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

Contents

1. Expose dataset refinement

Estimating detailed finger motion is extremely challenging from monocular RGB images, especially so when the "pixel real estate" per hand is small. Since ExPose[4] is a frame-wise method, it is not smooth over time. To create a dataset of sufficient quality for training generative models, we apply several post extraction refinements.

- Facial joints are omitted since actions are performed in neutral setting and provide no additional information.
- Due to self occlusion, finger motion of classes with dynamic hand movement like "make victory sign, thumb down" cannot be captured from all views. Hence actions performed by subject facing the camera are considered. Specifically, C1R1 and C3R2 in NTU dataset[14] and view 2 for HumanAct12[7]
- For temporal consistency we apply optimization based refinement. For every action sequence, given its ExPose extracted local body component $\mathcal{X}_l = \{[X_l^{(1)}, X_l^{(2)}, \dots X_l^{(p)}]_t\}$ and local hand component $\mathcal{X}_h = \{[X_h^{(1)}, X_h^{(2)}, \dots X_h^{(p)}]_t\}$ and where $[X_{l/h}^{(i)}]_t \in \mathbb{R}^{J\times 6}$, i.e. a 6-D rotation representation of J joints, for i - th person, at timestep t ($1 \leqslant t \leqslant T$). We optimise the temporal smoothing loss:

$$\mathcal{L}_{smooth} = \sum_{i=1}^{P}\sum_{t=1}^{T-1} \lambda_{body}||X_l^t - X_l^{t-1}||_2^2 + \lambda_{hand}||X_h^t - X_h^{t-1}||_2^2 \quad (1)$$

  Here $\lambda_{global} = 1$ and $\lambda_{hand} = 0.1$. To optimize we apply gradient descent for 100 iterations at learning rate of 10.
- After decoupling the motion using inverse kinematics, for all the classes in NTU and HumanAct12 dataset we noticed that the finger joints exhibit rotation with only a single degree of freedom. Therefore their motion is constrained to a single axis. For each finger joint, first the rotation is converted from 6D to angle axis representation. Next we convert the axis of all rotations to polar co-ordinate and calculate the median of these axis of rotations. We fix this median as the axis of rotation for the entire dataset and calculate the angle of rotation for each sample at each timestep. To calculate this angle, we initialize it with the original angle and optimize the L2 loss between with the original 6D representation and predicted 6D representation. Similar to previous step we use gradient descent for optimization.
- The global trajectory information is sourced from the original Kinect sequence of NTU dataset [8] with hip as the body root joint and wrists as the hand root joints.
- For two person sequences, the correspondence between individuals in the Kinect 3D sequence and the counterparts in RGB-based ExPose is established on a per-timestep basis by matching their respective 'facing direction', defined as the vector normal to the plane comprising of root (hip) joint and the two shoulder joints.
- Since multi-person contacts cannot be captured from

Expose, classes with two person hand interactions such as hugging, support somebody, touch pocket are omitted.
- We refer to this newly derived datasets as NTU-Xpose and HumanAct12-Xpose.

Video example to show the importance of preprocessing the dataset can be found in `Videos/Importance-of-preprocessing-expose`

2. Additional Details

2.1. Dedicated Body and Hand Components

Video example to demonstrate the importance of finger joints in the perception of an action can be found in `Videos/Importance-finger-Joint-Motion`. Finger joints have a lower degree of freedom. Therefore, despite having a smaller number of joints, body joints can dominate the finger joints within the action dynamics, which would result in finger joints being under represented. To mitigate this problem, we decouple finger and body components. Video example to appreciate this can be found in `Videos/Effect-of-decoupled-body-hand-components`.

2.2. ST-Block

Fig. 1 shows the detailed diagram for our ST-Block. To demonstrate the importance of key architectural components, we further show videos at `Videos/Effect-of-ST-block`. The videos `effect-st-block-<dataset>-<label>.mp4`, show how Multi Head Temporal Self-Attention helps capturing precise action representation for datasets captured at a very low frame rate. Specifically, for HumanActs12 dataset which is captured at a very low frame rate. Before addition of the Multi Head Temporal self-Attention, the model fails to generate action sequences with fine movement, as the CNN based spatial encoder alone is not sufficient to capture fine movements. But after addition of the multi head temporal self-attention layer, it learns to produce better results.

`ST-Block-captures-within-class-diversity-better.mp4` shows how the temporal encoder module helps tackling large within-class diversity. An example from Human3.6m dataset is shown in the video since it has very high within-class diversity. Before addition of the temporal encoding module in the ST-Block, the model would generate static skeleton sequence. But after addition of the temporal module, it learns to represent the within-class diversity.

2.3. Importance of each loss function

All the losses we employ are crucial for natural looking generation. 'Ablations' in Sec. 5 of the main paper summarizes trends by removing each loss function.

- **Sequence length loss**: Removing this loss causes the maximum drop in scores and overall quality. The sequence length loss encourages realistic action durations for generated sequences. Without this loss, after some frames, random joint values are generated for relatively shorter duration actions, causing undesirable visual artifacts.
- **3D loss**: In general, 3D loss ensures that parent-child relationship of kinematic skeleton tree is maintained correctly. Qualitatively, removing 3D loss causes symmetric bones such as left and right thigh (connecting hip and knee joints) to interchange. This dramatically affects visual appearance of the action.
- **6D loss**: This is the main reconstruction loss for our model. Using only 3D loss leads to generation of mean pose. Other works (e.g. ACTOR[17]) report similar artifacts in absence of this loss.
- **Global trajectory loss**: This loss encourages proper relative distance between multiple actors and global locomotion of actors.

2.4. Mesh Rendering

The animated mesh was generated using blender and Three.js. Using the official SMPL blender addon, we first generated the skinned neutral mesh and exported it as an fbx file from blender. Next, in order to rig the animation to the mesh, we imported the mesh file as well as the rotation data in a Three.js scene. We then extracted each of the 52 joints required and rotated them across the specified action time sequence thereby generating the animation.

Mesh rendered sequences across various datasets can be found at `Videos/Comparisons-with-baselines-and-ground-truth/<dataset>`. Note that since Human3.6[11] skeleton is different from SMPL-X skeleton (See Fig. 3), some version of motion retargeting would be required to deform SMPL-X mesh. To avoid additional artifacts from motion retargeting, we show skeleton sequences for generations on Human3.6 dataset.

2.5. Sequence Length Analysis

To compare sequence length distribution, 100 samples are selected from ground truth and generated from DSAG for each action class of NTU-Xpose-Single-person. Figure 2 shows that for different action classes, on average, DSAG generates similar sequence length to the ground truth data.

2.6. Hyperparameters

Table 1 shows our choice of hyperparameters for training on different datasets. Please note that the hyperparameter

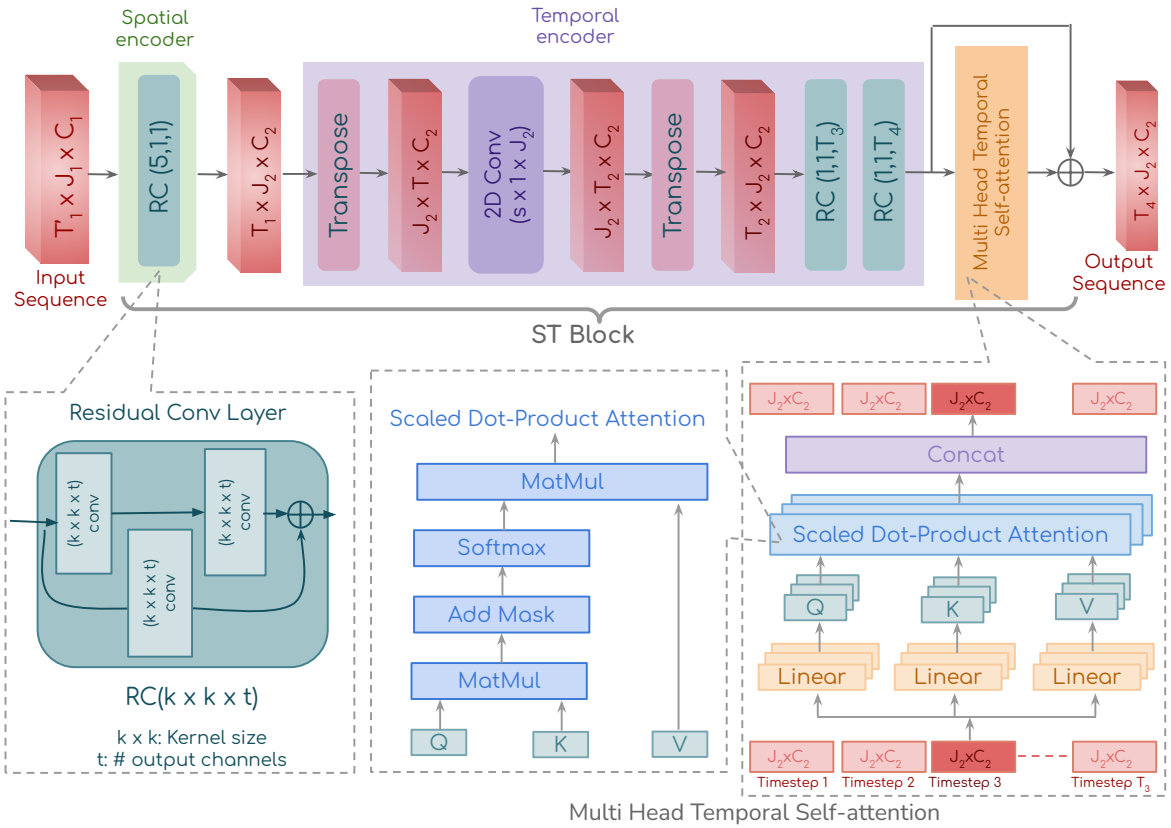

Figure 1: Detailed diagram for the ST-block (see Sec. **??**). Residual convolution (bottom left, shaded green) is applied to process the spatio-temporal information. Multi-head self-attention (bottom right, orange) is used to incorporate the global temporal dependency.

Dataset	λ_{KL}	λ_{len}	λ_{global}	λ_{6D}^{hand}	λ_{6D}^{body}	λ_{3D}^{hand}	λ_{3D}^{body}
NTU-VIBE[14]	Cyclic Annealing	1	2	-	10	1	1
NTU-Xpose[14]	0.1	5	1	1	10	1	1
HumanAct12[24]	0.01	1	-	-	50	1	1
HumanAct12-Xpose[24]	0.001	1	-	1	10	1	1
UESTC[12]	0.005	-	-	-	50	1	1
Human3.6m[11]	0.005	-	0.5	-	50	1	1

Table 1: Details of hyperparameters for training DSAG on different datasets. For NTU-VIBE λ_{KL} is obtained from a cyclic annealing schedule [6]

.

for KL-Divergence loss λ_{KL} is obtained from a cyclic annealing schedule [6]. For Human3.6m dataset, we divide each sequence into segments of 256 timesteps. For single person (NTU-Xpose-Single-person) and multiple person (NTU-Xpose-Multi-person), same set of hyper parameters are used.

3. Experiments

3.1. Baselines

MUGL:[8] proposed a variational autoencoder based method for large scale generation of single and multi-person actions of variable duration. It uses decoupled modules for local and global trajectory modeling, where the local pose

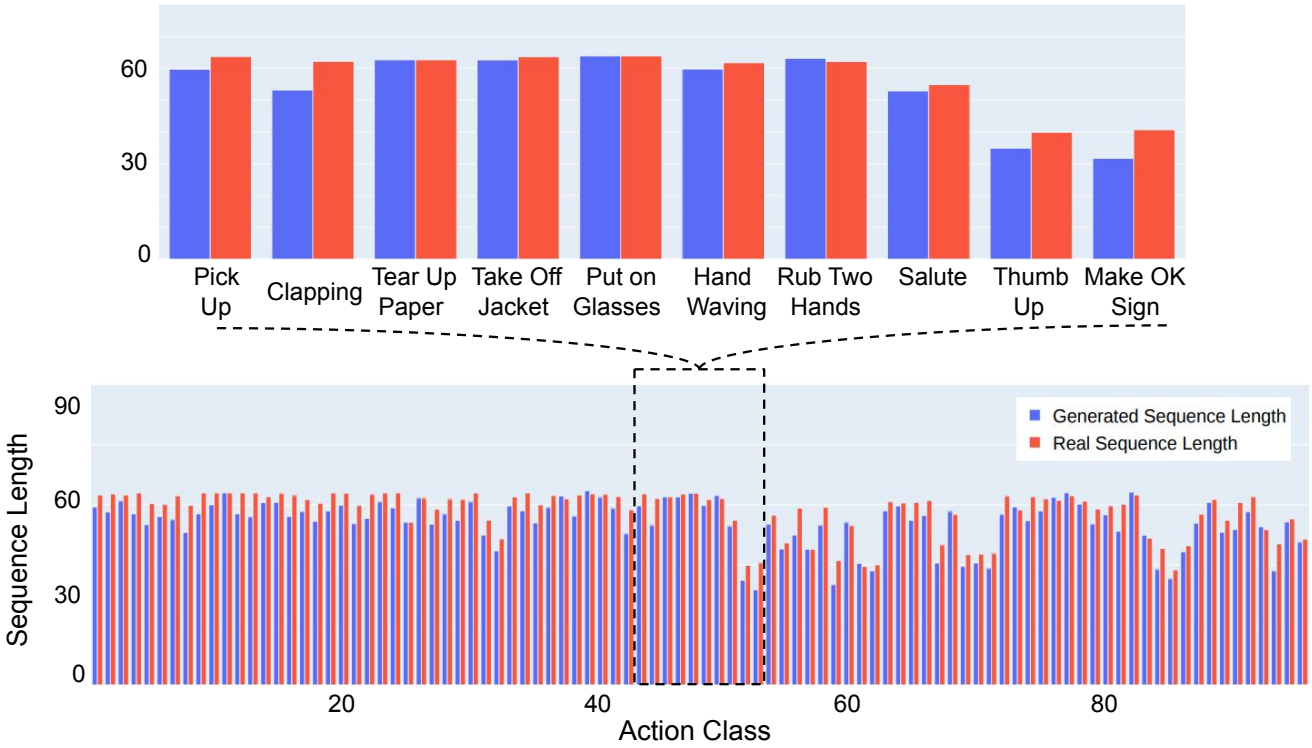

Figure 2: Figure shows comparison of class-wise mean sequence length of the real and generated sequences. DSAG is able to capture the non-identical duration of different classes.

Method	max # actors	# classes	Variable Duration	Joint config	# Datasets	# Parameters (K)	Avg. Inference Time
VAE-LSTM[9]	1	1	✗	B	1	$1001 - 1466$	$0.21 - 0.22$
SA-GCN[23]	1	10	✗	B	2	$18528 - 63101$	$52.92 - 289.41$
action2motion[7]	1	$12 - 13$	✗	B	3	$462 - 556$	$46.32 - 53.60$
ACTOR[17]	1	$12 - 40$	✓	B	3	$14918 - 14981$	$5.94 - 5.96$
Kinetic-GAN[5]	1	$10 - 120$	✗	B	2	$3643 - 4336$	$9.90 - 9.94$
MUGL[8]	2	120	✓	B	1	$108 - 3906$	$0.31 - 0.56$
	2	$12 - 120$	✓	$B + F$	4	$349 - 2863$	$0.45 - 0.98$

Table 2: A comparative summary of the baseline approaches and . B, F in Joint Config column represent Body and Finger Joints respectively. Since # Parameters and average inference time vary across different datasets, a range of values is reported. Avg. Inference time is in milliseconds.

sequence is represented using joint rotation and the global trajectory is represented using 3D positions of the root joint in each timestep. Even though this method is able to scale upto a large number of action classes, it fails to generate full body(including finger joints). This method doesn't generalize on other datasets.

Kinetic-GAN[5]: proposes a GCN based conditional-GAN framework for *fixed duration* large scale action generation across 94 action classes of NTU dataset [14]. However, Kinetic-GAN is trained only on *single person* NTU-RGBD Kinect sequences with 3D joint positions to represent pose. For two person classes, only one person is considered. Unlike our method, it uses 3D joint position to represent human pose tree, which fails to ensure consistent bone length through the action sequence. Since this method is highly sensitive to hyper-parameters, it fails to converge when trained on other datasets.

ACTOR:[17] generates *variable duration*, *single-person* mesh-based motion sequences conditioned on class label using a conditional variational autoencoder along with a transformer-based framework. To construct the baselines for single person generation, we trained the model on sin-

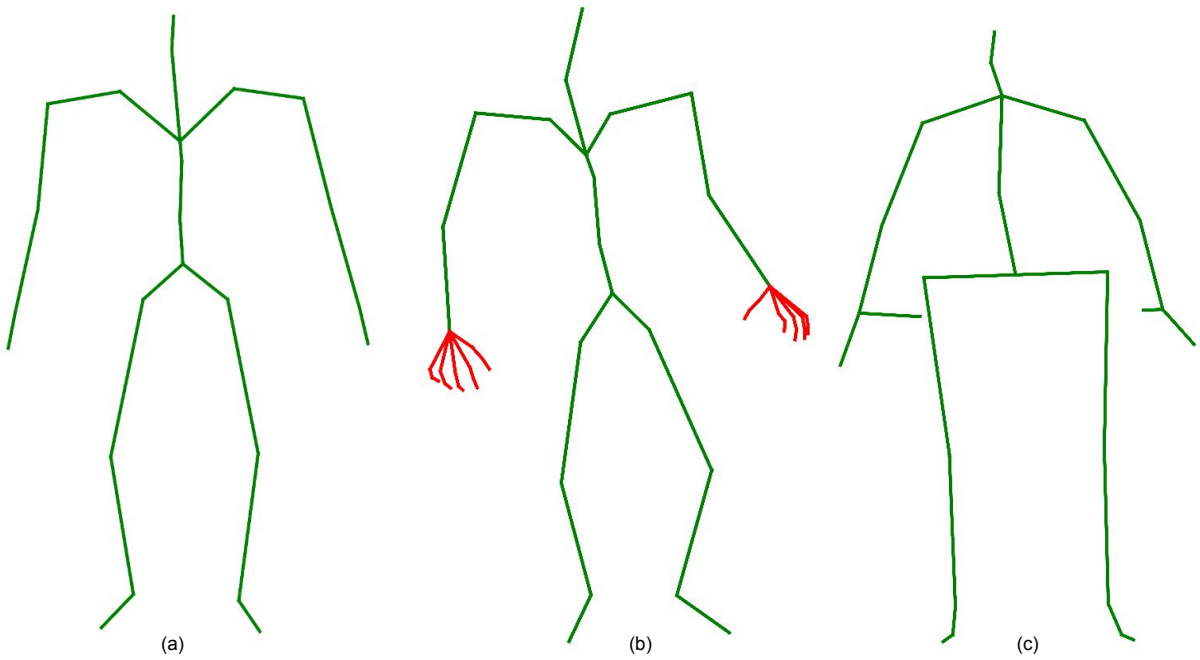

Figure 3: Configuration of different skeleton structures DSAG is trained on (a) indicates the skeleton structure followed by NTU-VIBE, UESTC and HumanActs12 datasets. (b) indicates the skeleton structure followed by NTU-Xpose and HumanActs12-Xpose datasets. (c) shows the structure of Human3.6m dataet.

gle person subset of NTU datasets (NTU-VIBE-Single-person, NTU-Xpose-Single-person) and HumanAct12-Xpose datasets. For NTU-VIBE-Single-person we used the same settings as defined for their experiements on NTU dataset, except changing the number of classes. For NTU-Expose-Single-person and HumanAct12-Xpose we replace SMPL with SMPL-X [16] to generate finger motion. For HumanAct12 and UESTC, we used pretrained ACTOR models available as baseline. The method is confined to a small number of action classes. Since this method requires SMPL mesh parameters to optimize, we were unable to train it on Human3.6m dataset since Human3.6m dataset doesn't contain mesh parameters.

SA-GCN:[23] proposes a self-attention based graph convolution backed conditional GAN based method conditioned on class labels. We modify SA-GCN, originally trained on 2D action sequences, for our 3D generative setting. But, this method is designed for generating human action sequence in 2-D coordinate space. Moreover when this method is trained to generate samples on 3-D coordinate space, the generated samples orientation rotate randomly in 3-D space. Another drawback is that it covers a small set action classes. This model fails to generate good quality sample when number of action classes are increased, especially the action class which involve leg movement.

Action2motion:[7] uses conditional VAE to generate human motion sequence from action class. Similar to our approach, action2motion represents the pose tree via 3D rotation. This method is also meant for small number of action classes (13 classes) and fails to generate good quality samples beyond that. This method fails to scale to a large number of action classes.

VAE-LSTM:[9] uses a LSTM based VAE to generate human action sequences modulated by control signals. For constructing our baseline, we modify VAE-LSTM to incorporate class conditioning instead of control signal. This method works well for smaller datasets containing human locomotion activities, controlled by control signals. This method fails to model activities at large scale.

3.2. Metrics

To quantify the naturalness, realism and diversity of generated sequences, we show results using five popular generative quality metrics. For all the metrics, smaller the score, better the generative quality.

Direct sample space metrics: Maximum Mean Discrepancy (MMD) captures similarity between generated and test set sample distributions [19, 23, 1, 8]. Since it is directly computed on the 3D joint space, there is no need of an external feature classifier. We employ two variants of MMD – MMD-A and MMD-S for evaluation. The base similarity is measured on a per-timestep basis for MMD-A and on

a per-sequence basis for MMD-S. Empirically, these direct sample space metrics have been found to correlate better with generation quality [8].

- **MMD-A:** For 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 [2].
- **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 $\mathcal{K}(g, g) + \mathcal{K}(e, e) - 2\mathcal{K}(g, e)$.

Feature space metrics: These include Fréchet Inception Distance (FID) [10], Diversity Score (DS) [7] and Multimodality Score (MS) [7]. These metrics use feature representations obtained from a pretrained skeleton action classifier. Despite their popularity, these approaches often correlate poorly with generation quality since the base classifier involves pose distorting preprocessing during feature extraction which affects representation quality. To avoid this issue, we use CTR-GCN [3], a state-of-the-art classifier which does not perform any such preprocessing. Details about training CTR-GCN can be found in Sec. 3.4

Fréchet Inception Distance (FID) [10]: FID measures the quality of generation by comparing feature distribution of generated samples and ground truth samples (extracted from a pretrained classifier). FID score is given as:

$$FID = ||\mu_r - \mu_f||_2 + tr(\Sigma_r + \Sigma_f - 2(\Sigma_r\Sigma_f)^{\frac{1}{2}}) \quad (2)$$

where μ_r, μ_f indicate the mean of ground truth and generated feature vectors. Σ_r, Σ_r indicate covariance matrix of ground truth and generated feature vector.

Diversity Score (DS) [7]: This measures the variance of the generated samples across all the action classes. From a set of randomly generated samples from multiple action classes, two subsets of the same size are sampled. Their respective motion feature representations $\{v_1, v_2, ..., v_n\}$ and $\{\hat{v}_1, \hat{v}_2, ..., \hat{v}_n\}$ are obtained. The diversity of the generated motion is defined as:

$$DS = \frac{1}{n}\sum_{i=1}^{n} ||v_i - \hat{v}_i||_2 \quad (3)$$

Following protocol of Action2motion[7], diversity score is deemed better if the score for the generated samples is closer the score on the test set. Instead of separately showing both the values we report the absolute difference between the generated and ground truth diversity scores.

Multimodality Score (MS) [7]: Unlike Diversity Score, Multimodality Score calculates variance of generated samples within a reference action class. For c-th action ($1 \leqslant c \leqslant C$), we create two generated sample subsets of size n_c. The feature representations for these two sets $\{v_{c,1}, ..., v_{c,n_c}\}$ and $\{\hat{v}_{c,1}, ..., \hat{v}_{c,n_c}\}$ are obtained. The MS score is defined as:

$$MS = \frac{1}{Cn_c}\sum_{c=1}^{C}\sum_{i=1}^{n_c} ||v_{c,i} - \hat{v}_{c,i}||_2 \quad (4)$$

For Multimodality score we follow the same protocol as that for Diversity score.

3.3. Drawbacks of using ST-GCN

Previous methods [7, 17] use ST-GCN[22] which involves a pose distorting preprocessing step. Each subject is translated such that the root joint of the first origin is glued to origin. Hip and spine joints are aligned to y-axis and shoulder joints are aligned to x-axis is translated to origin and the remaining person's are transformed

As pointed out by MUGL[8]. This step leads to following drawbacks:

- Locomotion (e.g. 'walking towards', 'walking apart') and multi-person interaction (e.g. 'kicking') classes are distorted.
- Using only shoulder and hip joints does not consistently provide the desired orientation normalisation.

Removing the preprocessing seems a possible solution. However, MUGL[8] showed a significant drop (20%) in classifier performance, making the resulting feature representations unreliable. Please view videos in Supplementary folder `Videos/Effect-of-gcn-classifier-preprocessing` to better understand the visual effect of classifier preprocessing.

3.4. CTR-GCN Implementation

To overcome the above mentioned limitations we used CTR-GCN [3], state of the art skeleton based action recognition model to calculate feature space based metrics. CTR-GCN was preferred as it doesn't require any additional preprocessing of data.

We updated the provided code with the required skeleton structure and hyperparameters in the config file as per the dataset being used. The feature vectors correspond to the resultant vector formed before passing to the fully connected layer. We constructed 4 models of CTR-GCN,

- **NTU-VIBE-Single-Person:** Skeleton structure used here is of SMPL [15] which is trained on 94 single person classes of NTU VIBE dataset. We used this model to extract feature vectors of the generated samples by all the baseline models trained on NTU-VIBE-Single-person, HumanAct12 [24] and UESTC [12] datasets.

- **NTU-VIBE-Multi-person:** This is similar as above but is trained on all 120 classes of NTU VIBE dataset with the hyperparameter, "num_person" set as 2. We used this model to extract features of the generated samples by all the baseline models trained on NTU-VIBE-Mutli-person dataset.
- **NTU-Xpose-Single-person:** Skeleton structure used here is of SMPL-H [18] which is trained on 94 single person classes of NTU EXPOSE dataset. We used this model to extract features of the generated samples by all the baseline models trained on NTU-Expose-Single-person and HumanAct12-Expose dataset.
- **NTU-Xpose-Multi-person:** This is similar as above but is trained on all 120 classes of NTU EXPOSE dataset with the hyperparameter, "num_person" set as 2. We used this model to extract features of the generated samples by all the baseline models trained on NTU-Expose-Mutli-person dataset.
- **Human3.6m:** Initially we trained CTR-GCN on Human3.6 [11] dataset. But due to very high within class diversity, the accuracy on test was low (approx. 60%). Leading to exploding feature based evaluation metrics. Hence for consistancy and completeness we used CTR-GCN trained on NTU-VIBE-Single-person. To evaluate on methods trained on Human3.6 dataset we choose the common subset of joints between SMPL skeleton and Human3.6 skeleton (See Fig. 3). But due to domain shift, we notice exploding scores.

3.5. Qualitative Analysis

Additional qualitative results for HumanActs12-Xpose and UESTC can be found in Figure 4 and Figure 5 respectively. Additionally videos comparing with baselines can be found in `Videos/Comparisons-with-baselines-and-ground-truth/<dataset>`

Examples showcasing generation diversity can be viewed at `Videos/Diverse-samples` in Supplementary. Our qualitative and quantitative results show that DSAG is able to generate diverse samples for all datasets. For UESTC, although DSAG is able to generate diverse samples, it outperforms all methods except ACTOR[17]. Similar trend can be seen in Diversity Score in Table 3 of main paper.

3.6. Quantitative Analysis

Details of Quantitative analysis can be found in Table 3 in the main paper. Additionally, class-wise visualization of evaluation metrics on each method across datasets can be found in `Code/Class-wise-Metric-Visualisation`.

4. Application - Motion Prediction

Although is designed for action generation, we repurpose the model for long-term motion prediction. Motion prediction is a related task where in a small initial action sequence is used to condition the generation of the full version. We use Human3.6M dataset for this experiment. We train the autoencoder version of and finetune the resulting model via a curriculum learning schedule for motion prediction. Following standard protocol [20], we conduct evaluation for prediction of 560 milliseconds (14 frames) and 1000 milliseconds (25 frames) with mean rotation error as the metric. The results in Table 3 show that we outperform all the baselines on average for long-term motion prediction.

Problem Formulation: For this task, the objective is to predict a sequence of 'future' frames $\mathcal{X}_f$ of length t_f given an initial frame sequence $\mathcal{X}_l$ of length t_l as input.

Overview: We follow a seq2seq approach where the local and global body encoder modules take previous frames $\mathcal{X}_l$ as input to create local body encoding f_l and global body encoding f_g. These encoding are concatenated and passed to the decoder to predict a set of future frames $\tilde{\mathcal{X}_f}$. Note no class label conditioning is required in our setup. We also do not use hand modules since finger motion is not available.

Optimization: Since it is not a generative model, we utilize only the reconstruction losses:

$$\mathcal{L} = \mathcal{L}^{\text{rec}}_{\text{local}} + \lambda_{global}\mathcal{L}^{\text{rec}}_{\text{global}} \tag{5}$$

Where $\mathcal{L}^{\text{rec}}_{local}$ loss is a combination of losses on local body component 6D space ($\mathcal{L}_{6D}$) and 3D space ($\mathcal{L}_{3D}$), i.e. $\mathcal{L}^{\text{rec}}_{local} = \lambda_{6D}\mathcal{L}_{6D} + \lambda_{3D}\mathcal{L}_{3D}$. λ_{6D} and λ_{3D} are hyperparameters. We use the same set of hyperparameters which can be found in Table 1.

Training: To initialize the weights, we first pretrain an autoencoder version of to reconstruct input human motion sequence, i.e. to reconstruct $\mathcal{X}_i$ of length t_i. During training the reconstruction loss is computed over the predicted and future ground truth sequences $\mathcal{X}_f$. During the training we use curriculum learning, where initially we predict a 5 frames and gradually increase the number of predicted frames in multiples of 5. The remaining frames predicted by the decoder are discarded.

Related Work Among the recent successful approaches, ConvSeq2Seq [13] uses CNNs to capture long term spatial and temporal correlations. LTD [21] uses a DCT (discrete fourier transform) representation to encode pose. HRI [20] uses an attention based feed-forward network to capture the repetitive nature of human actions.

Evaluation: Following standard protocol [20], we conduct evaluation for prediction of 560 milliseconds ($t_f = 14$ frames) and 1000 milliseconds ($t_f = 25$ frames) with mean rotation error as the metric. The results in Table 3 of main

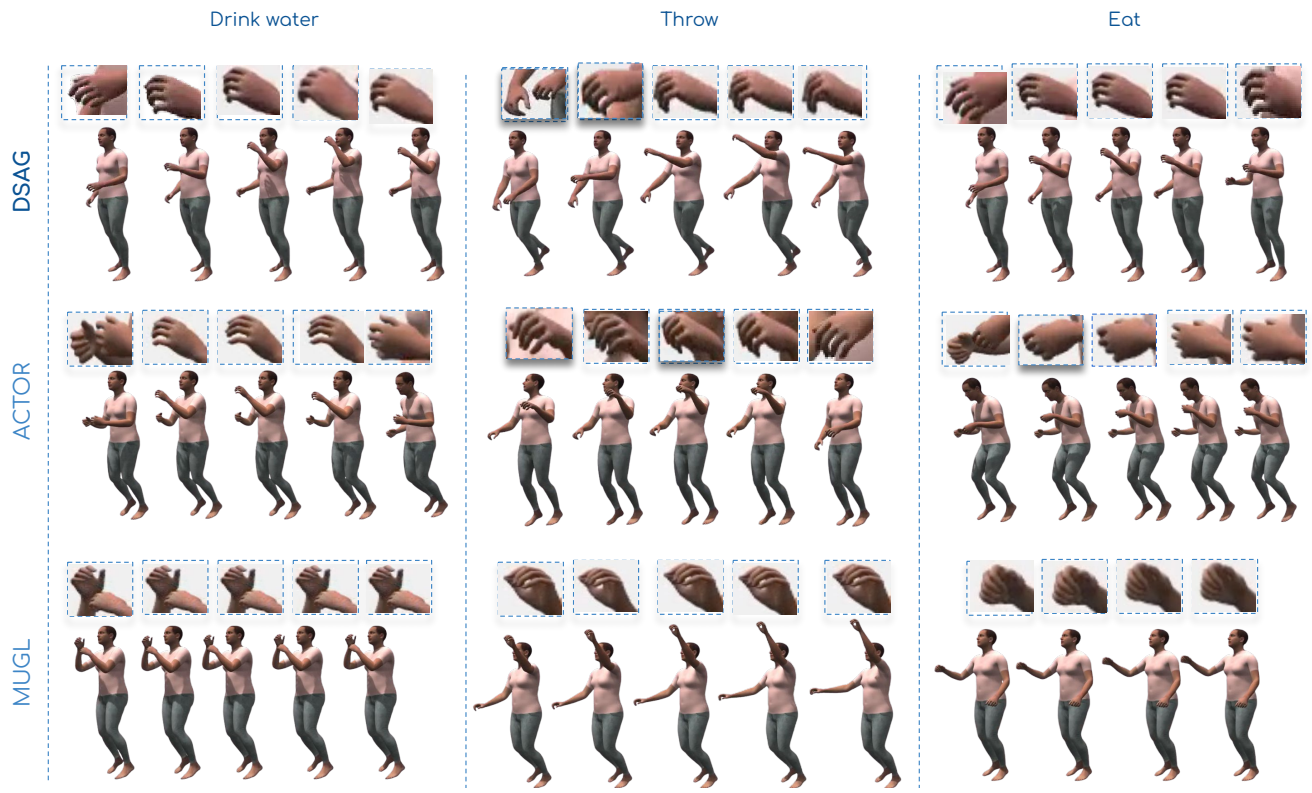

Figure 4: Visual comparison of generated single-person action sequence snapshot renderings across models trained on HumanAct12-Xpose dataset.

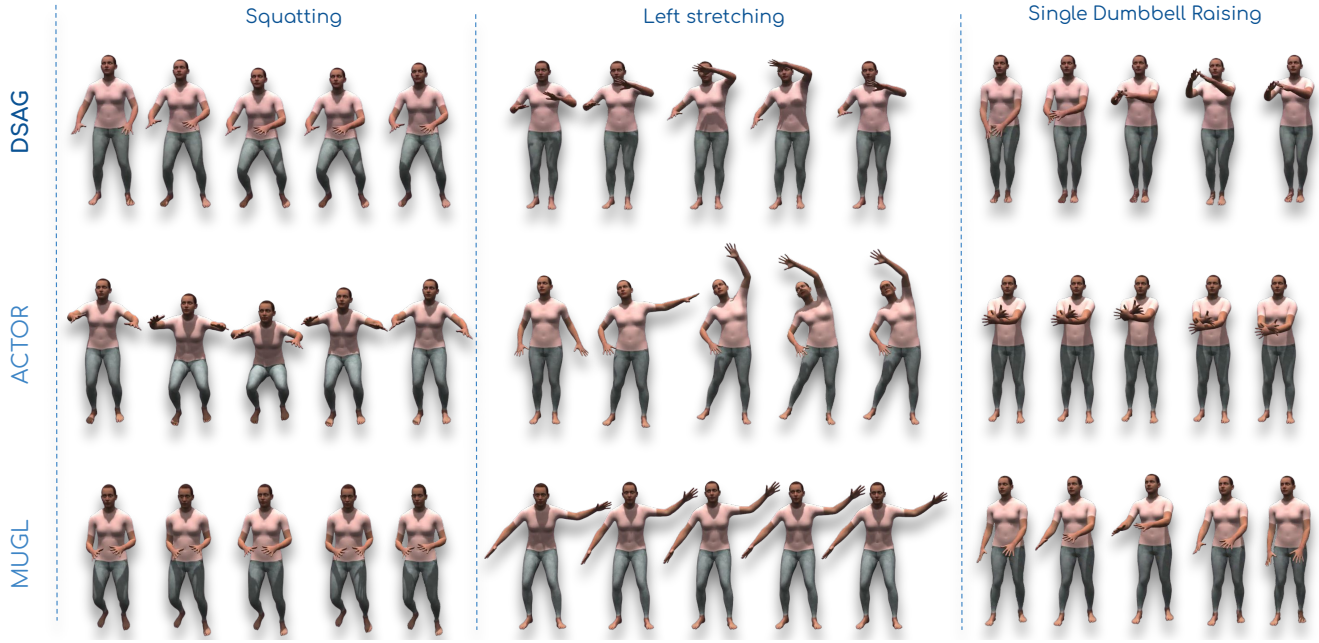

Figure 5: Visual comparison of generated single-person action sequence snapshot renderings across models trained on UESTC dataset.

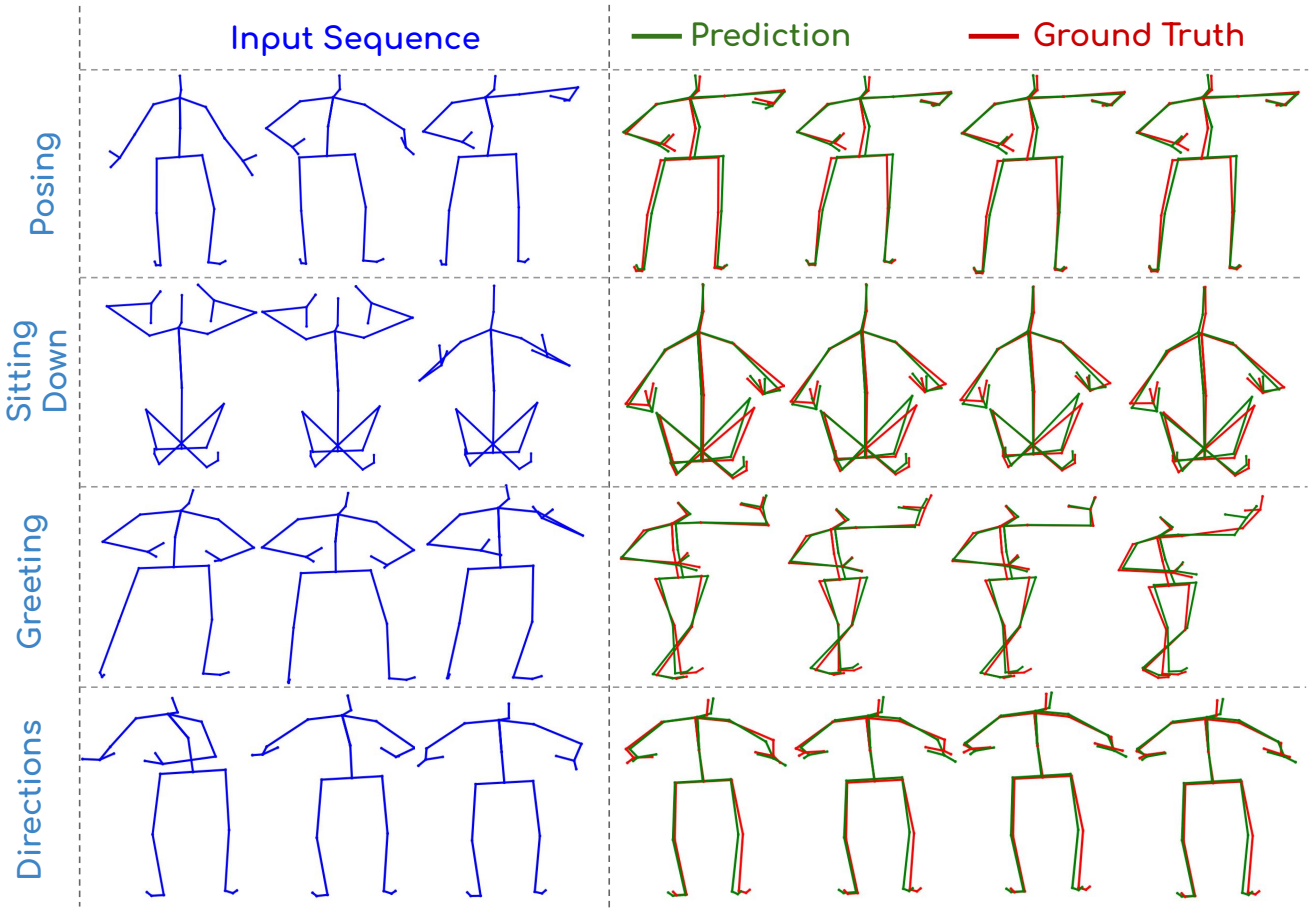

Figure 6: Long term motion prediction results for 1000ms. Left side shows the initial input sequence $\mathcal{X}_l$. The red and green skeleton on the right show the ground truth $\mathcal{X}_f$ and the predicted $\tilde{\mathcal{X}}_f$ sequences respectively. The close overlap between the predicted and ground truth validates the low rotation error achieved by DSAG

paper show that we outperform all the baselines on average for long-term motion prediction, which clearly shows the effectiveness of on long term motion prediction.

Fig. 6 shows qualitative results on predicting 1000ms across various classes of H3.6 dataset. Even though DSAG is designed for generation, the close overlap between between the predicted and ground truth sequences shows that it can be adapted for other related tasks such as prediction.

5. Failure Cases

Failure cases of our model can be found at: `Videos/Failure-Cases`. Our model generate poor quality sequences for classes with unreliable training data. This is particularly the case for actions involving close interaction between two people which occludes the perceived pose structure (e.g. "Hugging"). Action classes with poor hand joint estimation, e.g. "Make Victory Sign", also have poor quality generated sequences. DSAG also suffers from the foot sliding problem, i.e. sometimes the movement of the foot joints are not in sync with the movement in the global trajectory. It happens due to the absence of any dedicated method to handle explicitly, such as, foot contact loss.

6. Potential Social Impact

Although pose based skeleton representation doesn't provide any appearance based identifiable information about the actor performing an activity. The trademark of iconic activities like 'Gangnam style' should not infringed and should be credited appropriately. Future works which build upon our approach should keep cultural nuances of how gestures and actions are interpreted while choosing to expand the action vocabulary. Another undeniable fact is to acknowledge the harmful environmental effects of training such large scale deep-learning models. Lastly, people should avoid adopting this method to generate fake data for

	Walking		Eating		Smoking		Discussions		Directions		greeting		Phoning		Posing	
Model	560	1000	560	1000	560	1000	560	1000	560	1000	560	1000	560	1000	560	1000
convSeq2Seq[13]	0.87	1.00	0.86	1.24	0.98	1.67	1.42	2.03	1.00	1.44	1.73	1.90	1.66	2.05	1.95	2.63
LTD-10-25 [21]	0.65	0.67	0.76	1.12	0.87	1.57	1.33	1.70	0.84	1.26	1.43	1.59	1.45	1.65	1.62	2.42
LTD-10-10 [21]	0.69	0.77	0.76	**1.10**	0.88	1.58	1.27	1.75	0.90	1.35	1.47	1.59	1.49	1.74	1.61	2.55
HRI[20]	**0.59**	**0.64**	**0.74**	**1.10**	0.86	1.58	1.29	1.63	**0.81**	1.27	1.47	**1.57**	1.41	1.68	**1.60**	**2.32**
	1.16	1.66	0.86	1.22	**0.82**	**1.02**	**1.10**	**1.34**	0.90	**1.11**	1.48	1.83	**0.96**	**1.20**	1.65	2.81
	Purchases		Sitting		Sitting Down		Taking Photo		Waiting		Walking Dog		Walking Together		Average	
Model	560	1000	560	1000	560	1000	560	1000	560	1000	560	1000	560	1000	560	1000
convSeq2Seq[13]	1.68	2.50	1.31	1.72	1.45	1.98	1.09	1.32	1.68	2.45	1.73	2.04	0.82	1.29	1.35	1.82
LTD-10-25[21]	1.42	2.21	**1.08**	**1.45**	1.26	1.87	0.85	1.06	1.55	2.29	1.52	1.84	0.70	1.16	1.15	1.59
LTD-10-10[21]	1.47	2.27	1.12	1.52	1.17	1.67	**0.81**	**1.05**	1.57	2.37	1.58	1.86	0.65	**1.16**	1.16	1.62
HRI[20]	1.43	2.22	1.16	1.55	1.18	1.70	0.82	1.08	1.54	2.30	1.57	1.82	**0.63**	**1.16**	1.14	1.57
	0.91	**1.08**	1.30	1.61	**0.83**	**1.19**	0.91	1.70	**0.88**	**2.11**	**1.28**	**1.75**	1.17	1.31	**1.13**	**1.41**

Table 3: Comparisons for long-term motion prediction on 15 action categories of Human3.6M dataset. We show quantitative comparison on prediction of 560 millisecond(14 frames) and 1000 millisecond(25 frames).

unlawful use.

References

[1] Battan, N., Agrawal, Y., Rao, S.S., Goel, A., Sharma, A.: Glocalnet: Class-aware long-term human motion synthesis. In: WACV. pp. 879–888 (2021) 5

[2] Chang, Y.W., Hsieh, C.J., Chang, K.W., Ringgaard, M., Lin, C.J.: Training and testing low-degree polynomial data mappings via linear svm. Journal of Machine Learning Research **11**(48), 1471–1490 (2010), http://jmlr.org/papers/v11/chang10a.html 6

[3] Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., Hu, W.: Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13359–13368 (2021) 6

[4] Choutas, V., Pavlakos, G., Bolkart, T., Tzionas, D., Black, M.J.: Monocular expressive body regression through body-driven attention. In: European Conference on Computer Vision (ECCV) (2020), https://expose.is.tue.mpg.de 1

[5] Degardin, B., Neves, J., Lopes, V., Brito, J., Yaghoubi, E., Proença, H.: Generative adversarial graph convolutional networks for human action synthesis. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1150–1159 (2022) 4

[6] Fu, H., Li, C., et al.: Cyclical annealing schedule: A simple approach to mitigating KL vanishing. arXiv (2019) 3

[7] Guo, C., Zuo, X., et al.: Action2motion: Conditioned generation of 3d human motions. ACMMM (2020) 1, 4, 5, 6

[8] Gupta, D., Maheshwari, S., Sarvadevabhatla, R.K.: Mugl: Large scale multi person conditional action generation with locomotion. In: WACV (2022) 1, 3, 4, 5, 6

[9] Habibie, I., Holden, D., et al.: A recurrent variational autoencoder for human motion synthesis. In: BMVC (2017) 4, 5

[10] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a nash equilibrium. arXiv (2017) 6

[11] Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: 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). https://doi.org/10.1109/TPAMI.2013.248 2, 3, 7

[12] Ji, Y., Xu, F., Yang, Y., Shen, F., Shen, H.T., Zheng, W.S.: A large-scale varying-view rgb-d action dataset for arbitrary-view human action recognition. arXiv preprint arXiv:1904.10681 (2019) 3, 6

[13] Li, C., Zhang, Z., Sun Lee, W., Hee Lee, G.: Convolutional sequence to sequence model for human dynamics. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018) 7, 10

[14] Liu, J., Shahroudy, A., et al.: Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. TPAMI **42**(10), 2684–2701 (2020) 1, 3, 4

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

[16] Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2019) 5

[17] Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3D human motion synthesis with transformer VAE. In: ICCV (2021) 2, 4, 6, 7

[18] Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM Trans. Graph. **36**(6) (nov 2017). https://doi.org/10.1145/3130800.3130883, https://doi.org/10.1145/3130800.3130883 7

[19] Tolstikhin, I.O., Sriperumbudur, B.K., Schölkopf, B.: Minimax estimation of maximum mean discrepancy with radial kernels. NIPS pp. 1938–1946 (2016) 5

[20] Wei, M., Miaomiao, L., Mathieu, S.: History repeats itself: Human motion prediction via motion attention. In: ECCV (2020) 7, 10

[21] Wei, M., Miaomiao, L., Mathieu, S., Hongdong, L.: Learning trajectory dependencies for human motion prediction. In: ICCV (2019) 7, 10

[22] Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: AAAI (2018) 6

[23] Yu, P., Zhao, Y., et al.: Structure-aware human-action generation. In: ECCV. pp. 18–34 (2020) 4, 5

[24] Zou, S., Zuo, X., Qian, Y., Wang, S., Xu, C., Gong, M., Cheng, L.: 3d human shape reconstruction from a polarization image. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020) 3, 6