IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 6 June 2024, accepted 24 June 2024, date of publication 28 June 2024, date of current version 16 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3420433

RESEARCH ARTICLE

Music-Driven Synchronous Dance Generation Considering K-Pop Musical and Choreographical Characteristics

SEOHYUN KIM[1] **AND KYOGU LEE**[1,2], **(Senior Member, IEEE)**
[1]Music and Audio Research Group, Department of Intelligence and Information, Seoul National University, Seoul 08826, Republic of Korea
[2]Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul 08826, Republic of Korea

Corresponding author: Kyogu Lee (kglee@snu.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIT) under Grant RS-2023-00219429.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board of Seoul National University under Application No. 2212/002-003, and performed in line with the Declaration of Helsinki.

ABSTRACT Generating dance movements from music has been considered a highly challenging task, as it requires the model to comprehend concepts from two different modalities: audio and video. However, recently, research on dance generation based on deep learning has been actively conducted. Existing dance generation researches tend to focus on generating dances in limited genres or for single dancer, so when K-pop music that mixes multiple genres was applied to existing methods, they failed to generate dances of various genres or group dances. In this paper, we propose the K-pop dance generation model in an autoregressive manner, a system designed to generate two-person synchronous dances based on K-pop music. To achieve this, we created a dataset by collecting videos of multiple dancers simultaneously dancing to K-pop music and dancing in various genres. Generating synchronous dances has two meanings: one is to generate a dance that goes well with the input music and dance when both are given, and the other is to simultaneously generate multiple dances that match the given music. We call them secondary dance generation and group dance generation, respectively, and designed the proposed model, which can perform both two generation methods. In addition, we would like to propose additional learning methods to make a model that better generates synchronous dances. To assess the performance of the proposed model, both qualitative and quantitative evaluations are conducted, proving the effectiveness and suitability of the proposed model when generating synchronous dances for K-pop music.

INDEX TERMS Synchronous dance generation, K-pop group dance generation, autoregressive model, multi-step learning.

I. INTRODUCTION

Dance has been used in various human rituals, social communication, and entertainment from ancient times to the present. Additionally, music has always been closely associated with dance for all purposes. In modern times, dance is used for various reasons, but music is still one of the most important elements of dance. The rhythm, beat, and melody of music are important elements in dance because they are suitable for expressing dance, and one of the reasons why dance genres have become more diverse is because the genres of music have become more diverse. In particular, there are cases where a music genre becomes a dance genre, examples of which include techno, disco, and house. However, in some cases, dance does not necessarily correspond to a specific traditional music genre, as seen in geographical dance genres such as K-pop. K-pop includes rather a variety of dance genres and it's hard to define K-pop with a few combination of classical genres. Also, K-pop tends to prioritize diversity performed by several people rather than solo dance. When multiple people dance at the same time, some sections may show the same dance and others may show different dance, depending on the choreographer's choice.

The associate editor coordinating the review of this manuscript and approving it for publication was Jolanta Mizera-Pietraszko.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

The early models generating dance were rule-based. These methods had the drawback of not presenting creative and entirely new dances, instead, they demonstrated research success in suggesting similar dances within the existing database [7], [9]. Subsequently, methods using deep learning began to emerge and switched the paradigm of the dance generation. Due to the use of deep learning, research on generating natural dance movements that match given music has increased more than before. However, some of the existing studies have the disadvantage that only certain dance genres can be performed, or that it is difficult to generate dance when the composition of the music becomes complicated. In this paper, we aim to generate various dance movements using the characteristics of K-pop music. Existing dance generation research had limitations in not being able to respond well to the K-pop genre, and it was even more difficult to generate K-pop group dances. The proposed model can generate group dances that match K-pop music by considering the musical and choreographical characteristics of K-pop.

The proposed model can perform two types of synchronous dance generation: One is to generate a dance that goes well with the given music and dance when a pair of music and dance is given, and the other is to simultaneously generate multiple dances that go well with the given music. We call them secondary dance generation (SDG) and group dance generation (GDG), respectively. The model architecture was designed so that both SDG and GDG methods can be performed within one proposed model, and the desired generation method can be selected by changing model inputs as needed. In addition, we would like to propose additional learning methods such as *postnet* and *multi-step learning* to learn a model that generates better synchronous dance.

To evaluate the performance of the proposed model, quantitative and qualitative evaluations were performed. For quantitative evaluation, metrics such as Fréchet Inception Distance (FID), diversity (DIV), and beat alignment score (BAS) were utilized to assess the quality and diversity of the dance and the synchronization with the music. Through this, it was confirmed that the proposed model generates dance more suitable for K-pop music compared to baseline models. Qualitative evaluation was conducted through user evaluation, verifying the coherence between generated dances and the effectiveness of the proposed training methods.

The main contributions of this study can be summarized as follows. First, we built a K-pop dataset with a total length of 15.8 hours consisting of various dances and musics. Second, the proposed model, which generates synchronous dances, reflects the dance characteristics of K-pop based group dances. Third, we designed the dance generation model in an autoregressive manner, considering temporal and unique characteristics of K-pop music. Last but not least, we increased the diversity of dance movements to generate dance motions harmonious with each other but not identical.

II. RELATED WORK

This section describes related work, with subsections organized around cross-modality generation, dance generation, and dance datasets.

A. CROSS-MODALITY GENERATION

Recently, the role of deep learning is gradually expanding in the process of creating something new. In early stage, deep learning based research focused on tasks dealing with a single modality, such as upscaling image quality from low to high or enhancing the sound quality. However, the world often requires handling much more complex tasks and most tasks in reality often involve multiple modalities in a complex manner. Similarly, a human brain integrates and infers information, utilizing the multiple senses and intelligence to make judgments. Thus, in the recent deep learning researches, there has been continuous research on cross-modality models that learn networks using interactions between various modalities, much like human senses. In particular, many recent studies deal with models from a variety of fields whose inputs and outputs are in different domains. For example, the model proposed in Vinyals et al. [61] generates text for images, Cheng et al. [64] suggests a method that simultaneously considers video and audio for scene recognition, and Radford et al. [44] focuses on learning text and images together on a large-scale dataset, creating a model that enables interaction between modalities. In this way, the field of cross-modality generation emphasizes interactions between different modality. Recently, in this trend of research, there has been growing interest in generating dance by leveraging the relationship between music and dance data. The studies related to the dance generation are introduced in the next section.

B. DANCE GENERATION

The first dance generation studies began in the late 20th century, and at this time, rather than research on generating creative dance, research was conducted on how to find the dance chunk that best matches music among various dance chunks stored in a database. Therefore, to find a dance chunk sequence that matches the music sequence, they adopted a rule-based method such as the Viterbi algorithm, and tried to obtain a more natural result by using a post-processing method to smooth out connections between discontinuous dance chunks. However, as mentioned above, not only can creative results not be obtained, but it is difficult to provide results in variation [7], [62], [68], [69].

With the surge in cross-modality research within the realm of deep learning, there has been a significant shift in dance generation research. Rather than relying on heuristic algorithms, recent research aims to create a model that can continuously generate a wider variety of dance movements beyond those that exist in the database. These studies are similar to motion generation studies, but aim to generate a series of sequential movements that look like dance based on a given music. Music input is a key feature of dance

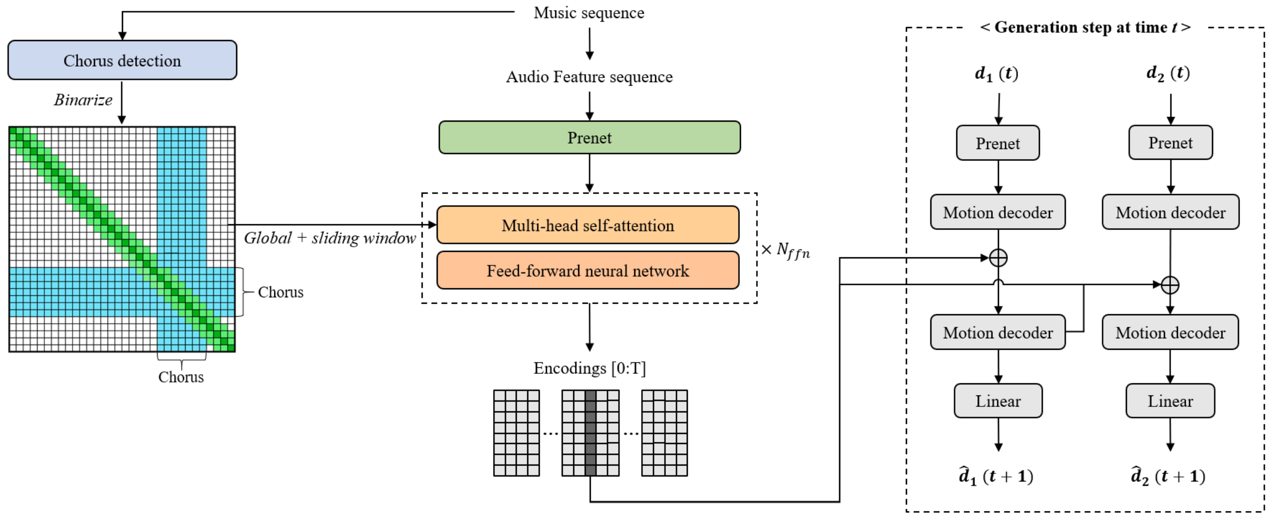

FIGURE 1. **The overview of entire model architecture. The model is based on an autoregressive *seq2seq* model, containing a music encoder and two motion decoders. To consider the characteristics of K-pop, chorus detection is preceded before the encoding processes. Two motion decoders each generate the lead dance and the secondary dance.**

generation models, and most dance generation models except for Wu et al. [52] that bidirectionally generate music and dance using cross-modal transformer or Okamura et al. [36] that generate dance using onomatopoeic input, use audio as input. In previous research, various deep learning models, including CNNs [10], [11], [12], RNNs [13], [14], LSTMs [17], [18], [19], [20], [21], [22], GNNs [16], [42], GANs [25], [26], [27], Transformers [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [59], Diffusion [39], [40], [41], and GPT [43], have been proposed to generate dance based on music input. LSTM and Transformer models have been particularly popular, as they excel in handling sequential data, which is essential for choreographing movements over time [19], [28]. In particular, recent studies using transformer or diffusion-based models have raised the possibility that the generated motion images may contain a variety of natural dance movements [37], [40].

C. DANCE DATASET

In the early stages of dance generation research, there were no commonly used datasets, and techniques for estimating poses from videos were not developed to a usable level. Therefore, motion capture modules were often used to collect data. Wallace et al. [13] attached 43 markers to the human body, generating data for 22 joints. Ye et al. [23] created a motion capture dataset by recording 41 joint positions to collect dance genres such as waltz, tango, cha-cha, and rumba, and segmented dances based on temporal annotations to serve as dance unit data. However, the number of dance videos uploaded to Internet platforms such as YouTube, TikTok, and Bilibili is increasing exponentially, and as pose estimation technology develops, the number of papers that collect pose data directly from videos is increasing [12], [25], [26], [32]. Meanwhile, the AIST dataset, which provides 10 genres of dance that includes music, became most commonly used in dance generation research after its release. This led to the creation of the AIST++ dataset, a 3D dataset based on AIST, which is widely used in both 2D and 3D forms, including papers such as [28], [37], [51], and [52]. Although these datasets have the advantage of offering clear high-quality dance footage, they suffer from the drawback of having limited data for each category. Additionally, some studies have categorized dance genres based on singers. For example, there is a study in which data on Michael Jackson was collected as a single dance genre [42].

III. DATASET

We had difficulty leveraging existing datasets to assess the connection between music and dance. This is because if the model learns only a specific dance or music genre, the composition of the movements appears simple or repetitive, and it becomes difficult to create movements in music that falls outside the scope of the genre. Therefore, the existing datasets were not suitable for generating K-pop dances containing various dance genres.

To collect K-pop data, we collected videos combining music and dance from easily accessible online video platforms such as YouTube and TikTok. By using the keyword ‘K-pop dance’ for searches, we gathered videos that captured in-the-wild dance images paired with audio. After retrieval, each dance was acquired sequentially using a multi-person tracking method [66]. Additionally, each dance was examined empirically to ensure that it had consistent joint values that did not change with other movements over time. We then performed two-dimensional pose estimation on the video using OpenPose [56] to obtain the human pose. In the case of OpenPose, it proved to be a suitable pose extraction method, especially in situations where frames were blurry or movements overlapped, as it effectively identified poses in such conditions. The coordinates were extracted separately for the body, face, hands, and feet, but we specifically used only 15 coordinates corresponding to key body parts

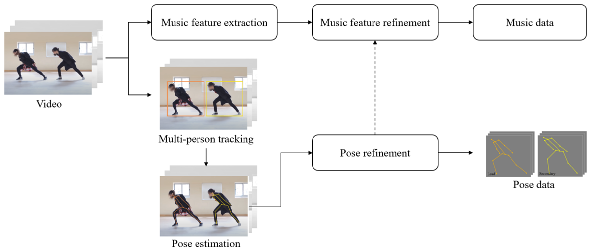

FIGURE 2. This figure depicts the data preprocessing process. We obtained audios and images from the in-the-wild videos. We adopted both multi-person tracking algorithm and pose extraction algorithm to extract motion information from the group dance videos. After, extracted information is refined to cope with false detections. The time interval extracting music features is adjusted to match that of the pose data and saved them as individual paired data.

from the 25 coordinates of two-dimensional torso values. This simplification was done with the purpose of reducing the complexity of the model while preserving the motion information of the data we are trying to learn as much as possible.

However, as OpenPose alone generates incomplete and inaccurate data, we employed several refining methods. Firstly, we performed corrections for the core coordinates extracted by OpenPose. If a coordinate had an inaccurate value, we compared it with the frames before and after, filling in empty frames and interpolating for values to ensure the smooth connection of dance sequences through coordinate values. Secondly, frames with incorrect values are refined in the audio sequence and dance sequence. For various reasons, extracted audio feature frames or dance frames may contain empty or incorrect values. For example, there are cases where two dancers overlap in the video and pose estimation was not performed correctly, or there is music but the corresponding dance video is lost for editing reasons. In the former case, an interpolate value is found that naturally connects the values in the previous and subsequent frames, while maintaining the distance between each joint to prevent an unrealistic human form. In the latter case, since there is absolutely no information about one modality, the corresponding frame with a value in another modality is also removed from the data. Lastly, in videos uploaded by individuals, the aspect ratio of the dancers included in the video is often adjusted for various reasons, so a process of normalizing is necessary. To do this, we calculate the average aspect ratio of dancers across all videos included in the dataset. Then, the proportions of people in all images are corrected to match the calculated aspect ratio.

In terms of audio data, we extracted it from videos using Librosa [55] and ffmpeg. The videos were provided by various uploaders on YouTube and TikTok, leading to variations in both video quality and audio recording methods. In particular, regarding audio recording, recorders embedded in different camera models show different characteristics, the quality and the volume of recordings are not consistent. Therefore, normalization methods are required to adjust and standardize these variations. We used ffmpeg to ensure consistency by adjusting basic parameters including the sampling rate, bitrate and so on, and then used Librosa to various extract audio features. The sampling rate was set at 15,360Hz, bitrate at 192kbps, and hop-size at 1024 to align the audio and video frames. The extracted audio features included mel-frequency cepstral coefficients (MFCC), MFCC delta, constant-Q chromagram, tempogram, onset strength. We concatenated all these features into a single vector, associating it with the dance features.

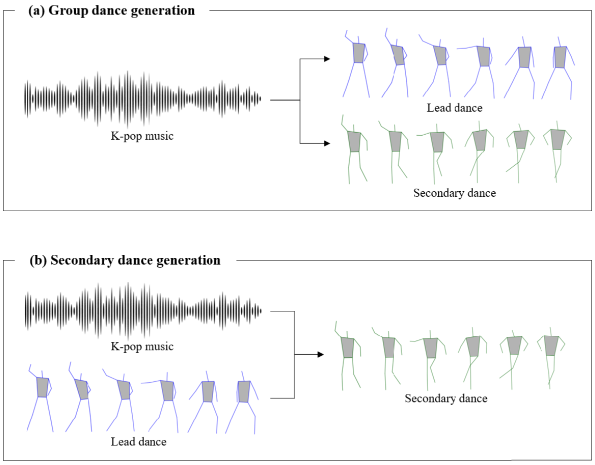

FIGURE 3. This figure presents two types of proposed synchronous dance generation methods. The group dance generation generates lead and secondary dances simultaneously corresponding to the input music, whereas the secondary dance generation generates only secondary dance from music and lead dance inputs.

As a result, 320 videos were collected, and all videos were divided into 1-minute long videos for learning. All video chunks less than 1 minute in length were excluded from the dataset. Finally, 950 one-minute videos containing audio and dance images were obtained, equivalent to a total of 15.8 hours. For comparison, the AIST++ dataset contains a total of 5.2 hours of data [37]. To maintain consistent quality, we scaled the resolution to 720p and set the frame rate to 30 FPS. We divided the dataset into training and testing, and randomly selected 900 video clips for training and 50 video clips for testing.

IV. METHOD

The *seq2seq* models can be broadly classified into two types. One is to perform sequence generation in an autoregressive manner, and the other is to perform in a non-autoregressive manner. In this paper, we propose an autoregressive method that can take into account the temporal characteristics and correlation between dance and input music when generating dance movements. The proposed model is designed to be able to perform both of the following two tasks: (a) 'GDG' that creates multiple dances simultaneously from music, (b) 'SDG' that creates partner dances that match the music and the lead dancer's dance. To this end, this chapter covers not only the structural aspects of the proposed model, but also the learning and inference methods.

A. OBJECTIVE

Listed music features are extracted from the raw audio signal. The dance motion sequences used for training are uniformly sampled at a frame rate of N_f frames per second. The length of the music chunk is determined at the same interval as the motion frames, and during audio feature extraction, the hop length is set to be the same as the frame length. If the sampling rate of the music is denoted as sr, then there are sr/N_f samples included in each hop length. Based on this hop length, all audio features can be extracted as follows:

$$M_i = \{m_i(1), m_i(2), \ldots, m_i(T)\}, \quad i = \{0, \ldots, N_{\text{feat}}\} \quad (1)$$

N_{feat} represents the number of types of extracted audio features, T is the number of music chunks determined by the hop length, therefore the audio feature extraction interval has the same value as the time between frames of motion.

The objective of this study is to find an autoregressive model $g(\cdot)$ that, given the audio feature sequence M as described above, predicts the dance motion sequences for the lead dance ($d_1 = \{d_1(1), d_1(2), \ldots, d_1(T)\}$) and the secondary dance ($d_2 = \{d_2(1), d_2(2), \ldots, d_2(T)\}$). In other words, we seek to find the model deriving D from M, in other words, $g(M) = D = \{d_1, d_2\}$.

The proposed model g is composed of three modules: a music feature encoder that encodes the music feature sequences M into music feature encodings, a dance encoder that encodes the dance motion at the previous timestep, and a dance decoder that utilizes the results from both modules to predict the dance motion at each timestep. We designed the details of each module taking into account the fact that the music feature sequence is considerably longer than a typical sequence, that there are two pairs of motion sequences to be predicted, and that it must reflect the characteristics of K-pop.

B. MODEL ARCHITECTURE

1) MUSIC FEATURE ENCODER

We first turn the raw music signal into music feature sequences M as described above. To facilitate convergence during the training phase and enhance generalization, this feature vector is first passed through a prenet. The prenet plays a role in improving generalization through non-linear transformations, dropout. In this model, the prenet is composed of a combination of linear transformations, ReLU activation functions, and dropout [50].

Self-attention mechanisms are known for being suitable structures to extract encodings by considering context in long sequences. They are particularly effective in encoding multi-scale hierarchical structures, such as music features [72], making them well-suited as the fundamental structure of an encoder. However, due to the nature of the self-attention mechanism, in order to perform encoding, attention must be calculated between every element and every other element in the sequence, resulting in a computational complexity of $O(n^2)$ [54]. It means, there is a significant challenge in terms of computational complexity for the music feature sequence due to its notably longer length compared to a typical sequence [30]. Various studies have been conducted to enable transformer to operate on long sequences, and several solutions have been proposed.

In the Longformer [24], several attention patterns are suggested, and the proposed model utilizes a combination of sliding window and global window attention patterns. Firstly, the sliding window effectively enables the encoding of local characteristics in music features. K-nearest neighbor elements are exploited to calculate the attention, rather than using the entire sequence. As a result, while the context range is limited, we can encode the continuous, temporal characteristics of the music using low computational complexity.

However, K-pop music, compared to other genres, is significantly influenced by the impact of the chorus, and the dance in K-pop music is also greatly influenced by the music's chorus. During the chorus, all dancers tend to perform similar dance, and the overall style of the dance is heavily influenced by the musical features of the chorus in the music. In the process of encoding the musical feature sequence using only a sliding window, it is not possible to fully capture the unique characteristics of the music, especially its prominent features in the chorus. To address this, it is necessary to estimate the chorus region in the music and ensure that the musical features from the estimated region are consistently included in the calculation of attention for all music samples. This idea can be implemented by introducing the concept of a global window. Then, musical feature encodings can be extracted that consider both the temporal characteristics of the music and the unique features of K-pop by using a kernel that combines sliding window and global window concepts.

To extract the chorus region from the music signal, we employ a self-attention convolution(SA-conv) based automatic chorus detection algorithm [67]. One of the representative methods for chorus detection involves using the self-similarity matrix (SSM). The SSM visualizes the similarity between a specific music chunk and chunks at different times in a two-dimensional matrix, allowing us to identify how similar segments are repeated across the entire music [71], [76]. However, this method relies on handcrafted rules to determine the chorus region ultimately, making it less robust in terms of its reliance on rule-based judgments of where the chorus region is. The SA-conv approach performs encoding through a multi-scale network that considers the hierarchical structure of music and passes the encoding through SA-conv to derive probability curves representing the presence of the chorus. Ultimately, the chorus region C is determined by applying adaptive thresholding to the curve. As explained above, the music chunks included in C are then utilized for the global window, and concurrently applied with the sliding window for attention calculation. In other words, this means that the amount of attention calculation varies depending on the length of the section determined as a chorus, and it is essential to adjust the length of this section to a computable level. By setting the adaptive threshold

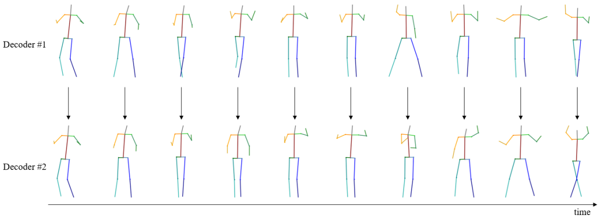

FIGURE 4. This figure is a sample skeleton resulting from the proposed model and shows the roles of the two decoders. Since the intermediate outputs of decoder #1 is used during the operation of decoder #2, it can be seen that the secondary dance is synchronized with the lead dance. On the other hand, the lead dance generated from decoder #1 generates dance motions that matches the music without any other restrictions.

to an appropriate level, it does not simply ensure that the entire chorus section is well detected, but only the minimum essential section is set as the chorus section, so that all essential information is included when calculating attention, but the amount of calculation is minimized.

2) DANCE MOTION ENCODER AND DECODER

We aimed to base our model on an autoregressive structure, encoding dance motions at each timestamp and predicting the next dance motion at the next timestamp. To achieve this, we propose a structure that efficiently predicts the next dance motion from previous dance motions while maintaining temporal context. Like the music encoder, it has the effect of improving generalization during training by passing through Prenet first.

Subsequently, a recurrent model is employed to encode the joint coordinates of the input motion. The recurrent module well expresses the concept of an autoregressive process and has a chain structure that can work properly. All human movements are very closely related to previous temporal movements, and are similar to the operation method of the recurrent model in that they move with a large context [30, 84]. Therefore, for encoding and decoding human motion, we utilize a recurrent model, specifically employing the GRU structure, which is simple yet capable of maintaining long-term memory [63]. The encoded motion information, concatenated with the musical feature encoding m_t at the corresponding timestamp, can be used to predict the next motion through the motion decoder. In equation form, it is expressed as follows:

$$h_{\mathrm{enc},1}(t) = \mathrm{GRU}\left(\mathrm{prenet}(d_1(t-1)), h_{\mathrm{enc},1}(t-1)\right) \tag{2}$$

$$h_{\mathrm{dec},1}(t) = \mathrm{GRU}\left(h_{\mathrm{enc},1}(t) \oplus m_t, h_{\mathrm{dec},1}(t-1)\right) \tag{3}$$

$$\hat{d_1}(t) = h_{\mathrm{dec},1}(t) \cdot W_1 + h_1 \tag{4}$$

h_{enc} represents the hidden state of the motion encoder, h_{dec} represents the hidden state of the motion decoder, W and h denote the weight and bias of the linear layer, respectively. The symbol $\oplus$ represents the concatenation operation. Here, $h_{\mathrm{enc}}(t)$ itself represents the predicted motion at timestamp t, and the linear layer serves the role of further refining it. The number 1 indicates that these are all encoding/decoding modules for the lead dance. Likewise, the next number 2 represents the module associated with the secondary dance. Subsequently, in section c.1, the postnet elaborates on this in detail.

The motion encoders of the lead dance and secondary dance share the same structure, but there are some differences in the motion decoder. This is because when predicting the movements of a secondary dance, the movement information of the lead dance must be incorporated. For the secondary dance's motion decoder, in addition to music feature encoding, we also concatenate the hidden state $h_{\mathrm{dec}}(t)$ of the lead dance's motion decoder. Then, it is done to integrate the lead dance's motion information during decoding. The formula is as follows:

$$h_{\mathrm{enc},2}(t) = \mathrm{GRU}\left(\mathrm{prenet}(d_2(t-1)), h_{\mathrm{enc},2}(t-1)\right) \tag{5}$$

$$h_{\mathrm{dec},2}(t) = \mathrm{GRU}\left(h_{\mathrm{enc},2}(t) \oplus m_t \oplus h_{\mathrm{dec},1}(t), h_{\mathrm{dec},2}(t-1)\right) \tag{6}$$

$$\hat{d_2}(t) = h_{\mathrm{dec},2}(t) \cdot W_2 + h_2 \tag{7}$$

C. CONVERGENCE ENHANCING TRAINING METHODS

The model is fundamentally trained to infer the dance motion sequences d_1 for the lead dance and d_2 for the secondary dance simultaneously from the music feature encoding. The proposed model performs as both a model predicting group dance motion sequences that harmonize with the music and a model predicting the dance sequence for the secondary dance that complements the dance of the lead dance when given both the dance and the music. To better accomplish the objectives of the proposed model during the training and inference of the dance generation model, several mechanisms have been incorporated.

1) POST-PROCESSING NETWORK

We predict joint coordinate values through the dance decoder of our model and then induce refinement of the decoding results by adding a simple post-processing network(postnet). In the proposed model, a straightforward linear layer is employed as the postnet. To ensure the postnet has an effect on refining the results, both the decoding results before and after the postnet are utilized in the loss function. Therefore, during training, the loss for a specific dance is represented as follows:

$$L_{\mathrm{recon}} = L(d_{gt}, \hat{d_{dec}}) + L(d_{gt}, \hat{d_{postnet}}) \tag{8}$$

The term $\hat{d_{dec}}$ represents the decoding result, and $\hat{d_{postnet}}$ is the result obtained by passing $\hat{d_{dec}}$ through the postnet. By using a loss that simultaneously compares both predicted values with the ground truth, we encourage the decoder to calculate values closer to the ground truth. At the same time, this setup guides the postnet to further refine the results, processing them to values closer to the ground truth.

2) MULTI-STEP LEARNING FOR SDG

As the dance encoder and decoder operate based on recurrent units, teacher forcing is inherently employed during training

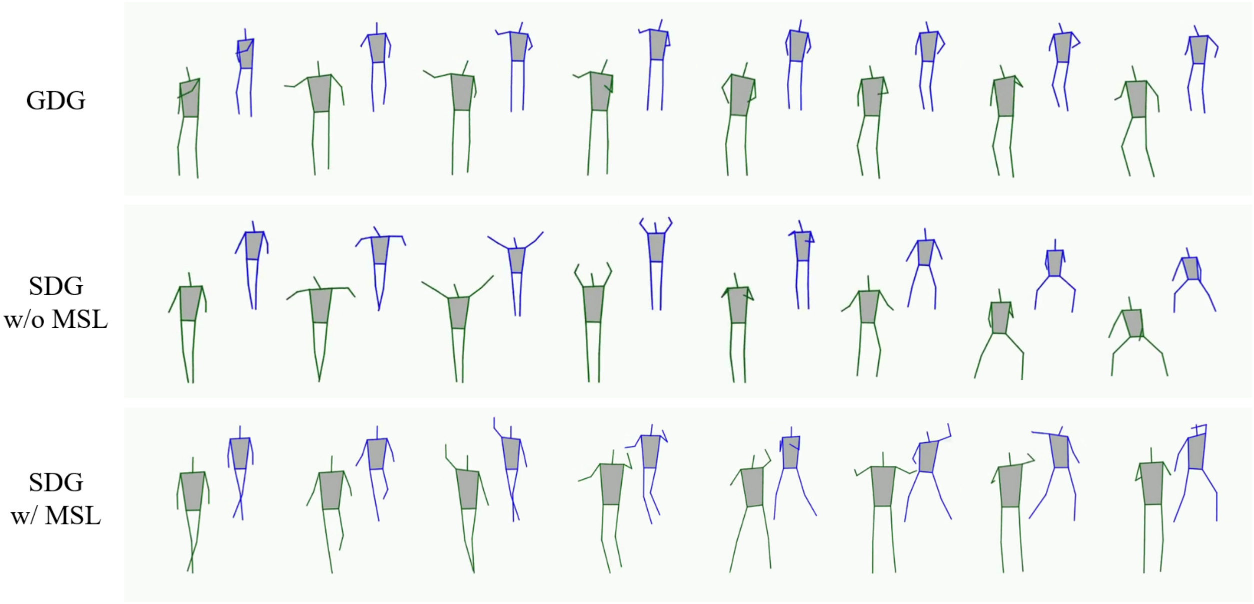

FIGURE 5. This figure shows example of dance sequences of GDG, SDG with MSL, and SDG without MSL, respectively, from top to bottom.

to prevent convergence problems caused by error propagation in the recurrent module. In other words, during training, the ground truth value of dance motion at each timestep t is used as input to predict the dance motion at the next timestep $(t + 1)$. In this training setup, the model is well-suited for learning objectives that involve predicting two dance motion sequences that harmonize with the music feature sequence. However, there is a tendency for the model not to align well with the objective of predicting the dance motion sequence for the secondary dancer that complements both the music feature sequence and the reference dance motion sequence input.

To address this, a multi-step teacher forcing approach is introduced. During training, up to a specific iteration n_{iter}, teacher forcing is applied to both the lead and secondary dance's dance motion learning. Beyond n_{iter} in the training phase, teacher forcing is applied only to the lead dance's dance motion learning, and for the secondary dance, the prediction from the previous timestep is used as input to predict the motion for the next timestep. Up to a certain iteration, both the lead and secondary dance's encoder/decoder are trained with teacher forcing to generate natural and music-harmonious dance movements. After determining that the model has reached a certain level of learning, starting from the n_{iter} iteration, the training is adapted to enable the model to generate partner movements that harmonize with the reference dance without explicit ground truth during actual inference. This training strategy allows the proposed model to be utilized for both objectives effectively.

V. EXPERIMENTS

A. SETTINGS

In the music feature encoder, the self-attention based encoder consists of a self-attention layer with 8 attention heads, and the hidden dimension is 1024. Three self-attention based encoding layers are stacked, denoted as $N_{fft} = 3$. The dimension of query, key and value in self-attention are all set to 64. In prenet, the dimension of the linear layer is 256 and the dropout ratio is 0.1. For the sliding window, the receptive field length was set to 100 samples. The prenet in the motion encoder is similar to the prenet in the encoder and has 256 dimensions. Additionally, the dropout rate is 0.1 and every GRU in the motion encoder and decoder has 512 cells. For training, we adopted the Adam optimizer with a batch size of 16 and a learning rate of 10^{-4}. The L2 loss is used as the reconstruction loss and is adopted for both before-postnet output and after-postnet output as mentioned in subsection ***postnet***. When training the model, multi-step learning(MSL) is adopted to increase the performance of the SDG, as detailed in subsection ***Multi-step learning for SDG***.

B. QUANTITATIVE EVALUATION

This study aims to confirm the motion diversity and motion quality of the generated dance and the dance-music relationship through evaluation. Thirty music clips were randomly selected from the K-pop test dataset. Single dance generation models such as 'Dance Revolution', 'Bailando', 'FACT' and 'EDGE' were used as baselines, and inference was performed using pre-trained models [30], [37], [59], [75]. At this time, in order to compare the baseline and the proposed model in same human form, the pose results of 'Bailando', 'FACT' and 'EDGE' were projected from three dimensions to two dimensions and the results of the proposed model were separated into single dance sequence.

1) FRÉCHET INCEPTION DISTANCE

Fréchet Inception Distance(FID) is used to evaluate the quality of a generated dance by measuring how close

TABLE 1. The quantitative evaluation comparing the proposed model with conventional single dance generation models in perspective of FID, DIV, and BAS. In order to compare equally with the baseline, the lead dance and the secondary dance in GDG, SDG, SDG without MSL are considered as individual dances. TF indicates performances of Transformer-based proposed model, presented in ablation study section. The results are better when FID is smaller, and DIV and BAS are larger.

	FID	DIV	BAS
Dance revolution	70.43	3.48	0.2040
Bailando	59.72	5.71	0.2134
FACT	43.58	5.98	0.2209
EDGE	34.75	7.44	0.2311
GDG-lead	23.26	8.06	0.2276
GDG-sec	25.80	8.29	0.2318
SDGw/oMSL-sec	14.57	8.20	**0.2380**
SDG-sec	**7.45**	**9.24**	0.2362
GDG-sec(TF)	23.64	8.43	0.2340
SDG-sec(TF)	7.22	9.22	0.2381

the distribution of dance results produced by a dance generation model is to the distribution of the actual original dance [70]. Following the approach of previous studies, we converted the results generated by the synchronous dance generation model into single dance sequences and compared their quality as single dances. We classified the separated single dance sequences into the lead dance (GDG-lead) and secondary dance(GDG-sec) of the GDG, the secondary dance(SDG-sec) of the SDG with MSL, and the secondary dance(SDGw/oMSL-sec) of the SDG without MSL according to the model's name and whether the dance was lead or secondary. The SDG without MSL is only used in evaluations to see the effectiveness of MSL. The FID was then calculated by comparing it to the ground truth of the pose data used.

Table 1 shows the results of quantative evaluations. Quantitative evaluation of the proposed methods, including GDG, SDG, and SDG without MSL, shows better results than the baseline methods. This proves that the dance generated by the proposed methods has statistically more similar characteristics to the actual K-pop dance than the dance generated by the baseline methods and produces high-quality dances. In addition, as mentioned earlier, K-pop group dances sometimes perform similar dances or different dances. GDG usually produces pair dances that harmonize with the input music, and dancers often dance similar to each other.

However, the pair dances of GDG differs from the characteristics of actual K-pop dances, which can act as a relative disadvantage in quantitative evaluation. The distribution of SDG without MSL dances was similar to the distribution of actual K-pop dances than GDG, but not better than SDG. This is because the secondary dance generated by SDG without MSL is very similar to the lead dance of SDG without MSL. On the other hand, SDG-sec generates a dance that matches the presented lead dance, but it does not always perform a dance similar to the lead dance. Therefore, SDG tend to be closer to K-pop dances on average than GDG and SDG without MSL. Furthermore, SDG generates more life-like dances because the lead dance cue enters the dance generation process. Due to these complex factors, SDG shows significantly better scores than other proposed methods. GDG-lead and GDG-sec showed similar results, which suggests that, as mentioned earlier, the two dances are composed of similar dances.

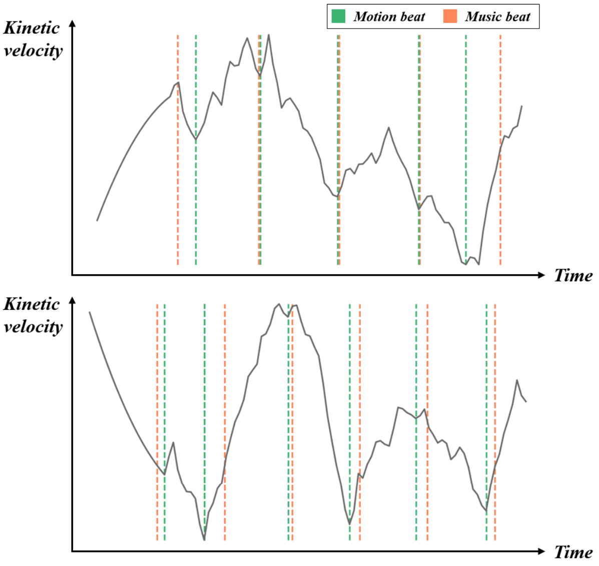

FIGURE 6. This figure visually depicts dance motion beats and music beats. Each green line represents the motion beat corresponding to the dance motion, and the orange line represents the music beat. The closer the positions of two beats are to each other, the higher the BAS result.

2) DIVERSITY

To assess how diverse the dance results produced by this model are in terms of motion, we decided to calculate motion diversity using the methodology used in previous studies [37]. Similar to FID, we wanted to see the motion diversity of isolated single dance sequences in the model results. This was done by calculating the average Euclidean distance in the feature space for all pairs of tasks tested. In terms of diversity, as shown in Table 1, the proposed methods show higher diversity than the baseline methods. This means that the proposed model learned with various K-pop dance datasets can generate more diverse K-pop dance sets compared to baseline models. In particular, SDG can generate dances based on more information than GDG because it receives an additional cue for lead dance, and it also performs higher than GDG in terms of the diversity of the generated dances. Furthermore, the reason why SDG shows better results than SDG without MSL is because the similarity between the lead dances of SDGw/oMSL and the SDGw/oMSL-sec is greater than the similarity between the lead dances of SDG and the SDG-sec.

3) BEAT ALIGN SCORE

Beat Align Score(BAS) evaluates how well the beats extracted from the music match the beats of the dance movements [59]. For this purpose, onset strength, a music feature representing the beat, was extracted using librosa and used as the music beat. First, the number of frames was adjusted so that the dance beat and the music beat

TABLE 2. **The qualitative evaluation results. Three proposed models are compared: GDG, SDG, and SDG without MSL. 25 evaluators assessed 9 videos on the Likert scale in perspectives of harmony of dance and music, harmony between dances, naturalness of dance, similarity between dances, and dance-like movements or not. Since the proposed scheme generates both lead and secondary dances, the questionnaire was designed to evaluate individual dance and group dance performance separately in some questions.**

	harmony with music		harmony btw. dances	similarity btw. dances	naturalness	dance-like movements	
	individual	group				individual	group
GDG	**4.13**	4.16	4.01	4.15	4.20	**4.21**	4.35
SDG w/o MSL	4.11	**4.24**	**4.12**	**4.23**	**4.41**	4.17	**4.36**
SDG w/ MSL	3.91	4.05	3.73	3.89	4.08	4.12	4.28

could be compared equally in time, and then the motion beat was created using a method of measuring changes in the joint direction and speed of the dance movement sequence. Then, the BAS for the entire sequence was calculated by comparing the motion beats with the beats extracted from the music. As shown in Table 1, the BAS of the proposed methods is mostly better than that of the baseline methods. As an exception, the BAS of EDGE is greater than that of GDG-lead, but is lower than the results of the remaining proposed models. This means that most of the proposed model generates dance with the motion beat more suitable for K-pop music. The motion beat is a part where the tendency of the kinetic velocity changes and the dance movement changes rapidly, and is considered the starting point of the detailed motion. This means that the motion beat of dance generated by the proposed model is distributed at a similar time to the beat position of K-pop music, and it means that a dance that is more coherent with K-pop music in time can be generated. Additionally, in the BAS index, it can be seen that the difference in scores between the proposed methods is not significant, which means that proposed methods sufficiently generate dances suitable for K-pop music compared to the baseline methods.

C. QUALITATIVE EVALUATION

In order to evaluate the aspect of group dance and the artistic aspect of dance, a survey was conducted on 25 evaluators who had no prior knowledge of the dance routine of K-pop music. Participants watched a video featuring two dancers and then rated the video on several items. Videos were randomly selected from a pool of generated dance sequences. Nine 10-second videos were evaluated on the Likert scale for the following items: ‘harmony with music’, ‘harmony between dances’, ‘similarity between dances’, ‘naturalness’, and ‘dance-like movements’. In the case of the ‘harmony with music’ and ‘dance-like movements’, the evaluation may differ between when the dance is presented as a group dance and when it is presented as an individual dance. Therefore, for the above two items, the individual sections that showed and evaluated only the generated dance and the group sections that showed and evaluated the generated secondary dance and lead dance simultaneously were presented separately. Additionally, to confirm the effectiveness of MSL applied to SDG, evaluators looked at the results of SDG without MSL along with GDG and SDG. All videos used for evaluation were randomly sampled from the 1-minute dance sequences generated using the K-pop test set.

The evaluators gave answers to the five items presented above, and the results are as follows. In the items of harmony with music and dance-like movements, The group tends to show good scores on average. This is presumed to be because the secondary dance generated by the proposed methods harmonizes well with the lead dance and shows a higher level of completion when presented together. In addition, GDG and SDG without MSL tend to generally have high scores in most items. This is because the secondary dance generated by the two proposal methods generally produces a dance similar to the lead dance, and people are generally considered to feel more dance-like and harmonious when watching a unified group dance. When evaluating the similarity between dances, the evaluators gave high scores to GDG and SDG without MSL, indicating that the second dance generated by the two proposed methods performed a similar dance to the lead dance. The item that evaluates the degree to which a pair of dances match is similarity between dances, and the item that asks how coherent the two dances are is harmony between dances. Nevertheless, the evaluators gave similar scores for similarity between dances and harmony between dances. This proves that the more similar the dances are, the more people consider them to be harmonious. The evaluators evaluated the naturalness and smoothness of the generated dance through the ‘naturalness’ item. All three proposed methods showed high naturalness. However, SDG without MSL received a particularly higher evaluation, because the evaluation reflects the advantage of SDG in generating the dance using more information, i.e. the lead dance cue, and also because the secondary dance generated without MSL is more similar to lead dance.

D. ABLATION STUDY

1) EFFECTIVENESS OF THE CHORUS SECTION AS A GLOBAL WINDOW

In this section, we evaluate the effectiveness of a method for calculating attention using the chorus section in music as a global window. For assessment, an ablation study was conducted comparing the results obtained when applying sliding window with global attention and when not applying global attention to sliding window. Evaluators look at choreography created using two parts within one piece of music and evaluate the consistency of the two choreography. Among the two music chunks, the first is selected as the part corresponding to the chorus, and the second is selected as the non-chorus part after the previously presented chorus. Choreography that matches the two music chunks is created in two ways: when attention is calculated using only a sliding window, and when attention is calculated up to global attention. Two pairs of choreography are presented to the evaluator and asked which of the two is more consistent. The results showed a preference rating of 0.72 when both sliding window and global attention were used, compared to 0.28 when only sliding window was used. This shows that using the samples of the chorus part as a global window for music encoding rather than using only a sliding window is a better way to encode the style of the entire song in K-pop. No matter which part of the song is used to generate the choreography, it can be seen that when global attention is used, it shows a high correlation with the choreography generated in other parts.

2) PERFORMANCE COMPARISON ACCORDING TO DECODING MODULE CHANGES

The purpose of this section is to analyze the effects of replacing a GRU-based decoder with a Transformer-based decoder in the proposed model. This section will detail both the benefits and drawbacks of this architectural change, examining aspects such as model performance, computational efficiency, and handling of long-term dependencies.

For comparison, we adopted a Transformer-based decoder composed of masked multi-head self attention, multi-head attention and position-wise feed-forward neural network. For secondary dance generation, one more decoder with the same structure was designed, and the normalized value of the multi-head attention output of the first decoder was applied as the key and value of the multi-head attention of the second decoder. As a result, a double decoder structure like the proposed technique was implemented using a transformer-based decoder. Transformer-based decoder is composed of a stack of 6 identical layers and all dimensions are set to 512.

One of the most significant advantages of using a Transformer decoder is its superior ability to capture long-term dependencies. As shown in Table 1, the Transformer-based model outperforms the GRU-based model in perspective of FID and BAS. This means that the Transformer-based model can generate a dance with a motion velocity similar to the beat of music, which means that the temporal, hierarchical structure of the created dance is closer to a real-world dance. In other words, when a dance motion frame is generated autoregressively, it has a close relation with previously created frames, which means that the Transformer-based model has higher long-term dependency. This is achieved through the self-attention mechanism, which allows the model to consider all positions in the input sequence simultaneously, as opposed to GRU's sequential processing. On the other word, GRU loses a lot of long-term context because it considers only the latest prediction, but the Transformer-based decoder shows better long-term dependency because it uses all past predictions for calculation. This capability translates to smoother and more coherent dance motion estimation that better align with the temporal structure of the whole dance.

GRU-based proposed model takes about 23 minutes per epoch while Transformer-based model takes about 35 minutes per epoch. Given a similar computational environment and dataset, the training time for a Transformer decoder can be significantly higher than a GRU due to the deeper architecture and more operations. For the inference process, it takes approximately 17 seconds to generate 10 seconds dance sequences while the Transformer-based model takes 28 seconds to generate the same length of dance sequence. It is required to perform as many decoding steps as the number of frames to generate, and this autoregressive nature slows down the inference time of both models.

Transformers generally perform better with larger datasets due to the extensive parameterization. For optimal performance, Transformers require comprehensive and extensive datasets, making data collection and preprocessing more critical. This reliance on larger datasets can be a limitation in scenarios where data is scarce. GRUs can often perform adequately with smaller datasets, benefiting from their simpler structure and fewer trainable parameters.

In the case of the K-pop dance dataset, the consistency between the data is low and various genres are included, so a relatively larger amount of data is required for training. When the size of the dataset is not large enough, a GRU-based architecture is more suitable considering training time and resources. But when the size of the dataset is sufficiently large, a Transformer-based architecture can produce better results. The amount of data used for training in this paper is not a large amount considering the nature of K-pop, so it can be seen that the performance indicators between the two models are not significantly different. If additional data is collected or a large open dataset is created in the future, the Transformer-based model is expected to show much higher performance.

Replacing a GRU-based decoder with a Transformer-based decoder in an autoregressive seq2seq dance generation model offers significant advantages in capturing long-term dependencies, improving training efficiency, and generating contextually coherent sequences. However, these benefits come at the cost of increased computational complexity, memory requirements, and dependency on larger datasets. Therefore, the choice between GRU and Transformer decoders should be informed by the specific requirements and limitations of the use case, balancing performance gains with computational practicality.

VI. CONCLUSION

In this paper, we propose the K-pop dance generation model in an autoregressive manner. We aimed to design a system designed to generate two-person synchronous dances based on K-pop music. To accomplish this, we collected various K-pop genre dance videos from online video platforms for training models. With collected videos, We trained a model that can perform both synchronous dance generation methodologies: SDG and GDG.

Additionally, we proposed learning tricks, *postnet* and *MSL*, to make the model better generates synchronous dances. Both quantitative and qualitative evaluations were conducted to evaluate the performance of the proposed model. From the quantitative evaluation,it was confirmed that the proposed method generates dance more suitable for K-pop music compared to conventional methods. A qualitative evaluation was conducted through user evaluation of various items to verify the coherence between the generated dances and the effectiveness of the proposed training method.

We still believe that there is room for improvement in this study. Because the purpose of the paper was different from other studies, it was difficult to use a 3D open dataset such as AIST++, so the research was conducted based on a 2D dataset. However, the method presented in this paper can be applied as is even when using 3D datasets. Therefore, if it is possible to secure a large number of 3D K-pop group choreography datasets or 2D to 3D conversion algorithm based on the sophisticated depth estimation scheme, the proposed technique can be expanded to generate 3D dances.

REFERENCES

[1] G. Kassing, *History of Dance: An Interactive Arts Approach*. Champaign, IL, USA: Human Kinetics, 2007.

[2] G. Madison, F. Gouyon, F. Ullén, and K. Hörnström, "Modeling the tendency for music to induce movement in humans: First correlations with low-level audio descriptors across music genres," *J. Experim. Psychol., Human Perception Perform.*, vol. 37, no. 5, pp. 1578–1594, 2011.

[3] B. Burger, M. R. Thompson, G. Luck, S. Saarikallio, and P. Toiviainen, "Influences of rhythm- and timbre-related musical features on characteristics of music-induced movement," *Frontiers Psychol.*, vol. 4, p. 183, Jan. 2013.

[4] E. G. Schellenberg, A. M. Krysciak, and R. J. Campbell, "Perceiving emotion in melody: Interactive effects of pitch and rhythm," *Music Perception*, vol. 18, no. 2, pp. 155–171, Dec. 2000.

[5] H. Thomas, *The Body, Dance and Cultural Theory*. New York, NY, USA: Palgrave Macmillan, 2003.

[6] D. Kim, D.-H. Kim, and K.-C. Kwak, "Classification of K-pop dance movements based on skeleton information obtained by a Kinect sensor," *Sensors*, vol. 17, no. 6, p. 1261, Jun. 2017.

[7] M. Lee, K. Lee, and J. Park, "Music similarity-based approach to generating dance motion sequence," *Multimedia Tools Appl.*, vol. 62, no. 3, pp. 895–912, Feb. 2013.

[8] S. Gentry and E. Feron, "Modeling musically meaningful choreography," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, vol. 4, 2004, pp. 3880–3885.

[9] F. Ofli, Y. Demir, Y. Yemez, E. Erzin, A. M. Tekalp, K. Balcı, İ. KızoĞlu, L. Akarun, C. Canton-Ferrer, J. Tilmanne, E. Bozkurt, and A. T. Erdem, "An audio-driven dancing avatar," *J. Multimodal User Interfaces*, vol. 2, no. 2, pp. 93–103, Sep. 2008.

[10] W. Zhuang, C. Wang, J. Chai, Y. Wang, M. Shao, and S. Xia, "Music2Dance: DanceNet for music-driven dance generation," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 18, no. 2, pp. 1–21, May 2022.

[11] J. Lee, S. Kim, and K. Lee, "Listen to dance: Music-driven choreography generation using autoregressive encoder–decoder network," 2018, *arXiv:1811.00818*.

[12] G. Sun, Y. Wong, Z. Cheng, M. S. Kankanhalli, W. Geng, and X. Li, "DeepDance: Music-to-dance motion choreography with adversarial learning," *IEEE Trans. Multimedia*, vol. 23, pp. 497–509, 2021.

[13] B. Wallace, C. P. Martin, J. Torresen, and K. Nymoen, "Towards movement generation with audio features," 2020, *arXiv:2011.13453*.

[14] X. Ren, H. Li, Z. Huang, and Q. Chen, "Self-supervised dance video synthesis conditioned on music," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 46–54.

[15] T. Tang, J. Jia, and H. Mao, "Dance with melody: An LSTM-autoencoder approach to music-oriented dance synthesis," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 1598–1606.

[16] A. Aristidou, A. Yiannakidis, K. Aberman, D. Cohen-Or, A. Shamir, and Y. Chrysanthou, "Rhythm is a dancer: Music-driven motion synthesis with global structure," *IEEE Trans. Vis. Comput. Graphics*, vol. 29, no. 8, pp. 3519–3534, 2022.

[17] O. Alemi, J. Françoise, and P. Pasquier, "GrooveNet: Real-time music-driven dance movement generation using artificial neural networks," *Networks*, vol. 8, no. 17, p. 26, 2017.

[18] N. Yalta, S. Watanabe, K. Nakadai, and T. Ogata, "Weakly-supervised deep recurrent neural networks for basic dance step generation," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.

[19] W. Zhuang, Y. Wang, J. Robinson, C. Wang, M. Shao, Y. Fu, and S. Xia, "Towards 3D dance motion synthesis and control," 2020, *arXiv:2006.05743*.

[20] Z. Wang, J. Jia, H. Wu, J. Xing, J. Cai, F. Meng, G. Chen, and Y. Wang, "GroupDancer: Music to multi-people dance synthesis with style collaboration," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 1138–1146.

[21] X. Zhang, Y. Xu, S. Yang, L. Gao, and H. Sun, "Dance generation with style embedding: Learning and transferring latent representations of dance styles," 2021, *arXiv:2104.14802*.

[22] S. Wu, S. Lu, and L. Cheng, "Music-to-dance generation with optimal transport," 2021, *arXiv:2112.01806*.

[23] Z. Ye, H. Wu, J. Jia, Y. Bu, W. Chen, F. Meng, and Y. Wang, "ChoreoNet: Towards music to dance synthesis with choreographic action unit," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 744–752.

[24] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," 2020, *arXiv:2004.05150*.

[25] Y.-F. Huang and W.-D. Liu, "Choreography cGAN: Generating dances with music beats using conditional generative adversarial networks," *Neural Comput. Appl.*, vol. 33, no. 16, pp. 9817–9833, Aug. 2021.

[26] H.-Y. Lee, X. Yang, M.-Y. Liu, T.-C. Wang, Y.-D. Lu, M.-H. Yang, and J. Kautz, "Dancing to music," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[27] A. Bhattacharya, M. Paranjape, U. Bhattacharya, and A. Bera, "DanceAnyWay: Synthesizing beat-guided 3D dances with randomized temporal contrastive learning," 2023, *arXiv:2303.03870*.

[28] G. Valle-Pérez, G. E. Henter, J. Beskow, A. Holzapfel, P.-Y. Oudeyer, and S. Alexanderson, "Transflower: Probabilistic autoregressive dance generation with multimodal attention," *ACM Trans. Graph.*, vol. 40, no. 6, pp. 1–14, Dec. 2021.

[29] J. Li, Y. Yin, H. Chu, Y. Zhou, T. Wang, S. Fidler, and H. Li, "Learning to generate diverse dance motions with transformer," 2020, *arXiv:2008.08171*.

[30] R. Huang, H. Hu, W. Wu, K. Sawada, M. Zhang, and D. Jiang, "Dance revolution: Long-term dance generation with music via curriculum learning," 2020, *arXiv:2006.06119*.

[31] J. Kim, H. Oh, S. Kim, H. Tong, and S. Lee, "A brand new dance partner: Music-conditioned pluralistic dancing controlled by multiple dance genres," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3480–3490.

[32] B. Li, Y. Zhao, Z. Shi, and L. Sheng, "DanceFormer: Music conditioned 3D dance generation with parametric motion transformer," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 2, pp. 1272–1279.

[33] N. Le, T. Pham, T. Do, E. Tjiputra, Q. D. Tran, and A. Nguyen, "Music-driven group choreography," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 8673–8682.

[34] B. Feng, T. Ao, Z. Liu, W. Ju, L. Liu, and M. Zhang, "Robust dancer: Long-term 3D dance synthesis using unpaired data," 2023, *arXiv:2303.16856*.

[35] K. Gong, D. Lian, H. Chang, C. Guo, Z. Jiang, X. Zuo, M. B. Mi, and X. Wang, "TM2D: Bimodality driven 3D dance generation via music-text integration," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 9942–9952.

[36] M. Okamura, N. Kondo, T. Fushimi, M. Sakamoto, and Y. Ochiai, "Dance generation by sound symbolic words," 2023, *arXiv:2306.03646*.

[37] R. Li, S. Yang, D. A. Ross, and A. Kanazawa, "AI choreographer: Music conditioned 3D dance generation with AIST++," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13381–13392.

[38] R. Li, J. Zhao, Y. Zhang, M. Su, Z. Ren, H. Zhang, Y. Tang, and X. Li, "FineDance: A fine-grained choreography dataset for 3D full body dance generation," 2022, *arXiv:2212.03741*.

[39] S. Yao, M. Sun, B. Li, F. Yang, J. Wang, and R. Zhang, "Dance with you: The diversity controllable dancer generation via diffusion models," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023, pp. 8504–8514.

[40] S. Yang, Z. Yang, and Z. Wang, "LongDanceDiff: Long-term dance generation with conditional diffusion model," 2023, *arXiv:2308.11945*.

[41] Q. Qi, L. Zhuo, A. Zhang, Y. Liao, F. Fang, S. Liu, and S. Yan, "DiffDance: Cascaded human motion diffusion model for dance generation," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023, pp. 1374–1382.

[42] J. P. Ferreira, T. M. Coutinho, T. L. Gomes, J. F. Neto, R. Azevedo, R. Martins, and E. R. Nascimento, "Learning to dance: A graph convolutional adversarial network to generate realistic dance motions from audio," *Comput. Graph.*, vol. 94, pp. 11–21, Feb. 2021.

[43] H. Zhuang, S. Lei, L. Xiao, W. Li, L. Chen, S. Yang, Z. Wu, S. Kang, and H. Meng, "GTN-bailando: Genre consistent long-term 3D dance generation based on pre-trained genre token network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[44] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2021, pp. 8748–8763.

[45] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," 2021, *arXiv:2104.13921*.

[46] J. Xu, S. De Mello, S. Liu, W. Byeon, T. Breuel, J. Kautz, and X. Wang, "GroupViT: Semantic segmentation emerges from text supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18113–18123.

[47] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10674–10685.

[48] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2021, pp. 8821–8831.

[49] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," 2016, *arXiv:1609.03499*.

[50] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," 2017, *arXiv:1703.10135*.

[51] G. Aggarwal and D. Parikh, "Dance2Music: Automatic dance-driven music generation," 2021, *arXiv:2107.06252*.

[52] S. Wu, Z. Liu, S. Lu, and L. Cheng, "Dual learning music composition and dance choreography," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 3746–3754.

[53] H. Zhu, M.-D. Luo, R. Wang, A.-H. Zheng, and R. He, "Deep audio-visual learning: A survey," *Int. J. Autom. Comput.*, vol. 18, no. 3, pp. 351–376, Jun. 2021.

[54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[55] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in Python," in *Proc. 14th Python Sci. Conf.*, vol. 8, 2015, pp. 18–25.

[56] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1302–1310.

[57] B. Han, Y. Li, Y. Shen, Y. Ren, and F. Han, "Dance2MIDI: Dance-driven multi-instruments music generation," 2023, *arXiv:2301.09080*.

[58] B. Qin, W. Ye, Q. Yu, S. Tang, and Y. Zhuang, "Dancing avatar: Pose and text-guided human motion videos synthesis with image diffusion model," 2023, *arXiv:2308.07749*.

[59] L. Siyao, W. Yu, T. Gu, C. Lin, Q. Wang, C. Qian, C. C. Loy, and Z. Liu, "Bailando: 3D dance generation by actor-critic GPT with choreographic memory," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11040–11049.

[60] S. Tsuchida, S. Fukayama, M. Hamasaki, and M. Goto, "AIST dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing," in *Proc. ISMIR*, 2019, vol. 1, no. 5, pp. 1–10.

[61] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3156–3164.

[62] T. Kim, S. I. Park, and S. Y. Shin, "Rhythmic-motion synthesis based on motion-beat analysis," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 392–401, 2003.

[63] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," 2014, *arXiv:1406.1078*.

[64] Y. Cheng, R. Wang, Z. Pan, R. Feng, and Y. Zhang, "Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 3884–3892.

[65] R. Fan, S. Xu, and W. Geng, "Example-based automatic music-driven conventional dance motion synthesis," *IEEE Trans. Visualizat. Comput. Graph.*, vol. 18, no. 3, pp. 501–515, Apr. 2011.

[66] P. Sun, J. Cao, Y. Jiang, Z. Yuan, S. Bai, K. Kitani, and P. Luo, "DanceTrack: Multi-object tracking in uniform appearance and diverse motion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20961–20970.

[67] Q. He, X. Sun, Y. Yu, and W. Li, "Deepchorus: A hybrid model of multi-scale convolution and self-attention for chorus detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 411–415.

[68] L. Kovar, M. Gleicher, and F. Pighin, "Motion graphs," in *Proc. SIGGRAPH*, 2002, pp. 723–732.

[69] A. Safonova and J. K. Hodgins, "Construction and optimal search of interpolated motion graphs," in *Proc. SIGGRAPH*, 2007, p. 106.

[70] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. NeurIPS*, 2017.

[71] M. Cooper and J. Foote, "Automatic music summarization via similarity analysis," in *Proc. Int. Soc. Music Inf. Retr.*, 2002, pp. 81–85.

[72] C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, N. Shazeer, I. Simon, C. Hawthorne, A. M. Dai, M. D. Hoffman, M. Dinculescu, and D. Eck, "Music transformer," 2018, *arXiv:1809.04281*.

[73] A. M. Lamb, A. G. A. P. Goyal, Y. Zhang, S. Zhang, A. C. Courville, and Y. Bengio, "Professor Forcing: A new algorithm for training recurrent networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29. Curran Associates, 2016.

[74] M. Shi, K. Aberman, A. Aristidou, T. Komura, D. Lischinski, D. Cohen-Or, and B. Chen, "MotioNet: 3D human motion reconstruction from monocular video with skeleton consistency," *ACM Trans. Graph.*, vol. 40, no. 1, pp. 1–15, 2020.

[75] J. Tseng, R. Castellon, and K. Liu, "EDGE: Editable dance generation from music," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 448–458.

[76] S. Gao and H. Li, "Popular song summarization using chorus section detection from audio signal," in *Proc. IEEE 17th Int. Workshop Multimedia Signal Process. (MMSP)*, Oct. 2015, pp. 1–6.

SEOHYUN KIM received the M.S. degree in digital contents and information studies from the Music and Audio Research Group, Seoul National University, Seoul, South Korea, where she is currently pursuing the Ph.D. degree. Her research interests include deep learning, audio-based multimodal research, and natural language processing.

KYOGU LEE (Senior Member, IEEE) received the Ph.D. degree in computer-based music theory and acoustics from Stanford University, Stanford, CA, USA. Currently, he is a Professor with the Department of Intelligence and Information, Seoul National University, Seoul, South Korea, and is leading the Music and Audio Research Group. His research interests include signal processing and machine learning techniques applied to music and audio.

• • •