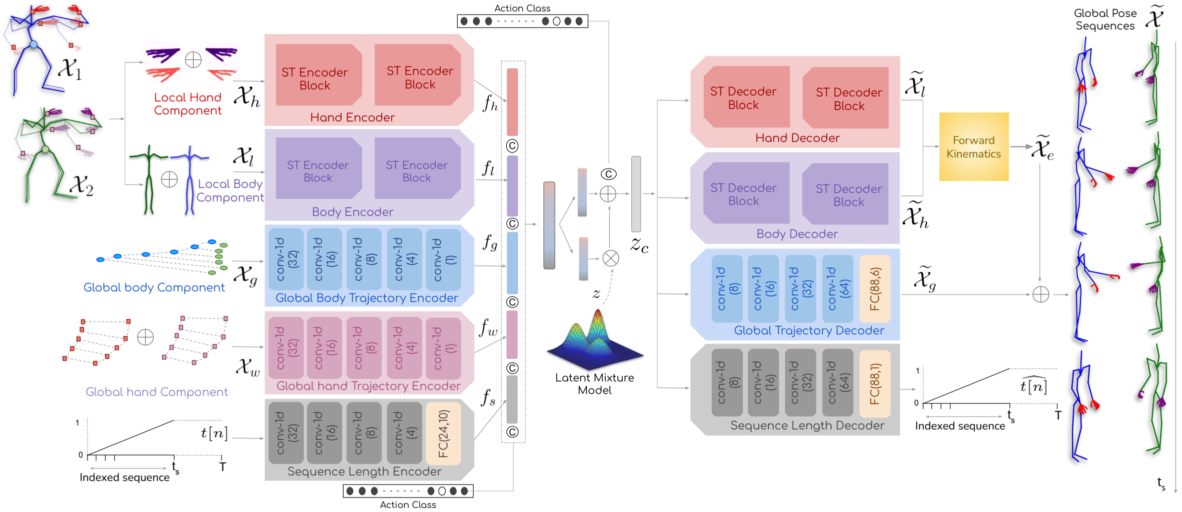

Figure 2: Architecture of DSAG showing dedicated action sequence components at local level (hand:$\mathcal{X}_h$ and body:$\mathcal{X}_l$), global level (hand:$\mathcal{X}_g$ and body:$\mathcal{X}_w$) and action duration $t[n]$. The local components are encoded using a series of novel ST-blocks (Fig. 3, Sec. 3.3.1). A series of 1D convolutions with swish [21] activation is used for encoding other components. The decoder components map the latent representation to the generated class-conditioned action sequence. $\mathcal{X}_1$ and $\mathcal{X}_2$ represent the actors. The blue and green dots at the torso of each actor indicate the shared origin of the action sequence's local component. The red and purple squares represent wrist joints' 3D coordinate global trajectories. Refer to Sec. 3 for details.

tion generation. We provide additional details on these components below.

Local body component ($\mathcal{X}_l$): This is obtained by translating the root joint of each time step's kinematic pose tree to the global origin (Fig. 2). The kinematic pose tree is represented using joint rotations, which is integrated via forward kinematics. This avoids problems such as unconstrained bone length and motion beyond articulation range. Each joint's rotation is represented as continuous 6D rotation [27]. We denote local pose sequences comprising the action as $\mathcal{X}_l = \{[X_l^{(1)}, X_l^{(2)}, \ldots X_l^{(p)}]_t\}$ where $[X_l^{(i)}]_t \in \mathbb{R}^{J \times 6}$, i.e. a 6-D rotation representation of J joints, $1 \leqslant t \leqslant T$. For single-person sequences, the reference sequence is duplicated P times for consistent processing. Also, finger joints are not included within the sequences. Instead, they are represented separately.

Global body component ($\mathcal{X}_g$): The global trajectory for the first actor is comprised of 3D root joint position sequence of the action. For other actors, it is represented using relative displacement of their respective root joints from the first actor's counterpart. Let the first actor's root node global trajectory be $G^{(1)} = [g_1, g_2, \ldots]$ where $g_i \in \mathbb{R}^3$. Let the relative displacement sequence for the j-th actor's root node ($1 < j \leqslant P$) be $D^{(j)} = [d_1, d_2, \ldots]$ where $d_i \in \mathbb{R}^3$. Thus, the global trajectory for j-th actor's root node is $G^{(j)} = G^{(1)} + D^{(j)}$. Note that $G^{(1)}$ and $D^{(j)}, 1 < j \leqslant P$ together comprise the global component $\mathcal{X}_g$.

Local hand component ($\mathcal{X}_h$): In the raw pose representation, fingers have more joints (30) than rest of the body (22). The finger joints tend to have a high degree of spatiotemporal correlation which is often not captured adequately by a monolithic pose representation. Also, finger joints have a lower degree of freedom compared to body joints. Therefore, despite the body having a relatively lower number of joints, gross body dynamics can potentially dominate the action representation, causing finger movements to be under-represented. To mitigate this effect, we introduce dedicated representations for hand joints (fingers). Similar to local body pose, we maintain a local *hand pose* representation $\mathcal{X}_h$. This consists of the wrist rooted finger joint kinematic tree of each hand translated to global origin. We denote this component as $\mathcal{X}_h = \{[X_h^{(1)}, X_h^{(2)}, \ldots X_h^{(p)}]_t\}$ where $[X_h^{(i)}]_t \in \mathbb{R}^{(2 \times J) \times 6}$, i.e. 6-D rotation representation of J joints of both the hands, $1 \leqslant t \leqslant T$ - see Fig. 2.

Global hand component ($\mathcal{X}_w$): Similar to global body trajectory $\mathcal{X}_g$ described previously, we maintain a global *hand* trajectory $\mathcal{X}_w$ comprising of the 3D wrist joint trajectories. The global 3D positions of two wrist joints (left wrist and right wrist) of all the actors are concatenated for each time step to obtain hand trajectory representation $\mathcal{X}_w \in \mathbb{R}^{T \times (2 \times P) \times 3}$.

Temporal Duration ($t[n]$): The temporal duration of an action is represented as a non decreasing sequence. Specifically, a sequence of length t_s is represented as:

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

, where T is the maximum possible sequence length. Note that $t[n]$ is a normalized non-decreasing sequence of length T whose values lie

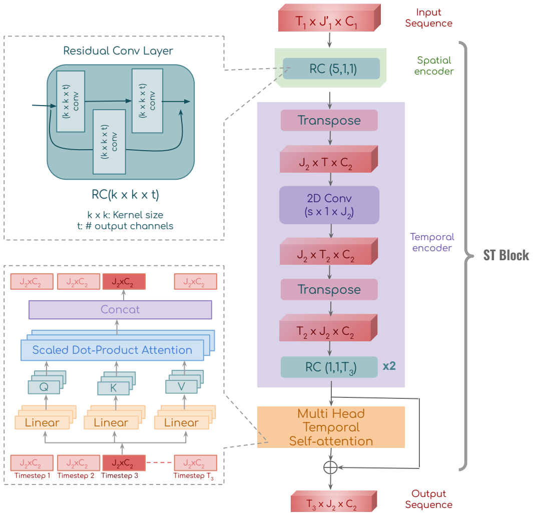

Figure 3: ST encoding block (see Sec. 3.3.1). Residual convolution (top left, shaded green) is applied to process the spatio-temporal information. Multi-head self-attention (bottom left, orange) is used to incorporate the global temporal dependency.

in the range $[0, 1]$ (see Fig. 2).

3.2. Conditional Generative Model

Our choice of generative model is an extended version of VAE known as Gaussian Mixture Variational Encoder (GM-VAE) [4], where the latent surrogate z is sampled from a Gaussian mixture model. The action class is used to condition the generation process (see Fig. 2).

Our approach modifies MUGL [6] which was originally designed for large-scale, variable duration and multi-person action generation. MUGL does not generalise across different datasets which contain actions captured at different frame rates and exhibit high within-class diversity. A chief reason is that MUGL's encoder and decoder blocks decouple spatial and temporal components. To mitigate this, we introduce SpatioTemporal (ST) Blocks (Sec. 3.3.1) which (i) help represent rapid localized movements involving hand or leg joints and low frame-rate action sequences better (ii) help capture subtle movements in joints. MUGL also copes poorly when additional finger joints are present. Our ST-Blocks, coupled with our newly introduced dedicated local and global representations for finger joints, overcome this shortcoming of MUGL as well. Next, we describe the architecture of our model's encoder and decoder modules.

3.3. Encoder Modules

The encoder modules consist of dedicated architectural blocks which map each of the action sequence representation components, e.g. Local body component, Global body component, Local hand component (Sec. 3.1), to corresponding feature representations.

3.3.1 Local Pose Encoders

The Local body component $\mathcal{X}_l$ is processed by a series of ST (spatiotemporal) blocks and then flattened to obtain the corresponding embedding f_l. Similarly, the Local hand pose sequence $\mathcal{X}_h$ is processed by a series of ST blocks and the flattened to obtain f_h.

ST (Spatiotemporal) Block: This block is a novel inclusion and provides two key capabilities: tackling large within-class diversity and enabling precise action representation for datasets captured at a very low frame rate. The block consists of three processing stages (Fig. 3). The first stage is a *spatial encoder* which takes the action sequence $\mathcal{X}_{in} \in \mathbb{R}^{T_1 \times P_1 \times J_1 \times C_1}$ as input. For the sake of simplicity we consider $P_1 \times J_1$ as a single entity, i.e., consider $J_1^{'} = P_1 \times J_1$. Next, the input $\mathcal{X}_{in} \in \mathbb{R}^{T_1 \times J_1^{'} \times C_1}$ is mapped to a more compact, lower dimensional feature representation $\mathcal{X}_s \in \mathbb{R}^{T_1 \times J_2 \times C_2}$ for each timestep individually, with $J_2 < J_1$ and $C_2 < C_1$. Note that the processing focuses on the spatial components, leaving the timestep dimension unchanged. The key idea here is to compress each timestep by removing redundant joint information to representation an activity.

The second stage is a *temporal encoder* which processes $\mathcal{X}_s \in \mathbb{R}^{T_1 \times J_2 \times C_2}$ from previous stage's output by focusing on the temporal channel. $\mathcal{X}_s$ is transposed and small local filters of dimension $s \times 1$ are applied at each joint dimension. Essentially, for each joint component, this procedure applies the convolution filter within a small temporal neighborhood and captures subtle temporal movements. This is beneficial for representing very subtle actions (e.g. *walking*) and for sequences with very high within-class diversity found in some datasets (e.g. Human3.6M [10]). The post-convolution result is once again transposed to have the same channel ordering as $\mathcal{X}_s$. Next, the transposed output is transformed via a series of 2D convolutions to reduce the dimensionality along the temporal channel.

The third stage is the *Multi-head Temporal Self-attention* (MHTS) block which helps preserve long term global dependencies within the action sequence. The multi-head attentional processing [23] within this block enables efficient capture of intra and inter timestep correlation. This aspect is especially crucial for rapid localized movements involving hand or leg joints (e.g.*walking*, *greeting*) and for low frame-rate action sequences. The MHTS block processes the feature representations across the temporal dimension as a feature sequence (Fig. 3) and outputs a richer version of the sequence $\mathcal{X}_{out} \in \mathbb{R}^{T_4 \times J_2 \times C_2}$. In conclusion, the second and the third stage of the ST-block learns both the local and global temporal dependencies to model complex human motion.

3.3.2 Global pose and temporal encoders

As with local pose representations, feature representations f_g, f_w are obtained corresponding to the global body component $\mathcal{X}_g$ and global hand component $\mathcal{X}_h$ of the action sequence (see Fig. 2). However, the processing blocks are conventional (1D convolutions), reflecting the simpler nature of the global components. Similarly, the variable duration representation of the action sequence $t[n]$ is subjected to similar processing to obtain the corresponding representation f_s.

Finally, the component wise representations (local hand encoding f_h, local body component f_l, global body trajectory encoding f_g, global hand trajectory encoding f_w, sequence length encoding f_s) are concatenated and conditionally modulated with action class label c. During training, the result is transformed via linear layers to generate the parameters of the variational approximation distribution. A latent vector z is sampled from a mixture of K gaussian components and conditionally modulated with class label c. This representation is transformed via a linear layer to obtain the conditioned latent vector z_c.

3.4. Decoder Modules

Local Decoders: The body decoders contain components which are symmetrically opposite to the counterparts in the encoder. Complementary to the ST (spatiotemporal) encoder block, the ST decoder block comprises of a multi head temporal self-attention module, temporal decoder and spatial decoder block. The local body decoder takes the conditioned latent representation z_c as input and transforms it via the spatiotemporal decoder blocks to generate the local body pose sequence $\widetilde{\mathcal{X}}_l$ (see Fig. 2). Similarly, a series of ST decoder blocks take z_c as input to generate the local hand pose sequence $\widetilde{\mathcal{X}}_h$. The local body pose and hand pose are concatenated to obtain the full body pose as a 6D rotation representation $\widetilde{\mathcal{X}}_e \in \{[\widetilde{X}^{(1)}, \widetilde{X}^{(2)}, \dots \widetilde{X}^{(p)}]_t\}$, where $\widetilde{X}_t^{(i)} \in \mathbb{R}^{J \times 6}$. This representation is transformed via a forward kinematics module to obtain the full local 3D joint pose sequence $\widetilde{\mathcal{X}}_e \in \{[\widetilde{X}^{(1)}, \widetilde{X}^{(2)}, \dots \widetilde{X}^{(p)}]_t\}$, where $\widetilde{X}_t^{(i)} \in \mathbb{R}^{J \times 3}$

Global Trajectory Decoder: This module is responsible for generating the root trajectory information. It takes the conditioned latent z_c as input and gradually up-samples it in the temporal dimension. Finally, the generated sequence is transformed via a linear layer to generate the global trajectory $\widetilde{\mathcal{X}}_g \in \mathbb{R}^{T \times J \times 3}$. $\widetilde{\mathcal{X}}_g$ contains the global root position of the first person and relative displacements for rest of the $(P-1)$ persons. The global trajectory information $\widetilde{\mathcal{X}}_g$ is incorporated into local *full* pose tree $\widetilde{\mathcal{X}}_e$ to obtain the final generated sequence $\widetilde{\mathcal{X}}$. Note that there is no separate global hand trajectory decoder counterpart. The hand trajectory encoder merely enriches the action sequence representation with dedicated global wrist movement information during the encoding process.

Dataset	# Actions	# Joints	# Sequ- ences	Multi- person	Loco- motion	FPS	Finger joints
NTU-VIBE[6]	120	24	$114K$	✓	✓	33	✗
NTU-Xpose	**104**	**52**	$30K$	✓	✓	33	✓
HumanAct12[28]	12	24	$2K$	✗	✗	13	✗
HumanAct12-Xpose[28]	12	52	$2K$	✗	✗	13	✓
UESTC[11]	40	24	$25K$	✗	✗	33	✗
Human3.6M[10]	15	32	$2K$	✗	✓	50	✗

Table 1: A comparative summary of datasets.

Sequence Length Decoder: This decoder transforms z_c into a 1D non-negative sequence $\mathbb{S}$ of length T. A cumulative sum sequence is constructed from $\mathbb{S}$ to model the temporal progression. This cumulative sequence is passed through a sigmoid activation to obtain $\widehat{t[n]}$. The location of the first element for which $\widehat{t[n]} \geqslant \theta_s$ is considered the length of the sequence $\widehat{t_s}$. More precisely, $\widehat{t_s} = 1 + \underset{j}{\operatorname{argmin}}\ [t[j] \geqslant \theta_s], 0 \leqslant j < T$ where θ_s is a fixed threshold.

3.5. Optimization

The loss function for DSAG is a combination of reconstruction loss and KL-Divergence loss, defined as:

$$\mathcal{L} = (\mathcal{L}^{\text{rec}}_{\text{local}} + \lambda_{global}\mathcal{L}^{\text{rec}}_{\text{global}} + \lambda_{len}\mathcal{L}^{\text{rec}}_{\text{len}}) + \lambda_{KL}\mathcal{L}_{\text{KL}} \quad (1)$$

Here, $\mathcal{L}^{\text{rec}}_{local}$ loss is a combination of losses on local body component and local hand component, i.e. $\mathcal{L}^{\text{rec}}_{local} = (\lambda^{hand}_{6D}\mathcal{L}^{\text{hand}}_{6D} + \lambda^{hand}_{3D}\mathcal{L}^{\text{hand}}_{3D}) + (\lambda^{body}_{6D}\mathcal{L}^{\text{body}}_{6D} + \lambda^{body}_{3D}\mathcal{L}^{\text{body}}_{3D})$, where $\mathcal{L}^{\text{hand}}_{6D}$ and $\mathcal{L}^{\text{hand}}_{3D}$ are MAE losses on local hand component in 6D rotation space and 3D joint space respectively. The counterpart losses for local body component are $\mathcal{L}^{\text{body}}_{6D}$ and $\mathcal{L}^{\text{body}}_{3D}$. $\mathcal{L}^{\text{rec}}_{global}$ is the MAE loss for global body trajectory component. $\mathcal{L}^{\text{rec}}_{len}$ is MAE loss averaged over the indexed sequence length representation. $\mathcal{L}_{KL}$ is the KL-divergence loss. λ_{KL}, λ_{len}, λ_{global}, λ^{hand}_{6D}, λ^{hand}_{3D}, λ^{body}_{6D} and λ^{body}_{3D} are hyperparameters.

4. Experiments

4.1. Datasets (Table 1)

NTU-RGBD-120[15]: This consists of 114,480 24-joint pose sequences across 120 single *and multi-person* actions performed by 106 subjects captured from 32 camera setups. Since the original skeleton representation in dataset doesn't contain dense finger joints and has inconsistent bone length across the sequence, we obtain 52-joint 3D pose sequences from RGB video frames using ExPose [2], a state of the art method for estimating full body pose, including finger articulation. To create a dataset of sufficient quality for training generative models, only action sequences where the subject is facing the camera are considered. Multi-person classes involving contact between the subjects are removed. Since Expose [2] is a frame based method, to mitigate temporal incoherency, additional refinement is performed by optimizing for temporal smoothness and forcing finger joints to have a single axis of rotation across each sequence. Finally, all 94 single person classes and 10 multi-person classes are considered totalling 30K sequences.

Apart from this dataset, we also use NTU-VIBE dataset [6]. This dataset contains relatively coarse grained local pose sequences obtained using VIBE [13], a video-based pose estimator, on RGB videos from the NTU-RGBD dataset [15].

HumanAct12[28]: This consists of 1191 sequences across 12 action categories which are mostly performed in-place. This dataset is challenging due to its relatively lower frame-rate compared to other datasets. As with NTU dataset, we use ExPose [2] estimation on the RGB video frames and refer to the created dataset as HumanAct12-Xpose.

UESTC[11]: This dataset contains 25k 24-joint sequences from 40 aerobics and exercise actions performed in-place. For the full body sequences, we use the dataset provided by Petrovich et al. [19]. We use the official cross-subject split to separate train and evaluation datasets.

Human3.6M[10]: This dataset consists of $2k$ 3D human pose sequences of 32 joints sourced from 5 female and 6 male subjects, 4 different viewpoints and with 15 actions. Despite the small number of categories, this dataset is challenging due to very high intra-category diversity in the sequences.

We use the available 3D mocap datasets (which lack detailed finger joints) for UESTC and Human3.6M dataset since RGB video data is not available for full body pose estimation using ExPose. Additional details regarding the datasets can be found in Table 1.

4.2. Implementation

The number of ST-Blocks in local encoders and decoders is set to 2 and the convolutional stride in temporal encoder component s is set to 3 (Sec. 3.3.1). The number of latent Gaussian components in GMVAE equals the number of action categories in each dataset. We train DSAG using Adam optimizer with an initial learning rate of 0.015, a learning rate scheduler with a decay rate of 0.5 and step size 10. For NTU, UESTC and Human3.6M datasets, we reduce frame-rate by subsampling the sequence by a factor of 4. During generation, we apply bicubic interpolation on the timestep dimension to match the ground truth data's frame rate. We omit subsampling for HumanAct12 dataset due to its very low capture frame rate (Table. 1). While training DSAG on Human3.6M and UESTC datasets, we remove the local hand encoder/decoder and wrist joint encoder since these datasets do not contain a significant number of finger joints. For HumanAct12 and UESTC, we remove global body trajectory encoder/decoder because these datasets contain only in-place activities without locomotion. For single person actions from NTU, we retain only the first person's generated sequence. We conduct all experiments on cluster machines with Intel Xenon E5 2640 v4 and Nvidia GeForce GTX Ti 11GB GPUs with Ubuntu 16.04 OS, with code written using python-3.7 and pytorch-0.4.

4.3. Metrics and Evaluation

To quantify the realism and diversity of generated sequences, we use five popular generative quality metrics. The first two – MMD-A and MMD-S – are based on Maximum Mean Discrepancy [22] and directly computed in 3D joint space [6, 3]. Empirically, these direct sample space metrics have been found to correlate better with generation quality [6]. The other three include Fréchet Inception Distance (FID) [8], Diversity Score (DS) [5] and Multimodality Score (MS) [5]. These metrics use features from a pretrained skeleton action classifier. As pointed out by MUGL[6], despite their popularity, these approaches often correlate poorly with generation quality since the base classifier used for feature extraction involves pose distorting preprocessing which affects representation quality. To mitigate this issue, we use CTR-GCN [1], a state-of-the-art classifier which does not perform any such preprocessing. We report the absolute difference between the generated and ground truth DS and MS scores. For all the metrics, smaller the score, better the generative quality.

To compute performance scores, we uniformly generate 300 samples per action class for NTU dataset variants and UESTC. Since Human3.6M and HumanAct-12 datasets have less samples, we generate same number of samples as those present in the test set for each class. For comparison with baseline methods which generate only single person actions, we train DSAG with replicated single person sequences for consistency.

5. Results

We compare DSAG against five representative baselines – MUGL [6], ACTOR [19], SA-GCN [25], action2motion [5] and VAE-LSTM[7].

The results in Table 2 demonstrate that DSAG outperforms state-of-the-art methods across multiple datasets and quality measures. Comparing the results on Xpose variants for DSAG and MUGL [6], the benefits of a full body model and having dedicated representations for local and global hand components can be clearly seen. This is also evident from the trends in class-wise FID scores as shown in Fig. 5. Extending the comparison, the feature enrichment provided by self-attention module (Sec. 3.3.1) benefits HumanAct12 setting which contains low frame rate sequences. Similarly, the dedicated temporal module benefits Human3.6M which contains sequences with high intra-class diversity. These

	NTU-VIBE-Single-Person					NTU-Xpose-Single-Person				
Model	MMD-A	MMD-S	FID	DS	MS	MMD-A	MMD-S	FID	DS	MS
ACTOR[19]	$0.87^{\pm0.12}$	$0.57^{\pm0.09}$	$119.33^{\pm15.59}$	$3.19^{\pm0.05}$	$1.68^{\pm0.11}$	$0.41^{\pm0.09}$	$0.30^{\pm0.07}$	$131.25^{\pm22.26}$	$3.16^{\pm0.07}$	$1.04^{\pm0.20}$
MUGL[6]	$0.34^{\pm0.12}$	$0.17^{\pm0.01}$	$152.44^{\pm36.61}$	$6.02^{\pm0.05}$	$9.85^{\pm0.55}$	$0.44^{\pm0.08}$	$0.39^{\pm0.06}$	$157.29^{\pm69.27}$	$4.78^{\pm0.05}$	$1.09^{\pm0.45}$
SA-GCN[25]	$0.68^{\pm0.12}$	$0.43^{\pm0.02}$	$179.19^{\pm13.29}$	$3.11^{\pm0.06}$	$1.58^{\pm0.18}$	$0.68^{\pm0.06}$	$0.54^{\pm0.07}$	$208.82^{\pm127.07}$	$6.77^{\pm0.10}$	$1.62^{\pm0.20}$
action2motion[5]	$0.57^{\pm0.11}$	$0.52^{\pm0.03}$	$161.12^{\pm17.96}$	$2.43^{\pm0.03}$	$1.50^{\pm0.13}$	$0.57^{\pm0.09}$	$0.42^{\pm0.09}$	$153.95^{\pm22.74}$	$4.18^{\pm0.07}$	$0.92^{\pm0.20}$
VAE-LSTM[7]	$1.11^{\pm0.17}$	$0.54^{\pm0.01}$	$127.39^{\pm11.82}$	$2.23^{\pm0.05}$	$3.23^{\pm0.19}$	$0.61^{\pm0.21}$	$0.59^{\pm0.04}$	$192.37^{\pm64.08}$	$5.32^{\pm0.51}$	$1.91^{\pm0.23}$
DSAG	$\mathbf{0.31^{\pm0.14}}$	$\mathbf{0.15^{\pm0.02}}$	$\mathbf{111.58^{\pm9.35}}$	$\mathbf{2.01^{\pm0.07}}$	$\mathbf{1.42^{\pm0.11}}$	$\mathbf{0.23^{\pm0.04}}$	$\mathbf{0.22^{\pm0.03}}$	$\mathbf{89.93^{\pm19.32}}$	$\mathbf{2.28^{\pm0.03}}$	$\mathbf{0.88^{\pm0.22}}$
	NTU-VIBE-Multi-Person					**NTU-Xpose-Multi-Person**				
MUGL[6]	$0.45^{\pm0.15}$	$0.36^{\pm0.03}$	$145.63^{\pm33.28}$	$6.12^{\pm0.07}$	$8.54^{\pm0.21}$	$0.24^{\pm0.03}$	$0.21^{\pm0.09}$	$154.79^{\pm32.14}$	$6.56^{\pm00.09}$	$3.44^{\pm0.27}$
DSAG	$\mathbf{0.42^{\pm0.17}}$	$\mathbf{0.13^{\pm0.01}}$	$\mathbf{114.53^{\pm21.23}}$	$\mathbf{3.31^{\pm0.02}}$	$\mathbf{4.18^{\pm0.14}}$	$\mathbf{0.15^{\pm0.02}}$	$\mathbf{0.28^{\pm0.03}}$	$\mathbf{115.51^{\pm26.28}}$	$\mathbf{3.20^{\pm0.06}}$	$\mathbf{2.17^{\pm0.34}}$
	HumanAct12					**HumanAct12-Xpose**				
Model	MMD-A	MMD-S	FID	DS	MS	MMD-A	MMD-S	FID	DS	MS
ACTOR[19]	$0.58^{\pm0.13}$	$0.22^{\pm0.01}$	$89.11^{\pm15.05}$	$1.55^{\pm0.08}$	$0.19^{\pm0.04}$	$0.32^{\pm0.10}$	$0.26^{\pm0.08}$	$135.26^{\pm21.28}$	$2.48^{\pm0.05}$	$0.14^{\pm0.26}$
MUGL[6]	$0.51^{\pm0.12}$	$0.24^{\pm0.07}$	$88.98^{\pm11.10}$	$1.32^{\pm0.04}$	$0.32^{\pm0.03}$	$0.41^{\pm0.09}$	$0.36^{\pm0.04}$	$149.27^{\pm56.84}$	$4.11^{\pm0.05}$	$0.1.27^{\pm0.64}$
SA-GCN[25]	$0.99^{\pm0.18}$	$0.25^{\pm0.03}$	$105.93^{\pm16.33}$	$1.63^{\pm0.04}$	$0.34^{\pm0.05}$	$0.61^{\pm0.10}$	$0.53^{\pm0.04}$	$273.62^{\pm92.54}$	$5.88^{\pm0.16}$	$2.83^{\pm0.10}$
action2motion[5]	$0.49^{\pm0.14}$	$0.31^{\pm0.10}$	$107.40^{\pm44.72}$	$\mathbf{0.27^{\pm0.04}}$	$0.19^{\pm0.05}$	$0.44^{\pm0.09}$	$0.35^{\pm0.06}$	$150.47^{\pm23.48}$	$2.44^{\pm0.07}$	$0.13^{\pm0.03}$
VAE-LSTM[7]	$1.14^{\pm0.12}$	$0.50^{\pm0.05}$	$114.02^{\pm13.18}$	$3.60^{\pm0.08}$	$0.14^{\pm0.07}$	$0.64^{\pm0.09}$	$0.39^{\pm0.11}$	$191.28^{\pm97.64}$	$4.19^{\pm0.13}$	$1.25^{\pm0.11}$
DSAG	$\mathbf{0.45^{\pm0.02}}$	$\mathbf{0.15^{\pm0.04}}$	$\mathbf{62.51^{\pm12.56}}$	$0.42^{\pm0.06}$	$\mathbf{0.06^{\pm0.02}}$	$\mathbf{0.23^{\pm0.04}}$	$\mathbf{0.16^{\pm0.04}}$	$\mathbf{84.95^{\pm8.65}}$	$\mathbf{1.83^{\pm0.03}}$	$\mathbf{0.06^{\pm0.04}}$
	Human3.6M					**UESTC**				
Model	MMD-A	MMD-S	FID	DS	MS	MMD-A	MMD-S	FID	DS	MS
ACTOR[19]	–	–	–	–	–	$0.43^{\pm0.10}$	$0.32^{\pm0.08}$	$91.55^{\pm24.37}$	$\mathbf{2.02^{\pm0.07}}$	$0.63^{\pm0.14}$
MUGL[6]	$0.66^{\pm0.06}$	$0.39^{\pm0.04}$	$355435.23^{\pm56273.61}$	$289.32^{\pm4.55}$	$33.32^{\pm4.55}$	$0.41^{\pm0.03}$	$0.39^{\pm0.02}$	$84.51^{\pm12.78}$	$3.17^{\pm0.05}$	$0.21^{\pm0.06}$
SA-GCN[25]	$1.47^{\pm0.20}$	$1.34^{\pm0.06}$	$13881.73^{\pm5904.40}$	$\mathbf{41.60^{\pm1.30}}$	$\mathbf{2.79^{\pm0.29}}$	$0.59^{\pm0.01}$	$0.44^{\pm0.01}$	$102.09^{\pm20.24}$	$4.43^{\pm0.06}$	$0.30^{\pm0.08}$
action2motion[5]	$0.57^{\pm0.06}$	$0.44^{\pm0.03}$	$472879.05^{\pm13885.83}$	$175.75^{\pm2.10}$	$20.40^{\pm0.45}$	$0.37^{\pm0.07}$	$0.23^{\pm0.05}$	$94.52^{\pm21.55}$	$2.32^{\pm0.06}$	$0.24^{\pm0.12}$
VAE-LSTM[7]	$1.12^{\pm0.03}$	$0.80^{\pm0.00}$	$\mathbf{11452.63^{\pm6973.80}}$	$83.82^{\pm1.15}$	$6.69^{\pm1.19}$	$0.59^{\pm0.03}$	$0.48^{\pm0.05}$	$95.60^{\pm13.76}$	$3.50^{\pm0.04}$	$0.24^{\pm0.08}$
DSAG	$\mathbf{0.49^{\pm0.07}}$	$\mathbf{0.24^{\pm0.06}}$	$518342.21^{\pm97481.51}$	$283.18^{\pm3.59}$	$29.84^{\pm1.18}$	$\mathbf{0.21^{\pm0.04}}$	$\mathbf{0.19^{\pm0.05}}$	$\mathbf{77.65^{\pm17.17}}$	$2.34^{\pm0.06}$	$\mathbf{0.12^{\pm0.06}}$

Table 2: Model comparison in terms of generative quality scores. See Sec. 5 for details.

Architectural Component	Ablation Details	MMD-A	MMD-S	FID	DS	MS
Spatio-temporal Module	Use one ST block	$0.17^{\pm0.04}$	$0.31^{\pm0.05}$	$169.24^{\pm27.21}$	$6.88^{\pm0.03}$	$3.98^{\pm0.22}$
	Use three ST blocks	$0.16^{\pm0.03}$	$0.29^{\pm0.05}$	$124.45^{\pm27.91}$	$3.15^{\pm0.07}$	$3.02^{\pm0.41}$
	Change the order of spatial and temporal block	$0.19^{\pm0.01}$	$0.31^{\pm0.03}$	$129.54^{\pm29.21}$	$6.79^{\pm0.03}$	$3.24^{\pm0.27}$
	Remove self-attention	$0.18^{\pm0.02}$	$0.29^{\pm0.05}$	$156.50^{\pm27.04}$	$6.95^{\pm0.05}$	$4.29^{\pm0.25}$
	Add self-attention before ST Block	$0.15^{\pm0.04}$	$0.28^{\pm0.03}$	$121.95^{\pm29.22}$	$3.97^{\pm0.03}$	$3.01^{\pm0.11}$
Hand Module	Dedicated modules for finger joints absent	$0.26^{\pm0.05}$	$0.39^{\pm0.04}$	$184.82^{\pm41.21}$	$6.41^{\pm0.09}$	$4.11^{\pm0.38}$
	Remove global hand trajectory encoder	$0.16^{\pm0.03}$	$\mathbf{0.27^{\pm0.03}}$	$119.36^{\pm33.24}$	$\mathbf{3.11^{\pm0.04}}$	$3.31^{\pm0.27}$
VAE	Unimodal Gaussian	$0.95^{\pm0.07}$	$0.50^{\pm0.03}$	$167.26^{\pm32.49}$	$6.95^{\pm0.05}$	$4.29^{\pm0.28}$
Optimization	No 3D loss	$0.64^{\pm0.06}$	$0.39^{\pm0.03}$	$179.54^{\pm42.09}$	$6.15^{\pm0.04}$	$4.11^{\pm0.19}$
	No rotation loss	$0.51^{\pm0.08}$	$0.37^{\pm0.02}$	$186.65^{\pm61.32}$	$5.98^{\pm0.04}$	$4.53^{\pm0.31}$
	No global trajectory loss	$0.21^{\pm0.02}$	$0.32^{\pm0.05}$	$142.61^{\pm33.89}$	$4.55^{\pm0.03}$	$3.97^{\pm0.32}$
	No sequence length loss	$0.91^{\pm0.05}$	$0.49^{\pm0.06}$	$198.22^{\pm46.43}$	$6.92^{\pm0.08}$	$5.01^{\pm0.34}$
DSAG (multi-person)		$\mathbf{0.15^{\pm0.02}}$	$0.28^{\pm0.03}$	$\mathbf{115.51^{\pm26.28}}$	$3.20^{\pm0.06}$	$\mathbf{2.17^{\pm0.34}}$

Table 3: Performance scores for DSAG ablative variants.

trends are also reflected in the qualitative comparison presented in Fig. 4. Since ACTOR [19] requires mesh parameters, it cannot be trained on Human3.6M for which only pose sequences are available. The high intra-class diversity of Human3.6M poses difficulty in training the action classifier for computing the FID and other feature-based scores, resulting in unnaturally high values seen in Table 4. Rendered videos, shortcomings of current approach and additional results can be found in project page.

Latent Embedding Analysis: The availability of multiple Gaussian latent components enhances modelling capacity by enabling some components to specialize for action categories. To characterize the mapping between latent components and action classes, we fix the action category and generate 100 samples per Gaussian component within the NTU-Xpose-Multi-Person setting. The component with the least MMD-A score (wrt test data) can be considered to maximally represent the action category. From the plot in Fig. 5(b), we see that only a handful of components tend to maximally represent multiple categories. Most of the components specialize for a single category.

Ablations: To examine the impact of design choices (action representations , architectural components, loss functions), we computed scores for ablative variants of DSAG trained on NTU-Xpose-Multi-Person. From the results in Table 3, we observe that performance degrades in the following settings: (i) too little or too many ST-blocks (ii) dedicated representations and modules for wrist rooted joints are absent (iii) order of spatial and temporal modules within ST-block is switched (iv) vanilla unimodal Gaussian VAE is used (v) either of the 3D or rotation loss functions are removed (vi) the sequence length loss is removed, and generated sequences are of fixed duration.

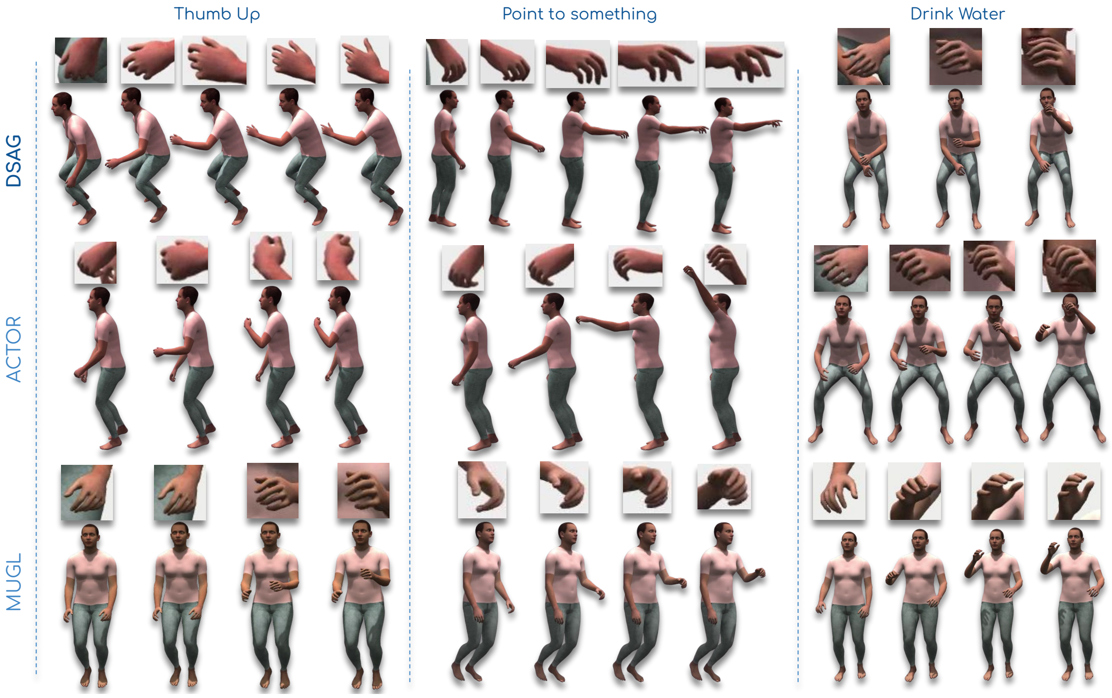

Figure 4: Visual comparison of generated single-person action sequence snapshot renderings across models trained on NTU-Xpose-Single-Person dataset. Note the varying duration of DSAG sequences. Also note that the examples for ACTOR [19] and MUGL [6] exhibit less finger and body movement compared to sequences from DSAG.

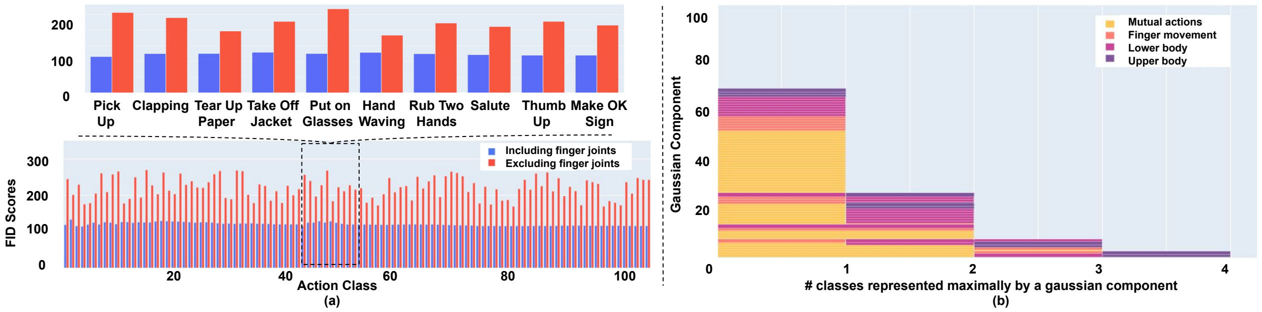

Figure 5: (a) Class-wise FID scores for DSAG and MUGL for all classes of NTU-Xpose-Multi-Person. Inset zoom shows classes for which finger movement dominates action dynamics. (b) Plot showing distribution of NTU-Xpose-Multi-Person classes in terms of maximally representative latent Gaussian components (Sec. 5).

6. Conclusion

DSAG is a controllable deep neural framework which generates full body multi person variable duration action sequences. Two key design choices enhance DSAG's scalability - dedicated local and global representations for encoding wrist-rooted joints (fingers) and spatiotemporal transformation blocks with multi-head self attention and specialized temporal processing. These choices enable DSAG to accommodate different body joint counts - some with fine-grained finger joints (24 - 52), a large range in frame rates (13 - 50 fps), global body movement (in-place, locomotion) and action categories (12 - 120), across multiple datasets (NTU-120, HumanAct12, UESTC, Human3.6m). The choices also enable improved quality generations, especially for actions characterized by subtle finger movements. Our experimental results demonstrate superiority of the proposed framework for action-conditioned fine-grained human motion synthesis at scale.

Acknowledgements

We acknowledge Qualcomm Innovation Fellowship's support to Debtanu Gupta and support from Department of Science & Technology, Govt. of India, via the IMSA project.

References

[1] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13359–13368, 2021.

[2] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision (ECCV)*, 2020.

[3] Bruno Degardin, João Neves, Vasco Lopes, João Brito, Ehsan Yaghoubi, and Hugo Proença. Generative adversarial graph convolutional networks for human action synthesis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1150–1159, 2022.

[4] Nat Dilokthanakul, Pedro AM Mediano, et al. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv*, 2016.

[5] Chuan Guo, Xinxin Zuo, et al. Action2motion: Conditioned generation of 3d human motions. *ACMMM*, 2020.

[6] Debtanu Gupta, Shubh Maheshwari, and Ravi Kiran Sarvadevabhatla. Mugl: Large scale multi person conditional action generation with locomotion. In *WACV*, 2022.

[7] Ikhsanul Habibie, Daniel Holden, et al. A recurrent variational autoencoder for human motion synthesis. In *BMVC*, 2017.

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a nash equilibrium. *arXiv*, 2017.

[9] Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *ACM ToG*, 36:1–13, 07 2017.

[10] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014.

[11] Yanli Ji, Feixiang Xu, Yang Yang, Fumin Shen, Heng Tao Shen, and Wei-Shi Zheng. A large-scale varying-view rgb-d action dataset for arbitrary-view human action recognition. *arXiv preprint arXiv:1904.10681*, 2019.

[12] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[13] Muhammed Kocabas, Nikos Athanasiou, et al. VIBE: Video inference for human body pose and shape estimation. In *CVPR*, 2020.

[14] Zimo Li, Yi Zhou, et al. Auto-conditioned recurrent networks for extended complex human motion synthesis. In *ICLR*, 2018.

[15] Jun Liu, Amir Shahroudy, et al. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *TPAMI*, 42(10):2684–2701, 2020.

[16] Matthew Loper, Naureen Mahmood, et al. Smpl: A skinned multi-person linear model. *SIGGRAPH Asia*, 34(6):248:1–248:16, 2015.

[17] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[18] Dario Pavllo, David Grangier, et al. Quaternet: A quaternion-based recurrent model for human motion. In *BMVC*, 2018.

[19] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3D human motion synthesis with transformer VAE. In *ICCV*, 2021.

[20] Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. BABEL: Bodies, action and behavior with english labels. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 722–731, June 2021.

[21] Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions. *CoRR*, abs/1710.05941, 2017.

[22] Ilya O Tolstikhin, Bharath K Sriperumbudur, and Bernhard Schölkopf. Minimax estimation of maximum mean discrepancy with radial kernels. *NIPS*, pages 1938–1946, 2016.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[24] Sijie Yan, Zhizhong Li, et al. Convolutional sequence generation for skeleton-based action synthesis. 2019.

[25] Ping Yu, Yang Zhao, et al. Structure-aware human-action generation. In *ECCV*, pages 18–34, 2020.

[26] Zhengyou Zhang. Microsoft kinect sensor and its effect. *IEEE MultiMedia*, 19:4–12, 2012.

[27] Yi Zhou, Connelly Barnes, et al. On the continuity of rotation representations in neural networks. *CVPR*, 2019.

[28] Shihao Zou, Xinxin Zuo, Yiming Qian, Sen Wang, Chi Xu, Minglun Gong, and Li Cheng. 3d human shape reconstruction from a polarization image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.