A Unified 3D Human Motion Synthesis Model via Conditional Variational Auto-Encoder *

Yujun Cai[1], Yiwei Wang[2], Yiheng Zhu[6], Tat-Jen Cham[1], Jianfei Cai[3], Junsong Yuan[5], Jun Liu[7], Chuanxia Zheng[1], Sijie Yan[8], Henghui Ding[1], Xiaohui Shen[6], Ding Liu[6], Nadia Magnenat Thalmann[4]
[1]Nanyang Technological University, Singapore.
{yujun001,ding0093,chuanxia001}@e.ntu.edu.sg, astjcham@ntu.edu.sg
[2]National University of Singapore wangyw_seu@foxmail.com
[3]Monash University, Australia jianfei.cai@monash.edu, [4] University of Geneva Thalmann@miralab.ch
[5]State University of New York at Buffalo University, Buffalo, NY, USA jsyuang@buffalo.edu
[6]ByteDance Research {yiheng.zhu,shenxiaohui,liuding}@bytedance.com
[7]SUTD, Singapore jun_liu@sutd.edu.sg, [8]The Chinese University of Hong Kong yysijie@gmail.com

Abstract

We present a unified and flexible framework to address the generalized problem of 3D motion synthesis that covers the tasks of motion prediction, completion, interpolation, and spatial-temporal recovery. Since these tasks have different input constraints and various fidelity and diversity requirements, most existing approaches only cater to a specific task or use different architectures to address various tasks. Here we propose a unified framework based on Conditional Variational Auto-Encoder (CVAE), where we treat any arbitrary input as a masked motion series. Notably, by considering this problem as a conditional generation process, we estimate a parametric distribution of the missing regions based on the input conditions, from which to sample and synthesize the full motion series. To further allow the flexibility of manipulating the motion style of the generated series, we design an Action-Adaptive Modulation (AAM) to propagate the given semantic guidance through the whole sequence. We also introduce a cross-attention mechanism to exploit distant relations among decoder and encoder features for better realism and global consistency. We conducted extensive experiments on Human 3.6M and CMU-Mocap. The results show that our method produces coherent and realistic results for various motion synthesis tasks, with the synthesized motions distinctly adapted by the given action labels.

*This research is supported by Institute for Media Innovation, Nanyang Technological University (IMI-NTU) and the National Research Foundation, Singapore under its International Research Centres in Singapore Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. This research is also supported in part by Monash FIT Start-up Grant and SenseTime Gift Fund, National Science Foundation Grant CNS1951952 and SUTD project PIE-SGP-AI-2020-02.

1. Introduction

Generating realistic and plausible human body animation with specified actions has been a widely explored but challenging task in computer vision and graphics [29, 4]. To synthesize smooth and natural motions, traditional methods [30, 32] rely on the availability of complex pose specifications, which are time-consuming and expensive to obtain.

Recent deep learning approaches [5, 60, 56, 55, 17, 20, 61, 58] have investigated generating plausible human motions. However, since different motion synthesis tasks have different goals and expectations (as seen in Figure 1), many approaches are either restricted to one type of motion synthesis task or use different methods to address the various tasks. For example, much work [6, 14, 37, 63] is focused on the motion prediction task, typically adopting recurrent neural network (RNN) architectures to predict future frames sequentially, with new ones dependent only on previously generated frames. Although performing well in motion prediction, these approaches are not directly suited for generalizing to other motion synthesis tasks such as motion completion, interpolation, and spatial-temporal recovery, as shown in Figure 1, for which both forward and backward dependencies should be exploited. Moreover, many methods [5, 60, 56] are focused on minimizing the reconstruction error between the ground truth and generated motion sequences, while less considering motion diversity and human-likeness, which are also significant for realistic generation. Furthermore, in precise motion animation, it is

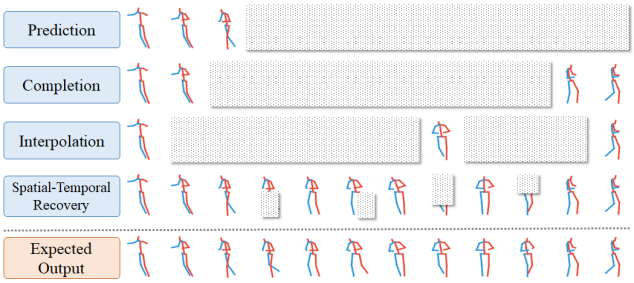

Figure 1. Our unified framework handles different 3D motion synthesis tasks, generalizing to several existing problems such as prediction, completion, interpolation, and spatial-temporal recovery. Given an arbitrarily masked pose series, our method synthesizes a full consecutive sequence without ignoring prior information from the pre-defined keyframes.

highly desirable for a user to be able to influence the type of action in the generated series, while few existing solutions are able to precisely manipulate the semantic information of motion series and generalize well to different synthesis tasks.

The above observations impel us to find a unified architecture for various 3D motion synthesis tasks, where we can generate realistic and meaningful results under different circumstances. In support of our ambition to flexibly incorporate semantic guidance for precise control of generated motions, we further explore the ability to influence the action types of the generated series in this work. To deal with various input conditions within a single model, we uniformly treat any arbitrary input as a masked motion series. The visible parts are considered as the input condition or constraint, while the masked regions are the places targeted for automatic generation. To accommodate diverse possible solutions that are consistent with the given observed frames, we propose a framework incorporating Conditional Variational Auto-Encoder (CVAE) for estimating a latent distribution of the missing regions, from which we can sample and synthesize multiple plausible results. To encourage congruency between estimated and ground truth distributions, we introduce two parallel but linked branches during training. As seen in Figure 2, the bottom branch uses ground truth to obtain the prior distribution of the missing regions and rebuild the original motion series. The upper branch, acting as the inference branch, takes the visible constraints to estimate the conditional latent distribution and samples diverse results from this distribution.

In addition to the CVAE-based framework for generalized motion synthesis, we also allow more precise manipulation of motion styles. In particular, we propose an Action-Adaptive Modulation (AAM) to subsume the given semantics into the generation process. To further enhance the realism and global consistency of the generated series, we introduce a cross-attention mechanism between encoder and decoder features to capitalize on the relationships between input and output poses, irrespective of temporal distances.

In summary, the main contributions of this work are:

- We propose a unified CVAE-based framework to handle various motion synthesis tasks such as motion prediction, completion, interpolation, and spatial-temporal recovery while meeting different input constraints, different fidelity and diversity requirements.
- We introduce an Action-Adaptive Modulation (AAM) that is able to control semantic motion styles of the generated series.
- We design a cross-attention mechanism that exploits long-term context information to enhance the realism and global consistency of synthesized sequences.

We conducted quantitative and qualitative assessments on the widely-used Human3.6M and CMU-Mocap datasets. Experiments show our approach outperformed existing pose synthesis methods, generating realistic and plausible motion series conditioned on various input constraints.

2. Related Work

Motion Synthesis is a general term that includes several tasks, such as motion prediction, completion and interpolation. The gamut of work in human motion synthesis includes using statistical [12, 39], learning-based [55, 17, 20, 61, 16] and physics-based [2, 46, 42] methods in both computer vision and graphics. Here we mainly focus on the most relevant learning-based approaches.

Motion Prediction A typical problem is motion prediction [14, 11, 6, 34, 3, 15], which typically refers to the task of predicting future human motion given a short initial motion segment. Due to the inherent temporal nature of this task, many existing methods [14, 37, 47, 5, 63] resort to recurrent neural networks (RNN) for temporal modeling. While advancing the boundaries of motion prediction, most of these methods generate frames step-by-step, with new ones depending on previously generated frames. This creates a barrier to leveraging backward dependencies, making it hard to apply to more generalized motion completion and interpolation tasks. Moreover, many approaches [35, 23, 31, 34, 3, 15] deterministically predict pose sequences by minimizing the MSE loss between ground truth and the generated pose series, which encourages the prediction of one single optimal result, while there may exist multiple plausible solutions satisfying the given constraints.

To produce robust and dynamic motion synthesis results, generative models [5, 60, 56, 55, 17, 20, 61, 1] have been introduced. Barsoum *et al*. [5] combined the Seq2seq framework with a GAN for motion prediction, which is able to

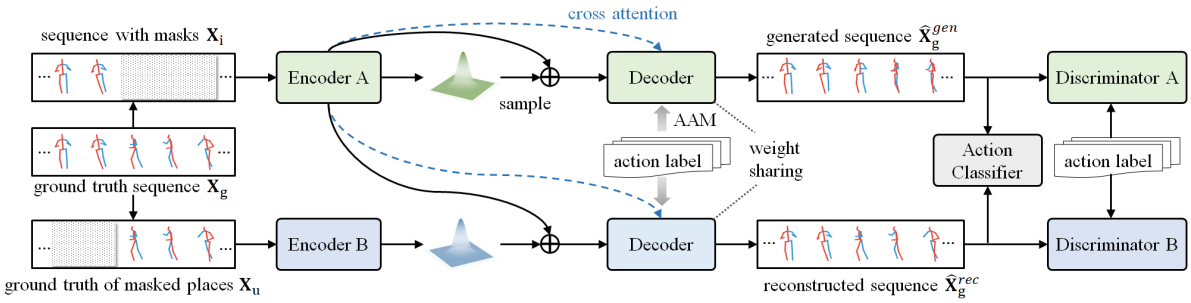

Figure 2. Schematic overview of our proposed network architecture. During training, two complementary parts, namely the visible partial regions $\mathbf{X}_i$ and the ground truth of the unknown missing $\mathbf{X}_u$, are simultaneously fed into the respective encoders to estimate the distribution of the missing regions over a latent space. The distributions are then used to sample the unseen embeddings to generate (upper-branch) and reconstruct (bottom-branch) the whole sequence. To further precisely manipulate the semantic guidance for the synthesized poses, we introduce the Action-Adaptive Modulation (AAM) in the decoding stage, followed by an action classifier to propagate the semantic information across the whole sequence. Moreover, a cross-attention mechanism is employed to enhance the consistency between the input and output features. During testing, the input keyframes only go through the upper-branch of the network without using the discriminator and classifier. The synthesized poses are combined with the input keyframes as the output of the network. Network A and B (*e.g.*, Encoder A and Encoder B) imply similar architectures without sharing weights.

generate multiple results by using different latent vectors drawn from a random distribution.

Motion Interpoloation and Completion For filling gaps of motion with specific key-frame constraints, many existing work utilize convolutional models [56, 25, 20, 64], recurrent models and [18, 7], accompanied with adverserial networks to provide consistent and plausible results. For instance, Harvey *et al*. [18] presented a transition generation technique that can serve as a new tool for 3D animators, based on adversarial recurrent neural networks. Henter *et al*. [19] proposed an autoregressive architecture for generating motion-data sequences based on normalizing flows. Cai *et al*. [7] proposed a two-stage GAN for skeleton motion synthesis, where the first stage initializes the best latent space for the input constraints, while the second stage generates temporal signals represented as latent vector sequences. To allow for bidirectional transforms between the latent and the skeleton spaces, Yan *et al*. [56] proposed a Convolutional Sequence Generation Network (CSGN), which transforms the skeleton sequence from a series of latent vectors sampled from a Gaussian process (GP) and utilizes Grpah Neural Networks [28, 52, 51, 53, 49, 54, 50] for pose synthesis. Though achieving local coherence in temporal space, we noticed that random sampling from a Gaussian process does not fully exploit the global context information, and also would not apply to the spatial-temporal recovery task. To address these issues, we introduce a novel CVAE-based framework which is capable of synthesizing meaningful pose series based on the global context, and more significantly can generalize to different tasks, such as motion prediction, completion, and spatial-temporal recovery.

3. Proposed Method

Given an arbitrarily masked pose sequence $\mathbf{X}_i \in \mathbb{R}^{T\times K}$, where T denotes the length of a sequence and K is the number of parameters describing each pose, our goal is to recover the missing regions $\hat{\mathbf{X}}_u$ and generate the full consecutive series $\hat{\mathbf{X}}_g = \{\mathbf{X}_i, \hat{\mathbf{X}}_u\}$. Here $\mathbf{X}_i = \mathbf{M} \odot \mathbf{X}_g$, and the ground truth complementary regions $\mathbf{X}_u = (1 - \mathbf{M}) \odot \mathbf{X}_g$. $\mathbf{M}$ refers to the binary mask applied to the sequence. As stated earlier, this problem is a generalization of the several existing motion synthesis tasks shown in Figure 1.

Unlike most existing motion prediction methods [14, 37, 47, 5] that resort to a temporal modeling architecture such as RNN, we propose a CVAE-based framework with two parallel branches during training, which is shown in Figure 2. Below, we discuss the individual components in detail.

Given the observed regions $\mathbf{X}_i$ in a pose series, we attempt to synthesize a plausible and realistic pose sequence $\hat{\mathbf{X}}_g = \{\mathbf{X}_i, \hat{\mathbf{X}}_u\}$ without losing the prior knowledge of the input key frames. To do so, we utilized a CVAE[44]-based framework, which estimates a parametric distribution of the unseen regions over a latent space, from which we can sample the latent vector $\mathbf{z}_u$ to generate the missing parts $\mathbf{X}_u$. Formally, this involves a variational lower bound of the conditional log-likelihood of the observation:

$$\begin{aligned}\log p(\mathbf{X}_u|\mathbf{X}_i) \geq &- \mathbf{KL}(q_\psi(\mathbf{z}_u|\mathbf{X}_u)||p_\phi(\mathbf{z}_u|\mathbf{X}_i)) \\ &+ \mathop{\mathbb{E}}_{q_\psi(\mathbf{z}_u|\mathbf{X}_u)}[\log(p_\theta(\mathbf{X}_u|\mathbf{z}_u, \mathbf{X}_i))],\end{aligned} \tag{1}$$

where $\mathbf{KL}$ is the Kullback-Leibler divergence, $\mathbf{z}_u$ is the sampled latent vector, $q_\psi(\mathbf{z}_u|\mathbf{X}_u)$ is the posterior sampling function, $p_\phi(\mathbf{z}_u|\mathbf{X}_i)$ is the conditional prior, and $p_\theta(\mathbf{X}_u|\mathbf{z}_u, \mathbf{X}_i)$ is the likelihood. Note that the distributions q_ψ, p_ϕ and p_θ can be parameterized by deep neural

networks, and we refer readers to the supplementary file for the detailed derivations.

One issue with Eq. (1) is that the second term only takes the sampling from $q_\psi(\mathbf{z}_u|\mathbf{X}_u)$ during training, which is encoded by the ground truth of the unseen regions $\mathbf{X}_u$. This may not be optimal for the synthesis during testing, where we only have visible regions $\mathbf{X}_i$. To mitigate the gap between training and testing, inspired by [44], we modify Eq. (1) by sampling from both $q_\psi(\mathbf{z}_u|\mathbf{X_u})$ and $p_\phi(\mathbf{z}_u|\mathbf{X}_i)$:

$$\begin{aligned}\log(p(\mathbf{X}_u|\mathbf{X}_i)) \geq &- \mathbf{KL}(q_\psi(\mathbf{z}_u|\mathbf{X}_u)||p_\phi(\mathbf{z}_u|\mathbf{X}_i))+ \\ &\mu \mathop{\mathbb{E}}_{q_\psi(\mathbf{z}_u|\mathbf{X}_u)}[\log(p_\theta(\mathbf{X}_u|\mathbf{z}_u,\mathbf{X}_i))]+ \\ &(1-\mu) \mathop{\mathbb{E}}_{p_\phi(\mathbf{z}_u|\mathbf{X}_i)}[\log(p_\theta(\mathbf{X}_u|\mathbf{z}_u,\mathbf{X}_i))].\end{aligned} \tag{2}$$

where $0 \leq \mu \leq 1$ is a tradeoff parameter.

Network Architecture. Figure 2 gives an overview of our CVAE-based statistical framework with two parallel paths, each of which consists of an encoder and a decoder and the decoder networks share identical weights. In particular, for the upper path (also the test path), the partially visible pose sequence $\mathbf{X}_i$ is used to infer the latent distribution $p_\phi(\mathbf{z}_u|\mathbf{X}_i)$ of the unseen sites, from which we can sample the latent vector $\mathbf{z}_u$ and generate the plausible full pose series. For the lower path, the latent distribution $q_\psi(\mathbf{z}_u|\mathbf{X}_u)$ is encoded from the ground truth of the unseen regions $\mathbf{X}_u$. By combining the features extracted from the visible regions and the sampled latent vector containing information of the missing parts, the goal of this path is reconstructing the original pose sequence $\mathbf{X}_g = \{\mathbf{X}_i, \mathbf{X}_u\}$. Moreover, to ensure that the synthesized data fit in the training set distribution, both paths are deployed with an adversarial learning network to facilitate high-quality generation.

3.1. CVAE-based Statistical Framework

3.2. Action-Adaptive Modulation

When synthesizing human body animations with the given constraints, a typically desired capability would be: *can we manipulate the motion styles of the generated series with certain semantic guidance such as action labels?* To tackle this issue, inspired by [24, 40, 65] that adjust the activation in normalization layers for image style transformation, we propose to deploy an Action-Adaptive Modulation (AAM) during the decoding stage (as shown in Figure 4), followed by an action classifier to enhance the distinctive action-related features. The detailed design of the AAM can be found in Figure 3, where the modulation parameters γ and β of the normalization layers are learned from the given action labels. To further utilize sequential order, we additionally insert a positional embedding into the action labels, assigning each frame with a unique encoded value.

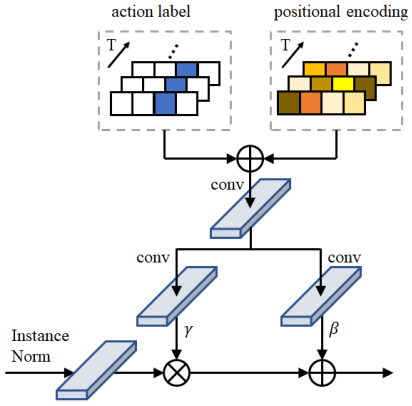

Figure 3. Action-Adaptive Modulation (AAM) is applied to the normalization layers in the decoding stage. Particularly, we first encode the pre-defined action label by adding a positional embedding to each frame. The element-wise summed results are then convolved to produce the per-frame modulation parameters γ and β as the learned scale and bias of the normalization.

Specifically, let $\mathbf{h}^i$ be the activations of the i-th layer of a deep convolutional network for a batch of N samples. $\mathbf{L} \in \mathbb{D}^T$ is the semantic label of each sample, where $\mathbb{D}$ is a set of integers denoting the one-hot action labels. T^i is the length of the sequence in the layer, and C^i is the number of channels in the activation map. The modulated action value at site ($n \in N, c \in C^i, t \in T^i$) is given by:

$$\gamma^i_{c,t}\left(\mathsf{PE}\left(\mathbf{L}\right)\right)\frac{h^i_{n,c,t}-\mu^i_{n,c}}{\sigma^i_{n,c}} + \beta^i_{c,t}(\mathsf{PE}(\mathbf{L})), \tag{3}$$

where $h^i_{n,c,t}$ is the activation before normalization, PE is the positional embedding function following [45], $\gamma^i_{c,t}(\mathsf{PE}(\mathbf{L}))$ and $\beta^i_{c,t}(\mathsf{PE}(\mathbf{L}))$ are learned modulation parameters of the normalization layer, and $\mu_{n,c}$ and $\sigma_{n,c}$ are mean and standard deviation of the activation in channel c of sample n.

3.3. Cross-Attention Mechanism

One potential limitation for the current framework is that it mainly relies on the convolutional architecture to propagate local signals progressively through the data. This approach, however, does not fully exploit the long-range dependencies of the distant features, as well as the consistency between the given visible regions and the synthesized poses. To address this issue, inspired by the Transformer network [45] that employs global attention to model long-term dependencies, we apply a cross-attention mechanism between the encoder and decoder features across scales. Specifically, we first perform a multi-head self-attention to capture the long-term dependencies between frames. Then we introduce cross attention between the encoder and decoder features, to enhance the correlations between the input constraints and the output series.

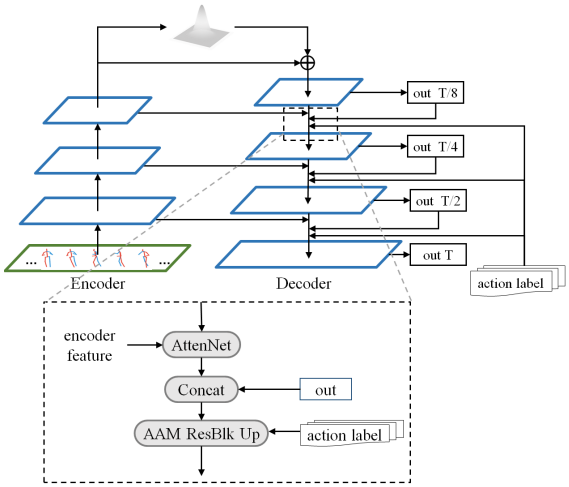

Figure 4. Illustration of the encoder and decoder architectures, which facilitate the effective process and consolidation of the features across scales. For clarity, we also depict a building block (bottom dashed box) showing the detailed architecture of each decoding stage.

3.4. Training

The following losses are used in training.

Distributive Regularization. In Eq. (2), the KL divergence term can be interpreted as regularizing the learned conditional prior $p_\phi(\mathbf{z}_u|\mathbf{X}_i)$ to the posterior distribution $q_\psi(\mathbf{z}_u|\mathbf{X}_u)$. In particular, to simplify the computation, we define both $q_\psi(\mathbf{z}_u|\mathbf{X}_u)$ and $p_\phi(\mathbf{z}_u|\mathbf{X}_i)$ as Gaussian distributions, and the loss is given by

$$\mathcal{L}^g_{KL} = -\mathbf{KL}(q_\psi(\mathbf{z}_u|\mathbf{X}_u)||p_\phi(\mathbf{z}_u|\mathbf{X}_i)) \tag{4}$$

To enhance smooth training, we further use a traditional VAE [27] to model the missing regions as arising from a smooth Gaussian prior $p(\mathbf{z}_u)$:

$$\mathcal{L}^p_{KL} = -\mathbf{KL}(q_\psi(\mathbf{z}_u|\mathbf{X}_u)||p(\mathbf{z}_u)) \tag{5}$$

where $p(\mathbf{z}_u) = \mathcal{N}(0, \sigma^2(m)I)$, to allow greater latent prior variance when the number of missing regions m is larger.

Matching Loss. Analogous to the likelihood term in Eq. (2), we encourage accurate sequence reconstruction for the reconstructive path (the bottom path in Figure 2) with an L1 loss, given by

$$\mathcal{L}^r_m = ||\hat{\mathbf{X}}^{rec}_g - \mathbf{X}_g||_1 \tag{6}$$

where $\hat{\mathbf{X}}^{rec}_g$ is the reconstructed pose series and $\mathbf{X}_g$ is the ground truth of the whole sequence. Conversely, for the generative path (the upper path in Figure 2) we want to accommodate diversity and thus only enforce the appearance matching of the visible regions between the generated sequence $\hat{\mathbf{X}}^{gen}_g$ and the corresponding ground truth $\mathbf{X}_g$, using

$$\mathcal{L}^g_m = ||\mathbf{M} \odot (\hat{\mathbf{X}}^{gen}_g - \mathbf{X}_g)||_1 \tag{7}$$

where $\mathbf{M}$ denotes the visible mask of the input sequence.

Adversarial Loss. To facilitate high-quality generation, we further incorporate two discriminators D_A, D_B to respectively judge whether the generated and reconstructed series fit in with the dataset distribution. Inspired by [48], we apply a mean feature matching loss to enhance the reconstruction accuracy:

$$\mathcal{L}^r_{ad} = ||D_B(\hat{\mathbf{X}}^{rec}_g, \mathbf{L}) - D_B(\mathbf{X}_g, \mathbf{L})||_2 \tag{8}$$

where $D_B(.)$ is the output feature maps of the discriminator D_B and $\mathbf{L}$ is the corresponding action label. For the adversarial loss in the generative path, we adopt LSGAN [36, 62] to improve realism:

$$\mathcal{L}^g_{ad} = [D_A(\hat{\mathbf{X}}^{gen}_g, \mathbf{L}) - 1]^2. \tag{9}$$

Classification Loss. A classification loss $\mathcal{L}_c$ is introduced to facilitate the action-based motion synthesis. In particular, the classification network is pre-trained by the ground truth pose series. To train our model, we apply L1 norm to regularize the classification scores of the generated and reconstructed results:

$$\mathcal{L}_c = ||CL(\hat{\mathbf{X}}^{gen}_g) - CL(\mathbf{X}_g)||_1 + ||CL(\hat{\mathbf{X}}^{rec}_g) - CL(\mathbf{X}_g)||_1 \tag{10}$$

where $CL(.)$ are scores from the classification network.

Overall Loss. Combining all the above losses, we obtain the overall loss function $\mathcal{L}$:

$$\begin{aligned}\mathcal{L} =& \lambda_{KL}(\mathcal{L}^g_{KL} + \mathcal{L}^p_{KL}) + \lambda_{ad}(\mathcal{L}^r_{ad} + \mathcal{L}^g_{ad}) \\ &+ \lambda_m(\mathcal{L}^r_m + \mathcal{L}^g_m) + \lambda_c\mathcal{L}_c\end{aligned} \tag{11}$$

where λ_* are the tradeoff parameters. In our implementation, we set $\lambda_{KL} = 20$, $\lambda_{ad} = 1$, $\lambda_m = 20$, and $\lambda_c = 1$.

4. Experiments

4.1. Implementation Details

We implemented our method with the PyTorch framework. The 3D poses are represented in root-relative 3D joint locations without removing the global orientation during training and inference. During training, we set a learning rate of 10^{-4}, with a mini-batch size of 128 samples using the Adam optimizer [26]. The length of the pose series T is set as 128 for Human3.6M and 64 for CMU Mocap. Four heads are used in cross-attention mechanism. Note that our model can also be applied to data without action labels by removing the AAM and classifier modules, and not using action labels for the discriminator. Please refer to the supplementary file for more details of our implementation.

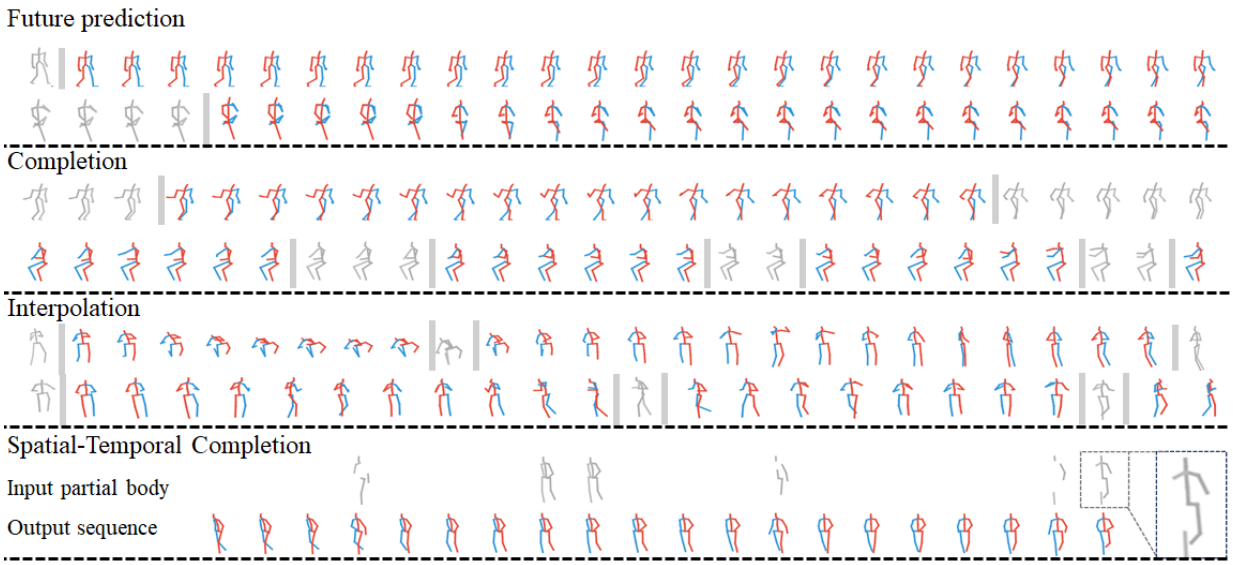

Figure 5. Qualitative examples of our proposed model for different motion-related tasks on Human3.6M and CMU-Mocap datasets. Gray poses are the pre-defined input frames (or partial body of some frames), while the red & blue skeleton sequences are the synthesized poses. Note that the input constraints can be flexibly set to arbitrary positions with varying densities.

4.2. Datasets

We evaluated our method on two publicly available datasets: Human3.6M [22] and CMU-Mocap[1].

Human3.6M. This dataset [22] is a large-scale and commonly used dataset for human motion synthesis and 3D pose estimation, which consists of 7 subjects performing a variety of actions with ground truth 3D pose annotations. In our experiments, in order to reduce redundant frames and encourage large motion variations, we subsample the video frames to 10 frames per second. The action classes we select are "Direction", "Sitting", "Sitting Down", "Walking", "Taking Photos", "Smoking" and "Eating". Following the standard setting in [10, 33, 38, 9, 13, 8], we take 5 subjects (S1, S5, S6, S7, S8) for training and 2 subjects (S9 and S11) for testing. The global translations and constant joints are excluded from our experiments.

CMU-Mocap. To show the generalization ability of our proposed method, we also evaluated our performance on the CMU mocap dataset (CMU-Mocap). In our experiments, we selected 6 action classes for evaluation, including "Basketball", "Jumping", "Washwindow", etc. We down-sample each sequence to 25 frames per second. The data processing is the same as that for Human3.6M.

4.3. Evaluation Metrics

Evaluating the quality of synthesized pose sequences is a difficult problem due to the diverse reasonable possibilities for each masked pose sequence. Previous methods either evaluate the distance between the ground truth and the generated results using protocols such as MPJPE (minimum mean per joint position error) [6, 14, 37, 63], or take distribution-based metrics, *e.g.* FID (Frechet Inception Distance), IS (Inception Score), to measure the generation fidelity [56, 7, 61]. In our paper, noting that the generation diversity tends to be smaller with more pre-defined input constraints but larger when it comes to generating long-range missing parts, we adopt both MPJPE and the distribution-based metrics (FID, IS, and Diversity) to quantitatively evaluate the synthesized samples, conditioned on the generation uncertainty of the motion series.

MPJPE. For completing sequences with pre-defined past and future frames, we assume that one of our 50 synthesized samples will be close to the ground truth, and select the single sample with the minimum MPJPE for the masked places. We apply a center mask of 30 frames to each sequence for the completion assessment.

FID, IS and Diversity. For long-range generations with very few initial input frames, since the sequences tend to have diverse possible futures in totally different trends, it is not appropriate to directly compare the synthesized results against the one ground truth. Instead, we mainly focus on whether the results are adequately realistic compared with real data, as well as diverse enough to generate multiple plausible solutions. In particular, following the evaluation metrics of [56] which are extended from the image/video generation problem [62, 59, 41, 7], we adopt the Frechet Inception Distance (**FID**) [21] that measures the statistical distance between the real and synthesized data in the feature

[1] Available at http://mocap.cs.cmu.edu/

Figure 6. Diverse actions generated from the same initial pose. By setting different semantic guidance, our model generates a variety of semantically meaningful results given the same input frame.

Table 1. Classification Accuracy (%) of different action classes on Human 3.6M using a classifier independently trained with ground truth 3D pose series. Our method with both action guidance and Action-Adaptive Modulation (AAM) considerably improves the classification accuracy by a large margin compared with the other two baselines.

Data	Sitting	Phoning	Walking	Takingphoto	Direction	SittingDown	Smoking	Eating	Average
Ours w/o action	0.28	0.45	0.82	0.10	0.30	0.87	0.32	0.47	0.45
Ours w/ action w/o AAM	0.41	0.62	0.91	0.51	0.63	0.90	0.57	0.66	0.65
Ours w/ action w/ AAM	**0.85**	**0.93**	**0.99**	**0.89**	**0.81**	**0.99**	**0.97**	**0.81**	**0.91**

Table 2. Quantitative results on Human3.6M with FID / IS / Diversity metrics to evaluate the long-term generation and MPJPE to assess the completion accuracy. The best results are marked in bold.

Method	$FID_g(\downarrow)$	$FID_a(\downarrow)$	IS(↑)	Diversity(↑)	MPJPE(↓)
HP-GAN [5]	272.3	365.2	3.48	0.05	–
Two-stage [7]	327.8	453.6	1.33	0.11	154.9
CSGN [56]	136.8	264.5	4.35	0.17	185.2
ours w/o action	88.2	192.3	6.12	**0.26**	115.6
ours w/ action	**62.5**	**111.3**	**6.87**	0.19	**97.3**

space, and the Inception Score (**IS**) [43] that analyzes class probabilities for each generated sequence over all classes. Additionally, we introduce a **Diversity** score that estimates the feature-based standard deviation of the multiple generated outputs with the same input condition. All generated motions are evaluated without removing the orientations. For fair and complete comparison, FID is further divided into two metrics: $\mathbf{FID}_g$ that generally assesses the distribution distances between the real and synthesized sequences over the whole dataset, and $\mathbf{FID}_a$ that calculates the average class-based statistical discrepancies. To get the classification scores and intermediate features, we follow [56] in training a ST-GCN-based classifier [57] with the ground truth training data. The intermediate feature used for FID and Diversity scoring is the feature map extracted from the last layer. To precisely analyze the motion patterns within a long-duration skeleton sequence, we cut the synthesized sequence into several short snippets, each 30 frames long, for the purpose of quantitative evaluation.

Action Classification. Apart from the above metrics to evaluate the generation quality, we also want to examine whether the synthesized samples are actually residing in the same motion styles with the given action labels. To this end, we employed the pre-trained ST-GCN [57] to compute the classification accuracy for each action class. Note that the higher the recognition accuracy, the better the generated samples follow the real class-specific motion patterns.

Table 3. Quantitative results on CMU-Mocap dataset. The best results are marked in bold.

Method	$FID_g(\downarrow)$	$FID_a(\downarrow)$	IS(↑)	Diversity(↑)	MPJPE(↓)
HP-GAN [5]	188.5	214.2	2.36	0.01	–
Two-stage [7]	386.4	426.9	1.02	0.02	162.2
CSGN [56]	146.7	223.5	3.75	0.04	188.3
ours w/o action	86.2	162.3	4.32	**0.08**	126.7
ours w/ action	**84.6**	**97.7**	**4.67**	0.05	**108.6**

Table 4. Impact of the cross attention mechanism on Human3.6M. For fair comparison, action labels are leveraged for both methods.

Method	$FID_g(\downarrow)$	$FID_a(\downarrow)$	IS(↑)	Diversity(↑)	MPJPE(↓)
ours w/o cross attention	71.3	120.1	6.43	0.18	104.4
ours w/ cross attention	**62.5**	**111.3**	**6.87**	0.18	**97.3**

Table 5. Perceptual user study to evaluate generation quality and motion style manipulation. A higher score indicates more realistic/style consistent results.

	HP-GAN [5]	Two-stage [7]	CSGN [56]	ours w/o action	ours w/ action
generation quality	3.13	1.98	2.86	3.91	**3.92**
style consistency	2.31	1.95	2.06	2.34	**4.12**

4.4. Results of Pose Series Generation

Quantitative Comparisons. We compared our method with the state-of-the-art approaches [5, 56, 7] that are able to provide multiple synthesized results on both Human3.6M and CMU-Mocap. Since our model can optionally be implemented with semantic guidance, we carried out experiments with two variants of our method: a) **ours w/o actions** that removes the AAM and classifier modules, and does not use action labels for the discriminator; b) **ours w/ action** that leverages the action labels for semantic manipulation. The methods compared include the RNN-based HP-GAN [5], the convolutional graph-based CSGN [56] that trans-

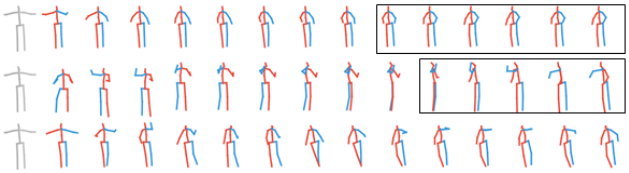

Figure 7. Visualization example of comparisons on Human 3.6M dataset. From top to bottom, we show the result of HP-GAN [5], CSGN [56] and our method. We see that the RNN-based HP-GAN gradually freezes to a static pose and CSGN may generate unnatural poses due to random sampling from Gaussian Process.

forms random vectors from Gaussian processes to pose series, and the two-stage generation model [7] that utilizes a single-to-sequence strategy. The quantitative results, measured by FID_a / FID_g / IS / Diversity / MPJPE for both the generation and completion tasks, are summarized in Tables 2 and 3. Since HP-GAN does not support the motion completion task, we only report the completion performance (MPJPE) of the Two-stage and CSGN models.

As seen in the tables, for both the completion and the long-range sequence generation, our method (**ours w/o action**) significantly surpassed the previous leading approaches on both Human3.6M and CMU-Mocap. Moreover, adding the action labels (**ours w/ action**) further improves the generation realism (FID/ IS) and the completion accuracy (MPJPE), since the generated sequences are more likely to follow the motion styles of the given action labels, leading to better consistency with the real data distributions. In addition, we noticed that compared to **ours w/ action**, our non-action method (**ours w/o action**) achieved higher diversity scores. This is expected, as generation diversity will be reduced by constraining to a motion style. For instance, a standing initial pose may lead to multiple possible future motions, such as walking, sitting, and running. If we specify the future action as walking, the possibilities will be limited accordingly.

Human Evaluation We conducted perceptual user studies to evaluate the generation quality and motion style manipulation, where 100 participants were shown the generated motion sequences from different methods, and asked to score these 1-5 for both generation quality and motion style consistency, where the higher the better. Five models (HP-GAN [5], Two-stage [7], CSGN [56], ours w/o action, ours w/actions) tested on Human3.6M dataset were used for evaluation. For fair comparison, each sequence was generated from the same input constraints with 10 initial frames. We then sampled 100 motion clips for each of the action classes from each model, and asked each participant to evaluate 20 sequences randomly selected from the question pool. Table 5 shows our method (ours w/o action, ours w/ action) led to better generation quality, while (ours w/ action) is effective in manipulating motion styles.

Qualitative Results. For qualitative analysis, we first provided some visual examples of our non-action method for various motion-related tasks on both Human3.6M and CMU-Mocap datasets. As can be seen in Figure 5, our model is able to produce realistic and plausible pose series, matching well with the miscellaneous input constraints.

Figure 6 gives a few visual results of our action-guided method with the same initial pose but different action labels. We can see that, by setting different semantic guidance, our method produces plausible future dynamics in totally different but semantically meaningful trends.

We also provide a visualization example to show how our result provides a higher-quality result compared to other approaches. As seen in Figure 7, the RNN-based HP-GAN led to a frozen state during long-term prediction, while CSGN occasionally generated unnatural poses due to their random sampling method. Conversely, our method produced plausible motions without freezing during long-range prediction.

4.5. Ablation Study

Impact of Action-Adaptive Modulation. To evaluate our proposed Action-Adaptive Modulation (AAM) for motion style manipulation, we conducted action classification experiments with the following baselines: a) **ours w/o action**; b) **ours w/ action w/o AAM**: directly concatenating the action label and latent features as input to the decoder without leveraging AAM; c) **ours w/ action w/ AAM:** our proposed method that utilizes AAM for semantic manipulation. The results are presented in Table 1. Compared to baseline-a, leveraging the action labels (baseline-b) unsurprisingly improved the classification accuracy from 45% to 65%. Moreover, using our proposed AAM further improved the accuracy to 91%, demonstrating the effectiveness of our proposed AAM for semantic manipulation.

Impact of the cross-attention mechanism. We also examined the impact of the cross-attention mechanism. As shown in Table 4, our method with the cross attention mechanism consistently outperformed the one without in all the evaluation metrics, clearly demonstrating its effectiveness.

5. Conclusion

We have proposed a unified CVAE-based model to handle various 3D motion synthesis tasks. Unlike existing methods, our framework enables automatic motion synthesis with flexible input constraints. To further manipulate the motion style of the generated series, we designed an Action-Adaptive Modulation (AAM) to propagate the semantic guidance through the whole sequence. We also introduced a cross-attention mechanism to improve realism and global consistency. Experimental results on two benchmark datasets demonstrated the superior performance of our proposed method.

References

[1] Kfir Aberman, Rundi Wu, Dani Lischinski, Baoquan Chen, and Daniel Cohen-Or. Learning character-agnostic motion for motion retargeting in 2d. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 2

[2] Ijaz Akhter and Michael J Black. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1446–1455, 2015. 2

[3] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. Structured prediction helps 3d human motion modelling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7144–7153, 2019. 2

[4] Brian F Allen and Petros Faloutsos. Evolved controllers for simulated locomotion. In *International Workshop on Motion in Games*, pages 219–230. Springer, 2009. 1

[5] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1418–1427, 2018. 1, 2, 3, 7, 8

[6] Judith Butepage, Michael J Black, Danica Kragic, and Hedvig Kjellstrom. Deep representation learning for human motion prediction and classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6158–6166, 2017. 1, 2, 6

[7] Haoye Cai, Chunyan Bai, Yu-Wing Tai, and Chi-Keung Tang. Deep video generation, prediction and completion of human action sequences. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 366–382, 2018. 3, 6, 7, 8

[8] Yujun Cai, Liuhao Ge, Jianfei Cai, Nadia Magnenat-Thalmann, and Junsong Yuan. 3d hand pose estimation using synthetic data and weakly labeled rgb images. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 6

[9] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 666–682, 2018. 6

[10] Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2272–2281, 2019. 6

[11] Yujun Cai, Lin Huang, Yiwei Wang, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Xu Yang, Yiheng Zhu, Xiaohui Shen, et al. Learning progressive joint propagation for human motion prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[12] Jinxiang Chai and Jessica K Hodgins. Constraint-based motion optimization using a statistical dynamic model. In *ACM SIGGRAPH 2007 papers*, pages 8–es. 2007. 2

[13] Haoshu Fang, Yuanlu Xu, Wenguan Wang, Xiaobai Liu, and Song-Chun Zhu. Learning pose grammar to encode human body configuration for 3d pose estimation. *arXiv preprint arXiv:1710.06513*, 2017. 6

[14] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4346–4354, 2015. 1, 2, 3, 6

[15] Partha Ghosh, Jie Song, Emre Aksan, and Otmar Hilliges. Learning human motion models for long-term predictions. In *2017 International Conference on 3D Vision (3DV)*, pages 458–466. IEEE, 2017. 2

[16] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029, 2020. 2

[17] Ikhsanul Habibie, Daniel Holden, Jonathan Schwarz, Joe Yearsley, and Taku Komura. A recurrent variational autoencoder for human motion synthesis. In *28th British Machine Vision Conference*, 2017. 1, 2

[18] Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. *ACM Transactions on Graphics (TOG)*, 39(4):60–1, 2020. 3

[19] Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. Moglow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020. 3

[20] Alejandro Hernandez, Jurgen Gall, and Francesc Moreno-Noguer. Human motion prediction via spatio-temporal inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7134–7143, 2019. 1, 2, 3

[21] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pages 6626–6637, 2017. 6

[22] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 6

[23] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 5308–5317, 2016. 2

[24] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 4

[25] Manuel Kaufmann, Emre Aksan, Jie Song, Fabrizio Pece, Remo Ziegler, and Otmar Hilliges. Convolutional autoencoders for human motion infilling. In *2020 International Conference on 3D Vision (3DV)*, pages 918–927. IEEE, 2020. 3

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[27] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 5

[28] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 3

[29] Lucas Kovar, Michael Gleicher, and Frédéric Pighin. Motion graphs. In *ACM SIGGRAPH 2008 classes*, pages 1–10. 2008. 1

[30] Jehee Lee, Jinxiang Chai, Paul SA Reitsma, Jessica K Hodgins, and Nancy S Pollard. Interactive control of avatars animated with human motion data. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pages 491–500, 2002. 1

[31] Kyoungoh Lee, Inwoong Lee, and Sanghoon Lee. Propagating lstm: 3d pose estimation based on joint interdependency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 119–135, 2018. 2

[32] Maxim Likhachev, Geoffrey J Gordon, and Sebastian Thrun. Ara*: Anytime a* with provable bounds on sub-optimality. *Advances in neural information processing systems*, 16:767–774, 2003. 1

[33] Jun Liu, Henghui Ding, Amir Shahroudy, Ling-Yu Duan, Xudong Jiang, Gang Wang, and Alex C Kot. Feature boosting network for 3d pose estimation. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):494–501, 2019. 6

[34] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *European Conference on Computer Vision*, pages 474–489. Springer, 2020. 2

[35] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9489–9497, 2019. 2

[36] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017. 5

[37] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2891–2900, 2017. 1, 2, 3, 6

[38] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2640–2649, 2017. 6

[39] Jianyuan Min, Yen-Lin Chen, and Jinxiang Chai. Interactive generation of human animation with deformable motion models. *ACM Transactions on Graphics (TOG)*, 29(1):1–12, 2009. 2

[40] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 4

[41] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 6

[42] Charles F Rose III, Peter-Pike J Sloan, and Michael F Cohen. Artist-directed inverse-kinematics using radial basis function interpolation. In *Computer Graphics Forum*, volume 20, pages 239–250. Wiley Online Library, 2001. 2

[43] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016. 7

[44] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems*, pages 3483–3491, 2015. 3, 4

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 4

[46] Ruben Villegas, Jimei Yang, Duygu Ceylan, and Honglak Lee. Neural kinematic networks for unsupervised motion retargetting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8639–8648, 2018. 2

[47] Borui Wang, Ehsan Adeli, Hsu-kuang Chiu, De-An Huang, and Juan Carlos Niebles. Imitation learning for human pose prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7124–7133, 2019. 2, 3

[48] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 5

[49] Yiwei Wang, Wei Wang, Yujun Ca, Bryan Hooi, and Beng Chin Ooi. Detecting implementation bugs in graph convolutional network based node classifiers. In *2020 IEEE 31st International Symposium on Software Reliability Engineering (ISSRE)*, pages 313–324. IEEE, 2020. 3

[50] Yiwei Wang, Wei Wang, Yuxuan Liang, Yujun Cai, and Bryan Hooi. Graphcrop: Subgraph cropping for graph classification. *arXiv preprint arXiv:2009.10564*, 2020. 3

[51] Yiwei Wang, Wei Wang, Yuxuan Liang, Yujun Cai, and Bryan Hooi. Curgraph: Curriculum learning for graph classification. In *Proceedings of the Web Conference 2021*, pages 1238–1248, 2021. 3

[52] Yiwei Wang, Wei Wang, Yuxuan Liang, Yujun Cai, and Bryan Hooi. Mixup for node and graph classification. In *Proceedings of the Web Conference 2021*, pages 3663–3674, 2021. 3

[53] Yiwei Wang, Wei Wang, Yuxuan Liang, Yujun Cai, and Bryan Hooi. Progressive supervision for node classification. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part I*, pages 266–281. Springer International Publishing, 2021. 3

[54] Yiwei Wang, Wei Wang, Yuxuan Liang, Yujun Cai, Juncheng Liu, and Bryan Hooi. Nodeaug: Semi-supervised node classification with data augmentation. In *Proceedings of the 26th*

ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 207–217, 2020. 3

[55] Zhenyi Wang, Ping Yu, Yang Zhao, Ruiyi Zhang, Yufan Zhou, Junsong Yuan, and Changyou Chen. Learning diverse stochastic human-action generators by learning smooth latent transitions. *arXiv preprint arXiv:1912.10150*, 2019. 1, 2

[56] Sijie Yan, Zhizhong Li, Yuanjun Xiong, Huahan Yan, and Dahua Lin. Convolutional sequence generation for skeleton-based action synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4394–4402, 2019. 1, 2, 3, 6, 7, 8

[57] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. *arXiv preprint arXiv:1801.07455*, 2018. 7

[58] Siyuan Yang, Jun Liu, Shijian Lu, Meng Hwa Er, and Alex C Kot. Skeleton cloud colorization for unsupervised 3d action representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 1

[59] Raymond A Yeh, Chen Chen, Teck Yian Lim, Alexander G Schwing, Mark Hasegawa-Johnson, and Minh N Do. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5485–5493, 2017. 6

[60] Ping Yu, Yang Zhao, Chunyuan Li, Junsong Yuan, and Changyou Chen. Structure-aware human-action generation. In *European Conference on Computer Vision*, pages 18–34. Springer, 2020. 1, 2

[61] Rui Zhao, Hui Su, and Qiang Ji. Bayesian adversarial human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6225–6234, 2020. 1, 2, 6

[62] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1438–1447, 2019. 5, 6

[63] Yi Zhou, Zimo Li, Shuangjiu Xiao, Chong He, Zeng Huang, and Hao Li. Auto-conditioned recurrent networks for extended complex human motion synthesis. In *International Conference on Learning Representations*, 2018. 1, 2, 6

[64] Yi Zhou, Jingwan Lu, Connelly Barnes, Jimei Yang, Sitao Xiang, et al. Generative tweening: Long-term inbetweening of 3d human motions. *arXiv preprint arXiv:2005.08891*, 2020. 3

[65] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5104–5113, 2020. 4