Action2Motion: Conditioned Generation of 3D Human Motions

Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, Li Cheng

1Department of Electrical and Computer Engineering, University of Alberta
2Physical Sciences Division, University of Chicago
3Graduate School of Arts and Sciences, Yale University
4School of Computer Science, University of Guelph

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

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. 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

ACM Reference Format:
Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng and Minglun Gong. 2020. Action2Motion: Conditioned Generation of 3D Human Motions. In Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3394171.3413635

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...
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 various applications, 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] (Sec.3.2) equipped with a Lie algebra pose representation (Sec.3.1). Possesses 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 a, our goal is to generate a motion – a 3D pose sequence, \( M = \{ p_1, ..., p_T \} \), of length \( T \). Here \( p_t \in \mathbb{R}^D \) 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 \( c^i_n = (x^{i}_n, y^{i}_n, z^{i}_n)^\top \) w.r.t coordinate system \( e_n \) will have coordinates \( c^i_m = (x^{i}_m, y^{i}_m, z^{i}_m)^\top \) w.r.t coordinate system \( e_m \) with

\[
\begin{bmatrix}
    c^i_m \\
    1
\end{bmatrix} =
\begin{bmatrix}
    R_n & d_n \\
    0 & 1
\end{bmatrix}
\begin{bmatrix}
    c^i_n \\
    1
\end{bmatrix},
\]

where \( R_n \in \mathbb{R}^{3 \times 3} \) is the rotation matrix, and \( d_n \in \mathbb{R}^3 \) is the translation vector along x-axis. Furthermore, \( d_n \) 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 \( SO(3) \times SO(3) \times \cdots \times SO(3) \), 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 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, \( so(3) \), a 3-dimensional vector space spanned by the elements of a \( 3 \times 3 \) skew-symmetric matrix \( \hat{W} \), as

\[
\hat{W} = \begin{pmatrix}
0 & -w_3 & w_2 \\
w_3 & 0 & -w_1 \\
-w_2 & w_1 & 0
\end{pmatrix},
\]

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

\[
w = \frac{1}{2 \sin \theta} \begin{pmatrix}
R(3, 2) - R(2, 3) \\
R(1, 3) - R(3, 1) \\
R(2, 1) - R(1, 2)
\end{pmatrix},
\]

where \( \theta = \arccos \left( \frac{I_{3 \times 3}^{\top} \cdot R \cdot I_{3 \times 3}}{2} \right) \) [23]. Since \( w \) is not uniquely mapped, we use the value with norm in range \([-\pi, \pi]\). Similarly, the inverse transformation is given by the exponential map \( \exp_{so(3)}(w) \) \( \to SO(3) \). 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 \( w_k \) the Lie algebra parameter vector of joint \( i \) in chain \( k \). The Lie algebraic parameters of a pose could be formalized as a vector \( p_{Lie} = (w_1^T, ..., w_{m_1}^T, ..., w_k^T, ..., w_{m_K}^T)^T \). Then the 3D location of joint \( J_k^i \) in chain \( k \) could be obtained by **forward kinematics**,

\[
J_k^i = \left[ \prod_{j=1}^{i-1} \exp(\hat{W}_j^T) \right] d_k^i + J_k^{0-1},
\]

Therefore, the Lie algebra parameters of a pose could be transformed to 3D coordinates vector through joint-wise forward kinematics \( \Gamma(p_{Lie}) : p_{Lie} \to p \), where \( p = (J_1^T, ..., J_{m_1}^T, ..., J_k^T, ..., J_{m_K}^T)^T \). Forward kinematics typically starts from a root joint \( J_0 \in \mathbb{R}^3 \) 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 \((J_0, ..., J_{0,T}) \in \mathbb{R}^{3 \times T} \) forms the body motion trajectory, with \( J_{0,t} \) being the root translation of pose at time \( t \).

Therefore, a human motion is a composition of three parts: motion trajectories, Lie algebraic parameters \( \phi(e, 3) \), and bone lengths. The Lie algebra of a motion is \( M_{Lie} = (p_{Lie}^1, ..., p_{Lie}^T) \), 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.
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 \( M = [p_1, ..., p_T] \), VAE aims to maximize the probability of \( M \) sampled from the learned model distribution. At time \( t \), RNN module \( p_\theta(p_t|p_{1:t-1}, z_{1:t}) \) predicts the current pose \( p_t \) having latent variables \( z_{1:t} \) and conditioned on previous states \( p_{1:t-1} \). We rely on a variational neural network \( q_\phi(z_t|p_{1:t}) \) to approximate the true unknown posterior distribution \( p_\theta(z_t|p_{1:t}) \). This way, the objective of maximizing the data likelihood over the real sequence could be achieved as the following variational lower bound:

\[
\log p_\theta(M) = \log \int p_\theta(M|z)p(z) \geq \mathbb{E}_{q_\phi(z_t|p_{1:t})} \log p_\theta(M|z) - D_{KL}(q_\phi(z_t|M) \parallel p(z)) = \sum_t \left[ \mathbb{E}_{q_\phi(z_t|p_{1:t})} \log p_\theta(p_t|p_{1:t-1}, z_{1:t}) - D_{KL}(q_\phi(z_t|p_{1:t}) \parallel p(z)) \right].
\]

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 \( p(z_t) \) in VAE is a Gaussian with unitary variance, \( N(0, I) \). 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 \( p_\phi(z_t|p_{1:t-1}) \) conditioned on previous steps \( p_{1:t-1} \). Therefore, the variational lower bound of the sequence could be re-written as

\[
\log p_\theta(M) \geq \sum_t \left[ \mathbb{E}_{q_\phi(z_t|p_{1:t})} \log p_\theta(p_t|p_{1:t-1}, z_{1:t}) - D_{KL}(q_\phi(z_t|p_{1:t}) \parallel p(z)) \right].
\]

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 $a$. 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 $p_t$. All encoders $E_q$ and decoders $D_e$ are modeled as linear fully connected layers. Prior network ($p_g$) and posterior ($q_g$) share the same structure, but with different parameter values of the GRUs and encoders. The generator, $p_g$, first produces the Lie algebraic parameters $\hat{p}_{Lie,t}$ and root translation $\hat{J}_0,t$ via linear layers and GRU$_t$, then obtains the 3D positions of skeleton $\hat{p}_t$ via the forward kinematics of Eq.(4). In training phase, the pose generation process at time step $t$ is

$$h_t = E_q(p_t, a, c_t), \quad c_t = \frac{t}{T}$$

$$(\mu_\phi(t), \sigma_\phi(t)) = \text{GRU}_\phi(h_t)$$

$$z_t \sim \mathcal{N}(\mu_\phi(t), \sigma_\phi(t))$$

$$g_t = E_u(p_{t-1}, a, c_t, z_t)$$

$$I_t = \text{GRU}_g(g_t)$$

$$(\hat{p}_{Lie,t}, \hat{J}_0,t) = D_e(I_t)$$

$$\hat{p}_t = \Gamma(\hat{p}_{Lie,t}, \hat{J}_0,t)$$

Similarly, the prior distribution $p_\phi(z_t|p_{1:t-1}, a, c_t)$ is formed as follows,

$$h_{t-1} = E_q(p_{t-1}, a, c_t), \quad c_t = \frac{t}{T}$$

$$(\mu_\phi(t), \sigma_\phi(t)) = \text{GRU}_\phi(h_{t-1}).$$

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

3.2.3 A mixed training strategy. As shown in Figure 2, in testing phase, output from the previous time step $\hat{p}_{t-1}$ is used as input to generate current pose, $\hat{p}_t$; while in training phase, ground truth data $p_{t-1}$ is employed in producing $\hat{p}_t$, a technique known as teacher forcing in sequence prediction. However, the discrepancy of curricula between training and testing compromises the stability of the trained model. An alternative approach is to apply the last generated pose $\hat{p}_{t-1}$ as input into current step in training, which also raises concern on divergence: as soon as the prediction $\hat{p}_t$ deviates from ground truth $p_t$, the error may escalate in follow-up steps $t + 1 : T$. In our context, a mixed strategy is adopted by randomly choosing among these two scenarios from a Bernoulli distribution $V \sim \text{Bernoulli}(p_{hit})$. Specifically, teacher forcing is chosen for the entire sequence $p_{1:T}$ if $V$ is 1, and vice versa. As a boundary condition, when creating the initial pose $\hat{p}_1$, its previous pose input $p_0$ for prior $q_\phi$ is a zero vector.

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

$$L_0,\phi,\psi = -\sum_{t=1}^{T} \left[ E_q(z_t|p_{1:t-1}, a, c_t) \log p_g(p_t|p_{1:t-1}, z_t, a, c_t) \right. \right.$$  

$$\left. - \lambda D_{KL} \left( q_\phi(z_t|p_{1:t-1}, a, c_t) \parallel p_g(z_t|z_{1:t-1}, a, c_t) \right) \right].$$

where $\lambda$ is a tuning parameter for trade-off between reconstruction error and KL-divergence. In our implementation, the reconstruction term reduces to an $\ell_2$ penalty between $\hat{p}_t$ and $p_t$; 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 90,999 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 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 $\{v_1, ..., v_{S_d}\}$ and $\{v'_1, ..., v'_{S_d}\}$ are extracted. The diversity of this set of motions is defined as

  \[
  \text{Diversity} = \frac{1}{S_d} \sum_{i=1}^{S_d} \| v_i - v'_i \|_2 .
  \]  

  $S_d = 200$ is used in experiments.

- **Multimodality:** 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 $S_t$, and then extract two subset of feature vectors $\{v_{c,1}, ..., v_{c,S_t}\}$ and $\{v'_{c,1}, ..., v'_{c,S_t}\}$. The multimodality of this motion set is formalized as

  \[
  \text{Multimodality} = \frac{1}{C S_t^2} \sum_{c=1}^{C} \sum_{i=1}^{S_t} \| v_{c,i} - v'_{c,i} \|_2 .
  \]  

  $S_t = 20$ 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,
worth noting that motion generation should not be a one-to-one mapping process. Instead, the generated motions are expected to be roughly human-like poses; on the other hand, temporally the pose correlations to prescribed action type. Two-stage GAN [3] generates a spatial points near the root joint, which shows the inefficacy of type. For conditional GRU, generated poses all collapse to a set of motions generated from different methods based on same action type. Act-MoCoGAN can generate smoother motions with reasonable human looks; 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.

### Table 1: Performance evaluation on HumanAct12 and NTU-RGB-D.

| Methods     | HumanAct12 | NTU-RGB-D |
|-------------|------------|-----------|
|             | FID\(\uparrow\) | Accuracy\(\uparrow\) | Diversity\(\rightarrow\) | Multimodality\(\rightarrow\) | FID\(\uparrow\) | Accuracy\(\uparrow\) | Diversity\(\rightarrow\) | Multimodality\(\rightarrow\) |
| Real motions | 0.092\(\pm\)0.007 | 0.997\(\pm\)0.001 | 6.853\(\pm\)0.053 | 2.449\(\pm\)0.038 | 0.031\(\pm\)0.004 | 0.999\(\pm\)0.001 | 7.108\(\pm\)0.048 | 2.194\(\pm\)0.025 |
| CondGRU     | 40.61\(\pm\)1.144 | 0.080\(\pm\)0.002 | 2.381\(\pm\)0.020 | 2.341\(\pm\)0.036 | 28.31\(\pm\)1.338 | 0.078\(\pm\)0.001 | 3.663\(\pm\)0.024 | 3.578\(\pm\)0.027 |
| Two-stage GAN | 10.48\(\pm\)0.089 | 0.421\(\pm\)0.006 | 5.960\(\pm\)0.049 | 2.805\(\pm\)0.036 | 13.86\(\pm\)0.911 | 0.202\(\pm\)0.003 | 5.328\(\pm\)0.039 | 3.490\(\pm\)0.027 |
| Act-MoCoGAN | 5.610\(\pm\)1.13 | 0.793\(\pm\)0.004 | 6.752\(\pm\)0.071 | 1.055\(\pm\)0.017 | 2.723\(\pm\)0.019 | 0.997\(\pm\)0.001 | 6.920\(\pm\)0.061 | 0.907\(\pm\)0.009 |
| Ours w/o Lie | 3.299\(\pm\)0.079 | 0.656\(\pm\)0.005 | 6.742\(\pm\)0.046 | 4.248\(\pm\)0.037 | 0.540\(\pm\)0.047 | 0.832\(\pm\)0.004 | 6.926\(\pm\)0.049 | 3.443\(\pm\)0.052 |
| Ours        | 2.458\(\pm\)0.079 | 0.923\(\pm\)0.002 | 7.032\(\pm\)0.038 | 2.870\(\pm\)0.037 | 0.330\(\pm\)0.008 | 0.949\(\pm\)0.001 | 7.065\(\pm\)0.043 | 2.052\(\pm\)0.030 |

### Table 2: Performance evaluation on CMU MoCap Dataset.

| Methods     | CMU MoCap | |
|-------------|-----------|-------------|
|             | FID\(\uparrow\) | Acc\(\uparrow\) | Div | MModality |
| Real motions | 0.065\(\pm\)0.006 | 0.93\(\pm\)0.002 | 6.13\(\pm\)0.079 | 2.72\(\pm\)0.066 |
| CondGRU     | 51.72\(\pm\)1.23 | 0.09\(\pm\)0.001 | 0.79\(\pm\)0.011 | 0.75\(\pm\)0.016 |
| Two-stage GAN | 14.34\(\pm\)0.17 | 0.17\(\pm\)0.003 | 4.41\(\pm\)0.064 | 1.62\(\pm\)0.024 |
| Act-MoCoGAN | 11.15\(\pm\)0.074 | 0.44\(\pm\)0.005 | 5.28\(\pm\)0.099 | 1.51\(\pm\)0.022 |
| Ours w/o Lie | 2.994\(\pm\)0.052 | 0.37\(\pm\)0.004 | 5.79\(\pm\)0.044 | 5.00\(\pm\)0.045 |
| Ours        | 2.885\(\pm\)0.116 | 0.68\(\pm\)0.003 | 6.50\(\pm\)0.061 | 4.12\(\pm\)0.056 |

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 types. For conditional GRU, generated poses all collapse to a set of 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 naturalness 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 HumanAct12 and 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.

4.2.5 Crowd-sourced Subjective Evaluation. In addition to the aforementioned 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 they 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 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 the 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 shed new lights on disentangled human 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 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 two or more people.

ACKNOWLEDGMENTS

This work is supported by the University of Alberta Startup-grant, the NSERC Discovery Grants including No. RGPIN-2019-04575, and the University of Alberta-Huawei Joint Innovation Collaboration grants.

REFERENCES

[1] Hyemin Ahn, Timothy Ha, Yunho Choi, Fhwyoon Yoo, and Songhwa 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 Ahoja 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 Chazaraouz, José Ramón Padilla-López, Pau Climent-Pérez, and Francisco Flórez-Revelo. 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 of human body motion. In Proceedings of the Conference on Computer Vision and Pattern Recognition-Workshops. 3581–3591.

[8] Fei Han, Brian Reily, William Hoff, and Hao Zhang. 2018. Adversarial geometry-aware human motion prediction. In Proceedings of the European Conference on Computer Vision (ECCV). 786–803.

[9] 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.

[10] JohnFo. 2019. A fast and easy way to train a human action recognition network. arXiv:1909.09410.

[11] Jakob Hossfeld, Jure Žabkar, and Blaz Dežman. 2018. Action prediction from video using a generative adversarial network. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 9661–9670.

[12] Qixing Huang, Angela S Lin, Lemeng Wu, Rodolfo Corona, Kevin Tai, Oqting Huang, and Raymond J Mooney. 2020. Language generation for music and motion. Vol. 2. Springer.

[13] Yunji Kim, Seonghyeon Nam, In Cho, and Seon Joo Kim. 2019. Unsupervised representation learning for motion generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 10004–10012.

[14] Sabina Kruijt, Luke Cadden-Brown, and Richard Bowden. 2020. Text2Sign: Towards Sign Language Production Using Neural Machine Translation and Generative Adversarial Networks. International Journal of Computer Vision 128 (2020), 891–908.

[15] Leilani Otoo, Moritz Hardt, and Alex Kot Chichung. 2019. NTU RGB+D 120: A Large-Scale Benchmark for Human Action Recognition with RGB-D Devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 7574–7583.

[16] Oriol Vinyals, Samy Bengio, Christian Lamboley, Matthew Illescas, and Quoc Le. 2015. Unsupervised sequence-to-sequence learning. In Proceedings of the 32nd International Conference on Machine Learning (ICML). 267–276.

[17] Tianfu Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. 2012. Mining actionlet keypoints for guiding class-conditional video prediction. In Twenty-Third International Joint Conference on Artificial Intelligence (IJCAI).

[18] Xu Zhu, Zhihang Wang, Zhong Zhan, Zhongshuai Li, and Cheng Zhang. 2020. Human Activity Recognition and Prediction using a Hierarchical Attention Network. In Proceedings of the 26th ACM international conference on Multimedia (ACM MM). 1598–1606.

[19] Zitao Zhou, Victor Zhao, Kai Gu, and Quanzheng Li. 2020. Action2Text: Learning neural action generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 3809–3819.

[20] Zhiyu Zhuang, Jie Gao, and Sebastian Tramel. 2019. A comprehensive survey of human action recognition from RGB-D videos. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 3592–3599.

A APPENDIX

A.1 Neural Network Architecture

Table 3 illustrates the architecture we used on dataset HumanAct12. For other two datasets, the dimension of input vector may vary according to the dimension of pose vector.
A.2 Implementation Details

Our approach is implemented by PyTorch. The output size of all encoder layers and decoder layer is set to 128 and 72, respectively. A two-layer GRU is used for generator, and one layer GRU is for both prior network and posterior network, all having the same hidden unit size of 128. The noise vector $z$ is 30 dimensional. Altogether, our learned model has about 450,000 parameters. The Adam optimizer is applied for training in all experiments, with learning rate of 0.0002, weight decaying of 0.00001, and default parameter values including $\beta_1 = 0.9$, $\beta_2 = 0.999$. We train our model with a mini-batch size of 128. To stabilize the training, the teacher forcing rate $p_{tf}$ is set to 0.6. Values of the above hyper-parameters are fixed throughout empirical evaluations across all datasets. Afterwards, we generate motion with length of 60, 100 and 60 on NTU-RGB-D, CMU MoCap and HumanAct12, respectively. The hyper-parameter $\lambda$ is a trade-off between reconstruction constraints and KL-divergence penalty, with its value in the above datasets setting to 0.1, 0.1 and 0.01, respectively. Empirically, a large lambda will increase the quality of generated motion but may decrease motion diversity; a flipped effect is held for a small lambda.

A.3 Exemplar Human Skeleton

Fig. 7 gives an exemplar of human skeleton anatomical structure. This skeleton contains 21 joints and 20 bones, and there are five kinematic chains in this skeleton which are

- spine: $\{j_0, j_1, j_2, j_3, j_4\}$,
- left arm: $\{j_3, j_5, j_6, j_7, j_8\}$,
- right arm: $\{j_3, j_9, j_{10}, j_{11}, j_{12}\}$,
- left leg: $\{j_6, j_{13}, j_{14}, j_{15}, j_{16}\}$
- right leg: $\{j_6, j_{17}, j_{18}, j_{19}, j_{20}\}$

A.4 Training Process Comparison

Fig. 8 gives an illustrative example of the training process comparison between our model with and without Lie algebra representation. Without Lie algebra representation, our model keep generating poses with obvious artifacts at iteration 10,000, and evolves to produce natural and stable walk motion within 500,000 iterations. Incorporation of Lie significantly cut down the number of iterations to 40,000. In addition, ours model with Lie algebra is ready to generate plausible pose sequences within 5,000 iterations.

A.5 Design of User Study

We conducted two user studies in our experiments, one is user preference(Fig.9(top)) survey and another is fake or real(Fig.9(bottom)). Practically, HTML is utilized to perform our survey for better user experience. In user preference survey, for each question, motions from different sources are mixed up with random order; 6 motion clips and the action type(e.g. eat) are shown to users; and users are asked to give their preference over these 6 motion clips. In fake or real survey, for each action type 6 motions sampled from our model and
real data (not necessarily equal distributed) are placed together, and users need to discriminate each motion as the generated or the real. Users are told to try not to make decision by comparing one with another.

A.6 Dataset Details

A.6.1 Improved NTU-RGB-D. In our experiments, the recent video 3d shape estimation method [15] is employed for re-annotating partial NTU-RGB-D dataset. As shown in Tab.4, 13 action types are picked from the original dataset, forming an refined dataset with 3902 motion clips in total. We evidence the feasibility of improving NTU-RGB-D dataset for motion generation task, and may further re-annotate larger proportion of the dataset in the future.

A.6.2 HumanAct12. Our dataset HumanAct12 is adopted from the dataset PHSPD proposed in [39]. The original dataset contains more dedicated multimodal human pose resources including polar images, RGB images, depth images, ultra images and 3d annotations. We select a subset of dataset PHSPD, and cut the long pose sequences...
Table 4: Statistics of improved NTU-RGB-D dataset.

| Action Label                        | Number of Motions |
|-------------------------------------|-------------------|
| Butt kicks (kick backward)          | 287               |
| Cheer up                            | 309               |
| Hand waving                         | 309               |
| Kicking something                   | 306               |
| Pick up                             | 306               |
| Running on the spot                 | 288               |
| Salute                              | 308               |
| Shake fist                          | 287               |
| Side kick                           | 288               |
| Sitting down                        | 306               |
| Squat down                          | 306               |
| Standing up (from sitting position) | 311               |
| Throw                               | 305               |
| **Entire Dataset**                  | **3902**          |

into smaller pieces which aligns with prescribed action types. The statistics of our dataset is given in Tab.5, where there are 1061 motion clips which are categorized into 12 action classes and 34 sub-classes. Our dataset only contains motions with 3D positions and corresponding action type annotation. But others modality (i.e, RGB images, etc.) are accessible in PHSPD via the alignment between our dataset and PHSPD.
Table 5: Statistics of dataset HumanAct12.

| Coarse-grained Label       | Fine-grained Label                        | Number of Motions | Total Number |
|----------------------------|------------------------------------------|-------------------|--------------|
| Warm up                    | Warm_up_wristankle                       | 25                |              |
|                            | Warm_up_pectoral                        | 49                |              |
|                            | Warm_up_eblowback                       | 43                |              |
|                            | Warm_up_bodylean_right_arm              | 26                | 215          |
|                            | Warm_up_bodylean_left_arm               | 24                |              |
|                            | Warm_up_bow_right                       | 24                |              |
|                            | Warm_up_bow_left                        | 24                |              |
| Walk                       | Walk                                     | 47                | 47           |
| Run                        | Run                                      | 50                | 50           |
| Jump                       | Jump_handsup                             | 54                | 94           |
|                            | Jump_vertical                           | 40                |              |
| Drink                      | Drink_bottle_righthand                  | 27                |              |
|                            | Drink_bottle_lefthand                   | 43                |              |
|                            | Drink_cup_righthand                     | 11                | 88           |
|                            | Drink_cup_lefthand                      | 3                 |              |
|                            | Drink_both_hands                        | 4                 |              |
| Lift_dumbbell              | Lift_dumbbell_righthand                 | 45                |              |
|                            | Lift_dumbbell_lefthand                  | 45                |              |
|                            | Lift_dumbbell_bothhands                 | 47                | 218          |
|                            | Lift_dumbbell_overhead                  | 43                |              |
|                            | Lift_dumbbell_bothhands_bend_legs       | 38                |              |
| Sit                        | Sit                                      | 54                | 54           |
| Eat                        | Eat_righthand                            | 33                |              |
|                            | Eat_lefthand                             | 25                | 77           |
|                            | Eat_pie/burger                          | 19                |              |
| Turn_steering_wheel        | Turn_steering_wheel                     | 56                | 56           |
| Phone                      | Take out phone, call and put back       | 28                | 61           |
|                            | Call with left hand                     | 33                |              |
| Boxing                     | Boxing_left_right                       | 26                |              |
|                            | Boxing_left_upwards                     | 39                | 140          |
|                            | Boxing_right_upwards                    | 41                |              |
|                            | Boxing_right_left                       | 34                |              |
| Throw                      | Throw_right_hand                        | 53                | 91           |
|                            | Throw_both_hand                         | 38                |              |
| Entire Dataset             | -                                        | -                 | 1191         |