Action2Motion: Conditioned Generation of 3D Human Motions

Chuan Guo[1] , Xinxin Zuo[1,4], Sen Wang[1,4], Shihao Zou[1], Qingyao Sun[2], Annan Deng[3]

Minglun Gong[4], Li Cheng[1]

[1]Department of Electrical and Computer Engineering, University of Alberta

[2]Physical Sciences Division, University of Chicago

[3]Graduate School of Arts and Sciences, Yale University

[4]School of Computer Science, University of Guelph

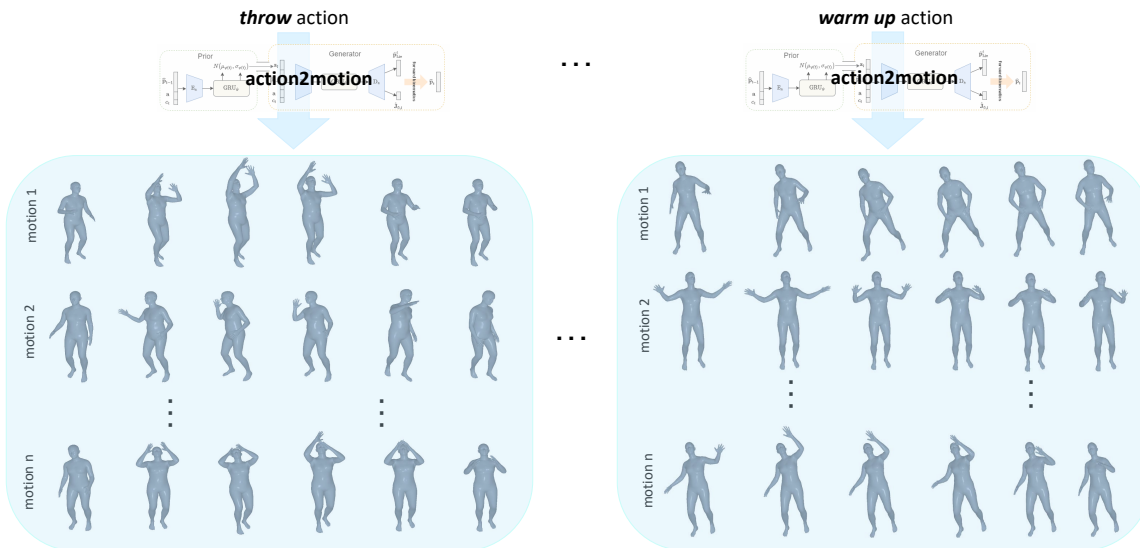{cguo2,lcheng5,szou2}@ualberta.ca,{xinxinzuo2353,wangsen}@gmail.com

Figure 1: Conditioned on an action category (such as throw, warm up), our approach can generate a diverse set of natural 3D human motions.

ABSTRACT

Action recognition is a relatively established task, where given an input sequence of human motion, the goal is to predict its action category. This paper, on the other hand, considers a relatively new problem, which could be thought of as an inverse of action recognition: given a prescribed action type, we aim to generate plausible human motion sequences in 3D. Importantly, the set of generated motions are expected to maintain its *diversity* to be able to explore the entire action-conditioned motion space; meanwhile, each sampled sequence faithfully resembles a *natural* human body articulation dynamics. Motivated by these objectives, we follow the physics law of human kinematics by adopting the Lie Algebra theory to represent the *natural* human motions; we also propose a temporal Variational Auto-Encoder (VAE) that encourages a *diverse* sampling of the motion space. A new 3D human motion dataset,

HumanAct12, is also constructed [1]. Empirical experiments over three distinct human motion datasets (including ours) demonstrate the effectiveness of our approach.

CCS CONCEPTS

• **Computing methodologies** → *Activity recognition and understanding*.

KEYWORDS

3D motion generation; Lie algebra; variational auto-encoder; 3D animation

1 INTRODUCTION

Looking at human is often a central theme in the images and videos of our daily life. In recent years, we have evidenced a surge of

---

[1]For more details, please visit our project website: https://ericguo5513.github.io/action-to-motion/

interests and brilliant progresses in video synthesis and prediction of future frames [6, 27, 30]. However, when coming to looking at human, it remains a significant challenge. This is, for example, evidenced in the efforts of many recent pixel-based video generation methods [3, 37] – the synthesized objects commonly possess strange appearances that deviate from photo-realistic human shapes, their motions are usually unsatisfactorily distorted. These observations pronounce the importance of properly representing human body poses and modeling their temporal articulations in automated video generation. It also inspires our investigation into the fundamental problem of action-conditioned generation of 3D human motions.

The problem of generating human motions is far from being trivial. The principal challenges are two folds: the generated motions should be sufficiently diverse to cover the broad range of ways individuals perform the same type of actions; each motion sequence is also expected to be visually realistic. Most existing efforts either work directly in 2D space, or require initial pose/motion. For instance, [37] generates 2D motions based on an action type and an initial pose in a deterministic way, and then synthesizes appearances in video frame by frame. In [3], a pose generator and motion generator are trained progressively with the help of GANs for human 2D motion generation. Other methods [13, 36] directly generate 2D human motions by VAEs or GANs. However, direct modeling of motions in 2D is inherently insufficient to capture the underlying 3D human shape articulations. Existing methods [3, 18, 36, 37] often approach human poses in terms of the joints' coordinate locations, which unnecessarily entangle the human skeletons and their motion trajectories, and introduce extra barriers in faithful modeling of human kinematics. To further complicate the matter, the variation of generated human dynamics could be severely limited by the initial priors [37].

To address the aforementioned challenges, we consider to generate 3D human motions without prior conditions of initial poses or motions. Instead, a novel framework is proposed that consists of a conditional temporal VAE based on Lie algebra representation. Inspired by [6], we leverage the posterior distribution learned from previous poses as a learned prior to gauge the generation of present pose; by tapping into the RNN implementation, this learned prior also encapsulates temporal dependency across consecutive poses. The proposed modeling design thus facilitates our approach to go beyond the stereotypical motion sequences and uncover varying interpolations of the articulation trajectories in the training set. For motion representation, the theory of Lie algebra has proven its effectiveness for articulate pose modeling in related tasks including pose estimation, action recognition, and motion prediction [8, 11, 20, 31]. Specifically, human pose is characterized as a kinematic tree based on physics principle of human full-body kinematics. There are multiple advantages of using Lie representation over the joint-coordinate representation: (i) Lie representation disentangles the skeleton anatomy, temporal dynamics and scale information; (ii) Lie representation faithfully encodes the anatomical constraints of skeletons, by following the physics law of forward kinematics; (iii) Lie algebra space is roughly an Euclidean space, thus the important linear algebra concepts deeply rooted in regression learning techniques could work again. As a by-product, the dimension of Lie algebra space naturally corresponds to the degrees of freedom (DoF), which is more compact comparing to

joint coordinates. In practice, the adoption of Lie algebra notably mitigates the trembling phenomenon prevailing in joint coordinates representation, facilitates the generation of natural, lifelike motions, and accelerates the learning process. Empirically, training with Lie algebra requires only one tenth of the number of iterations by training with joint coordinates in reaching equilibrium.

To summarize, our key contributions are three-fold: first, we aim to address a new problem of 3D motion sequence generation based on prescribed action categories. Our approach is to our knowledge the first in addressing such a problem; second, a novel Lie Algebra based VAE framework is proposed, capable of generating natural and diverse sets of human motions for prescribed action categories; and third, since the mainstream 3D human pose datasets are not directly applicable to our purpose, we curate own 3D human motion dataset. The NTU-RGB-D dataset [19] possesses large-scale 3D motions of various action categories; their pose annotation is severely inaccurate. We thus re-annotate this dataset with the help of a recent tool [15].

2 RELATED WORK

Multimodal 3D Human Dynamics Generation: There have been several prior endeavors in generating 3D human dynamics from various modalities, including audio, text and images. As audio signals do not contain explicit information of pose structure and motion dynamics, a common strategy is to learn the correlation between two modalities, and to obtain intrinsic representation of poses and motions. In [28], upper body gestures are generated based on speech signals. [29] employs LSTM based autoencoder to capture music-to-dance mapping, and [25] predicts body dynamics from violin and piano recital audios with LSTM model. A more recent work of [16] generates human dancing dynamics conditioned on music in a non-deterministic manner. However, due to the difficulties of collecting data, most of them are conducted in 2D space.

Motions could also be inferred from text, an emerging subject that is more related to our problem. Pioneer efforts such as [1, 18, 24] mainly resort to encoder-decoder RNN architecture for language-to-pose translation. The work of [2] learns a joint embedding space between sentences and human pose sequences. More recently, [26] applies more sophisticated neural translation network equipped with GANs for text-to-sign prediction.

The problem of action based human 2D motion generation is mostly related to our goal. Unfortunately there are few studies along this topic. Among them, [3] adopts two stage GAN framework to generate 2D human motion progressively. To our knowledge, our work is the first for action based 3D human motion generation.

Skeleton-based Human Pose Representations: Human pose representation plays a foundational role in modeling human motions. Proper representation may promote the realism of motions and foster performance robustness. Many existing methods directly utilize joints' coordinates to represent human skeletons [9, 12]. In [32], pair-wise relative positions of joints are used to represent human skeleton, while in [4], uninformative keypoints are further pruned during modeling. Another branch of human pose representation is part-based, which considers a skeleton as a connected set of segments. In [35], a human body is decomposed to five main parts, and a motion is parameterized by the temporal translations

and rotations of body parts. [7] represents a human skeleton with 3D joint angles, whereas the temporal information is modeled using dynamic time warping [21]. Finally, Lie group based methods such as [31], [11] and [20] characterize a human skeleton as kinematic chains, and represent 3D skeleton as a point in SO(3) or SE(3).

3D Human Motion Datasets: There have been several datasets targeting at human actions, with pose or 3D joint positions being marked either by stationary kinematic sensors or manual labors. CMU MoCap [5] and HDM05 [22] contain more than 100,000 poses and 2000 motions or pose sequences organized into multiple categories. However, the distribution of sequences among categories is highly unbalanced; the categories are often not necessarily aligned with typical action types. UTKinect-Action [33] and MSR-Action3D [17] possess more clearly defined action annotations; they however contain much less number of motions. NTU-RGB-D [19] is so far the largest human motion dataset, consisting of over 100,000 motions belonging to 120 classes. Yet the pose annotations are from MS Kinect that are inexact and noisy; the motions are temporally unstable. We are therefore motivated to curate an in-house 3D human action dataset, HumanAct12, as well as take effort in improving the pose annotation of NTU-RGB-D.

3 METHODOLOGY

Given an action category $\mathbf{a}$, our goal is to generate a motion – a 3D pose sequence, $M = [\mathbf{p}_1, ..., \mathbf{p}_T]$, of length T. Here $\mathbf{p}_t \in \mathbb{R}^D$ is the human pose at time t. Our framework, illustrated in Figure 2 and called **action2motion**, is essentially a conditional temporal VAE (Sec.3.2) equipped with a Lie algebra pose representation (Sec.3.1).

3.1 Disentangled Lie Algebra Representation

A human pose could be depicted as a kinematic tree consisting of five kinematic chains: spine and four limbs. With the help of Lie algebra theory, we are able to distill skeleton anatomical information, motion trajectories and bone lengths from the alternative representation of 3D joint locations.

Let $E = \{e_1, ..., e_N\}$ denote the set of oriented edges in a skeleton as oriented rigid body bones, with N being the number of bones. Each of the bones e_n is attached with its local coordinate system, with the bone itself orienting along the x-axis and the starting joint of the bone as origin. Let e_n and e_m be two consecutive bones along a kinematic chain, the coordinate transformation from e_n to e_m could be carried out through rotation and translation transformation matrices [11, 20, 34]. Mathematically, a joint with coordinates $\mathbf{c}_n^i = (x_n^i, y_n^i, z_n^i)^\top$ w.r.t coordinate system e_n will have coordinates $\mathbf{c}_m^i = (x_m^i, y_m^i, z_m^i)^\top$ w.r.t coordinate system e_m with

$$c_m^i = \begin{pmatrix} R_n & \mathbf{d}_n \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mathbf{c}_n^i \\ 1 \end{pmatrix}, \tag{1}$$

where $R_n \in \mathbb{R}^{3\times3}$ is the rotation matrix, and $\mathbf{d}_n \in \mathbb{R}^3$ is the translation vector along x-axis. Furthermore, $\mathbf{d}_n$ could be written as $(b_n, 0, 0)^\top$ with b_n denoting the bone length of e_n, a constant number over time.

Here, the rotation matrix R_n between two local coordinates is an element of Special Orthogonal Group SO(3), which is a matrix Lie group. Hence, excluding bone lengths, the relative geometry between e_n and e_m is a point in SO(3) and the whole skeleton is

represented as a point in SO(3)×SO(3)×...×SO(3), which is a matrix Lie group endowed with a differentiable manifold structure [11, 20]. Similarly, the motion could be characterized as a curve in Manifold. To carry out optimization in Manifold, we could engage the proper mathematical apparatus of Lie algebra or tangent space that could be regarded as a flat space, thus our familiar linear algebra computations could be utilized.

Lie algebra $\mathfrak{so}(3)$. The 3 by 3 identity matrix I_3 is an element of a SO(3) and is referred to as the identity element in this group [23]. The tangent space at the identity element I_3 of SO(3) is known as the Lie algebra space, $\mathfrak{so}(3)$, a 3-dimensional vector space spanned by the elements of a 3×3 skew-symmetric matrix $\hat{W}$, as

$$\hat{W} = \begin{pmatrix} 0 & -w_3 & w_2 \\ w_3 & 0 & -w_1 \\ -w_2 & w_1 & 0 \end{pmatrix}. \tag{2}$$

The association between a $R \in$ SO(3) and its Lie algebra vector $\mathbf{w} \in \mathfrak{so}(3)$ could be given by the logarithm map $\log_{\text{SO}(3)}$: SO(3) $\rightarrow \mathfrak{so}(3)$, as

$$\mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \frac{\theta}{2\sin\theta} \begin{pmatrix} R(3,2) - R(2,3) \\ R(1,3) - R(3,1) \\ R(2,1) - R(1,2) \end{pmatrix}, \tag{3}$$

where $\theta = \arccos \frac{trace(R)-1}{2}$ [23]. Since $\mathbf{w}$ is not uniquely mapped, we use the value with norm in range$[-\pi, \pi]$. Similarly, the inverse transformation is given by the exponential map $\exp^{\mathfrak{so}(3)}$: $\mathfrak{so}(3) \rightarrow$ SO(3). Please refer to [23] for more details.

Forward kinematics and motion trajectory. Let K denote the number of kinematic chains, m_k the number of joints in the k-th chain, and $\mathbf{w}_i^k$ the Lie algebra parameter vector of joint i in chain k. The Lie algebraic parameters of a pose could be formalized as a vector $\mathbf{p}_{\text{Lie}} = (\mathbf{w}_1^{1\top}, ..., \mathbf{w}_{m_1}^{1\top}, ..., \mathbf{w}_1^{K\top}, ..., \mathbf{w}_{m_K}^{K\top})^\top$. Then the 3D location of joint $\mathbf{J}_i^k$ in chain k could be obtained by **forward kinematics**,

$$\mathbf{J}_i^k = \left[\prod_{j=1}^{i-1} \exp(\hat{W}_j^k) \right] \mathbf{d}_i^k + \mathbf{J}_{i-1}^k. \tag{4}$$

Therefore, the Lie algebra parameters of a body pose could be transformed to 3D coordinates vector through joint-wise forward kinematics $\Gamma(\mathbf{p}_{\text{Lie}})$: $\mathbf{p}_{\text{Lie}} \rightarrow \mathbf{p}$, where $\mathbf{p} = (\mathbf{J}_1^{1\top}, ..., \mathbf{J}_{m_1}^{1\top}, ..., \mathbf{J}_1^{K\top}, ..., \mathbf{J}_{m_K}^{K\top})^\top$. Forward kinematics typically starts from a root joint $\mathbf{J_0} \in \mathbb{R}^3$ denoting the spatial translation of the entire human body, which is independent from the pose (i.e. Lie algebra parameters). Consider a motion with T consecutive poses, accordingly, the sequence $(\mathbf{J}_{0,1}, ..., \mathbf{J}_{0,T}) \in \mathbb{R}^{3\times T}$ forms the body motion trajectory, with $\mathbf{J}_{0,t}$ being the root translation of pose at time t.

Therefore, a human motion is a composition of three parts: motion trajectories, Lie algebraic parameters $\mathfrak{se}(3)$, and bone lengths. The Lie algebra of a motion is $M_{\text{Lie}} = (\mathbf{p}_{\text{Lie}}^1; ...; \mathbf{p}_{\text{Lie}}^T)$, with T being the number of poses. Note the number of effective Lie algebra parameters M_{Lie} is noticeably less than the number of elements in matrix M_{Lie}, since many bones rotate along less than three directions. At test stage, both Lie algebraic parameters M_{Lie} and motion trajectory are generated from our learned action2motion model. The skeletal bone lengths are acquired from typical real-life human bodies, which could be kept fixed during our motion generation
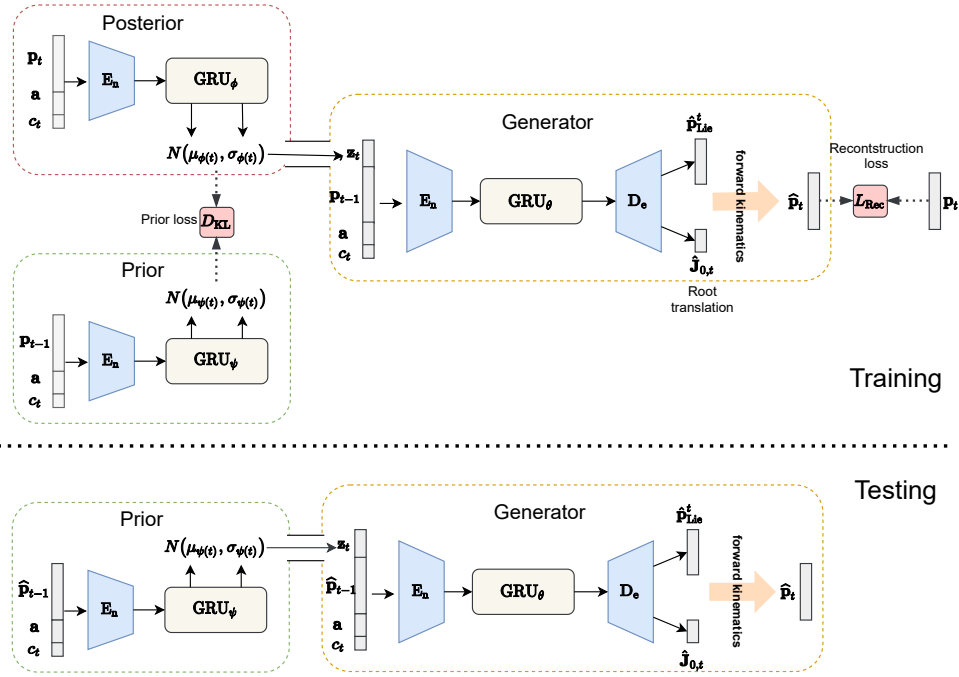
Figure 2: Overview of the proposed action2motion framework. (Top) Training phase: The input vector is a concatenation of action category a, time count c_t, and pose vector(p_t or p_{t-1}). The prior is obtained from the partial sequence of poses so far, $p_{1:t-1}$. KL-divergence is utilized to enforce posterior to be close to the prior. In generator, we first generate the Lie algebraic parameters and root translation of current pose, then produce the 3D joints positions by the physics law of forward kinematics (Sec3.1). (Bottom) Testing phase: A noise vector is sampled from the prior distribution, which kick-start the above-mentioned process in generate a 3D motion as sequence of poses.

process. The benefits are two-fold: (i) ensure the invariance of bone lengths in a motion, and (ii) enable the generation of motions with controllable body scale, i.e. by manually changing the bone length.

3.2 Conditional Temporal VAE

3.2.1 Preliminaries of temporal VAE. To generate a pose sequence, we would incorporate the Variational Auto-Encoder(VAE) unit into a recurrent module, with VAE handling the pose generation of each time-step, and RNN modeling the temporal dependence over time.

Formally, given a real motion or pose sequence $\mathbf{M} = [\mathbf{p}_1, ..., \mathbf{p}_T]$, VAE aims to maximize the probability of $\mathbf{M}$ sampled from the learned model distribution. At time t, RNN module $p_\theta(\mathbf{p}_t | \mathbf{p}_{1:t-1}, \mathbf{z}_{1:t})$ predicts the current pose $\mathbf{p}_t$ having latent variables $\mathbf{z}_{1:t}$, and conditioned on previous states $\mathbf{p}_{1:t-1}$. We rely on a variational neural network $q_\phi(\mathbf{z}_t | \mathbf{p}_{1:t})$ to approximate the true unknown posterior distribution $p_\theta(\mathbf{z}_t | \mathbf{p}_{1:t})$. This way, the objective of maximizing the data likelihood over the real sequence could be achieved as the following variational lower bound:

$$
\begin{aligned}
\log p_\theta(\mathbf{M}) &= \log \int_{\mathbf{z}} p_\theta(\mathbf{M}|\mathbf{z}) p(\mathbf{z}) \\
&\geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{M})} \log p_\theta(\mathbf{M}|\mathbf{z}) - D_{KL}(q_\phi(\mathbf{z}|\mathbf{M}) \parallel p(\mathbf{z})) \\
&= \sum_t \Big[\mathbb{E}_{q_\phi(\mathbf{z}_t|\mathbf{p}_{1:t})} \log p_\theta(\mathbf{p}_t|\mathbf{p}_{1:t-1}, \mathbf{z}_{1:t}) \\
&\quad - D_{KL}\Big(q_\phi(\mathbf{z}_t|\mathbf{p}_{1:t}) \parallel p(\mathbf{z}_t)\Big) \Big].
\end{aligned}
\tag{5}
$$

Here, the first term of the lower bounding function encourages the generated sample to be sufficiently close to the real sample; the second term penalizes the KL-divergence between prior and posterior distributions.

A simple form of the prior $p(\mathbf{z}_t)$ in VAE is a Gaussian with unitary variance, $\mathcal{N}(0, \mathbf{I})$. In practice, however, the prior distribution often vary with time. Take motions from the *walk* category for example, sometimes the pose variance could be small (e.g. roaming); sometimes could be large (e.g. sudden change of directions or velocity). The prior thus needs to be flexible enough to accommodate these variations. Intuitively, the prior of present time could be guessed given the context of previous time steps. Following [6], we parameterize the prior with a variational neural network $p_\psi(\mathbf{z}_t | \mathbf{p}_{1:t-1})$ conditioned on previous steps $\mathbf{p}_{1:t-1}$. Therefore, the variational lower bound of the sequence could be re-written as

$$
\begin{aligned}
\log p_\theta(\mathbf{M}) \geq \sum_t \Big[&\mathbb{E}_{q_\phi(\mathbf{z}_t|\mathbf{p}_{1:t})} \log p_\theta(\mathbf{p}_t|\mathbf{p}_{1:t-1}, \mathbf{z}_{1:t}) \\
&- D_{KL}\Big(q_\phi(\mathbf{z}_t|\mathbf{p}_{1:t}) \parallel p_\psi(\mathbf{z}_t|\mathbf{p}_{1:t-1})\Big) \Big].
\end{aligned}
\tag{6}
$$

The constraint between prior and posterior distribution further encourages temporal consistency.

3.2.2 Our Approach. Figure 2 depicts the architecture of our framework, which has three components: posterior network, prior network, and generator. The action category input is represented as an

one hot vector $\mathbf{a}$. In addition, a time counter $c_t \in [0, 1]$ is specifically used to keep record of where we are in the sequence generation progress, calculated as $\frac{t}{T}$. The rest of the input vector contains as a subvector the current pose $\mathbf{p}_t$. All encoders E_n and decoders D_e are modeled as linear fully connected layers. Prior network (p_ψ) and posterior (q_ϕ) share the same structure, but with different parameter values of the GRUs and encoders. The generator, p_θ, first produces the Lie algebraic parameters $\hat{\mathbf{p}}_{\mathrm{Lie}, t}$ and root translation $\hat{\mathbf{J}}_{0, t}$ via linear layers and GRU_θ, then obtains the 3D positions of skeleton $\hat{\mathbf{p}}_t$ via the forward kinematics of Eq.(4). In training phase, the pose generation process at time step t is

$$\mathbf{h}_t = \mathrm{E}_\mathrm{n}(\mathbf{p}_t, \mathbf{a}, c_t), \quad c_t = \frac{t}{T}$$

$$\left(\mu_\phi(t), \sigma_\phi(t)\right) = \mathrm{GRU}_\phi(\mathbf{h}_t)$$

$$\mathbf{z}_t \sim \mathcal{N}(\mu_\phi(t), \sigma_\phi(t))$$

$$\mathbf{g}_t = \mathrm{E}_\mathrm{n}(\mathbf{p}_{t-1}, \mathbf{a}, c_t, \mathbf{z}_t) \tag{7}$$

$$\mathbf{l}_t = \mathrm{GRU}_\theta(\mathbf{g}_t)$$

$$\left(\hat{\mathbf{p}}_{\mathrm{Lie}, t}, \hat{\mathbf{J}}_{0, t}\right) = \mathrm{D}_\mathrm{e}(\mathbf{l}_t)$$

$$\hat{\mathbf{p}}_t = \Gamma(\hat{\mathbf{p}}_{\mathrm{Lie}, t}, \hat{\mathbf{J}}_{0, t})$$

Similarly, the prior distribution $p_\psi(\mathbf{z}_t | \mathbf{p}_{1:t-1}, \mathbf{a}, c_t)$ is formed as follows,

$$\mathbf{h}_{t-1} = \mathrm{E}_\mathrm{n}(\mathbf{p}_{t-1}, \mathbf{a}, c_t), \quad c_t = \frac{t}{T}$$

$$\left(\mu_\psi(t), \sigma_\psi(t)\right) = \mathrm{GRU}_\psi(\mathbf{h}_{t-1}). \tag{8}$$

In testing phase, as real data $\mathbf{p}_t$ is not available, we instead sample $\mathbf{z}_t$ from the prior distribution $p_\psi(z_t | \cdot)$ to generate $\hat{\mathbf{p}}_t$.

3.2.3 A mixed training strategy. In our context, a mixed training strategy is adopted by randomly choosing whether or not using *teacher forcing* from a Bernoulli distribution $V \sim \mathrm{Bernoulli}(p_{\mathrm{tf}})$. Specifically, *teacher forcing* is chosen for the entire sequence $\mathbf{p}_{1:T}$ if V is 1, and vice versa. As a boundary condition, when creating the initial pose $\hat{\mathbf{p}}_1$, its previous pose input $\mathbf{p}_0$ for prior q_ψ is a zero vector.

3.2.4 Final Objective. Put together, our final objective function becomes

$$\mathcal{L}_{\theta, \phi, \psi} = -\sum_{t=1}^{T} \left[\mathbb{E}_{q_\phi(\mathbf{z}_t | \mathbf{p}_{1:t}, \mathbf{a}, c_t)} \log p_\theta(\mathbf{p}_t | \mathbf{p}_{1:t-1}, \mathbf{z}_{1:t}, \mathbf{a}, c_t) \right. \tag{9}$$
$$\left. - \lambda D_{KL}\left(q_\phi\left(\mathbf{z}_t | \mathbf{p}_{1:t}, \mathbf{a}, c_t\right) \| p_\psi\left(\mathbf{z}_t | \mathbf{z}_{1:t-1}, \mathbf{a}, c_t\right)\right) \right],$$

where λ is a tuning parameter for trade-off between reconstruction error and KL-divergence. In our implementation, the reconstruction term reduces to an ℓ_2 penalty between $\hat{\mathbf{p}}_t$ and $\mathbf{p}_t$; the model is also trained with the re-parameterization trick [14].

4 EXPERIMENTS

4.1 Datasets

The dataset of 3D human motions considered in our context is to possess multiple distinct action categories, with each action type containing a considerable amount of motions of diverse styles, and with proper pose annotations. Unfortunately the mainstreaming

datasets are not directly applicable. This has led us to revamp two existing datasets, NTU-RGB-D [19] and CMU MoCap with proper re-annotations, as well as constructing our in-house dataset, HumanAct12. Note in these datasets, the skeletons are all composed of 5 kinematic chains, and the root joint is located at pelvis.

NTU-RGB-D [19] originally contains 120 action types of 106 subjects. Its pose representation (3D joint positions) is from MS Kinect readout, which is known to unreliable and temporally inconsistent. We apply a state-of-art method [15] to re-estimate the 3D positions of 18 body joints (i.e. 19 bones) from the point cloud formed by aligning synchronized video feeds from multiple cameras. Note the poses are not necessarily matched perfectly to their true poses. It is sufficient here to be perceptually natural and realistic. A subset of 13 distinct actions are further chosen in our empirical evaluation, including e.g. *cheer up, pick up, salute*, constituting 3,900 motion clips.

CMU MoCap originally consists of 2,605 motion sequences, which however is not categorized by action type. Based on their motion descriptions, we identify 8 disparate actions, including *running, walking, jumping, climbing*, and manually re-organize 1,088 motions. Here each skeleton is annotated with 22 3D joints (19 bones). In practice, the pose sequences are down-sampled to a frequency of 12 HZ from 100 HZ.

HumanAct12 is the in-house dataset which is adopted from an existing dataset PHSPD [38, 39], consisting of 1,191 motion clips and 90,099 frames in total, with hierarchical action type annotations. To be specific, all motions are organized into 12 action categories, including *warm up, lift dumbbell*, and 34 subcategories including *warm up eblowback, lift dumbbell with right hand*). The fine-grained annotations offer more delicate details of the motions. In our experiments, we only use the coarse-grained action annotations. Compared to NTU-RGB-D, the 3D position annotations are more accurate, and the pose sequences are more stable. Compared to CMU MoCap, our HumanAct12has more organized action annotation, with more balanced number of motions per action. A body pose contains 24 joints (23 bones).

4.2 Experiment Results

4.2.1 Evaluation Metrics. Our quantitative evaluation is to examine the **naturality** and **diversity** of the generated 3D motions. Following [16], four metrics are considered for a comprehensive evaluation of the results: recognition accuracy, Frechet Inception Distance (FID) [10], diversity, and multimodality. FID is the most important indicator in our experiments: A **lower** FID suggests a better result. Note a result is claimed better than others on diversity and multimodality, only if its diversity and multimodality scores are **closer** to their respective values obtained from real motions. Finally, since there is no standard motion feature extractor, we train a standard RNN action recognition classifier for each dataset, and use its final layer as the motion feature extractor. Below are details of the four metrics:

- **Frechet Inception Distance** (FID): Features are extracted from 3,000 generated motions and real motions (obtained by sampling with replacement from the test set). Then FID is calculated between the feature distribution of generated motions vs. that of the

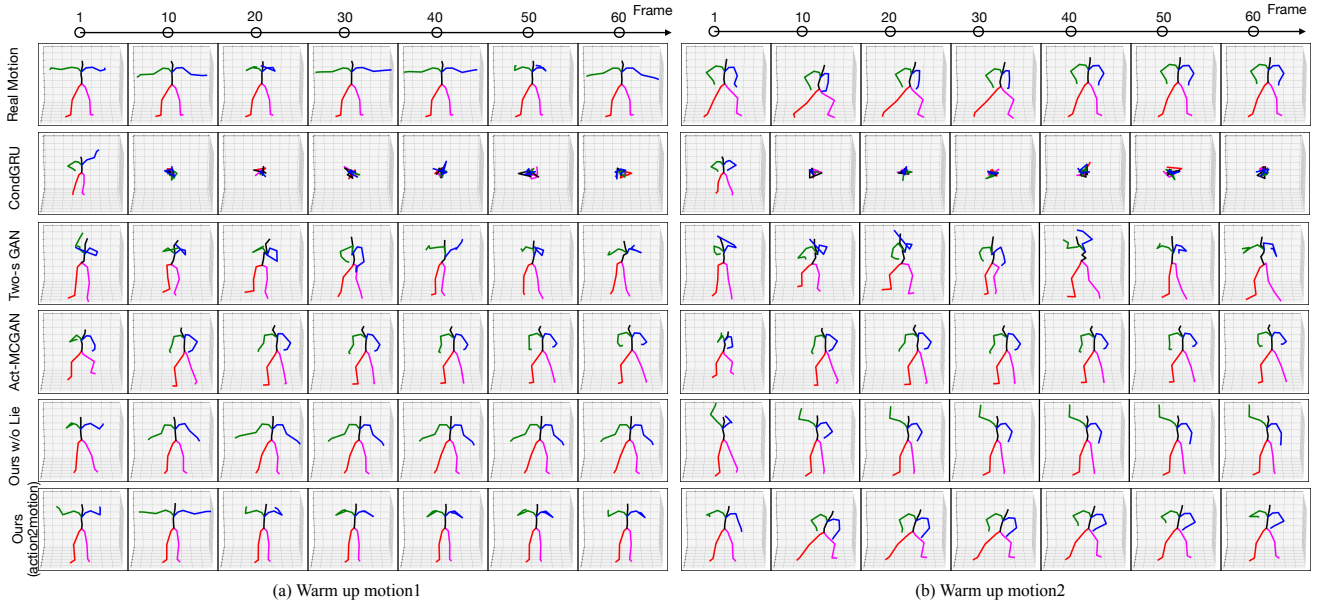(a) Warm up motion1   (b) Warm up motion2

Figure 3: Given action *warm up*, two motions are sampled from each motion source. In CondGRU, the first frame is required as part of the input; Nevertheless, in the remaining frames, all the joints collapse toward the root joints. Two-stage GAN, on the other hand, tends to produce jerking motions; the generated poses are also prone to perceivable distortion. Act-MoCoGAN oftentimes produce closely resembled pose sequences; its generated sequence also quickly reduce to stationary poses. Ours w/o Lie is able to generate motions with variation; meanwhile, its pose results often contain visible defects such as varying bone lengths and unnatural skeletons. Compared to the baselines, our approach (actoin2motion) excels in generating multiple distinct and realistic motions.

real motions. FID is an important metric widely used to evaluate the overall quality of generated motions.

- **Recognition Accuracy**: We use a pre-trained RNN action recognition classifier to classify the 3,000 motions, and calculate the overall recognition accuracy. The recognition accuracy indicates the correlation of the motion and its action type.

- **Diversity**: Diversity measures the variance of the generated motions across all action categories. From a set of all generated motions from various action types, two subsets of the same size S_d are randomly sampled. Their respective sets of motion feature vectors $\{\mathbf{v}_1, ..., \mathbf{v}_{S_d}\}$ and $\{\mathbf{v}'_1, ..., \mathbf{v}'_{S'_d}\}$ are extracted. The diversity of this set of motions is defined as

$$\text{Diversity} = \frac{1}{S_d} \sum_{i=1}^{S_d} \| \mathbf{v}_i - \mathbf{v}'_i \|_2 . \tag{10}$$

$S_d = 200$ is used in experiments.

- **Multimoldality**: Different from diversity, multimodality measures how much the generated motions diversify within each action type. Given a set of motions with C action types. For c-th action, we randomly sample two subsets with same size S_l, and then extract two subset of feature vectors $\{\mathbf{v}_{c,1}, ..., \mathbf{v}_{c,S_l}\}$ and $\{\mathbf{v}'_{c,1}, ..., \mathbf{v}'_{c,S_l}\}$. The multimodality of this motion set is formalized as

$$\text{Multimodality} = \frac{1}{C \times S_l} \sum_{c=1}^{C} \sum_{i=1}^{S_l} \left\| \mathbf{v}_{c,i} - \mathbf{v}'_{c,i} \right\|_2 . \tag{11}$$

$S_l = 20$ is used in experiments.

4.2.2 Comparison Methods. The problem of action conditioned 3D human motion generation is new. As a result there are few existing methods to compare with. Here, the following state-of-art methods from related areas are adapted for fair comparisons:

- **CondGRU**. We use conditional GRU as our deterministic baseline. Vanilla RNN model structure is considered for audio-to-motion translation in [25]. Here, we make minor modification that the model takes the condition vector and pose vector together as input at current step and output the pose vector for next step.

- **Two-stage GAN**. A two-stage GAN is proposed in [3] for action conditioned 2D human motion generation. A pose generator is first trained by WGAN, which is then plugged into a motion generator. The motion generator is learned to yield noise vector for pose generator at each time step, and a motion discriminator is employed to enforce temporal smoothness. Necessary modifications are made to this method to work in 3D.

- **Act-MoCoGAN**. MoCoGAN [30] is proposed for video generation, which produces a sequence of video frames from noise vectors and certain contents. By keeping the original architecture, we amend the video and image discriminators to motion and pose discriminators, respectively, for human dynamics generation.

- **Ours w/o Lie**. Here the Lie algebra representation in our action2motion approach is removed such that, the generator directly outputs the 3D position of joints. It is used to evaluate the effect of the proposed Lie algebra representation.

4.2.3 Qualitative Evaluation. Figure 3 compares the visual results of motions generated from different methods based on same action

Table 1: Performance evaluation on HumanAct12and NTU-RGB-D.(± indicates 95% confidence interval, and → means the closer to Real motions the better.)

Methods	HumanAct12				NTU-RGB-D			
	FID↓	Accuracy↑	Diversity→	Multimodality→	FID↓	Accuracy↑	Diversity→	Multimodality→
Real motions	$0.092^{\pm.007}$	$0.997^{\pm.001}$	$6.853^{\pm.053}$	$2.449^{\pm.038}$	$0.031^{\pm.004}$	$0.999^{\pm.001}$	$7.108^{\pm.048}$	$2.194^{\pm.025}$
CondGRU	$40.61^{\pm.144}$	$0.080^{\pm.002}$	$2.381^{\pm.020}$	$\mathbf{2.341^{\pm.036}}$	$28.31^{\pm.138}$	$0.078^{\pm.001}$	$3.663^{\pm.024}$	$3.578^{\pm.027}$
Two-stage GAN	$10.48^{\pm.089}$	$0.421^{\pm.006}$	$5.960^{\pm.049}$	$2.805^{\pm.036}$	$13.86^{\pm.091}$	$0.202^{\pm.003}$	$5.328^{\pm.039}$	$3.490^{\pm.027}$
Act-MoCoGAN	$5.610^{\pm.113}$	$0.793^{\pm.004}$	$6.752^{\pm.071}$	$1.055^{\pm.017}$	$2.723^{\pm.019}$	$\mathbf{0.997^{\pm.001}}$	$6.920^{\pm.061}$	$0.907^{\pm.009}$
Ours w/o Lie	$3.299^{\pm.079}$	$0.656^{\pm.005}$	$6.742^{\pm.046}$	$4.248^{\pm.037}$	$0.540^{\pm.047}$	$0.832^{\pm.004}$	$6.926^{\pm.049}$	$3.443^{\pm.052}$
Ours	$\mathbf{2.458^{\pm.079}}$	$\mathbf{0.923^{\pm.002}}$	$7.032^{\pm.038}$	$2.870^{\pm.037}$	$\mathbf{0.330^{\pm.008}}$	$0.949^{\pm.001}$	$\mathbf{7.065^{\pm.043}}$	$\mathbf{2.052^{\pm.030}}$

Table 2: Performance evaluation on CMU MoCap Dataset.

Methods	CMU MoCap			
	FID↓	Acc↑	Div	MModality
Real motions	$0.065^{\pm.006}$	$0.93^{\pm.002}$	$6.13^{\pm.079}$	$2.72^{\pm.066}$
CondGRU	$51.72^{\pm.123}$	$0.09^{\pm.001}$	$0.79^{\pm.011}$	$0.75^{\pm.016}$
Two-stage GAN	$14.34^{\pm.107}$	$0.17^{\pm.003}$	$4.41^{\pm.064}$	$\mathbf{1.62^{\pm.024}}$
Act-MoCoGAN	$11.15^{\pm.074}$	$0.44^{\pm.005}$	$5.28^{\pm.069}$	$1.51^{\pm.022}$
Ours w/o Lie	$2.994^{\pm.052}$	$0.37^{\pm.004}$	$5.79^{\pm.044}$	$5.00^{\pm.045}$
Ours	$\mathbf{2.885^{\pm.116}}$	$\mathbf{0.68^{\pm.003}}$	$6.50^{\pm.061}$	$4.12^{\pm.056}$

types. For conditional GRU, generated poses all collapse to a set a spatial points near the root joint, which shows the inefficacy of simple RNN models toward non-deterministic generative task. It is worth noting that motion generation should not be a one-to-one mapping process. Instead, the generated motions are expected to be *close* to the real motions in terms of their respective distributions.

We first investigate these methods on the motion naturality and correlations to prescribed action type. Two-stage GAN [3] generates roughly human-like poses; on the other hand, temporally the pose sequence typically contains jerking unnatural movements. Act-MoCoGAN can generate smoother motions with reasonable human poses; however, it tends to quickly descend into frozen poses even in a relatively short motion sequence. This may be partially attributed to the observation that learning sequential dependency across poses may go beyond the capability of standard GAN models. Ours w/o Lie algebra produces more natural motions that are also reasonable w.r.t. the given action type; there is still a notable gap though, when comparing to real motions. For instance, in the left sequence of Fig. 3, the human hands (i.e. blue and green lines) become longer from $t = 1$ to $t = 20$, and suddenly shrink in their lengths from $t = 20$ to $t = 60$. Consequently, the skeleton articulation is somewhat unnatural. Moreover, the latter portion of generated motions gradually freeze into fixed poses. The observed artifacts may be attributed to the nature of 3D coordinate representation that comes with strong entanglement and weak anatomical constraints. In contrast, our approach is capable of generating visually appealing motions that are also sensible to the given action type.

Diversity and multimodality are also important evaluation criteria. As observed from Figure 3, motions from Two-stage GAN are not well recognizable; Act-MoCoGAN suffers from low variations, which may due to the *mode collapse* issue of typical GAN methods.

In contrast, our action2motion is able to produce diverse motions. It highlights the merit of the proposed temporal conditional VAE on maintaining variations in sequential scenarios.

4.2.4 Quantitative Comparison. Evaluation results on HumanAct12 and NTU-RGB-D datasets are presented in Table 1, and the results over CMU MoCap dataset are shown in Table 2. For fair comparison, each experiment is repeated 20 times, and a statistical interval with 95% confidence is reported. Among the four metrics, FID is the most important indicator in evaluating the overall performance of a model. Meanwhile, recognition accuracy reflects the quality of models and the correlation with action type. Overall we have the following observations from the comparisons in Table 1 and Table 2 over four metrics and across three datasets. First, our approach, action2motion, clearly outperforms the comparison methods on FID. Deterministic methods such as CondGRU [25] is incapable of handling with such one-to-many generative task. The GANs employed in both two-stage GAN [3] and Act-MoCoGAN [30] help mitigating the problem, yet still not well enough. For recognition accuracy, our action2motion yields highest accuracy over HumanAct12and CMU MoCap dataset, which suggests the potential to generate highly recognizable motions and to capture the characteristics of action types. Act-MoCoGAN gains comparable accuracy. The explanation may be traced to the fact that Act-MoCoGAN has already incorporated the action classification loss in its training procedure. Meanwhile our action2motion reaches comparable or higher performance even without such explicit constraints during training; moreover, the use of Lie algebra representation significantly improves the recognition accuracy of generated motions over three datasets. The relative poor performances on CMU MoCap dataset may be attributed to high complexity of motions in this dataset.

The metrics of diversity and multimodality are important metrics complementary to FID and recognition accuracy. Note that for diversity and multimodality, the higher values are not necessarily better; instead the values are best to be close to those obtained from the set of real motions. Overall action2motion attains the closest values to real motions' for diversity and multimodality; Interestingly, compared to ours w/o Lie, introduce of Lie algebra representation lowers the abnormally high multimodality score while facilitating a proper level of diversity in all experiments. This, we believe, may come from the strength of action2motion in following the physics law, where Lie algebra representation enforces the production of valid motions from the motion kinematics manifold.
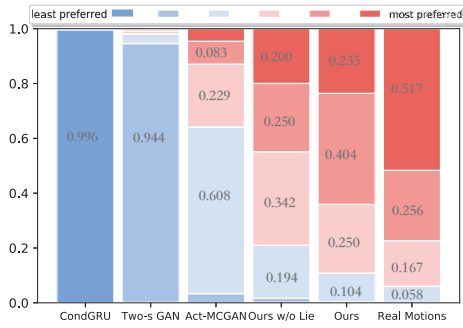
Figure 4: Preference results of generated motions. Bars [with] different color indicates the percentage of correspon[ding] preference of each source. For example, darkest blue ba[r in]dicates the percentage of motions ranked as least pref[erred] in each source.
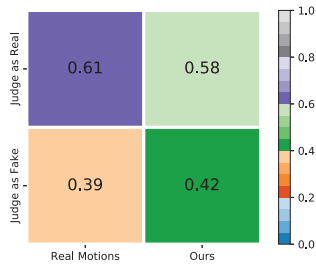

Figure 5: Human judge results of SMPL motions sampled from our method and real data in the testset.

4.2.5 Crowd-sourced Subjective Evaluation. In addition to the afore-mentioned objective evaluation, we also conduct visual cognitive evaluation: two user studies are conducted, involving 20 subjects of various age, gender, and race.

In the first user study, motions generated from multiple methods for the same action type are mixed together; users are then asked to rank their preference over these motions. The ranking is based on the visual perceptual quality of the generated motion, and how well it is matched with its action category. Fig. 4 illustrates the preference results. Compared to comparison methods, action2motion earns most appreciation from users; on the other hand, conditional RNN is clearly the least preferred choice, with two-stage GAN being the second least; Act-MoCoGAN is relatively more preferred, but still lacks positive attention from users. Ours w/o Lie starts to generate user friendly motions, with 20% motions are ranked at the first place by users. Action2motion further bridges the gap to real motions, with 64% generated motions being placed at top 2 positions by users. This human study solidly substantiates the capability of our approach toward synthesizing visually pleasing motions.

To further analyze the potential of our generated motions, we render a SMPL shape for each motion and ask users to discriminate whether it is real or fake (i.e. generated) of a known action type. Specifically, the second survey consists of 72 motions (i.e. half of them are generated, and half real) uniformly sampled from across all action types. As shown in Fig. 5, the motions from action2motion is visually only slightly inferior to real-life human motions: 58% of our generated motions are considered real by users, which is only

(a) Generated motion *Lift Dumbbell* with skeleton A(top) and skeleton B(bottom).

(b) Manipulating the motion of skeleton B through stretching length of *legs*(top), *arms*(middle) and *all bones*(bottom) by 1.25 times.
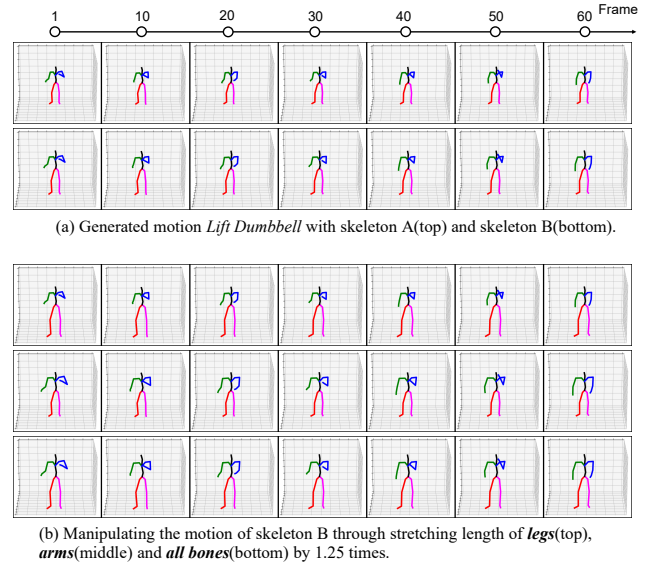
Figure 6: Examples of controllable motion generation. The upper sub-figure shows generating the same motion with different human skeletons. The lower sub-figure shows motions with varying scales of certain body parts.

3% lower than the number from real motions. The result suggests potentials of more interested downstream VR/AR applications, such as photo-realistic human video generation in gaming.

4.2.6 Controllable 3D Motion Generation. In Figure 6, we show the capacity of our approach in generating the same 3D motions with different human skeletons. In action2motion, this could be easily achieved by directly modifying bone lengths. In tradition methods, one could also vary the bone lengths of human motions by 3D rigid transformation. There are two advantages of our method. Firstly, it's easier to implement than transformation. Secondly, our method could change the length of arbitrary bones. Furthermore, our method sheds new lights on disentangled motion generation.

5 CONCLUSION AND OUTLOOK

Our paper looks at an emerging research problem of generating 3D human motions grounded on prescribed actions, and places special attention toward producing a diverse set of natural motions. This leads to the proposed action2motion, a conditional temporal VAE approach endowed with Lie algebraic representation. We also curate an in-house dataset and adapt two existing datasets to provide a suite of dedicated evaluation benchmarks. Extensive qualitative, quantitative, and subjective experiments demonstrate the effectiveness of our approach against the comparison methods. For future work, we plan to conduct more systematic investigation on a wider range of human actions, including those involving multiple people.

REFERENCES

[1] Hyemin Ahn, Timothy Ha, Yunho Choi, Hwiyeon Yoo, and Songhwai Oh. 2018. Text2action: Generative adversarial synthesis from language to action. In *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 5915–5920.
[2] Chaitanya Ahuja and Louis-Philippe Morency. 2019. Language2Pose: Natural Language Grounded Pose Forecasting. In *International Conference on 3D Vision (3DV)*. IEEE, 719–728.
[3] Haoye Cai, Chunyan Bai, Yu-Wing Tai, and Chi-Keung Tang. 2018. Deep video generation, prediction and completion of human action sequences. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 366–382.
[4] Alexandros Andre Chaaraoui, José Ramón Padilla-López, Pau Climent-Pérez, and Francisco Flórez-Revuelta. 2014. Evolutionary joint selection to improve human action recognition with RGB-D devices. *Expert systems with applications* 41, 3 (2014), 786–794.
[5] CMU. 2003. CMU graphics lab motion capture database. (2003).
[6] Emily Denton and Rob Fergus. 2018. Stochastic Video Generation with a Learned Prior. In *International Conference on Machine Learning (ICML)*. 1174–1183.
[7] Dariu M Gavrila, Larry S Davis, et al. 1995. Towards 3-d model-based tracking and recognition of human movement: a multi-view approach. In *International workshop on automatic face-and gesture-recognition*. Citeseer, 272–277.
[8] Liang-Yan Gui, Yu-Xiong Wang, Xiaodan Liang, and José MF Moura. 2018. Adversarial geometry-aware human motion prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 786–803.
[9] Fei Han, Brian Reily, William Hoff, and Hao Zhang. 2017. Space-time representation of people based on 3D skeletal data: A review. *Computer Vision and Image Understanding* 158 (2017), 85–105.
[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*. 6626–6637.
[11] Zhiwu Huang, Chengde Wan, Thomas Probst, and Luc Van Gool. 2017. Deep learning on lie groups for skeleton-based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 6099–6108.
[12] Mohamed E Hussein, Marwan Torki, Mohammad A Gowayyed, and Motaz El-Saban. 2013. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In *Twenty-Third International Joint Conference on Artificial Intelligence (IJCAI)*.
[13] Yunji Kim, Seonghyeon Nam, In Cho, and Seon Joo Kim. 2019. Unsupervised Keypoint Learning for Guiding Class-Conditional Video Prediction. In *Advances in Neural Information Processing Systems*. 3809–3819.
[14] Diederik P Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*.
[15] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. 2020. VIBE: Video Inference for Human Body Pose and Shape Estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*.
[16] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. 2019. Dancing to Music. In *Advances in Neural Information Processing Systems*. 3581–3591.
[17] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. 2010. Action recognition based on a bag of 3d points. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 9–14.
[18] Angela S Lin, Lemeng Wu, Rodolfo Corona, Kevin Tai, Qixing Huang, and Raymond J Mooney. 2018. generating animated videos of human activities from natural language descriptions. In *Proceedings of the Visually Grounded Interaction and Language Workshop at NeurIPS 2018*.
[19] Jun Liu, Amir Shahroudy, Mauricio Lisboa Perez, Gang Wang, Ling-Yu Duan, and Alex Kot Chichung. 2019. NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. *IEEE transactions on pattern analysis and machine intelligence* (2019).
[20] Zhenguang Liu, Shuang Wu, Shuyuan Jin, Qi Liu, Shijian Lu, Roger Zimmermann, and Li Cheng. 2019. Towards natural and accurate future motion prediction of humans and animals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 10004–10012.
[21] Meinard Müller. 2007. *Information retrieval for music and motion*. Vol. 2. Springer.
[22] Meinard Müller, Tido Röder, Michael Clausen, Bernhard Eberhardt, Björn Krüger, and Andreas Weber. [n.d.]. Mocap database hdm05. ([n. d.]).
[23] Richard M Murray, Zexiang Li, and S Shankar Sastry. 1994. *A mathematical introduction to robotic manipulation*. CRC press.
[24] Matthias Plappert, Christian Mandery, and Tamim Asfour. 2018. Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks. *Robotics and Autonomous Systems* 109 (2018), 13–26.
[25] Eli Shlizerman, Lucio Dery, Hayden Schoen, and Ira Kemelmacher-Shlizerman. 2018. Audio to body dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7574–7583.
[26] Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. 2020. Text2Sign: Towards Sign Language Production Using Neural Machine Translation and Generative Adversarial Networks. *International Journal of Computer Vision* 128 (2020), 891âĂŞ908.
[27] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2015. What makes tom hanks look like tom hanks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 3952–3960.
[28] Kenta Takeuchi, Dai Hasegawa, Shinichi Shirakawa, Naoshi Kaneko, Hiroshi Sakuta, and Kazuhiko Sumi. 2017. Speech-to-gesture generation: A challenge in deep learning approach with bi-directional LSTM. In *Proceedings of the 5th International Conference on Human Agent Interaction*. 365–369.
[29] Taoran Tang, Jia Jia, and Hanyang Mao. 2018. Dance with melody: An lstm-autoencoder approach to music-oriented dance synthesis. In *Proceedings of the 26th ACM international conference on Multimedia (ACM MM)*. 1598–1606.
[30] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. 2018. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 1526–1535.
[31] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. 2014. Human action recognition by representing 3d skeletons as points in a lie group. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 588–595.
[32] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. 2012. Mining actionlet ensemble for action recognition with depth cameras. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 1290–1297.
[33] Lu Xia, Chia-Chih Chen, and Jake K Aggarwal. 2012. View invariant human action recognition using histograms of 3d joints. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 20–27.
[34] Chi Xu, Lakshmi Narasimhan Govindarajan, Yu Zhang, and Li Cheng. 2017. Lie-X: Depth image based articulated object pose estimation, tracking, and action recognition on lie groups. *International Journal of Computer Vision* 123, 3 (2017), 454–478.
[35] Yaser Yacoob and Michael J Black. 1999. Parameterized modeling and recognition of activities. *Computer Vision and Image Understanding* 73, 2 (1999), 232–247.
[36] Xinchen Yan, Akash Rastogi, Ruben Villegas, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, Ersin Yumer, and Honglak Lee. 2018. Mt-vae: Learning motion transformations to generate multimodal human dynamics. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 265–281.
[37] Ceyuan Yang, Zhe Wang, Xinge Zhu, Chen Huang, Jianping Shi, and Dahua Lin. 2018. Pose guided human video generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 201–216.
[38] Shihao Zou, Xinxin Zuo, Yiming Qian, Sen Wang, Chi Xu, Minglun Gong, and Li Cheng. 2020. 3D Human Shape Reconstruction from a Polarization Image. In *Proceedings of the European Conference on Computer Vision (ECCV)*.
[39] Shihao Zou, Xinxin Zuo, Yiming Qian, Sen Wang, Chi Xu, Minglun Gong, and Li Cheng. 2020. Polarization Human Shape and Pose Dataset. arXiv:arXiv:2004.14899