

DanceIt: Music-Inspired Dancing Video Synthesis

Xin Guo, Yifan Zhao[✉], and Jia Li[✉], *Senior Member, IEEE*

Abstract—Close your eyes and listen to music, one can easily imagine an actor dancing rhythmically along with the music. These dance movements are usually made up of dance movements you have seen before. In this paper, we propose to reproduce such an inherent capability of the human-being within a computer vision system. The proposed system consists of three modules. To explore the relationship between music and dance movements, we propose a cross-modal alignment module that focuses on dancing video clips, accompanied by pre-designed music, to learn a system that can judge the consistency between the visual features of pose sequences and the acoustic features of music. The learned model is then used in the imagination module to select a pose sequence for the given music. Such pose sequence selected from the music, however, is usually discontinuous. To solve this problem, in the spatial-temporal alignment module we develop a spatial alignment algorithm based on the tendency and periodicity of dance movements to predict dance movements between discontinuous fragments. In addition, the selected pose sequence is often misaligned with the music beat. To solve this problem, we further develop a temporal alignment algorithm to align the rhythm of music and dance. Finally, the processed pose sequence is used to synthesize realistic dancing videos in the imagination module. The generated dancing videos match the content and rhythm of the music. Experimental results and subjective evaluations show that the proposed approach can perform the function of generating promising dancing videos by inputting music.

Index Terms—Pose sequence, music, video synthesis, dancing videos.

I. INTRODUCTION

IN AN imagined room of the human-being, the rhythm of Hip-hop music can be incarnated as a dynamic street dancer, and the music of *Swan Lake* usually leads to an elegant ballet dancer. Such a capability of cross-modal imagination is also a crucial foundation of the so-called creativity that makes human brains different from a computer. With the rapid development of computing devices and machine learning

technologies, a concern naturally arises: is it possible to reproduce such an inherent capability of the human-being within a computer vision system?

Keeping this in our mind, the main motif of our paper is to generate the natural dancing sequences along with the musical beats (as shown in Fig. 1), namely audio-inspired video synthesis. However, the dancing sequences to generate and musical audio are substantially two different modalities of media and generating from one modal to another is not a one-to-one mapping problem. Thus there exist two burning challenges to be solved: 1) how to generate a dancing frame corresponding to the musical beats; 2) how to make the video sequence vivid and natural as human acting?

Actually, many research efforts [1]–[4] have been done to explore the generation task across multiple modalities. Among these works, the most explored transformation is between text and other media formats. For example, Chang and Zhang [1] proposed to generate text descriptions from given video clips. Tan *et al.* [3] and Li *et al.* [4] focused on a reverse problem that intended to synthesize realistic images from texts.

Beyond these efforts in other modalities, there are few works focusing on this less-explored task. Shlizerman *et al.* [5] proposed a deep regression model to predict the sequence of playing violin and piano along with the specific musical note. However, this work mainly focuses on the 3D localization of a fixed animation model, which still remains a gap from the video in the wild scenarios. As the most relevant research of our work, Tang *et al.* [6] developed an end-to-end LSTM-autoencoder structure to predict dancing pose sequences corresponding to the music melody. Lee *et al.* [7] proposed a decomposition-to-composition network to align the dancing skeletons with musical beats. However, generating visual content with audio in the frequency domain is indeed an ill-posed problem. There may exist enormous possible dancing movements when hearing the same music piece. In addition, these generated sequences also face difficulties in some complex movements, *e.g.*, backside or turning around. Therefore in this paper, we propose to solve this problem by building correspondences across different modalities rather than learning one-to-one bijections, while the latter one usually leads to over-fitting in network training.

To solve this drawback and the first challenge as well, we resort to pose skeletons as intermediaries to bridge the connection between the audio and visual content (illustrated in Fig. 1). Instead of estimating specific poses aligned with each musical beat, we propose to learn the coefficient relationship between these two modalities. Thus a deep metric learning approach is developed to learn this co-efficiency when adopting *pose fragment* as the basic unit. To further solve the second challenge, we further propose to align different fragments with

Manuscript received May 26, 2020; revised December 5, 2020 and January 18, 2021; accepted May 20, 2021. Date of publication June 8, 2021; date of current version June 16, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61922006, in part by the Baidu Academic Collaboration Program, and in part by the CAAI-Huawei MindSpore Open Fund. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Junsong Yuan. (Corresponding author: Jia Li.)

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

Xin Guo and Yifan Zhao are with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China.

Jia Li is with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China, and also with the Peng Cheng Laboratory, Shenzhen 518000, China (e-mail: jiali@buaa.edu.cn).

Digital Object Identifier 10.1109/TIP.2021.3086082

1941-0042 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.

See <https://www.ieee.org/publications/rights/index.html> for more information.

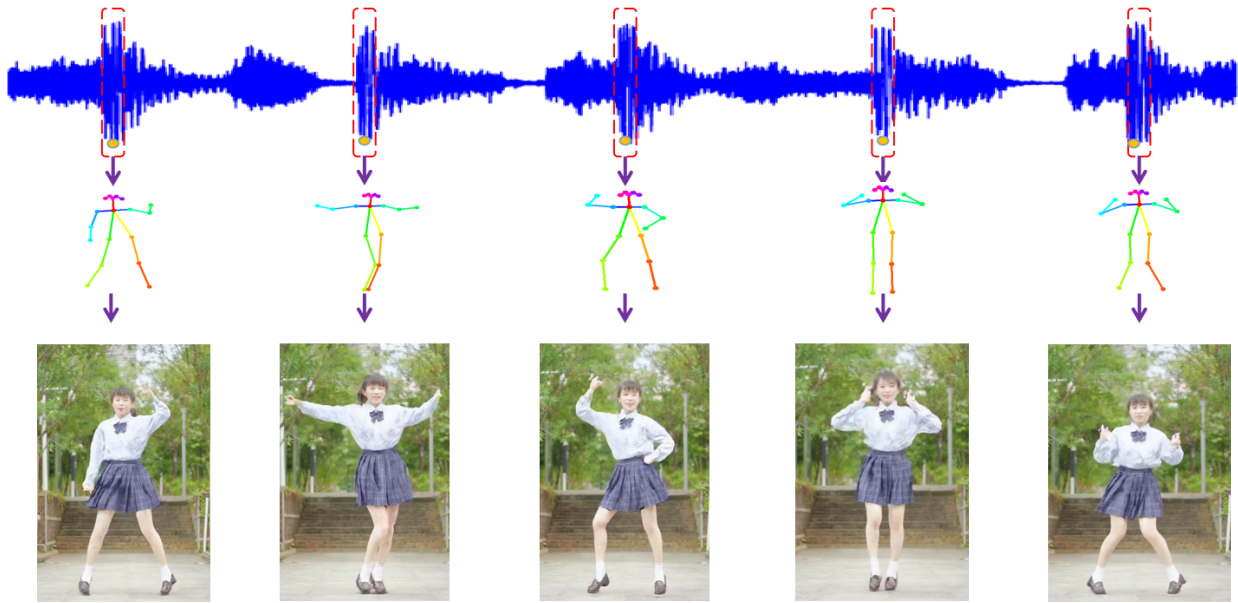


Fig. 1. The illustration of our motivation. Given a piece of input audio, our motif is to generate a dancing sequence automatically by taking the human pose as an intermediary. The representative pose and generated dancing frames corresponding to the audio peaks (in the red dotted box) are viewed in the second and third rows.

musical beats and refine the fragment transmissions, making the generated video vivid and natural.

Taken these two cues together, we tentatively propose a unified framework to synthesize the audio-inspired videos, which are leading by two alignment modules corresponding to the two main challenges. Given an input audio clip, the first cross-modal alignment module aims to find the correlations between audio clips and pose fragments. Taken the advantage of MFCC [8] and bidirectional LSTM [9], the audio clips are first encoded by the audio encoder to transform the audio from the frequency domain into low-dimension features. And the corresponding pose skeletons are encoded by a graph convolutional network to understand the sparse pose skeletons. Thus the pose and audio features from two different modalities can be embedded into one unified space, and then the correlations between them can be learned by minimizing the feature distances (see Fig. 6).

With the learned correlations, in the generation phase, we retrieve the best-matched pose fragment for each audio clip and concatenate them as an integral pose sequence. However, there usually exist sudden changes between two matched pose fragments and the pre-stored pose fragments may not perfectly match the audio beats, which makes the generated video seems unnatural. In order to solve this challenge, we propose a spatial-temporal alignment module that refines the pose sequences from two different perspectives. In the spatial domain, we first detect the sudden changes in the pose sequences and refine them by a Time Series Decomposition formula (TSD). In the temporal domain, we extract the musical beats (high-frequency points) as key cues and enforce the local maximum pose movements matching with the extracted beats. Hence the generated dancing video can show the notable pose movement (*e.g.*, raising arms) aligning with the musical beats.

After obtaining the natural dancing sequence, we follow the commonly-used pose-to-video generation to make a complete synthesis. Specially, we further adopt [10] with FaceGAN to generate a clean human face, which requires a short input video as the guidance. With the proposed two modules, experiments show that our method generates promising dancing videos with fewer consistency errors and best matches with human subjective evaluations.

Our main contributions are summarized as follows: 1) We propose a novel framework for synthesizing dancing videos from music sequences; 2) We propose a cross-modal alignment module to explore the relationship between music and dance clips; and 3) We propose a new method to align different pose fragments and align the rhythm of music and dance sequences.

The remainder of this paper is organized as follows: section II reviews related work and section III describes the proposed dataset. In section IV we describe our methods. We further conduct qualitative and quantitative experiments in section V. Finally, we conclude and discuss this paper in section VI.

II. RELATED WORK

A. Pose Estimation

Recent researches [11]–[14] tend to solve the pose estimation problem using deep learning techniques. Toshev and Szegegy [11] first applied the deep learning method to human pose estimation, namely deep pose, which achieves reliable performance on complex scenarios. Carreira *et al.* [15] proposed a self-correcting model (iterative error feedback) to solve the problem that feed-forward neural networks cannot model dependencies in the output space efficiently. In [16]–[19], researchers applied the graph model to the

neural network to enhance the prediction ability of the system. In [12], [20], [21], researchers built a multi-stage CNN regression model to expand the perceptive field of the network, so as to improve the predictive power of the model. Cao *et al.* [13] used the vector field to model the different limb structures of the human body, which solved the problem of wrong connection caused by the use of intermediate points of limbs. A pose grammar [22] was proposed to solve the problem of 3D human pose estimation. Ge *et al.* [23] adopted the point-cloud data and proposed to regress a low dimensional representation of the 3D hand. Moreover, Cai *et al.* [24] further proposed a depth-free testing model for 3D hand pose estimation.

B. Action Prediction

Due to the variability and randomness of movements, action prediction is a hard challenge. A large body of work [25]–[32] used LSTM to predict the next action based on a short video. Suwajanakorn *et al.* [2] built a hierarchical system to predict lips movements and synthesize Obama's speech videos. For the prediction of dancing actions, earlier works [33]–[35] proposed to learn the dancing motion based on hand-crafted features. For example, Fan *et al.* [34] proposed to synthesize the motion of virtual characters based on the music style and beats rate. Yu *et al.* [36] leveraged the self-attention mechanism to adaptively sparsify a complete action graph in the temporal space. Wang *et al.* [37] proposed to model smooth and diverse transitions on a latent space of action sequences with much lower dimensionality. Mao *et al.* [38] and Cai *et al.* [39] converted the motion into the frequency domain by using the discrete cosine transform. Tang *et al.* [6] proposed an LSTM-based autoencoder to mapping dancing skeletons based on the musical beat. Lee *et al.* [7] proposed a decomposition-to-composition network to align the dancing skeletons with musical beats. We will further elaborate on the differences and relations to these works in Section IV-E.

C. Action Recognition

Motivated the highly accurate pose estimation techniques, it has become a trend to use skeletons and joints to classify human action. Various methods of action recognition based on the skeleton were proposed during the last decade. Driven by the success of deep learning, the skeleton modeling method based on deep learning for action recognition has emerged. Some works [32], [40], [41] built an action recognition model based on RNN. Li *et al.* [42] and Ke *et al.* [43] built a model based on temporal CNNs in an end-to-end manner. Weng *et al.* [44] proposed a Spatial-Temporal Naive-Bayes Nearest-Neighbor for skeleton-based action recognition. Besides, Yan *et al.* [45] first applied GCN to the task of action recognition based on the skeleton, which achieves promising results in action recognition.

D. Image Synthesis

With regard to person image synthesis, recent studies [46]–[52] usually generated a character image based on a new pose. These works have achieved remarkable results in

generating the details of the image. Ma *et al.* proposed a new framework [51] and loss function [52], but the framework did not involve audio. In [53], the approach has shown that pose can be used as an effective supervisory signal for prediction and video generation. Chan *et al.* [10] built an interesting model to generate a target person dancing video with the same actions as the source person video. In this paper, we focus on the audio-inspired dancing video generation while adopting the pose skeletons as intermediaries for generation guidance.

E. Multi-Modal Studies

Many methods have achieved remarkable results in multi-modal studies. For example, Mori *et al.* [54] proposed a model to generate text from the image and describe the contents of the image. In [3], [4], and [55], all of them extracted text features to generate a single image. Moreover, the StoryGAN [56] was proposed to generate images from text to form a sequence of images that matches the content of the text. Liao *et al.* [57] explored the video manipulation driven by music. In addition, Chen and Huang [58] showed face recognition would achieve better results by adding audio input. Wang and Demirdjian [59] also built a complex system to prove that audio input is beneficial for body pose estimation. The system is motivated by MatchNet [60], which was aimed to extract the features of different objects to calculate their similarity.

III. DATASET

A. Dataset Collection

To synthesize the dancing imaginations from the corresponding videos, we first collect the dancing videos in the real scenarios, in which the moving dancers and audios can be obtained simultaneously. As the intermediate of the dancing descriptors, we thus collect the pose sequences in an automatic manner and meanwhile constructing the database of pose fragments. In addition, to simplify the problem formulation of generation, we tend to collect videos with only one dancer, thus the generation task of multiple people can be acquired by repeating this process.

1) *Audio Preprocessing*: The original audios from the real videos lie in the frequency domain, which is hard for extraction and subsequent processing. To obtain the aligned audio features with video frames, we adopt the MFCC algorithm [8] to transform the frequency signals to low-dimension vectors. Specifically, we compute the features¹ and choose the length of the windows as $1,000/video_{fps}$ with $fps = 24$, i.e., 41.66ms. The final audio feature dimension is 13 which can be easily embedded in our system.

2) *Pose Extraction*: As mentioned above, we need to extract the body keypoints from dancing videos in our system. Instead of the manual labeling of each video frame, we automatically obtain human keypoints by performing the OpenPose [61] algorithm. We find that there is an accidental detection error such as the local jitter (see, Fig. 2 (a)) in the continuous frames. To smooth the pose sequences and filter the out-liners,

¹https://github.com/jameslyons/python_speech_features

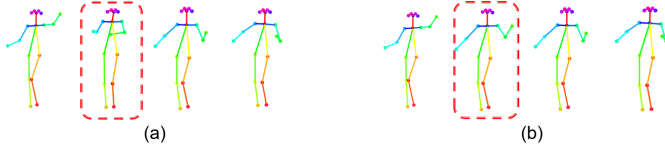


Fig. 2. The detection result after smoothing: (a) Jitter (in red dotted box) in the process of detecting body keypoints. (b) Results after smoothing.

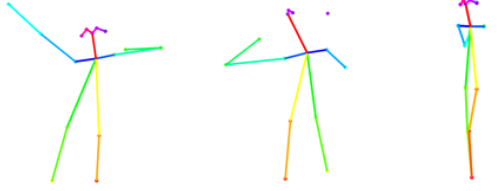


Fig. 3. Exemplar failures of keypoint detection, which can be automatically removed before constructing training data.

we conduct a simple linear interpolation among different pose frames. The smoothed results are shown in Fig. 2 (b). In addition, we automatically delete the incomplete detection (see, Fig. 3) following two rules: 1) it and the successive frames that follow it are too far from the previous frame, or 2) hand and foot nodes are not detected. After filtering the abnormal data, we segment the video clips as fragments with non-overlapping sliding windows, *i.e.*, 1s to 4s. We divide the extracted pose sequence into different fragments in seconds and then construct the fragments into the sequence database \mathbb{D} .

B. Dataset Statistics

1) *Data Scale*: We collect 122 dancing videos of females and 32 videos of males, including ACGN, robot, and hip-hop dance. Our collected videos range from 3 min to 5 min. A total of 9.0 hours of dancing videos is collected. We process all videos to 24 *fps*, extract frames into the dance image set and resize them to the same size. The total number of frames is 706,463.

2) *High Quality*: To construct a high-quality dataset, the dancing videos we collected are high-resolution (1920 × 1080) and have a stable fixed camera and bright lighting. Specifically, we select videos with high-quality music sound, noiseless interference, and solo performance. Fig. 4 shows example frames from videos and their corresponding keypoints.

3) *Abundant Diversity*: As shown in Fig. 5, we count the number of turns and the musical tempo in each dancing video. The tempo of the music and the number of turns vary significantly in each video. Moreover, the number of turns in all videos is greater than 1, and most of the number is greater than 5 (see, Fig. 5 (a)).

IV. THE APPROACH

A. Overview

Given an input audio clip \mathcal{A} and source video frames \mathcal{V}_s , the target of our synthesis framework is learning to generate target video frames \mathcal{V}_t , while adopting the corresponding



Fig. 4. Examples of our dataset: video frames in the dancing video and their corresponding pose keypoints.

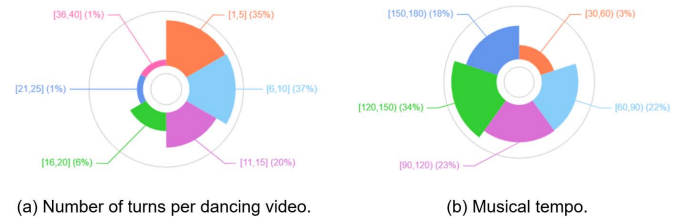


Fig. 5. Dataset statistics. We extract the tempo of each music and the number of turns in each video from the collected dance dataset. Then we present the distribution of the number of turns and musical tempo.

pose features \mathcal{P} as an intermediary. Let $\{\mathbf{x}_i^s = (\mathbf{v}_i^s, \mathbf{a}_i^s, \mathbf{p}_i^s) | \mathbf{x} \in \mathbb{S}\}_{i=1}^N$, which is a triplet of video, audio, and pose frames from the input source domain \mathbb{S} . And the triplet $\mathbf{x}^t \in \mathbb{T}$ of target domain is defined accordingly. Divided by the source and target domain, our framework can be constructed in two phases, which are shown in Fig. 6.

In the first matching phase, we take the generated \mathbf{x}^s as input and thus regularize the cross-modal alignment $\Phi_c(\cdot)$ to learn the correspondence between each \mathbf{p}_i^s and \mathbf{a}_i^s . Instead of direct learning the relation of specific identities in the limited source video, we expound this audio-to-video correspondence into a general condition by measuring the distance between audios and pose sequences. While pose features are usually regarded as the guidance for human video generation. This relationship learning process has the form:

$$\theta_c = \arg \min_{\theta} \Phi_c(\mathbf{a}_i^s, \mathbf{p}_j^s | \theta) \mathbb{1}(i = j) - \Phi_c(\mathbf{a}_i^s, \mathbf{p}_j^s | \theta) \mathbb{1}(i \neq j), \quad (1)$$

where $\mathbb{1}(\cdot)$ is the characteristic function and θ denotes the network parameters of cross-modal alignment module Φ_c . This objective regularizes audios and videos from the same frame with closer distances and vice versa.

With the learned relationship of videos and audios, the generation phase only takes audio \mathbf{a}^t from the target domain as input and retrieves the nearest pose sequence $\hat{\mathbf{p}}$ from the

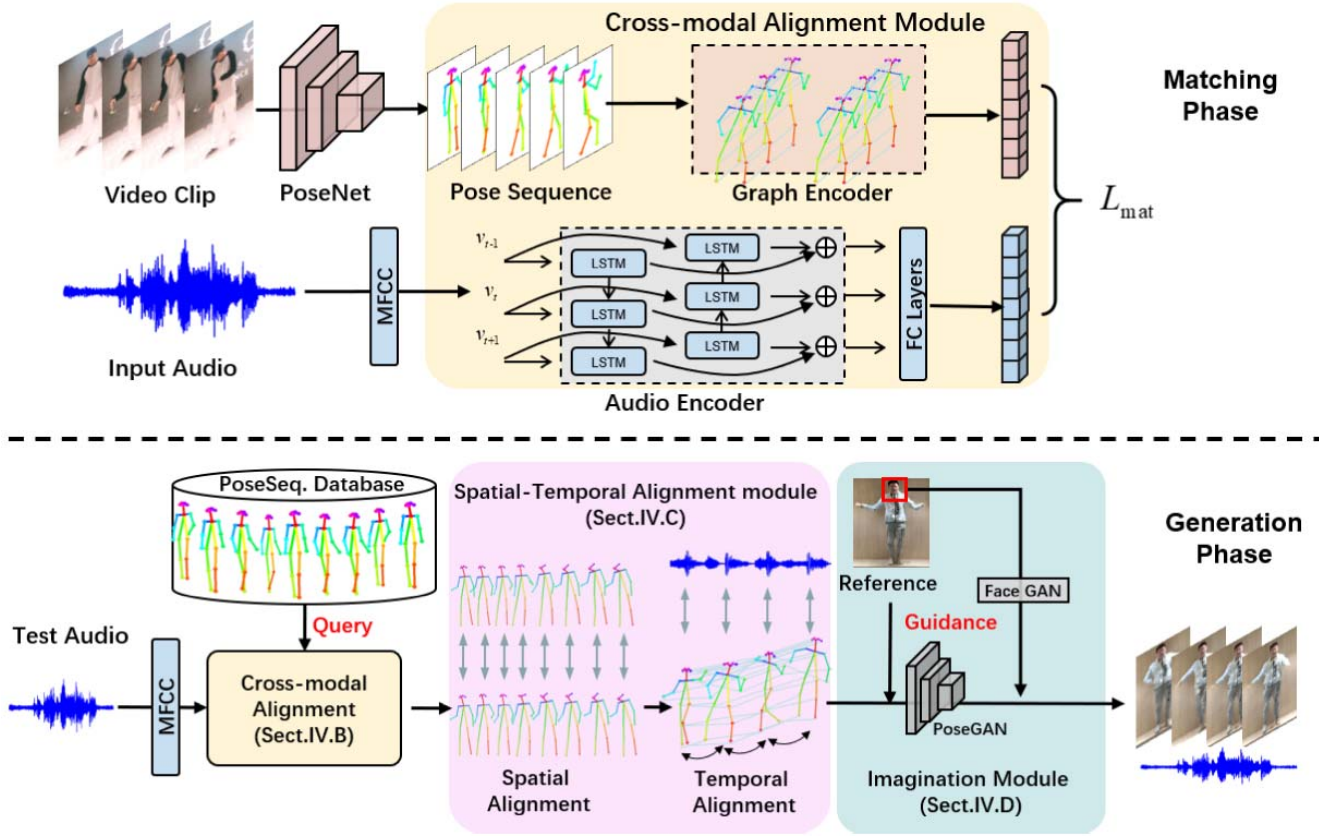


Fig. 6. The pipeline of the proposed framework. Matching Phase: given input video and audios, the matching phase learns the relationship between these two different modalities. Generation Phase: the spatial alignment is used to re-predict dance movements between discontinuous pose fragments. The temporal alignment aligns the beats of the music and pose sequences. The imagination module synthesizes the final dancing videos from the processed pose sequences.

constructed sequence database \mathbb{D} (described in Section III-A) from the source domain. It can be formally represented as:

$$\hat{\mathbf{p}} = \arg \min_{\mathbf{p} \in \mathbb{D}} \Phi_c(\mathbf{a}^s, \mathbf{p} | \theta_c). \quad (2)$$

The retrieved pose sequence $\hat{\mathbf{p}}$ is usually fragmented, causing the generated dancing movements unnatural. Hence we propose a spatial-temporal alignment module Φ_{st} , which refines the sequence from two different aspects. In the spatial domain, we detect the sudden changes among adjacent frames and smooth the results by an interpolation operation. In the temporal domain, we align the musical beats with substantial body movements, which makes the dancers dance rhythmically. After the alignment module Φ_{st} , the pose sequences are further encoded along with the guidance video clip $\mathbf{v}_g \in \mathbb{T}$ from the target domain to generate imagined dancing movements, which is supervised with an adversarial training process.

$$\mathbf{v}^t = \mathcal{G}(\Phi_{st}(\hat{\mathbf{p}}, \mathbf{a}^t), \mathbf{v}_g), \quad (3)$$

where \mathcal{G} denotes the generation network in Section IV-D. In the following subsections, we will elaborate the cross-modal alignment Φ_c in Section IV-B, spatial-temporal alignment module Φ_{st} in Section IV-C, and the imagination module in Section IV-D.

B. Cross-Modal Alignment

To measure the distance between audio \mathbf{a}^s and video frames \mathbf{v}^s , we first utilize the extracted pose sequences

in Section III and corresponded audio clips and fed these encoded features to the cross-modal alignment module (view in yellow in Fig. 6). However, these features are raw data that are simply transformed from the image or frequency domains, making the distance hard to measure. Hence we adopt the audio encoder and the pose encoder to embed different modalities into the same latent space and finally use a metric learning optimization to learn this correlation.

1) *Audio Encoder*: We first transform the audio with MFCC and process in Section III to a 13d vector per frame. Inspired by previous methods, which achieves satisfactory results to predict finger movements [5] and lip synthesizes [2], we adopt a bidirectional-LSTM (Bi-LSTM) and add a fully connected layer \mathbf{fc} to extract audio features. With the bidirectional information flow, our encoder can extract both forward and backward information to aggregate richer features from the audio. The encoding operation of the i th frame is:

$$\vec{h}_i = \text{LSTM}(\vec{\mathbf{a}}_i^s, \vec{h}_{i-1}, v_i), \quad (4)$$

$$\overleftarrow{h}_i = \text{LSTM}(\overleftarrow{\mathbf{a}}_i^s, \overleftarrow{h}_{i+1}, v_i), \quad (5)$$

$$\hat{\mathbf{a}}_i^s = \mathbf{fc}(\vec{h}_i, \overleftarrow{h}_i), \quad (6)$$

where v_i is the hidden layers and $\hat{\mathbf{a}}^s$ is the encoded output. The parameters of Bi-LSTM that we used are hidden state of 100, dropout of 0.1, learning rate of $1e-3$. The final audio feature is 16 dimensions.

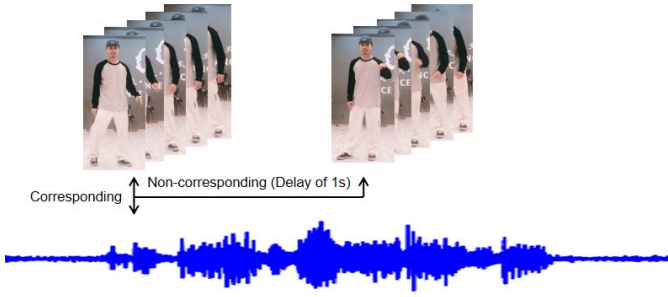


Fig. 7. Example of training data. Positive samples: corresponding audio with pose fragments. Negative samples: non-corresponding audio with delayed pose fragments.

2) *Pose Encoder*: To extract the context information of long-term frames, the generated pose sequences \mathbf{p}^s are then fed into a graph encoder. We adopt the ST-GCN [45] to exploit the relationship of different pose skeletons, which thus forms a holistic understanding of each pose fragment. The graph encoder extracts the temporal information g_{te} on 2d convolution kernels, and extracts their spatial information g_{sp} on graph convolution kernels:

$$g_{te} = \mathbf{W}_c \mathbf{p} + b_1, \quad (7)$$

$$g_{sp} = \Lambda^{-\frac{1}{2}} (\mathbf{A} + \mathbf{I}) \Lambda^{-\frac{1}{2}} g_{te} \mathbf{W}_g, \quad (8)$$

$$\hat{\mathbf{p}}_t = \mathbb{F}(g_{sp}) + b_2, \quad (9)$$

where \mathbf{W}_c and \mathbf{W}_g represent 2d convolution kernel and the graph convolution kernel respectively. \mathbf{A} is the adjacent matrix for the pose and $b_{\{1,2\}}$ are the learnable bias. Here $\Lambda^{ii} = \Sigma_j (\mathbf{A}^{ij} + \mathbf{I}^{ij})$. The embedded features of pose \mathbf{p}^s and audio \mathbf{a}^s are encoded with the same size of 16d.

After obtaining music features and corresponding dance features, we calculate the similarity by Euclidean distance. We propose a correlation matching loss \mathcal{L}_{mat} to regularize the similarity of features. The encoded pose feature $\hat{\mathbf{p}}$ and audio feature $\hat{\mathbf{a}}$ are attached with closer distance if they are from the same sequence, and vice versa. The overall loss function can be formulated as:

$$\mathcal{L}_{mat} = \|\hat{\mathbf{p}}_i - \hat{\mathbf{a}}_j\|_2^2 \mathbb{1}(i = j) + \|\hat{\mathbf{p}}_i - \hat{\mathbf{a}}_j - \epsilon\|_2^2 \mathbb{1}(i \neq j). \quad (10)$$

If $\hat{\mathbf{p}}$ and $\hat{\mathbf{a}}$ are relevant features, the loss function calculates the distance of two features and feeds it back to the network. On the irrelevant condition, we add a parameter ϵ to enlarge the distance of two features. However, only learning from the corresponding cases can easily lead to over-fitting. To solve this problem, we add a non-corresponding case to training data, where the pose fragments are delayed by a few seconds than MFCC features, which is illustrated in Fig. 7.

C. Spatial-Temporal Alignment

With the learned MatchNet $\Phi_c(\theta_c)$, each test audio could retrieve a sequence of pose fragments with the optimization in Eqn. (2). Although the retrieved pose fragments generate reliable movements, there exists a severe misalignment between different pose fragments, considering the fragments are discretely sampled in the construction of database \mathbb{D} .

Toward this end, it is desirable to further rearrange the sequence with a spatial-temporal alignment module to make the dancing sequence vivid.

1) *Spatial Alignment*: In the generation phase of Fig. 6, the audio guidance \mathbf{a}' from the target domain is encoded in the same way to get an audio vector. Instead of predicting the pose of each frame, we retrieve a most similar pose sequence \mathbf{p}' with the built database \mathbb{D} . However, it occurs abnormal movements among multiple pose fragments. Thus to solve this problem, we first adopt a spatial alignment to align the features being reasonable. The first step of spatial alignment is to detect the discontinuous frame in the retrieved pose sequences, whose skeletons show jitters with the previous frame. We define one frame as the discontinuous frame whose keypoint movements are larger than 10 pixels. Our basic idea is to refine these out-liners with a smooth movement function, which can be estimated with the whole dancing sequences.

In practice, the music melodies are composed of “beats”, which contain periodic signals but with dynamic changes in each period. Owing to the seq-to-seq design, here we introduce the temporal series decomposition for aligning and smoothing the possible non-continues frames between adjacent sequences. Taking this abnormal frame p_k as the central frame, we thus select the frames in nearby window $i \in [k - \omega_a/2, k + \omega_a/2]$ to form a Time Series Decomposition (TSD) algorithm.

With the defined sliding window, one notable issue is retaining the position of the start point and endpoint unchanged, which guarantees consistency after the overall refinement. Therefore we resort to a linear fitting operation $\mathcal{R} = \mathcal{F}(\mathbf{p})$ from the startpoint and endpoint of this sequence and $\mathcal{F}(\mathbf{p}_i)$ represents the i th value on this line. Hence we take the misalignment $\mathbf{d}_i = \mathbf{p}_i - \mathcal{F}(\mathbf{p}_i)$ for optimization. Once abnormal frames are detected, we conduct this smoothing operation inside each window $i \in [k - \omega_a/2, k + \omega_a/2]$.

Following the basic idea of TSD, we simply decompose each keypoint misalignment \mathbf{d} with the trend term M , the periodic term S and a random term γ . Taking one keypoint misalignment \mathbf{d}_i of the i th frame as an example, it has the form:

$$\mathbf{d}_i = S_i + M_i + \gamma_i. \quad (11)$$

We then perform differential operations $\varphi(\cdot)$ on \mathbf{d} to eliminate the periodic movement and obtain the trend function. The differential displacement $\hat{\mathbf{d}}_i$ without periodicity can be formulated as:

$$\hat{\mathbf{d}}_i = M_i(i|\alpha) = \alpha_0 + \sum_{m=1}^3 \alpha_m i^m, \quad (12)$$

where in each period T , the trend term M can be fitted with the learnable weight α of the cubic polynomial function.

The periodic term S is to find the minimum period T for $S_i = S_{i+T}$ and T is determined by volatility threshold $th = 5$ pixels.

$$S_i = \mathbb{E}(\mathbf{d}_i - \hat{\mathbf{d}}_i), \quad (13)$$

where $\mathbb{E}(\cdot)$ denotes the average expectation. The stationary random term $\gamma \sim \mathcal{N}(0, \sigma^2)$ is normally distributed. With this

Algorithm 1 Spatial Alignment Algorithm

Input : Target pose sequence \mathbf{p}^t : a sequence of body keypoint coordinates.
TSD parameters: $\{(M_i, S_i, \gamma_i), i = 0 \dots \omega_b\}$.

Output : Updated coordinate point: $\tilde{\mathbf{p}}$.

```

1 begin
2   Initialization:  $\varphi(\cdot)$  is the difference equation and
    $\xi(\cdot)$  calculates the length of the data in a period.
3   repeat
4     Make the linear fitting between the start and end
       points:
5      $\mathcal{R} = \mathcal{F}(\mathbf{p})$ 
6     Calculate the relative displacement of points in
       this window:
7      $\mathbf{d} \leftarrow \mathbf{p} - \mathcal{F}(\mathbf{p})$ 
8     Establish the spatial-temporal equation based on
       periodicity and trend:
9      $\mathbf{d}_i = M_i + S_i + \gamma_i$ 
10    Estimate and extract  $M_i$  and  $S_i$ :
11     $\hat{\mathbf{d}}_i = \varphi(\mathbf{d}_i)$ 
12     $M_i = \arg \min_{\alpha} (\hat{\mathbf{d}}_i - \sum_{m=0}^3 \alpha_m i^m)^2$ 
13     $S_i = \mathbb{E}(\mathbf{d}_i - \hat{\mathbf{d}}_i)$ 
14     $i \leftarrow i + \omega_a$ 
15  until  $i \geq \omega_b$ ;
16  Predict  $M^*$  and  $S^*$  of the abnormal window:
17   $M^* \leftarrow \mathbb{E}(M_i)$ 
18   $S^* = \arg \max_{S_i} (\xi(S_i))$ 
19  Make the linear fitting between the start and end
       points:
20   $\mathcal{R} = \mathcal{F}(\mathbf{p}^*)$ 
21   $\mathbf{d}^* = M^* + S^* + \gamma^*$ 
22   $\tilde{\mathbf{p}} \leftarrow \mathbf{d}^* + \mathcal{F}(\mathbf{p}^*)$ 
23 end

```

learnt decomposition, our target is using the previous ω_b ($\omega_b > \omega_a$) frames as reference frames and conduct this alignment per ω_a frames. We thus predict the TSD parameters of the next ω_a -length window (M^*, S^*, γ^*) and update the final coordinate point $\tilde{\mathbf{p}}$, which is illustrated in Alg. 1.

2) *Temporal Alignment*: Although the dancing sequences are well-smoothed by the spatial alignment, it seems unnatural when the musical beats (abrupt changes in audios) are not aligned with the large motions, *e.g.*, raising hands. Thus we adopt a temporal alignment to make the musical beats and motions consistent.

To make the sequences align with the music beats, we first extract the beat points of the target audio \mathbf{a}^t with the *librosa* library [62].

In the pose sequence, these pose beats are defined by:

$$\mu = \arg \max_j (|\tilde{\mathbf{p}}_j^t - \tilde{\mathbf{p}}_{j-1}^t|), \quad j \in [i, i + \omega_c]. \quad (14)$$

We select ω_c adjacent frames as a sliding window to search the local maximum to align. With the detected audio beats \mathbf{a}_μ , the ω_c frames can be re-organized as $[i, \mu]$ and $(\mu, i + \omega_c]$. For the preceding and the following frames of the current beat

point, we adopt cubic fitting interpolation to get the aligned pose $\hat{\mathbf{p}}^t$. This align function then has the form:

$$\alpha^* = \arg \min_{\alpha} \sum_{x=i}^{\mu} \left(\sum_{m=0}^3 \alpha_m x^m - \tilde{\mathbf{p}}_x^t \right)^2, \quad (15)$$

$$\hat{\mathbf{p}}^t = \alpha_0^* + \sum_{m=1}^3 \alpha_m^* x^m. \quad (16)$$

The optimization of $(\mu, i + \omega_c]$ is conducted in the same manner. After this temporal alignment, the musical beats and abrupt pose movements can be well aligned, which can serve as guidance for video imagination.

D. Imagination

Taken the optimized $\hat{\mathbf{p}}^t$ and video guidance \mathbf{v}_g from target domain \mathbb{T} , the final imagination is to generate reliable video frames. To achieve this goal, we adopt the pix2pixHD [63] as the generator for poseGAN. Video guidance \mathbf{v}_g is a 2-minute length video with the target character acting simply. In the stage of training, the goal of training generator is to generate a more realistic image and fools discriminator to regard as real. On the contrary, the purpose of training discriminator is to identify whether the image is from the real images or the generator synthesizing.

However, there still exist some abnormal regions in the generated images, especially in the face region. To enhance the representation of face regions, we adopt the FaceGAN [10] to enhance the facial region of the generated image. The purpose of the generator in the FaceGAN is to synthesis the detailed supplement of the face area between the synthesized images and the real images. We define the face region as the area near the nose position (*i.e.* 50×50 patches centered around the nose node) in the detected body pose, which is illustrated in Fig. 16. The generator of the FaceGAN outputs a residual gradient to help to synthesize the final face region. Additionally, in the process of training FaceGAN, the purpose of the discriminator is to distinguish the image synthesized by the generator or from the real images.

E. Discussions and Relations

To make a theoretical analysis with state-of-the-art models [6], [7], here we elaborate on three major differences as follows.

1) *Relationship Learning*: Both Lee *et al.*'s method [7] and Tang *et al.*'s method [6] are established based on the direct deduction from input music to dancing skeletons. However, direct mapping one modality to the other would usually lead to unreliable results. To this end, our work proposes a new perspective to explore the relationship between input audio sequence to a dancing sequence, namely seq-to-seq mapping. This guarantees the smoothness and rationality of the generated dancing videos.

2) *Audio Understanding*: Tang *et al.* [6] proposed to learn the skeleton movement with only extracting the “beat” frames to retain the motion alignments, which neglects the contextual information of input audio. To make a further improvement, Lee *et al.* [7] adopt a music style extractor to obtain the whole

artistic style of the audio content, *e.g.*, sadness or happiness. While in these researches, the contextual understanding of musical information can be easily neglected. By contrast, our proposed approach extracts the global information of the whole input audio, maintaining the audio content for the cross-modal relationship learning. We believe that understanding the whole audio content but not only the “beat” is necessary for the dancing video synthesis.

3) *Generation Targets*: The generation target of [7] and [6] are dancing skeletons, *i.e.*, human body keypoints. However, the final motif of music-inspired synthesis is to generate vivid and natural dancing videos. Hence, in our method, an imagination module is proposed to synthesize the dancing videos, and the skeletons play an intermediate role in the generation framework.

V. EXPERIMENTS

A. Experimental Setting

To quantitatively evaluate the generated dancing results of our approach, we conduct experiments from two aspects, *i.e.*, the quality of pose sequence and synthesized videos.² To evaluate the generation of pose sequences, we randomly select 80% of the data for training and the rest 20% for testing. In the processing of dataset assembling (Section III-A), we remove about 10% of the frames in the videos due to the extracted pose data that cannot be tackled. To synthesize dance videos in the imagination module, we collect 8 videos \mathbf{v}_g with a length of 120 seconds from different participants. We sample these generation videos in different natural scenarios, which have different backgrounds, clothing, and illuminations.

B. Implementation Details

Attributed to the light-weight implementation, our whole framework can be trained on a single consumer-level NVIDIA RTX 2070 GPU, 12 cores CPUs, and 16GB memory. It takes 35.1 hours for detecting pose keypoints and 0.5 hours for extracting the MFCC features of the music. The running time for 500 epochs of training of cross-modal system takes 2.9 hours (20.9s per epoch). After getting videos from the participants, we cost 12.0 hours to train the poseGAN to generate the target video image from the pose. In addition, it takes 0.7 hours to train the FaceGAN, which is used to strengthen facial areas. Since the shortest pose fragment is 1s and the number of frames is 24, we thus adopt Φ_{st} with ω_a of 8 and ω_b of 24 to conduct the experiments.

C. Comparison With State-of-the-Art

1) *Comparisons of Skeleton Alignment*: We adopt the official model of [7] and test our model and [7] with the same input test. Here we propose three metrics to evaluate the quality of generated skeletons.

a) *Beat alignment score*: A musical beat is defined as the maximum value of body movement in each sequence unit

²Some results can be found at <http://cvteam.net/projects/danceit/results.mp4>

TABLE I
QUANTITATIVE COMPARISONS OF OUR PROPOSED METHOD AND LEE *et al.* [7]. \mathcal{S}_{BA} DENOTES THE BEAT ALIGNMENT RATE. THE BEST PERFORMANCES ARE IN BOLD. \uparrow : THE HIGHER THE BETTER. \downarrow : THE LOWER THE BETTER

Methods	$\mathcal{S}_{BA}\uparrow$	MDD-H \downarrow	MDD-F \downarrow	SDD-H \downarrow	SDD-F \downarrow
Lee <i>et al.</i> [7]	60.05%	8.17	7.05	6.68	6.09
Ours (2s)	76.25%	6.23	5.67	6.81	5.99
Ours (3s)	67.93%	5.84	5.22	6.98	6.07
Ours (4s)	78.80%	5.63	5.17	6.67	5.97

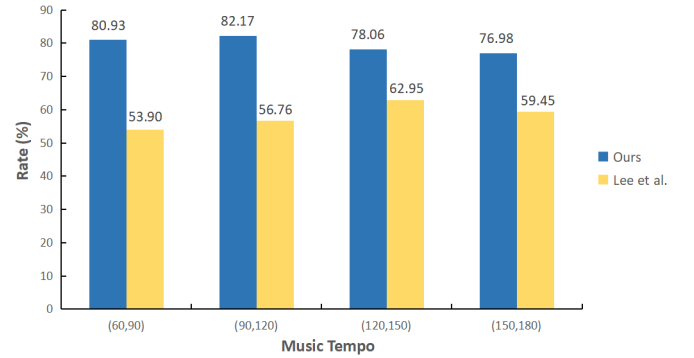


Fig. 8. Beat alignment score \mathcal{S}_{BA} under different musical beat environment. The tempo values are presented by beat number/per minute.

in Eqn. (14). Hence our aim is to calculate the matched beat number between the input audio \mathbf{a}_t^m and body movement μ :

$$\mathcal{S}_{BA} = \frac{\sum_{m=0}^n \mathbb{1}(\mathbf{a}_m^t = \mu_m)}{B_n}, \quad (17)$$

where $\mathbb{1}(\cdot)$ denotes the indicator function and B_n denotes the total number of the beat. The measurement \mathcal{S}_{BA} tends to be higher if there is a synchronization of input music and generated dancing movement.

b) Hand/Foot moving distribution distance (MDD):

Audio-inspired synthesizing methods usually generate different dancing movements when encountering the same input audio. Hence, we proposed to measure the skeleton movement between adjacent frames. *i.e.*, human beings would present similar *moving range* when hearing similar music. We measure the distance between hand and foot distributions between the groundtruth source \mathbf{p}^s and estimated pose $\hat{\mathbf{p}}^t$ as follows:

$$MDD_{\{h,f\}} = \frac{\text{KL}_{\{h,f\}}(\Delta \hat{\mathbf{p}}^t || \Delta \mathbf{p}^s) + \text{KL}_{\{h,f\}}(\Delta \mathbf{p}^s || \Delta \hat{\mathbf{p}}^t)}{2}, \quad (18)$$

where $MDD_{\{h,f\}}$ denotes the hand and foot moving distribution distance respectively and $\text{KL}(\cdot)$ represents the KL divergence scores. Δ denotes the difference between two adjacent frames, *i.e.*, the difference between \mathbf{p}_i and \mathbf{p}_{i-1} .

c) Hand/Foot spacing distribution distance (SDD):

Besides the movement of hand and foot keypoints, here we propose a new measurement to describe the spacing distance of one person, *e.g.*, the distance between one's feet. These measurements are defined similarly using Eqn. (18) to describe the discrepancy of adjacent frames.

Moreover, we exhibit the beat alignment score \mathcal{S}_{BA} in Fig. 8 under different musical beat environment, *i.e.*, beat

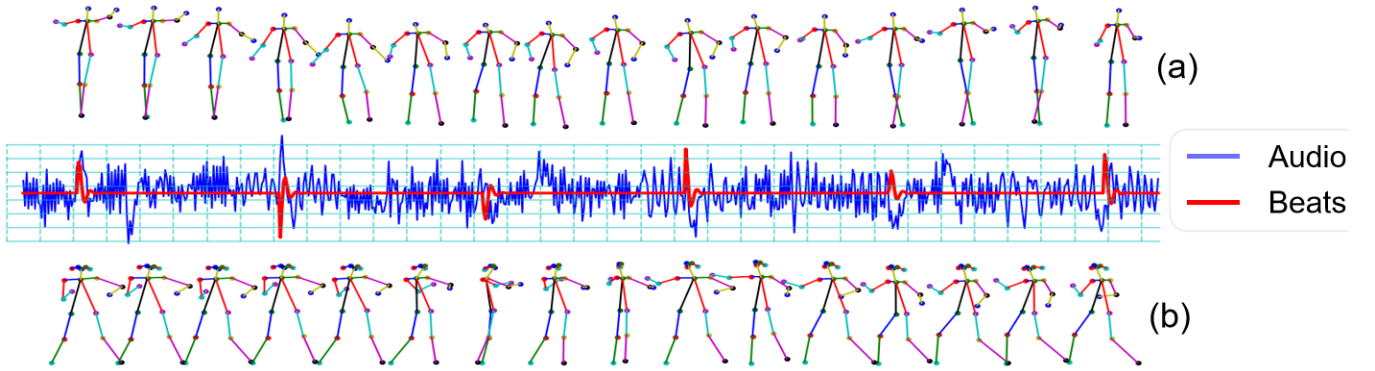


Fig. 9. Visualizations of beat alignment attribute. With the same input audio guidance, (a) presents the generated pose sequences of Lee *et al.* [7], and (b) presents the generated pose sequences by our method.

TABLE II
HAND AND FOOT MOVING DISTRIBUTION COMPARISONS WITH
STATE-OF-THE-ART. (X1,X2) REPRESENTS THE PIXEL
DISTANCE OF THE MOVING RANGE BETWEEN
ADJACENT FRAMES. THE VALUES ARE
SHOWN WITH THE PERCENTAGE OF
THE OVERALL DISTRIBUTION

Methods	(0,20)	(20,40)	(40,60)	(60,80)	(80,400)
Hand Moving Distribution					
Lee <i>et al.</i> [7]	91.0973	8.4076	0.4412	0.0322	0.0110
Ours	95.6649	4.2475	0.0867	0.0008	0
Foot Moving Distribution					
Lee <i>et al.</i> [7]	98.7880	1.0918	0.1022	0.0046	0.0026
Ours	98.7214	1.2640	0.0134	0.0012	0

number/per minute. The results indicate that under different beat rate conditions, our proposed method shows better alignment than the state-of-the-art model [7], especially in music with slow rhythm (*i.e.*, smaller than 120 beats/per minute).

2) *Comparisons of Moving Distribution*: One severe problem lies in the direct-mapping strategies, *e.g.*, [6] and [7], is the abnormal jittering. Considering the realistic dancing video, the movement of each body keypoint would present in a normal range. A large movement between adjacent frames (*i.e.*, 1/30s) is not encouraged. Here we present the distribution of hand/foot keypoint movements in Tab. II. It can be found that our model generates more smooth and progressive body movement while [7] performs large jittering movements. In addition, for moving distance larger than 80pixs (which is usually impossible for body motion in 1/30s), [7] shows some abnormal distributions in both hand and foot motions. Benefiting from the sequence to sequence mapping and the spatial-temporal alignment module, our approach generates notably better results.

3) *Visualization Comparisons*: To visually compare with state-of-the-art models [7], we exhibit four representative cases to verify the generation quality of our approach. With the same music guidance in the bottom (musical beats are colored in red line), the generated pose skeleton and our skeleton are presented at top and bottom respectively. We exhibit a representative case with the same input music in Fig. 9 our

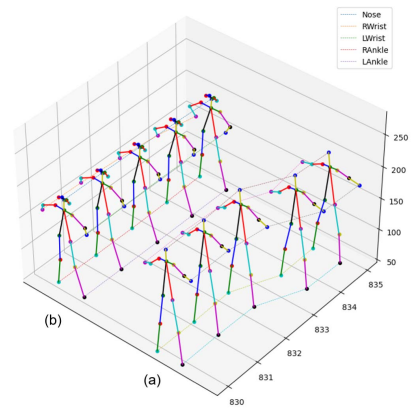


Fig. 10. 3D Visualizations of continuous movements. (a): skeletons generated by our proposed network. (b): skeletons generated by [7]. The three axes indicate the x , y pixel localization and frame i .

generated dancing skeleton has a strong ability for attaining rich body motions that are highly aligned with the musical beats. While the Lee *et al.* [7] in the first row shows slightly body movements but containing more abnormal jitterings. Moreover, it can be found that our generated pose sequences show continuous rhythmic dancing sequence while [7] shows slight changes. Benefiting from the cross-modal alignment and rich pose skeleton representation with head keypoints, our approach generates dynamic head poses while [7] are not endowed. Last but not least, in Fig. 10, considering the different mapping manners (*i.e.*, direct mapping and seq-to-seq mapping), [7] shows abnormal movement in the 834th frame, while our model shows continuous dancing skeleton sequences.

D. Performance Analysis

1) *Matching Phase*: The training and validation Euclidean distance of Eqn. (10) can be found in Tab. III. The matching system performs best with the combination of ($lr = 0.0001$, $dropout = 0.1$, $D_f = 16$), which achieves the minimal loss in the training set and validation set simultaneously. To test the ability to find the consistency of the music and pose clips, we measure the similarity of two features by the Euclidean

TABLE III
THE EUCLIDEAN DISTANCE (EUCLID.) AND CORRELATION ACCURACY (CORR.) OF DIFFERENT HYPER-PARAMETERS OF CROSS-MODAL ALIGNMENT. D_f DENOTES THE DIMENSION OF OUTPUT FEATURES

Method	Train Euclid.	Test Euclid.	Train Corr.	Test Corr.
$lr = 0.0005$	2.49	1.79	0.65	0.73
$lr = 0.001$	4.63	3.50	0.60	0.70
$lr = 0.01$	16.91	12.46	0.50	0.65
$dropout = 0.2$	3.79	2.62	0.60	0.67
$dropout = 0.4$	6.14	4.82	0.52	0.59
$dropout = 0.5$	8.19	6.51	0.46	0.54
$D_f = 4$	2.62	2.12	0.76	0.82
$D_f = 8$	2.43	2.30	0.75	0.82
$D_f = 32$	2.60	1.99	0.65	0.70
$D_f = 64$	3.24	2.80	0.57	0.58
Ours	2.37	1.48	0.77	0.83

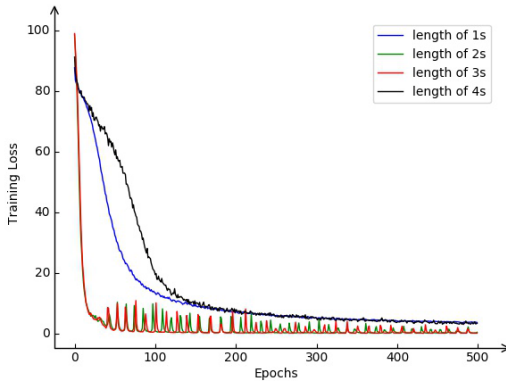


Fig. 11. Losses of our model when training different length of pose fragments. From left to right are different length pose fragments from 1s to 4s (The x-axis is the training epochs and the y-axis is the loss).

distance between them. To evaluate the matching accuracy from another perspective, we thus define the correlation accuracy as: i) the music and pose clip are corresponding if the distance between two features is less than 1, ii) the music and pose clip are non-corresponded if the distance is larger than 1.

The length of pose fragment is a key factor that affects the matching and generation quality. To test the feasibility of our framework, we experiment with different lengths of \mathbf{x}^s from 1s to 4s. It is obvious that all the lengths of pose fragments can converge in a short period of training time (see Fig. 11 for details). And the matching model with a short length of \mathbf{x}^s shows a more rapid converge than that of a large length. However, there exists a notable trade-off between the matching phase and the generation phase. In the spatial-temporal alignment module Φ_{st} of generation stage, we notice that the results $\hat{\mathbf{p}}^t$ are sometimes unstable when the length of \mathbf{x}^s is extremely short (e.g., 1s) and unable to achieve a complex movement, e.g., turning the body around. Thus to make the system consistent, we choose the length of pose fragment with more than 1s.

The other intuitive way to evaluate the matching results is the exhibition of skeleton movements. After the model Φ_{st} is trained, we experiment with different kinds of music \mathbf{a}^t to the

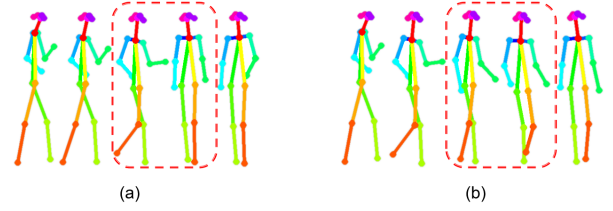


Fig. 12. The pose sequence after spatial alignment: (a) the discontinuity (in the red dotted box) between two adjacent pose fragments. (b) the optimized pose sequence.

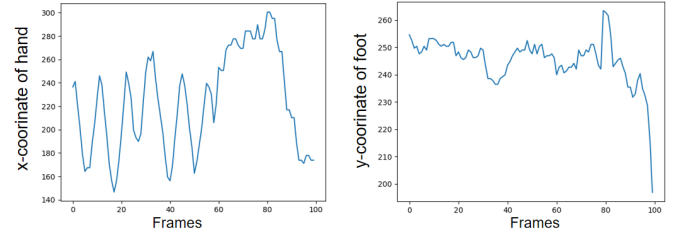


Fig. 13. The movement tendency of the dance. The left figure is the x-axis trend of the hand node and the right one is the y-axis trend of the foot node. The dance movements have strong periodicity in some periods (left) and not in others (right).

system and exhibit two typical \mathbf{p}^t at the same moment (see Fig. 15). We observe that various \mathbf{p}^t in different music closely matching the content of the music and the skeletons act in a steady motion without abrupt changes. Remarkably, our model has the potential to predict complex pose movements, which is extremely lacking in the end-to-end learning scheme. For example, there exists a backside action (top of Fig. 15) in the generated results.

2) *Generation Phase*: In the generation phase of pose sequence \mathbf{p}^t , we need to retrieve the adequate pose fragment for database \mathbb{D} . However, these fragments are not on the same scale owing that they are collected from different video frames. To solve this problem, we propose a normalization algorithm to unify the pose fragment into the same scale. For each selected pose fragment, we record the maximum distance between the nose keypoint and the feet keypoint. Then we resize the retrieved fragments as the same size as the target person. The ratios of different pose fragments are then used to expand the frame equally in the x and y axis.

a) *Spatial alignment*: We therefore visualize the spatial misalignment in Fig. 12 (a). It is obvious that there exists an abnormal movement in the retrieved pose sequence. As stated in the spatial alignment module, we extract two dance movements of hand keypoint and foot keypoint in Fig. 13 respectively. It can be concluded that the hand keypoint moves periodically and the foot keypoint moves nonperiodic. Based on this finding, we conduct the proposed spatial alignment with the trend term M_t and periodic term S_t of i th frame. The refined results with our spatial alignment module are shown in Fig. 12 (b), in which the human pose moves more stably and turns around in a more reasonable way.

b) *Temporal alignment*: To further align musical beats with pose movements, we use the temporal alignment for

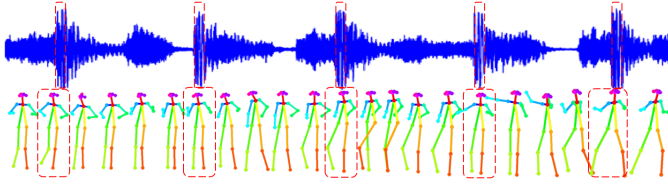


Fig. 14. The pose sequence after temporal alignment: matched dance movements at different musical moments. After the processing of the temporal algorithm, the rhythms (in the red dotted box) of the pose sequence and music are aligned.



Fig. 15. The predicted pose sequence of different music: top sequences matched to the music *Updown Funk*, bottom sequences matched to the music *Follow Me*. Our model has the potential to predict complex pose movement.



Fig. 16. Visualized results with FaceGAN: (a) the groundtruth image. (b) the generated image. (c) the gradient generated by the FaceGAN generator. (d) the image optimized by FaceGAN.

consistency, which is shown in Fig. 14. The processed music can be viewed in blue, in which the beats are viewed in the red dotted box. To align the pose sequences with musical beats, the red dotted box shows the most significant movement in the local window of ω_c frames. To be specific, we set the parameter ω_c as the interval length of the adjacent musical beats, which remains constant in each piece of music. In this manner, it can be concluded that the changing of movement are naturally aligned with the musical beats and meanwhile overall pose moving trends are greatly maintained. For example, at the first beat, the right foot of the skeleton moves a large step simultaneously.

c) Video imagination: After getting the final pose sequence \mathbf{p}^f , we synthesize the target person dancing video \mathbf{v}^f with the resolution of 512×512 in the imagination module, as shown in Fig. 17 (a). It can be seen that our model can be generalized the complex scenarios with higher light intensities. However, the face regions are usually blurry owing that our human beings pay much more attention to the face regions than other regions. Thus we adopt the FaceGAN in Fig. 16 to adopt the gradient map of the face region as input. With this local enhancement, the local regions are further concatenated with the holistic image, which is viewed in Fig. 17 (b)).



(a)



(b)

Fig. 17. The example of synthesizing target dance image: (a) the image of the target character generated from the pose. (b) the generated image with the face region enhanced by FaceGAN.

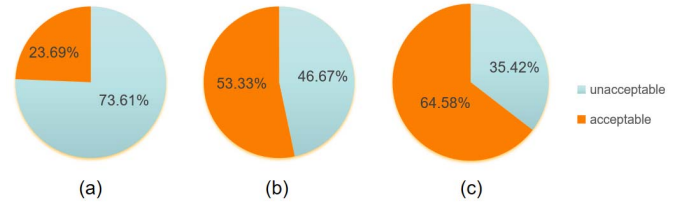


Fig. 18. User study. (a) The voting result of the dancing videos obtained without any optimization. (b) The voting result of the dancing videos obtained by linear interpolation. (c) The voting result of the dancing videos obtained after the processing of the spatial-temporal alignment module.

E. Time Efficiency

Our system consists of the cross-modal alignment, spatial-temporal alignment and imagination module. With the resolution of 512×512 images, the computation costs of our two alignment modules are 23.59GFLOps and the cost of the imagination module is 373.28GFLOps. In the process of generating dancing sequences, our proposed model generates 48.8 frames per second, which is relatively light-weight. In the process of final video synthesis, our proposed imagination module generates 8.59 frames per second.

F. Subjective Evaluation

To evaluate the quality of the generated dancing videos, we conduct a user study to measure the effectiveness of our proposed spatial-temporal alignment module. We evaluate two different versions of our method: 1) pose sequences without any optimization, 2) pose sequences with linear fitting interpolation. We develop a simple user interface to play these two videos with the same musical pieces and choose the option and the subsequent studies are conducted in the same setting.

To make a fair comparison, we randomly sample 20 video clips from the generated dancing videos of each version. Each participant watches the video clips and chooses one of the two options: 1) the synthesized dancing video is realistic, 2) the synthesized dancing video is unrealistic. As shown in Fig. 18 (a), without any optimization, 26.39% of the votes indicate the synthesized dancing videos are realistic, while 73.61% of them are unrealistic. The reason for these voting results is that the dance movement shows abnormal changes, which are unnatural to human subjective evaluations.



Fig. 19. Synthesized results of multiple targets: Images of different human dances generated in different music. By inputting different music and videos of different persons, our proposed approach is able to generate satisfactory results that match human subjective evaluations.

The new voting results obtained after the post-processing of the pose sequence with linear fitting interpolation are shown in Fig. 18 (b). 53.33% of participants vote that the synthesized dancing videos are realistic, while 46.67% do not. After performing the proposed alignment module, this new voting result obtained after spatial alignment is shown in Fig. 18 (c). 64.58% of participants vote that the synthesized dancing videos are realistic, while 35.42% participants do not.

G. Generalizations and Failure Modes

1) *Video Syntheses Under Different Scenarios*: Benefiting from the alignment modules and the pose normalization, our approach is able to provide synthesized videos in various scenarios. As shown in Fig. 19, the generated videos under different conditions show natural movements and clear dancing poses, which answers the question in Section I: our proposed method can generate reliable imaginations with the input of music pieces. For example, with the confusing background to the human bodies, our approaches generate stable pose sequences while keeping the background relatively static, especially in the last group of special costumes.

2) *Failure Modes*: Although our proposed method generates reasonable dancing videos, there are still some limitations in our system. As shown in Fig. 20 a), the generated arms are blurry in some complex scenarios, which forces us to improve the quality of the generator. However, our two alignment modules provide satisfactory results of dancing poses, which could be reliable guidance for the imagination module. In the second

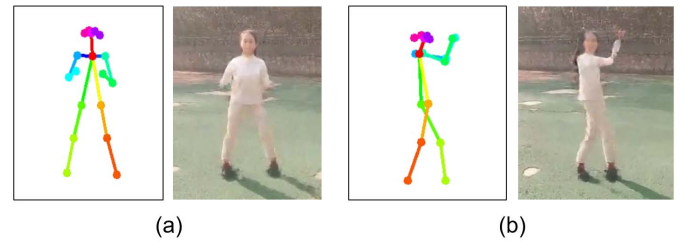


Fig. 20. The failure cases of synthesizing images: (a) The arms are blurry. (b) The face and arms are unnatural.

case of Fig. 20 b), when performing the complicated poses, the imagination still faces challenges in distinguishing the overlapping parts with similar visual appearances, which is left for our future work.

VI. CONCLUSION

In this paper, we focus on the audio-inspired video synthesis, which aims to generate reliable dancing videos corresponding to the input audio. To solve this important problem, we first collect a data set with musical audios and video clips, while combining the pose sequences at the same time. Based on this dataset, we make an attempt to align these two media with two alignment modules, *i.e.*, the cross-modal alignment module and the spatial-temporal alignment module. Considering the audio and video frames are two completely different modalities, we process these two media with different

encoders and take the pose as an intermediary to build the connections. Thus the relationship between the audio and pose sequences is established. Then this learned relationship helps the retrieval of pose fragments with given test audio, and we concatenate these fragments as the dancing guidance. However, there exists strong misalignment between the pose fragments and audio beats, we thus adopt the spatial-temporal alignment module to align the pose sequences with musical beats and changing steadily in a more natural manner. With the generated pose sequences as guidance, we finally generate the imagined dancers moving vividly aligning with the input music. Experimental evidence demonstrates that our method shows a reliable matching correlation of different types of media and generates realistic videos that match the human subjective evaluations.

ACKNOWLEDGMENT

The authors would like to express their sincere appreciation to Tengfei Shi, Weili Lu, Jing Guo, and Zhengdong Zhang for their volunteer work in providing the dancing videos used in this research.

REFERENCES

- [1] S.-F. Chang and D. Zhang, "System and method for extracting text captions from video and generating video summaries," U.S. Patent 7339992, Mar. 4, 2008.
- [2] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing Obama: Learning lip sync from audio," *ACM Trans. Graph.*, vol. 36, no. 4, p. 95, Jul. 2017.
- [3] F. Tan, S. Feng, and V. Ordonez, "Text2Scene: Generating compositional scenes from textual descriptions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6710–6719.
- [4] W. Li *et al.*, "Object-driven text-to-image synthesis via adversarial training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 12174–12182.
- [5] E. Shlizerman, L. Dery, H. Schoen, and I. Kemelmacher-Shlizerman, "Audio to body dynamics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7574–7583.
- [6] 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*, Oct. 2018, pp. 1598–1606.
- [7] H.-Y. Lee *et al.*, "Dancing to music," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 3586–3596.
- [8] F. Zheng, G. Zhang, and Z. Song, "Comparison of different implementations of MFCC," *J. Comput. Sci. Technol.*, vol. 16, no. 6, pp. 582–589, Nov. 2001.
- [9] S. Zhang, D. Zheng, X. Hu, and M. Yang, "Bidirectional long short-term memory networks for relation classification," in *Proc. 29th Pacific Asia Conf. Lang., Inf. Comput.*, 2015, pp. 73–78.
- [10] C. Chan, S. Ginosar, T. Zhou, and A. Efros, "Everybody dance now," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5933–5942.
- [11] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1653–1660.
- [12] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4724–4732.
- [13] 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. 7291–7299.
- [14] A. Newell, Z. Huang, and J. Deng, "Associative embedding: End-to-end learning for joint detection and grouping," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 2277–2287.
- [15] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, "Human pose estimation with iterative error feedback," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4733–4742.
- [16] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1799–1807.
- [17] W. Yang, W. Ouyang, H. Li, and X. Wang, "End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3073–3082.
- [18] L. Ge *et al.*, "3D hand shape and pose estimation from a single RGB image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10833–10842.
- [19] Y. Cai *et al.*, "Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 2272–2281.
- [20] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer*, Oct. 2016, pp. 483–499.
- [21] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1831–1840.
- [22] H.-S. Fang, Y. Xu, W. Wang, X. Liu, and S.-C. Zhu, "Learning pose grammar to encode human body configuration for 3D pose estimation," in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2018, vol. 32, no. 1, pp. 6821–6828.
- [23] L. Ge, Y. Cai, J. Weng, and J. Yuan, "Hand PointNet: 3D hand pose estimation using point sets," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8417–8426.
- [24] Y. Cai, L. Ge, J. Cai, N. Magnenat-Thalmann, and J. Yuan, "3D hand pose estimation using synthetic data and weakly labeled RGB images," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, May 11, 2020, doi: 10.1109/TPAMI.2020.2993627.
- [25] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, "Recurrent network models for human dynamics," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4346–4354.
- [26] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2891–2900.
- [27] I. Habibie, D. Holden, J. Schwarz, J. Yearsley, and T. Komura, "A recurrent variational autoencoder for human motion synthesis," in *Proc. Brit. Mach. Vis. Conf.*, London, U.K., Sep. 2017, pp. 1–12.
- [28] D. Holden, J. Saito, and T. Komura, "A deep learning framework for character motion synthesis and editing," *ACM Trans. Graph.*, vol. 35, no. 4, p. 138, 2016.
- [29] Z. Li, Y. Zhou, S. Xiao, C. He, Z. Huang, and H. Li, "Auto-conditioned recurrent networks for extended complex human motion synthesis," vol. 3, 2017, *arXiv:1707.05363*. [Online]. Available: <http://arxiv.org/abs/1707.05363>
- [30] J. Walker, K. Marino, A. Gupta, and M. Hebert, "The pose knows: Video forecasting by generating pose futures," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3332–3341.
- [31] Y.-W. Chao, J. Yang, B. Price, S. Cohen, and J. Deng, "Forecasting human dynamics from static images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 548–556.
- [32] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3D human action recognition," in *Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer*, 2016, pp. 816–833.
- [33] 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.
- [34] R. Fan, S. Xu, and W. Geng, "Example-based automatic music-driven conventional dance motion synthesis," *IEEE Trans. Vis. Comput. Graphics*, vol. 18, no. 3, pp. 501–515, Mar. 2012.
- [35] F. Ofli, E. Erzin, Y. Yemez, and A. M. Tekalp, "Learn2Dance: Learning statistical music-to-dance mappings for choreography synthesis," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 747–759, Jun. 2012.
- [36] P. Yu, Y. Zhao, C. Li, J. Yuan, and C. Chen, "Structure-aware human-action generation," in *Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer*, 2020, pp. 18–34.
- [37] Z. Wang *et al.*, "Learning diverse stochastic human-action generators by learning smooth latent transitions," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 12281–12288.
- [38] W. Mao, M. Liu, M. Salzmann, and H. Li, "Learning trajectory dependencies for human motion prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9489–9497.
- [39] Y. Cai *et al.*, "Learning progressive joint propagation for human motion prediction," in *Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer*, 2020, pp. 226–242.

- [40] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGD+D: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1010–1019.
- [41] W. Zhu *et al.*, "Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks," in *Proc. 13th AAAI Conf. Artif. Intell.*, Feb. 2016, pp. 3697–3703.
- [42] C. Li, Q. Zhong, D. Xie, and S. Pu, "Skeleton-based action recognition with convolutional neural networks," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2017, pp. 597–600.
- [43] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3D action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3288–3297.
- [44] J. Weng, C. Weng, and J. Yuan, "Spatio-temporal naive-Bayes nearest-neighbor (ST-NBNN) for skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4171–4180.
- [45] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. 32nd AAAI Conf. Artif. Intell.*, Feb. 2018, pp. 7444–7452.
- [46] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. Guttag, "Synthesizing images of humans in unseen poses," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8340–8348.
- [47] R. de Bem, A. Ghosh, T. Ajanthan, O. Miksik, N. Siddharth, and P. Torr, "A semi-supervised deep generative model for human body analysis," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, Sep. 2018.
- [48] R. D. Bem, A. Ghosh, A. Boukhayma, T. Ajanthan, N. Siddharth, and P. Torr, "A conditional deep generative model of people in natural images," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1449–1458.
- [49] D. Joo, D. Kim, and J. Kim, "Generating a fusion image: One's identity and Another's shape," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1635–1643.
- [50] C. Lassner, G. Pons-Moll, and P. V. Gehler, "A generative model of people in clothing," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 853–862.
- [51] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, "Pose guided person image generation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 406–416.
- [52] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz, "Disentangled person image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 99–108.
- [53] R. Villegas, J. Yang, Y. Zou, S. Sohn, X. Lin, and H. Lee, "Learning to generate long-term future via hierarchical prediction," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 3560–3569.
- [54] Y. Mori, H. Takahashi, and R. Oka, "Image-to-word transformation based on dividing and vector quantizing images with words," in *Proc. 1st Int. Workshop Multimedia Intell. Storage Retr. Manage.*, 1999, pp. 1–9.
- [55] G. Yin, B. Liu, L. Sheng, N. Yu, X. Wang, and J. Shao, "Semantics disentangling for text-to-image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2327–2336.
- [56] Y. Li *et al.*, "StoryGAN: A sequential conditional GAN for story visualization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6329–6338.
- [57] Z. Liao, Y. Yu, B. Gong, and L. Cheng, "Audeosynth: Music-driven video montage," *ACM Trans. Graph.*, vol. 34, no. 4, p. 68, 2015.
- [58] L. S.-H. Chen and T. S. Huang, "Joint processing of audio-visual information for the recognition of emotional expressions in human computer interaction," Ph.D. dissertation, Dept. Elect. Eng., Univ. Illinois Urbana-Champaign, Champaign, IL, USA, 2000.
- [59] S. B. Wang and D. Demirdjian, "Inferring body pose using speech content," in *Proc. 7th Int. Conf. Multimodal Interfaces (ICMI)*, 2005, pp. 53–60.
- [60] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg, "MatchNet: Unifying feature and metric learning for patch-based matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3279–3286.
- [61] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," 2018, *arXiv:1812.08008*. [Online]. Available: <http://arxiv.org/abs/1812.08008>
- [62] B. McFee *et al.*, "Librosa: Audio and music signal analysis in Python," in *Proc. 14th Python Sci. Conf.*, Austin, TX, USA, vol. 8, Jul. 2015, pp. 18–25.
- [63] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.



Xin Guo received the M.S. degree from the State Key Laboratory of Virtual Reality Technology and System, School of Computer Science and Engineering, Beihang University, in 2021. His research interests include computer vision and cross-modal learning.



Yifan Zhao received the B.E. degree from the Harbin Institute of Technology in July 2016. He is currently pursuing the Ph.D. degree with the State Key Laboratory of Virtual Reality Technology and System, School of Computer Science and Engineering, Beihang University. His research interests include computer vision and image understanding.



Jia Li (Senior Member, IEEE) received the B.E. degree from Tsinghua University in 2005, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences in 2011. He is currently a Full Professor with the School of Computer Science and Engineering, Beihang University, Beijing, China. Before he joined Beihang University in June 2014, he used to conduct research at Nanyang Technological University, Peking University, and Shanda Innovations. He is the author or coauthor of over 90 technical papers in refereed journals and conferences, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI), *IJCV*, IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), CVPR, and ICCV. His research interests include computer vision and multimedia big data, especially the understanding and generation of visual contents. He is a Senior Member of ACM, CIE, an CCF. He has been supported by the Research Funds for Excellent Young Researchers from the National Natural Science Foundation of China since 2019. In 2017, he was selected into the Beijing Nova Program and ever received the Second-Grade Science Award of Chinese Institute of Electronics in 2018. He received the two Excellent Doctoral Thesis Award from the Chinese Academy of Sciences in 2012 and the Beijing Municipal Education Commission in 2012. He received the First-Grade Science-Technology Progress Award from the Ministry of Education, China, in 2010.