A Brand New Dance Partner: Music-Conditioned Pluralistic Dancing Controlled by Multiple Dance Genres

Jinwoo Kim[1], Heeseok Oh[2], Seongjean Kim[1], Hoseok Tong[1] and Sanghoon Lee[*1]

[1]School of Electrical and Electronic Engineering, Yonsei University
[2]Department of Applied AI, Hansung University

Abstract

When coming up with phrases of movement, choreographers all have their habits as they are used to their skilled dance genres. Therefore, they tend to return certain patterns of the dance genres that they are familiar with. What if artificial intelligence could be used to help choreographers blend dance genres by suggesting various dances, and one that matches their choreographic style? Numerous task-specific variants of autoregressive networks have been developed for dance generation. Yet, a serious limitation remains that all existing algorithms can return repeated patterns for a given initial pose sequence, which may be inferior. To mitigate this issue, we propose MNET, a novel and scalable approach that can perform music-conditioned pluralistic dance generation synthesized by multiple dance genres using only a single model. Here, we learn a dance-genre aware latent representation by training a conditional generative adversarial network leveraging Transformer architecture. We conduct extensive experiments on AIST++ along with user studies. Compared to the state-of-the-art methods, our method synthesizes plausible and diverse outputs according to multiple dance genres as well as generates outperforming dance sequences qualitatively and quantitatively.

1. Introduction

Dance has long been considered as a universal language that can share emotions more effectively than words. Nowadays, many people share their life-log via short-form video apps such as TikTok and Youtube Shorts [27, 54]. However, dancing is a highly creative and artistic process, hence professional training is often followed to express a feeling of elegant and rhythmic own story in a short-form video. For this reason, the music-conditioned dance generation, despite significant progress, is a challenging task that should capture high kinematic complexity rhythmically.

Recently, deep autoregressive networks have been used to synthesize dance motions learning long-range dependencies with the input music. Most state-of-the-art (SOTA) methods [21, 29, 37, 38] exploit RNN or Transformer architectures and generate dance for a given initial pose sequence with music. While previous studies produce temporally coherence sequence according to music, we find that all existing algorithms remain severely limited diversity by extending a given initial pose sequence into a repeated patterns. Although skilled dancers and choreographers often repeat dance patterns, they try to subtly diversify their dance lines. So, the lack of diversity issue is critical to the dance synthesizing.

In this paper, we tackle the problem of the music-conditioned pluralistic dance generation synthesized by multiple dance genres. The key challenge of pluralistic dance generation is to produce perceptually realistic and various motions aligning to musical beats. To overcome this challenge, we propose a generic new approach that bridges the gap between music-conditional sequence-to-sequence learning and recent unconditional generative architectures via `Transformer Conditional GAN`. Specifically, we leverage the generative capability from transformer decoder, embedding both `conditional` and `stochastic` representations via self-attention module. By injecting a latent code and querying a certain duration of music and initial pose sequence, the proposed model enables the synthesis of diverse and consistent dance as shown in Figure 1.

However, to synthesize dance controlled by multiple dance genres, injecting latent code with conditions is an ineffective process in such multi-domain translation task, which can not be scalable to the increasing number of domains. As discussed in multi-domain image-to-image trans-

*Corresponding author.

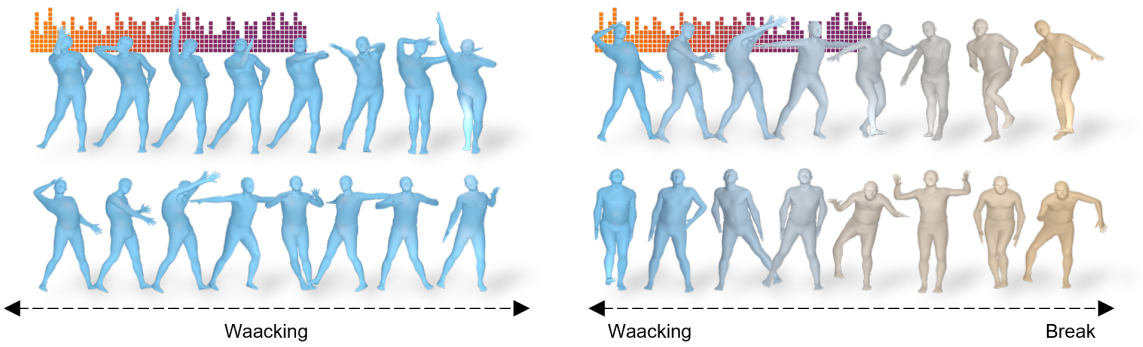

Figure 1. **Goal:** We learn a Music-conditioned transformer NETwork (MNET) to generate diverse dance motions following beats. Given music, our model not only generates diverse sequences within one dance genre (left two rows) but also synthesizes various dance genres during the music (right two rows).

lation literature [4,22,28,42,46], given `k` domains, `k(k-1)` generators are required to sufficiently handle translations between each and every domain, limiting their practical usage. To address the scalability, we employ two modules, a mapping network and a multi-task discriminator, to the sequence-to-sequence generative learning inspired by [10]. The mapping network learns to transform random Gaussian noise into each dance genre code, which is termed `style code` for a specific domain. For the multi-task discriminator, we take the role of the classifier to the transformer encoder, where the module performs per-style clssification as in the standard GAN setting. Considering multiple domains, both modules have multiple output branches. Finally, our generator learns to successfully synthesize diverse dance motions over multiple domains with a single model utilizing an adversarial framework.

The main contribution of this work is three-fold: (1) We newly introduce MNET, a novel Transformer-based conditional GAN framework, and train it to generate pluralistic dance motions by sampling from each latent representation of multiple dance genres. (2) We demonstrate that it is possible to learn to generate realistic and diverse dance motions which scalable to multiple dance genres in terms of visual quality and empirical metrics. (3) We present a comprehensive ablation studies of the architecture and loss components outperforming state-of-the-art performance on the AIST++ dataset, which contains 3D motions reconstructed from real dancers paired with music and multiple dance genres. Code will be available for research purposes. Project page

2. Related Work

Music to dance generation. Different from common motion synthesis [2, 3, 7, 17–19], dance generation includes its own challenges in that choreographed movements are extremely complex to animate. Thanks to the success of 2D human pose estimation [8], most earlier works have been studied in 2D pose context [16, 29, 32, 53, 55] by leveraging the huge amount of paired pose and music from dance videos available online. While 3D dance generation should capture high kinematic complexity for the dynamic synthesis. Various methods have been proposed to handle this task, where network architectures such as LSTMs [31, 61, 67], GANs [29, 57], and sequence-to-sequence methods [1,21] have been explored. Most recent approaches employ transformer-based architectures. TSMT [37] employs a two-stream motion transformer that computes the discrete representation of the output pose which degrades motion quality. FACT [38] present a full-attention based cross-modal transformer that adopts sequence-to-sequence learning to generate more realistic 3D dance sequence. Close to the previous works, DanceNet3D [35] also employ transformer architecture, but they further introduce a kinematic chain network that enables the model to adapt to the temporal locality of motions. In contrast to these prior works, our goal is to synthesize dance motions in a controlled way embedding both conditional and stochastic representation via transformer architecture.

Cross-modal sequence-to-sequence learning. Most existing works for cross-modal sequence-to-sequence generation task are often dominated by modeling between vision and text such as image/video captioning [25, 41, 59] and text-to-image generation [52, 62]. With the recent success of multi-head attention module, transformer-based approaches now the de-facto architecture achieving SOTA performance for many sequence-to-sequence learning tasks [13,23,36,56,64]. 3D dance generation innately requires to take into account the consistency between dance and music simultaneously. This framework is closely related to the research of learning a universal multi-modal generation task.

Multiple domain synthesis. Multiple domain synthesis has long been discussed especially in image-to-image translation aiming to learn a mapping between different visual domains. Most of the works inject a low-dimensional latent

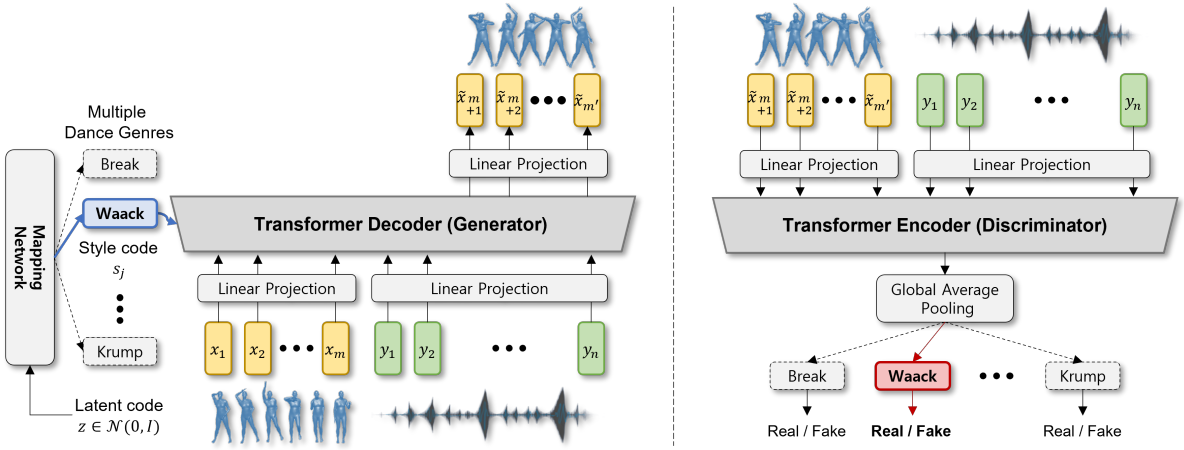

Figure 2. **Method overview:** We illustrate the generator (left) and the discriminator (right) of our transformer-based conditional GAN to generate diverse dance motions synthesized by multiple dance genres. The mapping network transforms a latent code into style code for multiple dance genres. The generator outputs long-range future motion by taking both sequences of seed motion and music piece as query and the style code as key and value. Given a sequence of real (fake) motions with a corresponding music piece, the discriminator distinguishes between real and fake motions from multiple domains.

code into the generator and map the sampled code between the two domains for various styles when generating images [22, 22, 28, 42, 45]. However, they are limited to scale the increasing number of domains. To address the multiple domain scalability, StarGAN [9] learns the mappings between all available domains using a single generator but is limited to learning a deterministic mapping per each domain. To get the diverse images across multiple domains, StarGAN2 [10] introduces a mapping network replacing its domain label with the specific style code sampled from random Gaussian noise. Inspired by the success of image-to-image translation, our work explores music-conditioned 3D dance generation considering multiple dance genres.

Transformer Generative Adversarial Networks. Transformer models have recently demonstrated exemplary performance in language tasks [6, 11, 58]. Vision transformer has increased interest in neural network models [12, 14, 39]. Several works utilize transformer as a generative adversarial training to learn data distribution from sampled latent noise which is completely free of convolutions [24, 33]. In this work, we adopt the strong representation capability of transformers to bridge the gap between music-conditional sequence-to-sequence learning and unconditional generative architectures for both conditional and stochastic representations.

3. Music-Conditioned Pluralistic Dancing

Problem definition. Suppose we have a dataset $\mathcal{D} = \{\mathbf{x}, \mathbf{y}\}$, where $\mathbf{x} = \{x_t\}_{t=1}^{m}$ is a sequence of movement, and $\mathbf{y} = \{y_t\}_{t=1}^{n}$ is a music clip. Previous dance generation methods attempt to reconstruct a sequence of future motion $\hat{\mathbf{x}} = \{\hat{x}_t\}_{t=m+1}^{m'}$ conditioned by a seed sample of motion $\mathbf{x}$ and a longer music sequence $\mathbf{y}$, where $n \gg m'$. This results in only a single solution with fixed dance genre. In contrast, our goal is to sample from $p(\hat{\mathbf{x}}|\mathbf{x}, \mathbf{y}, \mathbf{s})$, where $\mathbf{s}$ is an arbitrary style of multiple dance genres.

Motion and music representation. For 3D dance generation, we employ SMPL pose parameters [40] which represent 24×3 scalar values of joint rotations in the kinematic tree. We combine the 6-dim rotation matrix repersentation [30, 34, 65] for all 24 joints, along with a 3-dim global translation vector, resulting in a $x_t \in \mathbb{R}^{147}$ motion representation. For the music data, we follow the previous work to extract a $y_t \in \mathbb{R}^{35}$ representation which contains 1-dim envelope, 20-dim MFCC, 12-dim chroma, 1-dim one-hot peaks and 1-dim one-hot beats by employing the publicly available audio processing toolbox Librosa [43].

3.1. Conditional Transformer GAN

The goal is to generate diverse dance synthesized by multiple dance genres using transformer generator G given $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{s}$. Thus, the model can generate dance $\hat{\mathbf{x}}$ based on $G(\mathbf{x}, \mathbf{y}, \mathbf{s}_j)$ where j is the index of multiple domains. Here, we first present mapping network F generates a style code $\mathbf{s}$ of each dance genre. Then, we introduce our sequence-to-sequence architectures, transformer decoder G, chosen for pluralistic dance generation. Lastly, we introduce a transformer encoder D that takes the role of the multi-task discriminator. This process is depicted in Figure 2.

Mapping network. Motivated by existing work [10], mapping network generates a style code $\mathbf{s}_j = F_j(\mathbf{z})$ from given latent code $\mathbf{z}$ for all available dance genres, where j is index

of each domain. F consists of stacked MLPs with multiple outputs branches corresponding to all available domains. This makes F can produce diverse and scalable style code by sampling the latent code $\mathbf{z} \in \mathcal{Z}$.

Generator design. Designing a G based on the transformer architecture is a nontrivial task. A challenge is that the GAN training becomes highly unstable when coupled with a multi-head attention module that is hindered by high-variance gradients during adversarial training. Therefore, we empirically chose the architectural design and we discuss several baselines in Section 4.4. For our choice, we employ a transformer decoder as the G. Given the seed motion $\mathbf{x}$ and audio features $\mathbf{y}$, these are concatenated and fed as a query. In contrast, the style code $\mathbf{s}$ from the mapping network is fed as key and value. The transformer decoder outputs a sequence of $\hat{\mathbf{x}}$ by passing through a linear projection.

Discriminator design. Similar to the G, the transformer-based discriminator D takes the concatenated real (fake) motion $\mathbf{x}_{m+1:m'}$ ($\hat{\mathbf{x}}_{m+1:m'}$) and audio feature $\mathbf{y}$, but we employ transformer encoder as a discriminator. To allow the G to synthesize a dance sequence reflecting the style for all genres, the D is a multi-task discriminator, which consists of multiple output branches. Each branch D_j classifies the dance into being real and fake of its domain j.

3.2. Training

Overall, the architecture is trained using an adversarial loss as well as a number of additional losses. We present an ablation study regarding loss function in Section 4.4.

Adversarial loss. Given a seed motion $\mathbf{x}$ with its dance genre index j and corresponding music features $\mathbf{y}$, we sample a latent code $\mathbf{z} \in \mathcal{Z}$, and extract target style code $\mathbf{s}_j = F_j(\mathbf{z})$. The generator G takes an $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{s}_j$, to generate an output dance sequence $G(\mathbf{x}, \mathbf{y}, \mathbf{s}_j)$ via an original adversarial loss

$$\begin{aligned}\mathcal{L}_{ori} =& \mathbb{E}_{\mathbf{x}}[\log D_j(\mathbf{x}_{j,m+1:m'}, \mathbf{y})] + \\ & \mathbb{E}_{\mathbf{x},\mathbf{z}}[\log(1 - D_j(G(\mathbf{x}_j, \mathbf{y}, \mathbf{s}_j), \mathbf{y}))],\end{aligned} \tag{1}$$

where the $D_j(\cdot)$ has an objective to distinguish generated motion sequence from the real ones corresponding to the specific domain j. While the G learns to utilize style code $\mathbf{s}_j$ through mapping network F and generates the output that is indistinguishable from the real motion sequence of the domain j.

Our goal is not only to generate diverse dance motions but also synthesize sequence following the style code that represents multiple dance genres. Suppose we have a dance motion $\mathbf{x}_j$ representing style $\mathbf{s}_j$ as a seed motion, and we use this as an input of a generator to synthesize a dance motion $G(\mathbf{x}_j, \mathbf{y}, \mathbf{s}_i)$ using a different style $\mathbf{s}_i$. We found that the transformer conditional GAN focuses on the style of the seed motion sequence which sticks to the style variants according to $\mathbf{x}_j$, and ignores the input style variants $\mathbf{s}_i$. To further guarantee that the generated motion sequence properly preserves the domain-specific style regardless of seed motion $\mathbf{x}_j$, we employ style-focusing term

$$\begin{aligned}\mathcal{L}_{sty} = \mathbb{E}_{\mathbf{x},\mathbf{z}}[&\log D_j(G(\mathbf{x}_j, \mathbf{y}, \mathbf{s}_i), \mathbf{y}) + \\ &\log(1 - D_i(G(\mathbf{x}_j, \mathbf{y}, \mathbf{s}_i), \mathbf{y}))].\end{aligned} \tag{2}$$

By indirectly moving the output of each discriminator branch apart, the generator can focus more on style code, encouraging each domain to learn disentangled representations. Thus, the new adversarial loss is now can be computed by $\mathcal{L}_{adv} = \mathcal{L}_{ori} + \mathcal{L}_{sty}$

Appearance matching loss. The output of the model is the future motion sequence supervised by appearance matching loss using both pose parameters and vertex coordinates. As such, we use L2 loss between the ground-truth pose sequence $\mathbf{x}_{m+1:m'}$ and our prediction $\hat{\mathbf{x}}_{m+1:m'} = G(\mathbf{x}, \mathbf{y}, \mathbf{s}_j)$ as $\mathcal{L}_p = \mathbb{E}_{\mathbf{x}}\left[\left\|\mathbf{x}_{m+1:m'} - \hat{\mathbf{x}}_{m+1:m'}\right\|_2\right]$. For the pose parameters, we contain both the SMPL rotations and the global translations. For the global consistency, we further minimize the distance between the ground-truth and predicted vertices. Following [49], we integrate differentiable SMPL layer with a mean shape (i.e., $\beta = \vec{0}$) as a part of end-to-end framework to obtain the root-centered vertices of the mesh $\mathbf{v}_{m+1:m'}$ and $\hat{\mathbf{v}}_{m+1:m'}$. By minimizing the vertices L2 distance, we define vertex loss as $\mathcal{L}_v = \mathbb{E}_{\mathbf{v}}\left[\left\|\mathbf{v}_{m+1:m'} - \hat{\mathbf{v}}_{m+1:m'}\right\|_2\right]$. Finally, our appearance matching loss is given by $\mathcal{L}_{app} = \mathcal{L}_p + \mathcal{L}_v$.

Style diversity loss. To further encourage the generator G to produce pluralistic dance motions, we explicitly regularize G via diversity loss [10, 42]

$$\mathcal{L}_{div} = \mathbb{E}_{\mathbf{x},\mathbf{z}_1,\mathbf{z}_2,j}[\|G(\mathbf{x}, \mathbf{y}, \mathbf{s}_{j,1}) - G(\mathbf{x}, \mathbf{y}, \mathbf{s}_{j,2})\|_2], \tag{3}$$

where the target style codes $\mathbf{s}_{j,1}$ and $\mathbf{s}_{j,2}$ are produced by F conditioned on two random noise $\mathbf{z}_1$ and $\mathbf{z}_2$. The mapped sequences $G(\mathbf{x}, \mathbf{y}, \mathbf{s}_{j,1})$ and $G(\mathbf{x}, \mathbf{y}, \mathbf{s}_{j,2})$ in the same dance genre j are more likely to be collapsed into the same mode. By maximizing the $\mathcal{L}_{div}$ in the same genre, our model focuses more on variations of the input style code that contribute to the output diversity.

Overall loss. Thus, the overall loss can be summarized as follow:

$$\min_{G,F} \max_{D} \mathcal{L}_{adv} + \mathcal{L}_{app} - \lambda_{div}\mathcal{L}_{div}, \tag{4}$$

where λ_{div} is the importance of weighting determining the trade-off between diversity and realistic motion sequence (see Section B of the appendix). The weighting factors of the remaining loss terms are equally weighted in our experiments.

	Motion Plausibility				Generation Diversity			Motion-Music Cons.	User Study
	$FID_k \downarrow$	$FID_g \downarrow$	$FID_s \downarrow$	Acc. $\uparrow$	$Dist_{m,k} \uparrow$	$Dist_{m,g} \uparrow$	$Dist_{m,s} \uparrow$	BeatAlign $\uparrow$	MNET WinRate
AIST++	-	-	-	98.6	10.39	8.48	8.91	0.292	42.3%
Dancenet [66]	56.67	16.47	38.49	43.6	2.10	2.64	2.76	0.220	90.38 %
DanceRevolution [21]	42.93	14.85	26.53	72.9	3.82	3.31	2.45	0.215	84.17 %
FACT [38]	33.08	11.82	11.37	76.1	5.83	5.28	5.31	0.241	62.39 %
MNET (ours)	**29.52**	**9.36**	**7.90**	**83.7**	**6.93**	**6.77**	**6.32**	**0.246**	-

Table 1. **State-of-the-art comparison:** We compare to the three recent methods. Our model generates plausible motion sequences than other baselines in terms of FID, and better represents the style of the dance genres through the Acc. where the score evaluates the style consistency of the generated dance. Our model shows more diversified dance motions when conditioned on different music and more consistent results aligned with input music beat. $\downarrow$ A lower value is better. $\uparrow$ A higher value is better.

3.3. Implementation Details

In our experiments, we set the input of a seed motion sequence as $m = 120$ frames (2 seconds) and a music sequence as $n = 240$ frames (4 seconds) following the previous setting [38], where the two sequences are aligned on the first frame. The output of our generator is the future motion sequence with $m^{'} - m = 60$ frames supervised by the proposed losses. During inference, we continuously generate future motions in an auto-regressive manner. The seed motion is replaced with newly generated motion and the music is shifted 60 frames to feed into the generator at every step.

We use a 8-blocks transformer encoder for the discriminator and we increase the number of blocks to 12 for the transformer decoder in the generator. All the two transformers have 8 attention heads with $d = 512$ hidden sizes. We experimentally find that the increasing number of heads does not improve the performance during conditional GAN training, and we discuss it in Section B. Furthermore, we use a relative positional encoding for all transformer architecture instead of an absolute positional encoding [12, 58]. Following [20, 39] a relative position bias $B \in \mathbb{R}^{N_q \times N_{kv}}$ is included in computing self-attention

$$\text{Attention}(Q, K, V) = \text{SoftMax}(QK^T/\sqrt{d} + B)V, \quad (5)$$

where, $Q \in \mathbb{R}^{N_q \times b}$, $K, V \in \mathbb{R}^{N_{kv} \times b}$ are the query, key and value metrices, and N_q is the query sequence of length and N_{kv} is the key-value sequence of length, respectively.

We use the AdamW optimizer with a fixed learning rate of $\lambda = 1e^{-4}$ and all weights are randomly initialized. All our experiments are trained with 10 mini-batch size. The performance is sensitive to this parameter and we discuss it in Section B. The training finishes after $500k$ steps on 8 GPUs with an accumulated VRAM 96GB. If hardware permits, 16-bit precision training is enabled.

4. Experiments

4.1. Dataset and Baseline

AIST++ dataset. One of the biggest bottlenecks in the 3D dance generation approaches is the data problem. To mitigate this issue, recent works [16, 21, 29, 35] collected large amounts of dance videos on the Internet, and extracted 3D pose sequences with synchronized audios. However, most of the data is not publicly released and may not be reliable because of the 2D to 3D depth ambiguity. In contrast, AIST++ [38] is a large-scale 3D human dance motion dataset that is captured from calibrated multi-view videos. The dataset has a wide variety of 3D motions paired with music which contains 1408 sequences, 30 subjects, and 10 dance genres. All our experiments are performed on AIST++ dataset.

Baseline. For the comprehensive evaluations, we mainly compare our proposed method with FACT [38] which shows current SOTA results for 3D dance generation. Further, we employ SOTA 2D dance generation methods which are Dancenet [66] and DanceRevolution [21]. We adopt this with small modifications to generate 3D joint locations which enable the direct comparisons quantitatively and qualitatively. These models are re-trained until convergence following the same experimental settings proposed in each study using the AIST++.

4.2. Quantitative Comparisons

To evaluate our approach, we measure (1) motion plausibility, (2) generation diversity and (3) motion-music consistency, following the [21, 37, 38]. For all criteria, our model shows superior performance compared to baselines, as shown in Table 1.

Motion plausibility. We measure geometric and kinematic Fréchet Inception Distance (FID) for the motion plausibility. To measure the distribution of generated and ground-truth dances, the two well-designed motion feature extractors [44, 47] are employed that produces a kinetic feature $\mathbf{z}_k \in \mathbb{R}^{72}$ and a geometric feature $\mathbf{z}_g \in \mathbb{R}^{33}$. Furthermore, we train a style classifier on dance motions of 10 dance genres and utilize it to extract a style feature $\mathbf{z}_s \in \mathbb{R}^{512}$ for the given dances. We denote the FID based on these geometric, kinetic and style features as FID_k, FID_g and FID_s, respectively. Besides, we use style classifier to measure the prediction accuracy of the dance genres.

	Break ($\mathbf{s}_0$)	Pop ($\mathbf{s}_1$)	Lock ($\mathbf{s}_2$)	Middle hip-hop ($\mathbf{s}_3$)	LA style hip-hop ($\mathbf{s}_4$)	House ($\mathbf{s}_5$)	Waack ($\mathbf{s}_6$)	Krump ($\mathbf{s}_7$)	Street jazz ($\mathbf{s}_8$)	Ballet jazz ($\mathbf{s}_9$)
$\text{FID}_k \downarrow$	26.31	23.44	30.10	31.48	29.85	30.75	29.10	22.49	34.18	32.72
$\text{FID}_g \downarrow$	9.03	9.17	11.39	12.87	9.68	10.28	11.95	10.89	9.32	9.05
$\text{FID}_s \downarrow$	10.82	9.85	9.39	7.65	7.37	7.42	8.39	9.18	8.27	7.06
$\text{Dist}_{s,k} \uparrow$	5.11	5.79	3.68	3.97	4.71	4.75	4.16	6.56	3.66	5.17
$\text{Dist}_{s,g} \uparrow$	6.85	6.33	7.75	6.47	5.24	5.96	6.83	5.18	5.31	3.94
$\text{Dist}_{s,s} \uparrow$	3.75	3.19	4.68	5.35	5.17	3.79	2.35	2.97	3.62	4.81

Table 2. **Comparisons on individual dance genres.** We investigate motion plausibility and generation diversity quantitatively on individual dance genres. Our model synthesizes dance motions realistic and diverse for all dance genres. This indicates that the mapping network effectively separates all domains.

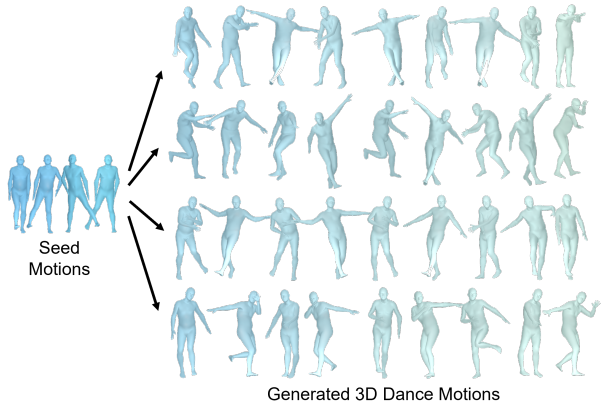

Figure 3. **Generation diversity:** We illustrate the generation diversity using different kinds of music. We fix latent code and select the same dance genre (Break) during iterative inference. We demonstrate that our model is capable of generating different ways by capturing musical change. More results can be found in the supplementary video.

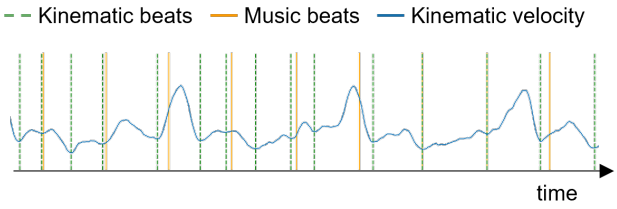

Figure 4. **Motion-music consistency:** We visualize an example of beat alignment between music and generated dance. The orange and green dashes on the graph indicate the extracted musical beats and kinematic beats, respectively. The kinematic beats are computed as the local minima of the kinetic velocity (blue line).

However, measuring motion plausibility is hard for the pluralistic dance generation, as our goal is to get diverse but reasonable solutions for the given conditions. The ground-truth dance is only one solution of many, and the metrics that are calculated between the real and generate sequence, are not measurable. Therefore, we generate 10 sequences for each 10 dance genres from 20 kinds of music, producing a total of 2000 samples. Similar to [63], we assume that our top 1 sample for each dance genre (ranked by the multi-task discriminator) is close to the original ground truth. We generate motion sequence with $T = 1200$ frames (20 seconds). As shown in Table 1, all FID scores are recorded significantly lower than our baselines which means that our generated samples are much closer to that of the real. Furthermore, the prediction accuracy of the dance genre achieves 83.7%.

Generation diversity. In contrast to the baseline methods, our model not only generates multiple sequences by various music but also produces multiple sequences for the same music by sampling the style codes. We compute these diversity using the average Euclidean distance in the feature space. To measure the diversity by various music, we generate 40 motion sequences and compute the distance employing geometric, kinematic, and style features which are denoted as $\text{Dist}_{m,k}$, $\text{Dist}_{m,g}$ and $\text{Dist}_{m,s}$, respectively. Our model obtains a higher score for all metrics which means that our model is more dependent on input music than the baseline approaches and thus provides multiple dance motions for different music clips of the same style. Table 1 shows results of the diversity by various music. Figure 3 visualizes dance motions by various music. Similar to music variation diversity, we calculate the diversity by style code from the same music using the geometric, kinematic, and style feature extractor. We generate 20 samples for each dance genre from randomly selected 30 music clips and compute distance in the three feature spaces, which are denoted as $\text{Dist}_{s,k}$, $\text{Dist}_{s,g}$ and $\text{Dist}_{s,s}$. Table 2 shows results of the motion plausibility and generation diversity quantitatively on individual dance genres. The recorded scores for all dance genres show similar values for all metrics which represent that the style code for each dance genre is properly disentangled.

Motion-music consistency. As a skilled choreographer moves rhythmically in accordance with the music beat, the outputs of a well-trained dance generation model require consistency between motion beat and music beat. We evaluate the motion to music consistency using the Beat Alignment Score introduced in [38]. The Beat Alignment Score is defined as the average distance between kinematic beat

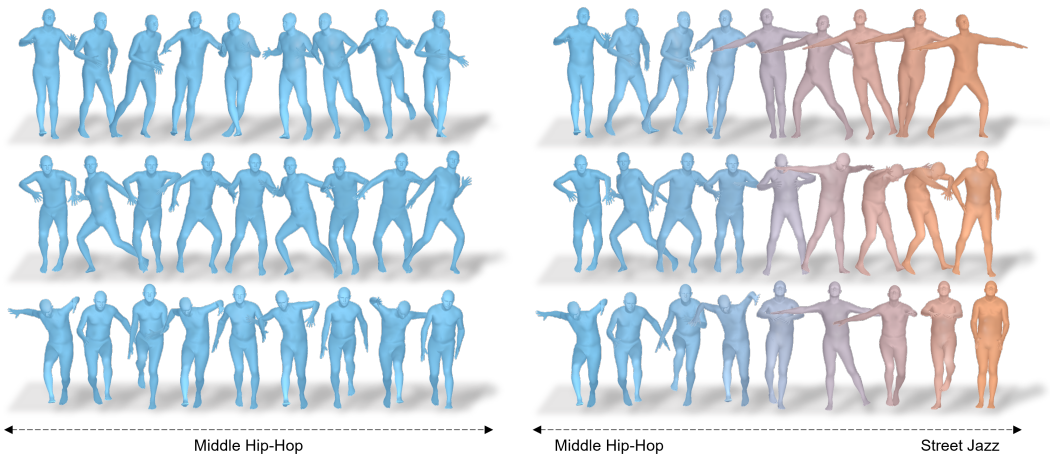

Figure 5. **Qualitative result:** We illustrate the diversity of our generations for the two aspects. The left visualizes outputs guided by latent codes. Note that the motions in each row share a music and dance genres with different latent codes. The right visualize outputs guided by style codes. Note that the motions in each row share music and latent codes but the style codes are differently selected by multiple branches of the mapping net during iterative inferences. More results can be found in the supplementary video.

and its nearest music beat

$$\text{BeatAlign} = \frac{1}{m}\sum_{i=1}^{m}\exp(-\frac{\min_{\forall t_j^y \in B^y}||t_i^x - t_j^y||^2}{2\sigma^2}), \quad (6)$$

where $B^x = \{t_i^x\}$ is the kinematic beats which are the local minima of the kinetic velocity, $B^y = \{t_j^y\}$ is the music beats which are extracted using Librosa [43] and σ is a normalize parameter where we set $\sigma = 3$ in all experiments. As shown in Table 1, our model shows superior music and motion consistency compared to these baselines. Further, we show the visualization of motion beat and music beat consistency in Figure 4.

4.3. Qualitative Results

Here, we visualize several examples from our generation for the two perspectives: latent-code guided generation and style-code guided generation. Then, we compare the user preferences of our methods with baseline approaches. We further discuss qualitative comparison in Section C.

Latent-code guided generation. The left of Figure 5 provides visualized examples guided by different latent codes. We show the 3 generations per dance genres. Our model takes 3 differently sampled noises through the mapping network and selects one dance genre among several branches, where each style code represents the same dance genre but is born in different latent codes. We demonstrate that the proposed model generates different ways to distinguish the different latent codes in the same dance genres.

Style-code guided generation. The right of Figure 5 provides visualized examples guided by different style codes. After the model is trained, we generate continuous motion in an auto-regressive manner at test time. To show the outputs synthesized by style codes, we fixed latent codes and change the style codes by selecting different branches of the mapping network during the repeated future motion generation process. By doing so, our generation can only focus on the variation of the style codes, whose domain-specific information is already taken care of by the mapping network. We observe that our method successfully renders distinctive styles sequentially across all dance genres.

User study. We compare the user preference of our method with baseline approaches. For user study, each subject is asked to select one between our results and one randomly selected counterparts for the question of "which person is dancing more plausible to the music?". 23 subjects participated in the user study. As shown in Table 1, our method obtains the majority of votes compared to all baselines. Further, it is noteworthy that the preference between AIST++ and generated motions is competitive.

4.4. Ablation Study

In this section, we verify the choice and effect of our contributions separately. We conduct the following ablation experiments in architecture design and loss study. The effectiveness is measured using the motion plausibility (FID_k), generation diversity ($\text{Dist}_{m,k}$), and motion-music consistency (BeatAlign).

Architecture design. For the dance generation, the attention-based approach (i.e., Transformer) has demonstrated advantages over several architectural designs such as a simple autoencoder or a GRU-based recurrent neural network in several previous studies [35, 50]. However, the challenge is that GAN training tend to unstable when coupled with transformer architecture [24, 33]. Here, we investigate the backbone of transformer-based GAN design

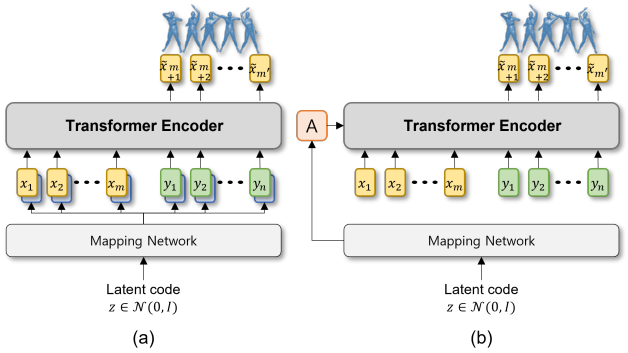

Figure 6. **Generator architecture:** The two plausible baseline architectures. (a) embeds style code by concatenating with input conditions. (b) replaces normalization with self-modulated layer norm in the transformer architecture.

in terms of both generator and discriminator.

Designing a generator based on the transformer is a non-tribial task. The challenge comes from properly embedding the latent code in the dense attention process. We first discuss two plausible baseline architectures, as shown in Figure 6. Both baselines employ a transformer encoder and generate motions from embeddings. Figure 6 (a) takes the style codes by concatenating with input conditions. Alternatively, Figure 6 (b) embeds the style code by replacing the layer norm [5] in transformer architecture with a self-modulated layer norm (SLN) [33]

$$\text{SLN}(\mathbf{h}_l, \mathbf{s}) = \gamma_l(\mathbf{s}) \odot \frac{\mathbf{h}_l - \mu}{\sigma} + \beta_l(\mathbf{s}), \tag{7}$$

where μ and σ are the mean and variance of the summed inputs within the layer, and γ_l and β_l compute learnable parameters controlled by the style code $\mathbf{s}$. These models are trained with the same setting as the proposed approach. Table 3 shows the quantitative performances under different generator architectures. We find that Figure 6 (a) works well but shows a large performance gap with our proposed generator. Figure 6 (b) underperforms other architectures due to training instability which means that SLN interacts poorly with self-attention. Besides, our generator works together with the motion-discriminator which captures the sequential motion using GRUs [26]. The results show the proposed method are compatible with both transformer-based and RNN-based discriminators.

Loss study. Here, we investigate the influence of the objective function in our transformer GAN. For all experiments, we fix the original adversarial loss and add the proposed loss functions in succession. As shown in Table 4, when using a single adversarial loss, our model is not sufficient to learn the high kinematic complexity (A) where the outputs produce motions with significant jitter. In contrast, when we use appearance loss (B and C), the performance improved significantly, especially for FID_g. This suggests that the appearance loss effectively constraints the pose space, but limits diversity by collapsing the sampled latent code into a similar space. We then improved this baseline by adding diversity loss (D). However, there is a trade-off between diversity and realistic motion sequence according to weighting parameters λ_{div} of diversity loss. We empirically determine the λ_{div} and discuss it in Section C. Finally, we introduce regularization to disentangle the style codes of different dance genres (E), improving results further.

Generator	**Discriminator**	$\text{FID}_k \downarrow$	$\text{Dist}_{m,k} \uparrow$	BeatAlign ↑
Figure 6 (a)	MNET	35.27	5.32	0.225
Figure 6 (b)	MNET	59.84	3.91	0.206
MNET	Motion-dis.	**29.40**	6.20	0.239
MNET	MNET	29.52	**6.93**	**0.246**

Table 3. **Ablation study of architecture design:** We compare plausible architectural designs regarding both generator and discriminator.

	$\text{FID}_k \downarrow$	$\text{Dist}_{m,k} \uparrow$	BeatAlign ↑
A Adversarial Loss $\mathcal{L}_{org}$	59.58	4.57	0.197
B + Pose Parameters $\mathcal{L}_p$	33.84	4.20	0.239
C + Vertex Coordinates $\mathcal{L}_v$	31.71	4.39	0.215
D + Diversity $\mathcal{L}_{div}$	29.88	6.71	0.207
E + Style-Focusing $\mathcal{L}_{reg}$	**29.52**	**6.93**	**0.246**

Table 4. **Ablation study of loss function:** We compare quantitative scores by adding loss functions with different configurations.

5. Conclusion

We propose a new Transformer-based GAN model to generate music-conditioned pluralistic dance motions synthesized by multiple dance genres, translating a motion of one dance genres to diverse motions of a target dance genres, and supporting multiple target dance genres. We provide a detailed discussion to assess different components of our proposed approach quantitatively and qualitatively. The experimental results show that our model can generate motions with rich styles across multiple domains, remarkably outperforming the baselines in terms of both automatic metrics and human evaluation. Currently, our model requires seed motion to generate future motion. Exploring how to generate diverse dance motions without seed motion is a more practical usage and exciting direction.

Acknowledgment. This work has supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1A2C3011697) and the Yonsei University Research Fund of 2021 (2021-22-0001).

References

[1] Hyemin Ahn, Jaehun Kim, Kihyun Kim, and Songhwai Oh. Generative autoregressive networks for 3d dancing move synthesis from music. *IEEE Robotics and Automation Letters*, 5(2):3501–3508, 2020. 2

[2] Emre Aksan, Peng Cao, Manuel Kaufmann, and Otmar Hilliges. Attention, please: A spatio-temporal transformer for 3d human motion prediction. *arXiv preprint arXiv:2004.08692*, 2, 2020. 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] Amjad Almahairi, Sai Rajeshwar, Alessandro Sordoni, Philip Bachman, and Aaron Courville. Augmented cyclegan: Learning many-to-many mappings from unpaired data. In *International Conference on Machine Learning*, pages 195–204. PMLR, 2018. 2

[5] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 8

[6] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 3

[7] 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. 2

[8] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019. 2

[9] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 3

[10] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020. 2, 3, 4

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3, 5

[13] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. Weakly supervised dense event captioning in videos. *arXiv preprint arXiv:1812.03849*, 2018. 2

[14] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021. 3

[15] William Falcon et al. Pytorch lightning. *GitHub. Note: https://github.com/PyTorchLightning/pytorch-lightning*, 3, 2019. 12

[16] Joao P Ferreira, Thiago M Coutinho, Thiago L Gomes, José F Neto, Rafael Azevedo, Renato Martins, and Erickson R Nascimento. Learning to dance: A graph convolutional adversarial network to generate realistic dance motions from audio. *Computers & Graphics*, 94:11–21, 2021. 2, 5

[17] 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. 2

[18] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)*, 35(4):1–11, 2016. 2

[19] Daniel Holden, Jun Saito, Taku Komura, and Thomas Joyce. Learning motion manifolds with convolutional autoencoders. In *SIGGRAPH Asia 2015 Technical Briefs*, pages 1–4. 2015. 2

[20] Han Hu, Zheng Zhang, Zhenda Xie, and Stephen Lin. Local relation networks for image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3464–3473, 2019. 5

[21] Ruozi Huang, Huang Hu, Wei Wu, Kei Sawada, Mi Zhang, and Daxin Jiang. Dance revolution: Long-term dance generation with music via curriculum learning. *arXiv preprint arXiv:2006.06119*, 2020. 1, 2, 5, 14

[22] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018. 2, 3

[23] Vladimir Iashin and Esa Rahtu. Multi-modal dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 958–959, 2020. 2

[24] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two transformers can make one strong gan. *arXiv preprint arXiv:2102.07074*, 1(3), 2021. 3, 7

[25] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 2

[26] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5253–5263, 2020. 8

[27] Kimerer LaMothe. The dancing species: how moving together in time helps make us human. *Aeon*, 2019. 1

[28] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018. 2, 3

[29] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. Dancing to music. *arXiv preprint arXiv:1911.02001*, 2019. 1, 2, 5

[30] Inwoong Lee, Doyoung Kim, Seoungyoon Kang, and Sanghoon Lee. Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1012–1020, 2017. 3

[31] Inwoong Lee, Doyoung Kim, and Sanghoon Lee. 3-d human behavior understanding using generalized ts-lstm networks. *IEEE Transactions on Multimedia*, 23:415–428, 2020. 2

[32] Juheon Lee, Seohyun Kim, and Kyogu Lee. Listen to dance: Music-driven choreography generation using autoregressive encoder-decoder network. *arXiv preprint arXiv:1811.00818*, 2018. 2

[33] Kwonjoon Lee, Huiwen Chang, Lu Jiang, Han Zhang, Zhuowen Tu, and Ce Liu. Vitgan: Training gans with vision transformers. *arXiv preprint arXiv:2107.04589*, 2021. 3, 7, 8

[34] 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. 3

[35] Buyu Li, Yongchi Zhao, and Lu Sheng. Dancenet3d: Music based dance generation with parametric motion transformer. *arXiv preprint arXiv:2103.10206*, 2021. 2, 5, 7

[36] Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. Entangled transformer for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8928–8937, 2019. 2

[37] Jiaman Li, Yihang Yin, Hang Chu, Yi Zhou, Tingwu Wang, Sanja Fidler, and Hao Li. Learning to generate diverse dance motions with transformer. *arXiv preprint arXiv:2008.08171*, 2020. 1, 2, 5

[38] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412, 2021. 1, 2, 5, 6, 12, 14

[39] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021. 3, 5

[40] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 3

[41] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7219–7228, 2018. 2

[42] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1429–1437, 2019. 2, 3, 4

[43] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, pages 18–25. Citeseer, 2015. 3, 7

[44] Meinard Müller, Tido Röder, and Michael Clausen. Efficient content-based retrieval of motion capture data. In *ACM SIGGRAPH 2005 Papers*, pages 677–685. 2005. 5

[45] Sanghyeon Na, Seungjoo Yoo, and Jaegul Choo. Miso: Mutual information loss with stochastic style representations for multimodal image-to-image translation. *arXiv preprint arXiv:1902.03938*, 2019. 3

[46] Anh-Duc Nguyen, Seonghwa Choi, Woojae Kim, and Sanghoon Lee. Graphx-convolution for point cloud deformation in 2d-to-3d conversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8628–8637, 2019. 2

[47] Kensuke Onuma, Christos Faloutsos, and Jessica K Hodgins. Fmdistance: A fast and effective distance function for motion capture data. In *Eurographics (Short Papers)*, pages 83–86, 2008. 5

[48] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019. 12

[49] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019. 4, 12

[50] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. *arXiv preprint arXiv:2104.05670*, 2021. 7

[51] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 12

[52] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, pages 1060–1069. PMLR, 2016. 2

[53] Xuanchi Ren, Haoran Li, Zijian Huang, and Qifeng Chen. Music-oriented dance video synthesis with pose perceptual loss. *arXiv preprint arXiv:1912.06606*, 2019. 2

[54] E Glenn Schellenberg, Ania M Krysciak, and R Jane Campbell. Perceiving emotion in melody: Interactive effects of pitch and rhythm. *Music Perception*, 18(2):155–171, 2000. 1

[55] Eli Shlizerman, Lucio Dery, Hayden Schoen, and Ira Kemelmacher-Shlizerman. Audio to body dynamics. In *Pro-*

ceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7574–7583, 2018. 2

[56] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473, 2019. 2

[57] Guofei Sun, Yongkang Wong, Zhiyong Cheng, Mohan S Kankanhalli, Weidong Geng, and Xiangdong Li. Deepdance: music-to-dance motion choreography with adversarial learning. *IEEE Transactions on Multimedia*, 23:497–509, 2020. 2

[58] 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. 3, 5

[59] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. 2

[60] Omry Yadan. Hydra - a framework for elegantly configuring complex applications. Github, 2019. 12

[61] Nelson Yalta, Shinji Watanabe, Kazuhiro Nakadai, and Tetsuya Ogata. Weakly-supervised deep recurrent neural networks for basic dance step generation. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019. 2

[62] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 2

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

[64] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8739–8748, 2018. 2

[65] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019. 3

[66] Wenlin Zhuang, Congyi Wang, Siyu Xia, Jinxiang Chai, and Yangang Wang. Music2dance: Dancenet for music-driven dance generation. *arXiv preprint arXiv:2002.03761*, 2020. 5

[67] Wenlin Zhuang, Yangang Wang, Joseph Robinson, Congyi Wang, Ming Shao, Yun Fu, and Siyu Xia. Towards 3d dance motion synthesis and control. *arXiv preprint arXiv:2006.05743*, 2020. 2