Computational Visual Media
https://doi.org/10.1007/s41095-0xx-xxxx-x

Research/Review Article

Let's All Dance: Enhancing Amateur Dance Motions

Qiu Zhou[1,#], Manyi Li[2,#], Qiong Zeng[1](✉), Andreas Aristidou[3], Xiaojing Zhang[1], Lin Chen[5], and Changhe Tu[1](✉)

© The Author(s)

Abstract Professional dancing is characterized by high impulsiveness, elegance, and aesthetic beauty. In order to reach the desired professionalism, it requires years of long and exhausting practice, good physical condition, musicality, but also, a good understanding of the choreography. Capturing dance motions and transferring them into digital avatars is commonly used in the film and entertainment industries. However, so far, access to high-quality dance data is very limited, mainly due to the many practical difficulties in motion capturing the movement of dancers, which makes it prohibitive for large-scale acquisitions. In this paper, we present a model that enhances the professionalism to amateur dance movements, allowing the movement quality to be improved in both the spatial and temporal domains. The model consists of a *dance-to-music alignment* stage responsible for learning the optimal temporal alignment path between the dance and music, and a *dance-enhancement* stage that injects features of professionalism in both the spatial and temporal domains. To learn a homogeneous distribution and credible mapping between the heterogeneous professional and amateur datasets, we generate amateur data from professional dances taken from the AIST++ dataset. We demonstrate the effectiveness of our method by comparing it with two baseline motion transfer methods via thorough qualitative visual controls, quantitative metrics, and a perceptual study. We also provide temporal and spatial module analysis to examine the mechanisms and necessity of key components in our framework.

Keywords Animation, Music-to-Motion Alignment, Dance Motion Enhancement, Dance Motion Analysis

1 Introduction

Dance is a performing art form that consists of purposeful, rhythmical, and well-patterned sequences of body movement that has aesthetic and often symbolic value [1]. Capturing dance motions and transferring them into avatars not only facilitates expressive film or animation production process, but also contributes to the conservation of cultural heritage and dance education. However, so far, access to high-quality dance data is limited. Most of the currently available motion capture repositories typically contain basic human movements, while only a limited number of dance-specific databases consist of prime dance movements performed by professionals [2, 3]. This is because professional dances are characterized by dynamic body language, high impulsiveness, elegance, smoothness, fluidness, and aesthetic beauty that usually require the performer to have long-term dance experience and skills, followed by extensive practice sessions, excellent physical condition, and acquaintance with the years of dance studies. This poses a practical challenge when capturing realistic and high-quality dance motions, which makes it restrictive for large-scale acquisitions, or on a regular basis [4]. To perform a professional dance, the performer should get familiar with *the content and rhythm of the choreography*, and achieve the *specific physical amplitude of the choreography with the appropriate energy and balance* [5]. On top of that, in order to achieve a satisfactory dance quality during the motion capture process, dancers have to repeat the performance many times to avoid mistakes.

In this paper, we present a technique that enhances professionalism to dance moves, allowing the movement quality to be improved in both the spatial and temporal

1 School of Computer Science & Technology, Shandong University, Qingdao, 266000, China. E-mail: zhouqiulv@gmail.com; qiong.zn@sdu.edu.cn; xj.zhang@mail.sdu.edu.cn; chtu@sdu.edu.cn.

2 School of Software, Shandong University, Jinan, 250101, China. E-mail: manyi_li163@163.com.

3 Department of Computer Science, University of Cyprus, Nicosia, 1678, Cyprus; CYENS Centre of Excellence, Nicosia, 1016, Cyprus E-mail: a.aristidou@ieee.org.

4 Qingdao Institute of Humanities and Social Sciences, Shandong University, Qingdao, 266000, China. E-mail: chenlin.spring@gmail.com.

Manuscript received: 2022-01-27; accepted: 2022-XX-XX

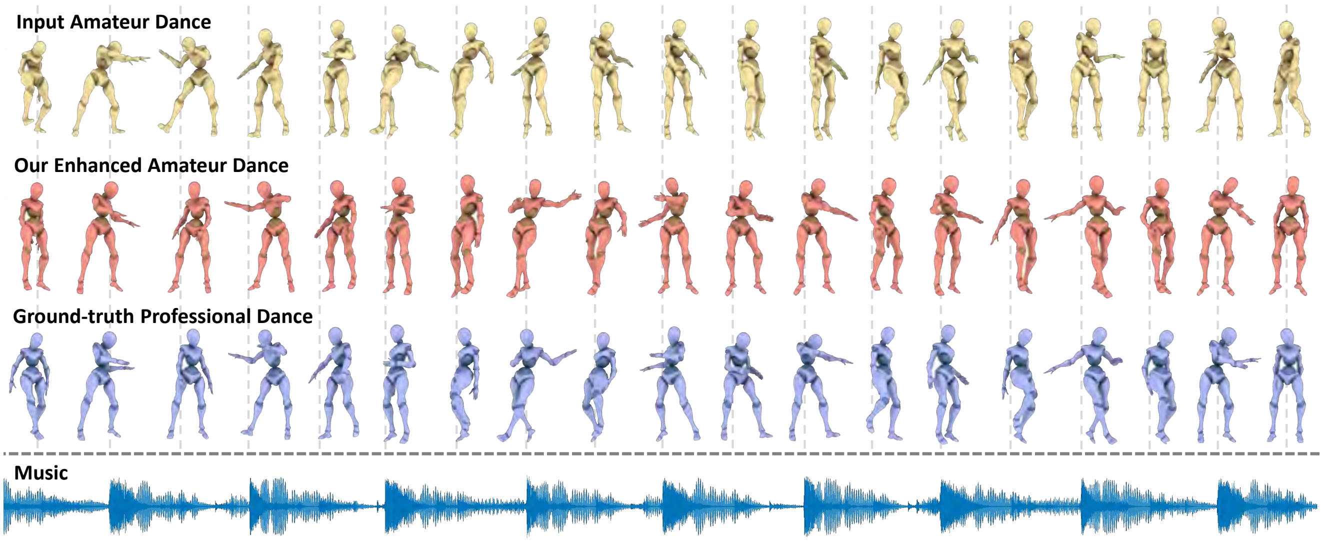


Fig. 1 Our approach enhances professionalism to dances performed by non-professional dancers. The first row shows the input amateur dance sequence, the second row shows our enhanced dance motion, and the third row illustrates the corresponding ground-truth professional dance. It can be observed that our results have similar temporal and spatial features with the ground-truth dance sequence.

domains, meeting the following key constraints: (i) the production of flowing and smooth dance moves, (ii) the expansion of the anatomical and physical amplitude of human movements, to meet the demanding restrictions of the choreography, and (iii) the well-synchronization of the movement to follow the rhythm of the music. In this way, our method reduces the need to hire professional dancers, facilitates the process of obtaining high-quality dance movements even by amateurs, enriches existing databases with professional data to enable better training of deep networks, and finally aligns dance motion data with a given audio file.

One obvious direction to deal with the challenge of enhancing professionalism on dance movements, is to leverage a deep style-transfer framework [6–10], by considering the amateur dances as the source style and the professional dances as the reference style. However, despite that style transfer algorithms is a possible way to handle this problem, it is not exactly the same. Professionalism is not a specific style, but a kind of evaluation metric. Dances with different styles might be seen as professional ones. Professional dance preparation, whatever the style, not only has specific anatomical and physical demands, but also requires artistic qualities, such as musicality, expression and distinct communication skills. On top of that, the existing style-transfer methods face the following two technical challenges: (i) they mainly focus on locomotions with well-defined styles, while different styles of motions have explicit changes in the whole sequences. In contrast to that, dances often contain highly-dynamic and heterogeneous movements, and the professional and non-professional dances may share a large number of similar poses but with a limited number of local changes, see Figure 1. Therefore, it is difficult to learn mappings between unpaired professional and non-professional dances, as state-of-the-art motion style transfer methods do [8, 9]; (ii) they mainly focus on music-free motions, with no explicit and deterministic control on the correlations between motion content and other external factors, such as the music rhythm. Even though existing methods may cause timing changes in motion based on the style differences hidden in the data, such changes are uniformly distributed over the temporal domain.

In this work, we propose a two-stage dance enhancement model that adds *professionalism* to existing dance motion, along with the release of a new dataset with paired professional and amateur dances that enables the training on the model. We define the "dance professionalism" term, and describe how it can be evaluated through various attributes, e.g., flow, amplitude and rhythm of a dance. In particular, we improve the quality of dance motions in both spatial and temporal domains, focusing on the following three professionalism properties: (i) the production of *fluent and smooth* movements; (ii) the *intensely physical amplitude* of the movements that are restricted by the poor physical condition of the amateur dancer; and (iii) the *temporal alignment* of the dance movements to a given music rhythm. First, our model estimates the temporal correlations between dance motions and music rhythm, followed by a temporal alignment and spatial motion enhancement process, under the guidance of the proposed professionalism metrics. The dance-to-music alignment stage consists of network that learns the affinity matrix between dance and music with attention mechanisms, and a classic dynamic time-warping module to infer an optimal temporal alignment path matrix. Secondly, the dance-enhancement-stage enables adjustment of the dance motion in both temporal and spatial domains, under the guidance of the optimal alignment path, a reconstruction loss, and a consistency loss. The reconstruction loss constraints the network to preserve the original motion content of the amateur dance, while adjusting it to be similar to the corresponding professional dance. The consistency loss preserves the temporal continuity of the enhanced dance motion and decreases temporal noises.

One of the most critical challenges we faced in this project is the lack of data for training our network. Professional and corresponding amateur dances may be different in various combinations of the above professionalism properties. Since the professionalism of a dance is independent of its choreography or style, the dances in a professional or amateur dance dataset may contain highly dynamic and heterogeneous movements. This makes it difficult to learn a homogeneous distribution for the professional or amateur dataset using existing methods, let alone a credible mapping between the two heterogeneous datasets. In addition, the mappings between professional dances and amateur ones are not deterministic. Therefore, before designing our network, we first introduce a key-pose based data augmentation scheme to generate amateur data from professional dances, taken from the AIST++ dataset [3]. The data augmentation scheme modifies the movements in all three professionalism metrics, and the constructed dataset contains many-to-one paired amateur and professional dances.

We demonstrate the effectiveness of our method by comparing it with two state-of-the-art motion transfer methods [6, 8] via thorough qualitative visual controls, quantitative metrics, and a perceptual study. Apart from using our synthesized amateur data, we additionally captured several dance sequences performed by amateur dancers, to further examine the generalizability of our method. User responses indicate that our method enhances amateur motion that cannot be easily distinguished from actual professional dances. In addition, we provide temporal and spatial module analysis via an ablation study to evaluate the mechanisms and necessity of key components in our framework.

The main scientific contributions of the paper are:

- We introduce the concept of enhancing professionalism in dance movements; we give a first definition of what dance professionalism is, and how a professional dance can be distinguished from an amateur.
- We design a novel two-stage deep learning framework that extracts meaningful features from motion inputs, in terms of the newly defined professionalism criteria, to improve the quality of dance motions. It integrates a reconstruction loss, to preserve the original content of the dance, and a consistency loss, to maintain the temporal coherency of the reconstructed motion.
- We propose a novel model designed to synchronize 3D dance motions with a reference audio under non-uniform and irregular misalignment.
- We present thorough evaluations, and an ablation study to examine the efficiency and necessity of our method.

2 Related Work

2.1 Dance Evaluation

Dance is an expressive form of performing art that consists of aesthetic movements of the body in a rhythmic way, usually to music, for the purpose of expressing an idea or emotion, releasing energy, or simply taking delight in the movement itself [5]. To professionally perform dance, the performers regularly attend long routine training, and have extensive experience in dance studies, choreography, and musicality, along with the excellent physical condition, which enable them to perform complex movements with extreme physical amplitude demands in some instances [11, 12]. Only a few works in the dance research community have identified qualitative factors for professional dances. For example, Neave *et al.* [13] and Torrents *et al.* [14] have reported their qualitative experiments that, kinematic parameters related to the amplitude of movement have high associations to the perception of dance beautification and aesthetics, while Park [15] investigated the correlation between dance professionalism and motion smoothness (measured in jerk-based quantitative measures). However, no explicit quantitative metrics have been proposed so far to completely evaluate dance professionalism.

In computer graphics, several interactive dance systems have been proposed to enhance dance learning and teaching [16–18]. Basically, these methods export dance movement features to enable comparisons between dances performed by professional dancers (teachers) and amateurs (students). For example, Chan *et al.* [19] implemented a self-learning dance system by visually comparing motion accuracy through Euclidean distance between the professional and amateur motions. Aristidou *et al.* [20] leveraged the well-known Laban Movement Analysis (LMA) [21] theories to introduce quantitative feature components that measure the quality characteristics between two dance motions. Although these movement measurements can be used to understand motion qualities and to compare similarities between dance motions, no metrics have been developed so far that explicitly measure the professionalism of dance motions.

2.2 Motion Style Transfer

One obvious way to add professionalism on existing motion sequences is to use methods that are based on the concept of motion style-transfer. These methods aim to transform the style of a reference motion to a source motion, while simultaneously preserving the original source motion content. Several previous approaches [22, 23] have been proposed in the literature to infer styles between motions by using hand-crafted features. For example, Tenenbaum and Freeman [22] explicitly separate style and content using asymmetric bilinear models. Aristidou *et al.* [23] built statistic correlations between the LMA features and emotions, and used such correlation to support interactive emotion-based motion transfer. However, Those methods explicitly construct common mappings between hand-crafted features and motions, and thus it is hard to be generalized in heterogeneous or large-scale dataset.

Machine Learning Based Techniques. To avoid the disadvantages of selecting hand-crafted features, researchers started to extract style information from large-scale paired data using machine learning techniques [24–26]. Brand and Hertzmann [24] introduced a style Hidden-Markov-Model (HMM) and minimized the information entropies to separate structure, style and accidental properties. Following their work, Hsu *et al.* [25] built dense correspondences between different motions with an iterative motion warping algorithm, and then proposed a linear time-invariant model to translate motion styles, while Xia *et al.* [26] proposed to learn local regression models. However, machine learning-based methods require explicit or implicit motion registration between the input and output motions, and therefore, they are limited to those styles and contents that exist in the training dataset; as a result, they cannot be well-generalized to new stylistic motions.

Deep Learning Based Techniques. In recent years, deep learning techniques have been greatly adopted to transfer motion styles [6–10, 27–29], enabling more efficient and effective performance on complex and even unpaired motions. For instance, Holden *et al.* [6] leveraged convolutional autoencoders [30] to learn hidden motion representations with paired input and output motions. The same authors in [7] further improved this model with an additional feed-forward neural network, and transformed motion style in the hidden motion space under the constraint of *Gram matrix* [31]. Later on, Aberman *et al.* [8] proposed a neural network to disentangle latent style and content codes, where the latent style code is used to modify the decoded content motion through an AdaIN [32] operator. Due to the introduction of a multi-style discriminator, this method can handle unpaired motions. Following their work, Wen *et al.* [10] recently proposed an unpaired and unsupervised motion style transfer method using a generative flow model. Despite their great progress, existing deep-learning-based methods mainly focus on locomotions with limited number of motion structures, and have no explicit control on music correlations. Different from locomotions, dance performed by professional dancers may contain heterogeneous motions with various choreographies (e.g., different motion poses and ordering of poses) and are well-synchronized to temporal rhythmic patterns. In that manner, our method deals with these challenges by simultaneously learning the intrinsic motion attributes and the motion-rhythm correlations that commonly appear in professional dances.

2.3 Music-driven motion synthesis

Many scholars have been working on methods for music-driven dance synthesis. Typical solutions leverage the graph-based framework [33–36]. As a pioneering work, Kim *et al.* [34] constructed a movement-transition-graph-based on extracted motion beats and synthesized new motions under kinematic and rhythmic constraints. More recently, the use of machine learning to synthesize music-driven dance motions has witnessed impressive progress [4, 37–41]. For instance, Lee *et al.* [38] proposed a decomposition-to-composition framework to generate 2D movements conditioned on a given music, under the guidance of learned correlations between music beat and dance units. Chen *et al.* [4] proposed a choreomusical embedding module to learn stylistic and rhythmic music-dance correspondences, and incorporated the embedding distances into the traditional graph-based dance synthesis framework. More recently, Aristidou *et al.* [41] introduced a music-driven neural framework that generates rich and diverse dance motions that respect the overall choreographic structure of a dance genre. However, music-driven dance synthesis learning methods heavily rely on high-quality dance motion data that are synchronized to a given audio, in order to be adequately trained. Since access to dance data made by professionals is not always possible, our method can be used to enrich databases using data from amateurs that have been artificially enhanced to look more visually appealing; our work simultaneously learns music-to-dance correspondences and leverage them to learn dance-to-dance correlations.

2.4 Audio Alignment

To enhance non-professional dances, an essential goal is to align dance motions to reference audio. Motion-audio synchronization aims to temporally align human motion dynamics to audio rhythms, which is a fundamental scheme to synthesize rhythmic human motions. Traditional motion-audio synchronization methods leverage hand-crafted 2D motion and rhythmic features, estimate their correspondences, and warp motions under the guidance of motion-rhythm matchings [35, 42–44]. Over the years, more attention has been devoted to the video-audio timing alignment problem. A common basic idea is to find optimal video-to-audio correspondences and use them to guide warping between visual-audio features, either using hand-crafted features [45, 46] or deep multi-modal features [47, 48]. Among those methods, Wang *et al.* [49] introduced two attention modules before the feature extraction stage to highlight important spatial and temporal regions. In contrast, instead of emphasizing on specific features, we introduce an integrated attention module to map correspondences between spatial and temporal motion features, and audio rhythms, without any hand-crafted elements or post-processing. Our model is the first to synchronize 3D dance motions with reference audio file under irregular and non-uniform misalignment.

3 Data Augmentation

An important challenge that needs to be addressed in this work is the lack of training data. The limited existing dance motion datasets [2, 3, 39] typically contain high-quality professional dances, but lack of corresponding non-professional ones. As a result, it is difficult to build correlations between paired professional and non-professional movements. Capturing realistic non-professional dances requires the amateur dancers to learn the original professional choreography, which sometimes is challenging and requires practice, and might be difficult to cover the large variability of the movements of the professional dancers. Instead, we propose a data augmentation scheme that artificially synthesizes random non-professional dances, by altering professional ones taken from the AIST++ dataset [3], both on their spatial and temporal domains.

3.1 The Definition of Dance Professionalism

Before we present our data augmentation schema, it is important to define first the criteria that distinguish a professional dance movement from an amateur. In that matter, we consulted expert choreographers, experienced dancers, and dance teachers, who pointed out the following key criteria:

- **Sense of rhythm:** Professional dancers can perfectly follow the beat of the music, while the amateur dancers often lose their synchronization and have difficulties in following the rhythm of the music.
- **Physical amplitude:** Professional dancers have excellent physical condition, which allows them to perform complex and dynamic movements, in some cases reaching the limits of the body. In contrast, non-professional dancers usually have difficulties in completing certain dance moves as they have limitations due to their poor physical condition (e.g., to extend their body to the limits, or perform the split etc.).
- **Motion quality:** The movements of a professional dancer are elegant, smooth, and the movement cycle is nicely completed. In contrast, the movements of an amateur dancer usually are not in balance, they abruptly start and end movements without fully completing the movement cycle (sharply movements), and they are a bit shaking (not smooth). All these result in amateurs demanding more effort than the professionals because they do not control their movement as experts do.
- **Concentration and consistency:** Amateur dancers usually focus on one part of the body (e.g., legs or arms) and somehow neglect the consistency of movements in other parts of the body (such as head, style, etc.). Note that, our method does not take this feature into account.
- **Choreography:** Professionals have a richer choreography in terms of the diversity of movements, compared to the amateurs who usually repeat the same movements several times. It is important to recall here that changing the dance choreography is out of the scope of this paper.

3.2 Generating Amateur Dance Movements

Amateur dancers have, in general, difficulties in synchronizing their movements on the music beat, to achieve certain physical amplitudes, and to perform controlled and smooth movements. Therefore, to enrich our database, we introduce a method that artificially alters professional dance movements, through random disturbances, to generate their corresponding amateur counterparts. It is important to note that our alternation approach needs to meet the following three conditions: (a) keep the choreography of professional dances unchanged; (b) *temporal disturbance*: alter the temporal alignment between

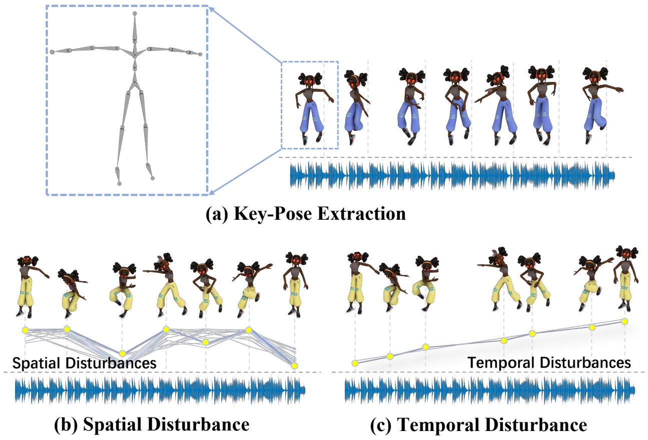


Fig. 2 Data Augmentation: the process for generating amateur dance movements. (a) The key-poses rendered on a girl's avatar with *blue* trousers, and our skeleton structure (top left) highlighted within a dashed rectangle. (b) The key-poses are spatially modified based on the spatial disturbance curves. Note that we highlight one curve for a specific joint with the spatial factors on key-poses (indicated by *yellow* dots). (c) The temporally modified key-poses and the accompanying temporal disturbance curves.

motion and music/rhythm; (c) *spatial disturbance*: change the physical amplitudes of motions.

In that manner, we propose a key-pose-based scheme that first extracts key poses based on the motion beat; then, it randomly generates spatial disturbance factors to limit or exaggerate the physical amplitudes of movements, and temporal disturbance factors to disrupt the music-motion synchronization. Finally, it computes the spatial and temporal disturbances between those key-poses using piece-wise linear interpolation [50], which are later used to modify the professional dances. Figure 2 shows an overview of our data augmentation method.

Motion Representation. We represent a dance motion $\mathbf{M}$ as a sequence of T skeleton poses. Each skeleton pose $\mathbf{P}$ is represented by $J = 21$ joint rotations that are organized in a hierarchical order and depicted by unit local positions [51] between parent-child joints, denoted as $\mathbf{P} \in \mathbb{R}^{(J-1)\times 3}$. Therefore, a dance motion can be denoted by $\mathbf{M} \in \mathbb{R}^{T\times(J-1)\times 3}$, where $T = 426$ to $2,878$ frames without being trimmed into short clips. Pose positions are then translated back to rotations via a Jacobian-based inverse kinematics solver [52]. Note that the root rotations and translations are discarded in the motion representation to avoid significant changes on the choreography.

Key-Pose Extraction. When learning to dance, it is usually easier for students to identify prominent changes of movements (such as pausing and turning). Based on that observation, we define the representative poses as those with changes on the direction of speed [46]. To facilitate the key-pose extraction, we first uniformly sample several poses in a certain time duration t. The time duration t is set to three seconds in our implementation. We then search for their nearest motion beats as the corresponding key-poses, where the motion beat is estimated by the minimum speed of all joint speeds at a certain frame. Since some neighboring key-poses may bring about rapid direction changes, we filter out those key-poses that have neighboring motion beats within 1 second.

Spatial Disturbance aims to disrupt physical amplitudes of the movements, limiting or exaggerating their intensity. Thus, we define the spatial factor $\mathbf{S}' \in \mathbb{R}^{N\times J}$ for the N selected key-poses to control the spatial disturbance on all the skeleton joints, and randomly generate corresponding values through an approximately inverse normal distribution:

$$S'_n = \tanh(s'_n * d) * \alpha + \beta, \tag{1}$$

where s'_n is the randomly generated spatial disturbance value for the key-pose n; α and β are used to control the shape of inverse normal distribution - in our implementation they equal to 1.1 and 1.3 respectively; d is a randomly generated binary parameter that enables or disenables the exaggeration of the pose, $d = 1$ indicates exaggeration is on. Note that all the joints in a specific frame share the same d value. Then S'_n is propagated to each frame of the entire sequence as $\mathbf{S} \in \mathbb{R}^{T\times J}$ $(T > N)$ via linear interpolation.

A straightforward way to apply the aforementioned spatial factor for motion disturbance is to directly multiply it with the rotations or positions of each joint. However, this may produce infeasible poses that violate physical/bone constraints. On the contrary, we interpolate the new pose (local position) between the current and a standard standing pose, under the guidance of the spatial factor, as follows:

$$\mathbf{p}'_{t,j} = |\mathbf{p}_{t,j}| \cdot \left(\frac{\mathbf{p}_{t,j}}{|\mathbf{p}_{t,j}|} \cdot S_{t,j} + \mathbf{u}_j \cdot (1 - S_{t,j})\right), \tag{2}$$

where $\mathbf{p}_{t,j}$ denotes the local position of the j-th joint at the t-th frame, $\mathbf{u}_j$ is the pre-defined direction for the joint j on the standard standing pose. To simplify the process, we define three key interpolation directions for the standing pose: an up direction $\mathbf{u}_j = (0, 0, 1)$ for joints on the spine, a down direction $\mathbf{u}_j = (0, 0, -1)$ for the rest of joints, and no modifications for the shoulder and waist joints.

Temporal Disturbance aims to disrupt the mappings between the dance motion and the corresponding musical rhythm. We define the temporal factor $\mathbf{Q} \in \mathbb{R}^N$ to control

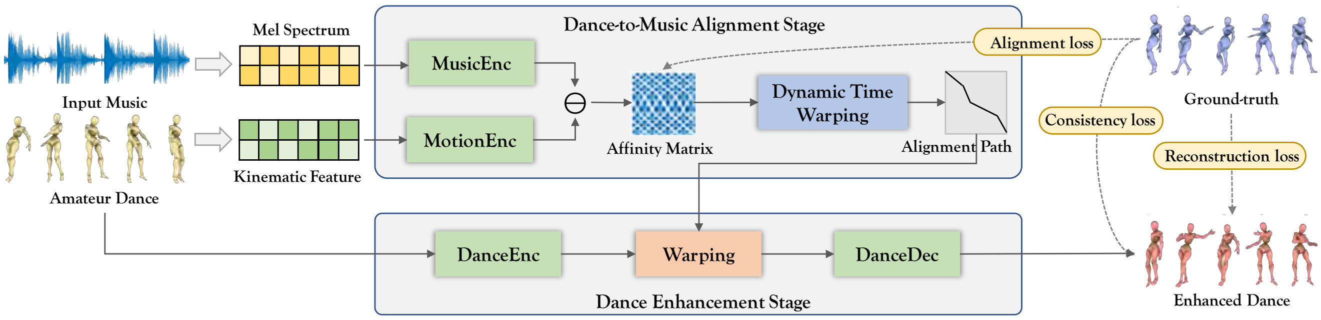


Fig. 3 Our two-stage dance professionalism architecture. The dance-to-music alignment stage learns the temporal alignment of the input dance motion to match the corresponding music, through a dynamic-time-warping operation on the encoded deep features of dance motion and music. In the dance enhancement stage, we first extract the hidden dance motion features to express the original motion content, which are then modified under the guidance of the temporal alignment matrix and further decoded into the enhanced dance motion under the constraints of a reconstruction and consistency loss.

the temporal disturbance at the N key-poses, and randomly generate the values to warp the original dance motion sequence according to Eq. 1. The parameters of α and β are equal to 50 and 0 respectively. We then move the key-poses to the new positions by shifting $\mathbf{Q}$ frames, where a negative Q_n means shifting backward and a positive Q_n indicates shifting forward. Note that in this process we need to check time crossings between key-poses and preserve the motion monotonicity. Finally we calculate the movements of the intermediate poses between every two neighboring key-poses through linear interpolation.

Using the aforementioned criteria, we constructed a large non-professional dataset of dances (that is paired to professional ones) with high variability. In our settings, we repeat the temporal and spatial disturbance four times, creating many-to-one paired amateur and professional dances, which is $4\times$ larger than the original AIST++ database.

4 Dance Professionalism Framework

Our framework, by taking as input a dance motion sequence performed by an amateur dancer, and its corresponding audio file, aims to enhance professionalism by considering the following three conditions – *keep the original choreographic content, generate fluent movements, and add physical amplitudes*. The enhancement is made at both the temporal and spatial domains; in the temporal domain, our framework aligns the amateur dance motion to music to achieve fluent and consistent motions, while in the spatial domain, it targets at amplifying the physical amplitudes of the input motion to match those of a professional dance, preserving though the original content of the dance's choreography. In that way, we have designed a two-stage deep framework: a dance-to-music alignment stage, and a dance enhancement stage. The dance-to-music alignment stage estimates the temporal mapping of the amateur dance, so as to match the input music; these estimates are later integrated into the dance enhancement stage to enable temporal warping of the encoded dance content features, which are later decoded to reconstruct a professional dance with the same choreographic content. Figure 3 illustrates the two-stage architecture of our framework, whose details will be described in the following sections.

4.1 Dance-to-Music Alignment Stage

The main premise of the dance-to-music alignment stage is to find the optimal alignment between the input music and dance sequences. Taking into account the highly complex correspondences between a dance motion and music, we propose to use auto-encoders to learn the cross-modal frame-to-frame mapping between the high-level motion and the music features extracted from the raw data.

High-Level Feature Extractor. For the input music signal with T frames, we compute the Mel-Scaled Spectrogram using the well-known *librosa* [53] audio analysis library, depicted as $\mathbf{G} \in \mathbb{R}^{T\times B}$ and B is the number of frequency bins. For the input T-frame dance motion that is depicted by position offset vectors $\mathbf{M} \in \mathbb{R}^{T\times(J-1)\times 3}$, we first calculate the corresponding joint positions, and then estimate the velocities and accelerations of each joint on x, y, z directions per frame, displayed as $\mathbf{K} \in \mathbb{R}^{T\times C}$. Note that C equals to $J \times (3 + 3)$ with $J = 21$ joints in the human skeleton.

Dance-to-Music Alignment Network. The dance-to-music alignment network is composed of two encoders – *MusicEnc* and *MotionEnc* – to map the music feature G and dance motion feature K to the corresponding latent feature sequences f_G and f_K, respectively. The two encoders share the same network

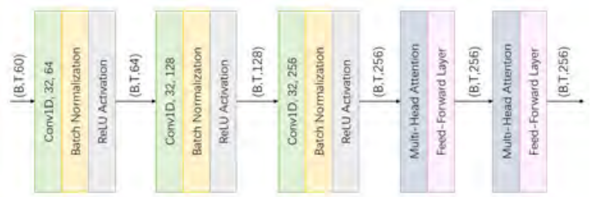


Fig. 4 The network architecture of the *MusicEnc*. The input music feature sequence is processed by three temporal convolution blocks, each containing a 1D-convolution layer, batch normalization and ReLu layer. Then it goes through two transformer blocks containing the multi-head self-attention and feed-forward layers to obtain the encoded feature sequence. The *MotionEnc* and *DanceEnc* share the same network architecture in our implementation.

architecture but with different weights. Following the two encoders, we compute the Euclidean distance between the frames of the two latent feature sequences to form an $T \times T$ affinity matrix, which is defined as:

$$F(i,j) = \|f_G(i) - f_K(j)\|_2^2 \tag{3}$$

where i is the index of the music frame, and j is that of the motion frame. Figure 4 shows the structure of our dance-to-music alignment network.

Specifically, for each encoder, the input sequence is first processed by three temporal 1D-convolution layers sequentially, and each of them is followed by a batch normalization and a ReLU layer. Considering the complex correlations between dance choreographies and their music correspondence, we leverage the attention mechanisms in a transformer network to learn contextualized dance-to-music information, providing adaptive local neighbors for both the dance and music encoders. In particular, we add a shallow transformer with two multi-head self-attention and feed-forward layers on the basis of the 1D-convolution layers, and thus obtain latent feature sequences f_G or f_K encoding temporal context information. Note that the self-attention layers in the transformer are biased towards the local neighbors of each frame by setting the attention mask matrix B_a:

$$B_a(i,j) = \begin{cases} 0, & |i-j| < \delta \\ -\infty, & \text{otherwise} \end{cases} \tag{4}$$

where δ is a parameter to control the neighbor size and equals to 50 in our implementation.

Dynamic Time Warping. The target now is to find the optimal alignment between the input dance and music, so that each dance frame can be matched to the music frame with minimal alignment distance. Under the guidance of the affinity matrix deduced from the dance-to-music alignment network, we perform dynamic programming [54, 55] to obtain an optimal alignment path matrix $\mathbb{W}$ between the latent dance and music features.

4.2 Dance Enhancement Network

The enhancement stage aims to modify amateur dances in such a way that they will look more professional in terms of physical amplitudes and dance-to-music synchronization. To achieve this goal, we leverage an auto-encoder network to modify the non-professional dances in latent feature space. Specifically, the non-professional dance sequence denoted by unit local positions is used to extract corresponding latent features via an encoder – *DanceEnc*. The latent dance features are then temporally warped under the guidance of the optimal dance-to-music alignment path, followed by a decoder to output the corresponding professional dance sequence.

The *DanceEnc* has similar implementation details to the *MusicEnc* and *MotionEnc* network. We warp the encoded feature sequence f_D by calculating the dot-product between the f_D and the alignment path matrix $\mathbb{W}$ obtained from the dynamic time warping module. The decoder is implemented as a three-layer MLP network to project the feature sequence to the final enhanced dance.

4.3 Training and Loss

The two stages in our framework are trained separately with different loss functions. In particular, the dance-to-music alignment network is trained using an *alignment* loss, while the dance enhancement network is trained with a *reconstruction* and a *consistency* loss. In this process, we leverage the optimal alignment path as a condition to modify the latent dance features; we use the ground-truth alignment path as an initial warping condition, and then fine-tune the network with the estimated alignment path. It is important to note here that the length of the input dance motions during training and testing can be arbitrary.

Alignment loss. We assume that the dance sequences and their corresponding music sequence are well-synchronized: a motion frame is well matched with its paired music frame. Therefore, we design the temporal alignment loss on the affinity matrix in a contrastive learning manner. To be more specific, for each music frame, we select the corresponding dance frame as the positive sample and a randomly selected frame as the negative one. Then, we compute the triplet loss on the latent features of the three frames as the alignment loss:

$$\begin{aligned} L_{triplet} = \sum_t^T [\left\| f_G(t) - f_K(\hat{\phi}(t))] \right\|_2^2 - \\ \|f_G(t) - f_K(r)\|_2^2 + a]_+, \end{aligned} \tag{5}$$

where f_G, f_K are the music feature and dance feature respectively, r is a randomly sampled frame index, $\hat{\phi}(t)$ is the index of the corresponding dance frame for the music frame t.

Reconstruction loss. To improve the physical amplitude of the movement in amateur dances to look more professional, we trained our network using paired amateur and professional data; our target is to force the enhanced amateur movements to be as close as possible to their corresponding professional ones. Therefore, we define a reconstruction loss that minimizes the local position error between the enhanced motion and the ground truth, given by the equation:

$$L_{recon} = \sum_{t}^{T} \sum_{j}^{J} |p_{t,j} - \hat{p}_{t,j}|, \tag{6}$$

where $p_{t,j}$ is the local position of joint j at frame t on the enhanced dance motion, and $\hat{p}_{t,j}$ is the corresponding local position on the ground-truth professional dance motion.

Consistency loss. To enforce temporal coherency on the enhanced dance, we introduce a consistency loss by measuring the error between the velocity of the enhanced dance and that of the corresponding ground-truth. Our consistency loss is mathematically described as:

$$L_{cons} = \sum_{t}^{T} \sum_{j}^{J} |v_{t,j} - \hat{v}_{t,j}|, \tag{7}$$

where v and $\hat{v}$ are the velocity of the enhanced and ground-truth dance motion, respectively.

5 Results and Discussion

In this section, we present the dataset used for training and testing in our method, the implementation details, and the evaluation metrics. We also demonstrate the efficiency of our framework in several experiments, a perceptual survey that evaluates its performance in terms of professionalism, realism, and dance-to-music synchronization, and an ablation study. Figure 1 shows a gallery of selected frames extracted from the input amateur motion (yellow), our result (red), and the ground-truth professional dance (blue). It can be observed that our method enhances professionalism to the input motion in such a way that its temporal and spatial features better match those of the ground-truth dance sequence. The quality of our enhanced dance animations may be examined in the supplementary video.

Dataset. The original AIST++ dataset [3] contains 1,408 sequences of 3D human dance motion represented as joint rotations along with root trajectories. Each sequence of dance motion is accompanied with the corresponding music, which is well-synchronized with the animation. Overall, the dataset consists of 10 dance genres with hundreds of different choreographies, providing high richness and varieties of dance contents. We follow the music-choreography data splits used in the original paper [3] for our network training and testing/validation. For each professional music-dance pair in the AIST++ dataset, we produced multiple amateur dance counterparts using our key-pose based dance synthesis algorithm (see Section 3), by controlling the range of their temporal and spatial disturbance factors. In total, we generated $3,680$ non-professional dances for the training, 80 for testing and 80 for validation.

Implementation Details. We implemented our framework in Pytorch and tested it on a 6-core PC with Intel i7 at 3.7GHz, 16GB RAM, and with NVIDIA Tesla P100 GPU. All networks in our framework were trained with a batch size of 64 and learning rate of 10^{-4}, and optimized by the Adam optimizer [56]. In total, it took about 12 training hours for the dance-to-music alignment network and 6 training hours for the dance enhancement network, on 4 NVIDIA Tesla P100 GPUs.

Evaluation Metrics. To the best of our knowledge, there are no quantitative metrics currently available to evaluate the professionalism of dances. Therefore, we used the *temporal alignment error (Time Error)*, the *Pose Error* and the *Fréchet inception distance (FID)*, as the evaluation metrics, and observed the temporal and spatial differences between the input motions, our enhanced dance motions, and the corresponding ground-truth professional dances. The three evaluation metrics are defined as:

- *Temporal alignment error (Time Error)* is the average distance between the indices of motion poses per music frame, in the optimal dance-to-music alignment path and the ground-truth alignment path.
- *Pose Error (PE)* measures the average Euclidean distance between joint positions on the specific poses in two motions sequences to be compared.
- *Fréchet inception distance (FID)* is introduced to measure how far the distribution of the enhanced dance is to that of the ground-truth professional one [4, 57]. We calculate FID based on the extracted kinematic features [3] of the enhanced and ground-truth professional dances.

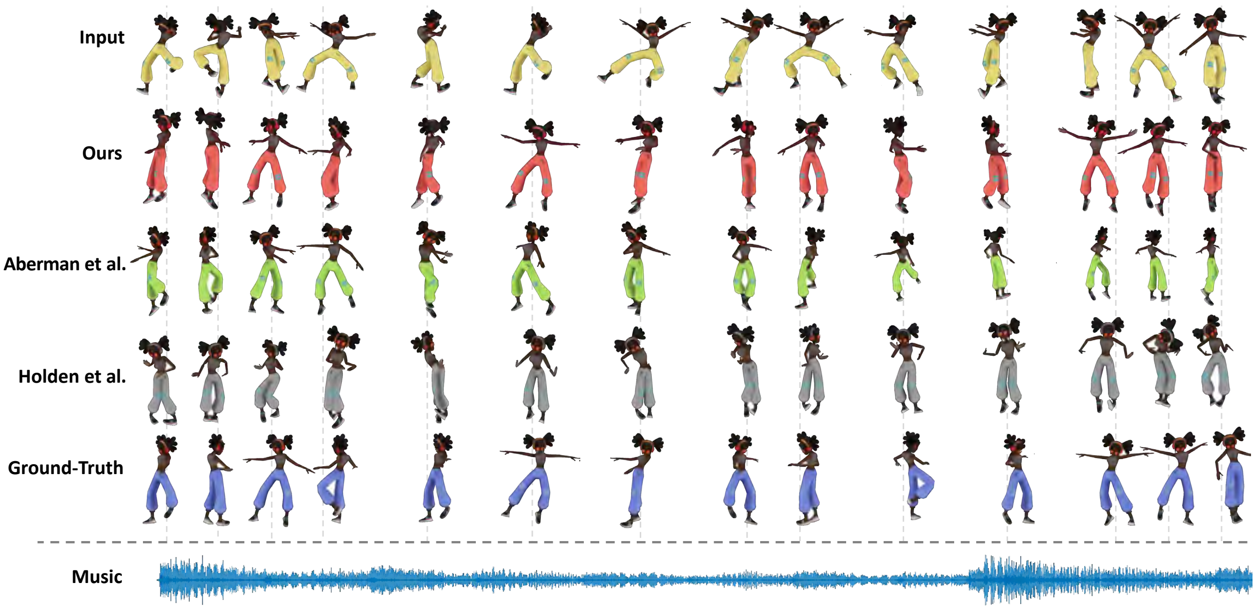


Fig. 5 Qualitative comparison of our method to the baseline methods [7, 8]. Each row shows a set of frames selected from the same music beat. It can be observed that our results are closer to the ground-truth, compared to the two alternatives, in terms of dance-to-music alignment and pose reconstruction. For an animated version, please refer to our supplementary video.

5.1 Evaluation

In this section, we evaluate the performance of our method with two baseline methods – the Holden's *et al.* [7] and the Aberman's *et al.* [8] methods – using the three aforementioned evaluation metrics. In addition, we conducted three perceptual studies to qualitatively evaluate: (a) the quality and realism of our artificially generated amateur dance motions; the quality and realism of our experimental results in enhancing professionalism on amateur movements using (b) our synthetically generated amateur dataset, and (c) real, motion-captured amateur dances. More details about our perceptual study can be found in our supplementary materials.

5.1.1 Comparisons

Baseline Methods. As far as we know, there are no other methods in the literature that deal with the dance enhancement problem. Thus, we compare the results of our approach with two state-of-the-art motion style transfer methods – Holden *et al.* [7] and Aberman *et al.* [8], which also use auto-encoders as the backbone network. Different from our problem, these methods take a content motion and a target style motion as input, and then generate an output motion by preserving the same but desired style of input content with the target motion. Note that, these methods do not take into consideration the music.

To adapt the two baseline methods to our problem, we conducted the following modifications: (1) Since they require motions to have the same length, we down-sample our dataset with the same length (400 frames); (2) We use our synthesized amateur dance together with the accompanying music as the content input, and randomly select another professional dance of the same genre as the target style motion for their network; (3) Note that Aberman's *et al.* [8] network is trained using unpaired motion data with a consistency loss, which equals to minimize a reconstruction error between the input and output content when the input content sequence and the style sequence are within the same style. To make their method applicable to paired motion data, we leverage the consistency loss to calculate the reconstruction error between the output of the network and the ground-truth professional dance. The Holden's *et al.* [7] method is trained with its original loss functions.

Qualitative Comparison. Figure 5 illustrates selected poses from the input dance motion sequence (in yellow), ours (in red), the two alternative methods (Aberman's *et al.* [8] in green, and Holden's *et al.* [7] in gray), and the ground truth (in blue). The music beats [53] are marked with gray dotted lines to indicate the temporal coherence. It can be observed that our method successfully produces good correspondences to the professional dance sequences, with satisfactory temporal alignment and spatial amplitudes. In contrast to our method, the two alternatives are not synchronized to the beat (since they are not designed for aligning dance-to-music), and their reconstructed poses are farther away than ours compared to the ground-truth movement.

Table 1 Quantitative evaluation of our results and the motion style transfer methods on the test set.

Methods	Pose Error↓	FID ↓
Holden *et al.* [7]	4.323	322.327
Aberman *et al.* [8]	3.156	36.456
Ours	**1.951**	**14.817**

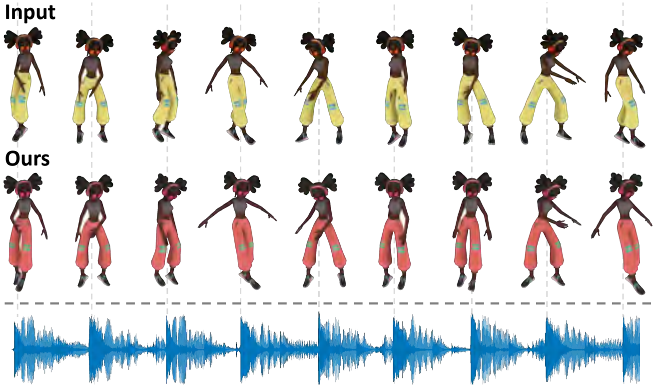


Fig. 6 In this example, we tested our network by setting a dance sequence performed by a professional dancer as input. It can be observed that the output motion remains natural and realistic, and it is similar to the original one.

Quantitative Comparison. For a quantitative evaluation, Table 1 reports the pose error and the FID metric. It can be observed that the quantitative metrics confirm our observations; the two baseline methods produce the worst results in terms of amplitudes compared to our results, having large Pose Error and FID score. However, since they are not designed to compute an explicit temporal alignment between dance and music, we do not consider, in this evaluation, the Time Error metric.

In addition, we use a professional dance sequence as input to the network to further evaluate the naturalness and realism of the output motion. As shown in Figure 6, the results confirm the efficiency of our method in generating natural movements, returning a movement that is realistic and perfectly aligned to the music beat.

5.1.2 Perceptual Study

Evaluation of our synthetic amateur dance dataset. We first conducted a perceptual study to evaluate the quality and realism of our synthetic amateur motions, and whether they can be considered as true amateur dances. For this task, we recruited, in total, 20 participants, 11 of whom were females and 9 were males. Each participant watched 28 pairs of side-by-side dance motions; on the right side, we showed amateur motions, which were either selected from our synthetic dataset (16 samples), or captured by amateur dancers who imitate professional dance moves (12 samples); on the left side of the video, we showed the corresponding ground-truth dance expert motions, so that the participants can take the professional motion as a reference to examine the quality of the amateur and synthetic motions.

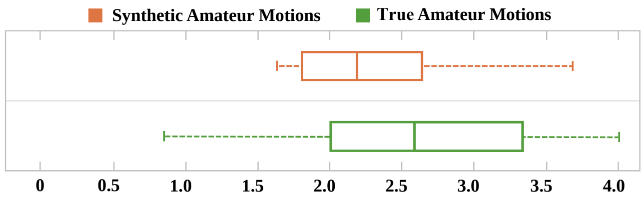
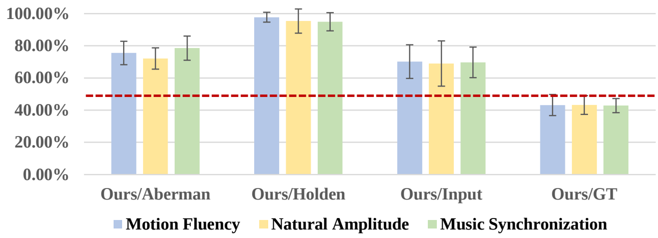


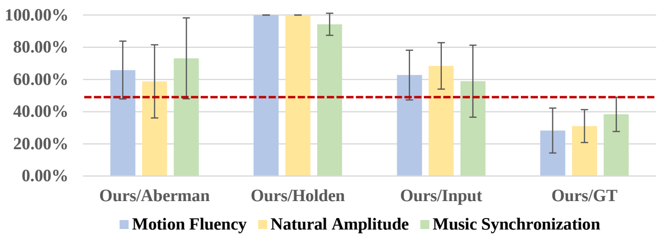
Fig. 7 Average score of evaluating whether the dance motions are captured from amateur dancers or have been algorithmically synthesized by computers. Red boxplot shows the average score of all the synthetic amateur motions, and green boxplot shows that of the true, motion-captured amateur motions.

The participants were asked to rate on a 5-likert scale (0 means that the motion is not performed by an amateur since there is too much computer-generated noise; 2 indicates that is hard to decide; and 4 means that the participant is strongly confident that the motion was performed by an amateur dancer) whether the presented motion on the right side has been captured from an amateur dancer, or generated by a computer algorithm. The scores are statistically analyzed to compare our synthesized motions and the true amateur motions. Figure 7 shows boxplots of the average score for the synthetic amateur motions and true amateur motions. Both cases have an average score between two and three, which indicates that it is hard for participants to discriminate whether the motions are computer-generated or not. However, it is important to mention here that our synthetic dataset may have some differences compared to the true, motion-captured data. Our synthetic amateur motions are generated by randomly setting disturbances in spatial and temporal spaces to imitate the amplitude and music synchronization of amateur dances, thus there may exist some motions with too exaggerated or limited movements. In addition, dances performed by amateurs may have lower consistency of the body parts and contain different choreographies compared with those performed by the expert; these differences have not been considered in our synthetic data generation process. Therefore, as expected, our synthetic amateur motions get a slightly lower score than the true amateur motions.

Professionalism Evaluation on Synthetic Data. We conducted a perceptual survey to evaluate the quality of our results when the synthetic dataset was used. We compared the results of our method with two baselines ([7, 8]), the input, and the ground truth, based on the following three professionalism aspects: (i) the *smoothness and fluency* of the dance motions; (ii) the *naturalism of the dance physical*



(a) Results from amateurs users.



(b) Results from dance expert users

Fig. 8 Professionalism evaluation on synthetic dance motions. Each group of bars indicates the average percentage of participants that voted for our results, with 95% confidence intervals. *Blue* bars show the average percentage of evaluations on the motion fluency; *yellow* bars show the average scores on the natural amplitude; *green* bars show the average scores on the dance-to-music synchronization. Note that, the bars that are higher than the red dotted line indicate that our results were preferred by the majority of the users.

amplitudes; and (iii) *the dance-to-music synchronization*. For this evaluation, we randomly selected seven motions, each of them from a different dance genre.

We recruited 20 participants: 15 of them were amateur dancers (with less than one year of dance experience) and five of them were expert dancers (with more than eight years of dance experience). Each participant was shown, in total, 28 pairs of dance motions; each pair includes one generated from our approach, and the other from the ground-truth dataset or generated using one of the two baselines. For each pair, the participants were asked to select the dance motion that is: (i) smoother and more fluent; (ii) the one that has more natural physical amplitudes; and (iii) the one that is better synchronized with the music, in three independent questions. Note that all the experimental dance motions were randomly ordered to avoid learning effects.

The answers were gathered together to quantify the overall professionalism of the dance motions. The statistics of the perceptual study are shown in Figure 8, which lists the average percentage of participants who preferred our results over the results of the two baseline methods, the input, and the ground-truth. It can be observed that our method received higher score compared to the two baselines and the input for all three aspects of professionalism, in the votes of both the amateur and expert observers. Apart from

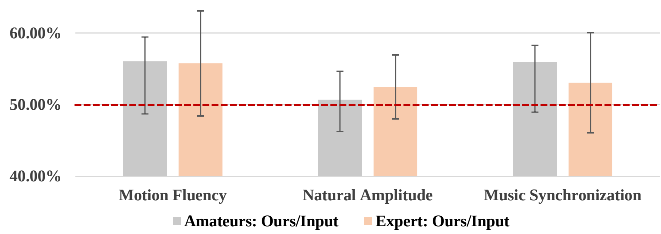


Fig. 9 Professionalism evaluation on true amateur dance motions. Each group of bars indicates the average percentage of participants that voted for our results, with 95% confidence intervals. *Gray* bars show the votes of the amateur users; *orange* bars show the votes of the expert users.

smoother and more natural motion, we believe that the better dance-to-music alignment plays an important role for these results. As expected, both the amateur and dance expert participants gave higher scores to the ground-truth motions, compared to ours.

Professionalism Evaluation on Real Motions. Finally, we used the 12 motion-captured dance sequences performed by amateur dancers, to further evaluate the performance of our method on real amateur data. The motion-captured dances were performed by amateur dancers, who imitated 12 ground-truth professional dances chosen from our testing dataset. In this survey, we recruited 20 participants, of whom 15 were amateur dancers, and five were expert dancers. Similar to the previous study, each participant was shown pairs of true amateur motions and our enhanced results with random ordering, and was asked to select the motion that is: (i) smoother and more fluent; (ii) more natural physical amplitudes; and (iii) better synchronization to the given music. Figure 9 presents the statistics of this study. Compared to the input amateur dances, our enhanced results were preferred by the most amateur and expert participants; note that our results scored significantly higher on the following two professionalism aspects: motion fluency, and music synchronization. As expected, our method performs worse on the true amateur dataset than the synthetic dataset, because our network has been explicitly trained using synthetic amateur data. An interesting future work is to enrich our training dataset with true amateur dances or to better simulate synthetic data so that they can better execute real amateur dance motions.

5.2 Ablation Study

To evaluate the contribution of the dance-to-music alignment stage and the necessity of each of its components, we conducted an ablation study; the study evaluates several variations of the proposed network, by removing or replacing key components with other alternatives. In more details,

Table 2 Comparing the results in the temporal alignment analysis and spatial enhancement analysis.

Methods	Time Error↓	Pose Error ↓	FID↓
Input	24.433	3.081	29.973
W/O Alignment	-	2.479	31.324
W/O DTW	21.177	2.366	20.392
ConvNet	18.547	1.998	13.595
Transformer	18.547	3.371	77,317
Conv.+Trans. (Ours)	18.547	1.951	14.817

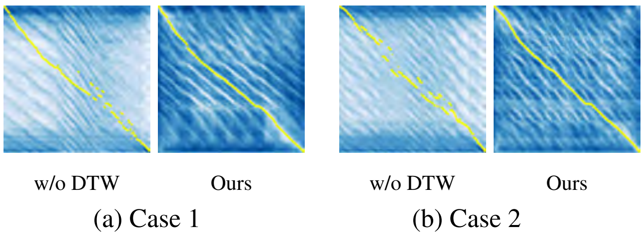


Fig. 10 Visualization of temporal alignment results. The temporal affinity matrix is encoded by a *Blue* colormap with *white* as low values and *blue* as high values. The optimal alignment path is illustrated as a *yellow* curve.

we assessed our network: (i) without integrating the the dynamic time warping module (W/O DTW); and (ii) without combining the dance-to-music alignment stage (W/O Alignment). Table 2 reports the results of the ablations study; for visual comparisons, please refer to our supplementary material.

Temporal-Alignment Stage. In the case that the temporal alignment stage is not integrated into our framework (W/O Alignment), we concatenate the input music and amateur dance, and set them as input to the dance encoder (*DanceEnc*), during the dance enhancement stage. The encoded latent features are directly fed into the decoder without warping. Table 2 lists the results of this setup; we can easily observe the necessity of having the temporal alignment stage, since the network cannot implicitly learn the temporal warping from the convolutional layers.

Dynamic-Time-Warping Component. To evaluate the effect of the dynamic-time-warping component, we use temporal attention mechanisms as an alternative to our learning alignment method. Without the dynamic-time-warping component, we built the optimal temporal alignment path for each music frame via selecting the motion frame with the maximum attention value in their affinity matrix. More details about how we have built and trained the affinity matrix can be found in the supplementary material. The results listed in Table 2 confirm that the performance of the attention-based implementation

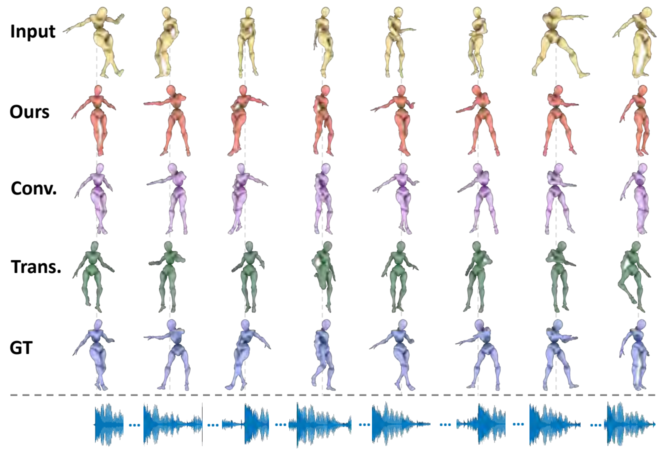


Fig. 11 Ablation study: visual comparison between our final configuration, and the case of using only *ConvNet* and *Transformer*. It can be clearly observed that our final structure produces dance poses that are more similar to the ground-truth, compared to the two alternatives.

(W/O DTW) is worse than the original implementation. To better demonstrate the results, we visualize the alignment results of our method and the attention-based implementation in Figure 10. In each visualization, the blue background illustrates the $T \times T$ temporal affinity matrix (rows show the motion frames, columns show the music frames), which is covered by the optimal alignment path indicated by a *yellow* curve. The *white* color indicates low alignment correspondence between the motion and music frame. It can be observed that the optimal alignment path produced by the W/O DTW is scattered, which means that the aligned poses between neighboring frames may have large changes and cause motion jitters. Different from that, our optimal alignment path is continuous and monotonous. This evaluation validates the necessity to have a separate dynamic time warping component for the temporal alignment.

Dance Enhancement Stage To examine the impact of our deep neural architectures, we implemented the dance enhancement network with two baseline structures – *ConvNet* and *Transformer*. The *ConvNet* encoder is composed of three Conv-BN-ReLU blocks. The encoder of *Transformer* is implemented as a shallow network with two attention-forward blocks, while ours (*ConvNet with a Transformer*) concatenates three Conv-BN-ReLU blocks and the two attention-forward blocks. The decoder for all three structures is implemented as the three-layer MLP. Note that, in this experiment we kept the same parameters, with regard to the dance-to-music alignment stage, for the three structures.

The last three rows in Table 2 show that our network's structure performs better than the *Transformer*. Compared

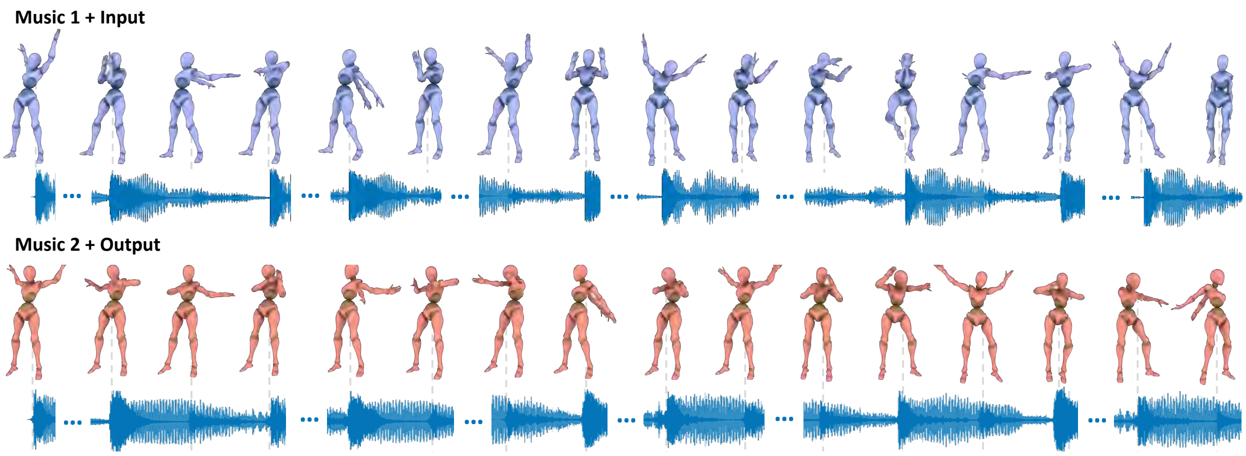


Fig. 12 Synchronizing the same dance to different audio files. The upper row shows the dance and its original music rhythm, while the bottom row shows the synchronized dance motion to a new audio file. Our method is capable of aligning a dance motion on audio files with different rhythms and beats, enabling data reuse. Please refer to our supplementary video for animated results.

with *ConvNet*, our performance is slightly better in pose error, but a bit worse in FID. Since pose error measures the pose similarity to the ground-truth per frame while the FID measures the overall kinematic feature distribution, we believe that the pose error metric is more important to evaluate visual effects. As shown in Figure 11, the enhanced poses produced by our structure are visually closer to the ground-truth, than using *ConvNet* or *Transformer* alone. This evaluation indicates that the convolution layers are essential for encoding dance features with temporal context information.

5.3 Application: Dance-to-Music Synchronization

One of the main features of our method is that it aligns 3D motion data with audio files under non-uniform and irregular misalignment. This feature enables some very interesting applications, where the same dance can be reused in audio files with different beats. Figure 12 shows an example of such dance-to-music synchronization. It can be observed that the input and output dance sequences share similar poses, but are temporally misaligned since they are artificially synchronized to music files played at different beats. To the best of our knowledge, there are no other works in the literature that do dance-to-music synchronization in 3D motion. Our approach enables data reuse, and puts the foundations to facilitate future development in this important application area.

6 Conclusion

In this paper, we have presented a deep learning framework that enhances professionalism on amateur dances, satisfying three main professionalism properties – *fluent dance movements*, *physical amplitude*, and *temporal alignment of dance and music*, without changing the content of the original choreography. The framework consists of a dance-to-music alignment stage and a dance-enhancement-stage, where the first stage learns an optimal temporal alignment path between the input dance and the accompanying music, and the second stage enhances the dance motion from both spatial and temporal domains. We have also presented a key-pose based dance augmentation scheme that artificially generates non-professional dance data, using the AIST++ [3] dataset. We demonstrate the effectiveness of our framework by comparing it with two baseline style transfer methods [7, 8] via a qualitative visual survey, quantitative metrics, and a perceptual study. We also presented a useful application that reuses existing dance motions files by synchronizing them with audio files that have a different rhythm.

Limitations and Future Work. In our study, we mainly focus on two attributes of dance professionalism – extending specific physical amplitude via spatial amplitude enhancement, and making fluent dances via temporal motion-rhythm synchronization. However, dance professionalism is also correlated with some other semantic attributes, such as *smoothness*, *energy*, *balance*, and *aesthetics*. In future work, it would be interesting to investigate those professionalism attributes, and design algorithms that emphasize semantics in dances, e.g., enhancing the *aesthetics* of the input dance. Besides, our framework modifies the input amateur dances based on their original content. No additional constraints have been considered for adding or deleting poses in the original amateur dances. Therefore, when there exist choreographic errors in the input amateur dances, or their choreography is not rich and diverse enough, our method fails to revise it. A possible future direction is to use motion motifs [58] to learn fine-grained mappings between the professional and non-professional dances, and to build a knowledge codebook for the dance enhancement, similar to the concept of [41]. Last but not least, our framework is built on paired professional and synthesized non-professional dance motion dataset. When the input amateur dance contains poses far away from the distribution of dances in the dataset, our method may produce unsatisfactory results. A future work is to enrich the synthesized non-professional dataset with real captured amateur dance motions, or design an unpaired dance enhancement approach by leveraging characteristics of different dance genres. We would also like to experiment with other pose representations, e.g., [59, 60], to avoid the use of IK to restore joint rotations, and avert potential rotation discontinuities caused by the network. Finally, our method can be used to improve e-learning applications e.g., for XR systems when users try to learn dance with a virtual avatar.

Acknowledgements

This research was supported by the grants of NSFC (No. 62072284), the grant of Natural Science Foundation of Shandong Province (No. ZR2021MF102), the Special Project of Shandong Province for Software Engineering (11480004042015), and internal funds from the University of Cyprus. The authors would like to thank Anastasios Yiannakidis (University of Cyprus) for capturing the amateur dances, and the volunteers for attending the perceptual studies. The authors would also like to thank the anonymous reviewers and the editors for their fruitful comments and suggestions.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

Electronic Supplementary Material

We provide a supplementary document describing details about our implementation and the perceptual study. We also provide an accompanying video clip to show visual comparisons of our method and the baseline methods.

References

[1] Hanna JL. The Performer-Audience Connection: Emotion to Metaphor in Dance and Society, 1983, Univ. of Texas Press.

[2] Aristidou A, Shamir A, Chrysanthou Y. Digital Dance Ethnography: Organizing Large Dance Collections. *ACM Journal on Computing and Cultural Heritage*, 2019, 12(4): 29:1–29:27.

[3] Li R, Yang S, Ross DA, Kanazawa A. AI Choreographer: Music Conditioned 3D Dance Generation with AIST++. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021, 13401–13412.

[4] Chen K, Tan Z, Lei J, Zhang SH, Guo YC, Zhang W, Hu SM. ChoreoMaster: Choreography-Oriented Music-Driven Dance Synthesis. *ACM Transactions on Graphics*, 2021, 40(4): 145:1–145:13.

[5] Butterworth J. Dance Studies: The Basics, 2011, Routledge Press.

[6] Holden D, Saito J, Komura T, Joyce T. Learning Motion Manifolds with Convolutional Autoencoders. In *SIGGRAPH Asia 2015 Technical Briefs*, 2015, 1–4.

[7] Holden D, Saito J, Komura T. A Deep Learning Framework for Character Motion Synthesis and Editing. *ACM Transactions on Graphics*, 2016, 35(4): 138:1–138:11.

[8] Aberman K, Weng Y, Lischinski D, Cohen-Or D, Chen B. Unpaired Motion Style Transfer from Video to Animation. *ACM Transactions on Graphics*, 2020, 39(4): 64:1–64:12.

[9] Dong Y, Aristidou A, Shamir A, Mahler M, Jain E. Adult2child: Motion Style Transfer Using CycleGANs. In *Motion, Interaction and Games (MIG)*, 2020, 13:1–13:11.

[10] Wen YH, Yang Z, Fu H, Gao L, Sun Y, Liu YJ. Autoregressive Stylized Motion Synthesis With Generative Flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, 13612–13621.

[11] Koutedakis Y, Sharp NCC. The Fit and Healthy Dancer, 1999, Wiley Press.

[12] Krasnow D, Chatfield SJ. Development of the "Performance Competence Evaluation Measure": Assessing Qualitative Aspects of Dance Performance. *Journal of dance medicine & science*, 2009, 13(4): 101–107.

[13] Neave N, McCarty K, Freynik J, Caplan N, Hönekopp J, Fink B. Male Dance Moves that Catch a Woman’s Eye. *Biology Letters*, 2011, 7(2): 7221–224.

[14] Torrents C, M C, Jofre T, Morey G, Reverter F. Kinematic Parameters that Influence the Aesthetic Perception of Beauty in Contemporary Dance. *Perception*, 2013, 42(3): 447–58.

[15] Park YS. Correlation Analysis between Dance Experience and Smoothness of Dance Movement by Using Three Jerk-Based Quantitative Methods. *Korean Journal of Sport Biomechanics*, 2016, 26(1): 1–9.

[16] Alexiadis DS, Kelly P, Daras P, O’Connor NE, Boubekeur T, Moussa MB. Evaluating a Dancer’s Performance Using Kinect-Based Skeleton Tracking. In *Proc. of the ACM International Conference on Multimedia*, MM ’11, 2011, 659–662.

[17] Raheb KE, Stergiou M, Katifori A, Ioannidis Y. Dance Interactive Learning Systems: A Study on Interaction Workflow and Teaching Approaches. *ACM Computing Surveys*, 2019, 52(3): 50:1–50:37.

[18] Chen HY, Cheng YH, Lo A. Improve Dancing Skills with Motion Capture Systems: Case Study of a Taiwanese High school Dance Class. *Research in Dance Educat.*, 2021: 1–19.

[19] Chan JC, Leung H, Tang JK, Komura T. A Virtual Reality Dance Training System Using Motion Capture Technology. *IEEE Trans. on Learning Technologies*, 2011, 4(2): 187–195.

[20] Aristidou A, Stavrakis E, Charalambous P, Chrysanthou Y, Himona SL. Folk Dance Evaluation Using Laban Movement Analysis. *ACM Journal on Computing and Cultural Heritage*, 2015, 8(4): 20:1–20:19.

[21] Laban R. The mastery of Movement (4 ed.), 2011, Dance Books Ltd.

[22] Tenenbaum J, Freeman W. Separating Style and Content. In *Advances in Neural Information Processing Systems (NIPS)*, volume 9, 1997, 662–668.

[23] Aristidou A, Zeng Q, Stavrakis E, Yin K, Cohen-Or D, Chrysanthou Y, Chen B. Emotion Control of Unstructured Dance Movements. In *Proc. of the ACM SIGGRAPH/EG Symp. on Computer Animation*, SCA '17, 2017, 10:1 – 10:9.

[24] Brand M, Hertzmann A. Style Machines. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 2000, 183–192.

[25] Hsu E, Pulli K, Popović J. Style Translation for Human Motion. *ACM Transactions on Graphics*, 2005, 24(3): 1082–1089.

[26] Xia S, Wang C, Chai J, Hodgins J. Realtime Style Transfer for Unlabeled Heterogeneous Human Motion. *ACM Transactions on Graphics*, 2015, 34(4): 119:1–119:10.

[27] Mason I, Starke S, Zhang H, Bilen H, Komura T. Few-shot Learning of Homogeneous Human Locomotion Styles. *Computer Graphics Forum*, 2018, 37(7): 143–153.

[28] Smith HJ, Cao C, Neff M, Wang Y. Efficient Neural Networks for Real-Time Motion Style Transfer. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 2019, 2(2): 13:1–13:17.

[29] Du H, Herrmann E, Sprenger J, Cheema N, Hosseini S, Fischer K, Slusallek P. Stylistic Locomotion Modeling with Conditional Variational Autoencoder. In *In Proc. of Eurographics - Short Papers*, 2019, 9–12.

[30] Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol PA. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *J. of Machine Learning Research*, 2010, 11(110): 3371–3408.

[31] Gatys L, Ecker A, Bethge M. A Neural Algorithm of Artistic Style. *Journal of Vision*, 2016, 16(12).

[32] Huang X, Belongie S. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, 1510–1519.

[33] Arikan O, Forsyth DA. Interactive Motion Generation from Examples. *ACM Transactions on Graphics*, 2002, 21(3): 483–490.

[34] Kim Th, Park SI, Shin SY. Rhythmic-Motion Synthesis Based on Motion-Beat Analysis. *ACM Transactions on Graphics*, 2003, 22(3): 392–401.

[35] Lee HC, Lee IK. Automatic Synchronization of Background Music and Motion in Computer Animation. *Computer Graphics Forum*, 2005, 24(3): 353–361.

[36] Shiratori T, Nakazawa A, Ikeuchi K. Dancing-to-Music Character Animation. *Computer Graphics Forum*, 2006, 25(3): 449–458.

[37] Tang T, Jia J, Mao H. Dance with Melody: An LSTM-Autoencoder Approach to Music-Oriented Dance Synthesis. In *Proceedings of the 26th ACM International Conference on Multimedia (MM)*, 2018, 1598–1606.

[38] Lee HY, Yang X, Liu MY, chun Wang T, Lu YD, Yang MH, Kautz J. Dancing to Music. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 32, 2019, 1–11.

[39] Tsuchida S, Fukayama S, Hamasaki M, Goto M. AIST Dance Video Database: Multi-genre, Multi-dancer, and Multi-camera Database for Dance Information Processing. In *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, 2019, 501–510.

[40] Zhuang W, Congyi Wang SX, Chai J, Wang Y. Music2Dance: DanceNet for Music-driven Dance Generation. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2022, 18(2).

[41] Aristidou A, Yiannakidis A, Aberman K, Cohen-Or D, Shamir A, Chrysanthou Y. Rhythm is a Dancer: Music-Driven Motion Synthesis with Global Structure. *IEEE Trans. Visual. & Comput. Graph.*, 2022, Early Access.

[42] Tadamura K, Nakamae E. Synchronizing Computer Graphics Animation and Audio. *IEEE MultiMedia*, 1998, 5(4): 63–73.

[43] Cardle M, Barthe L, Brooks S, Robinson P. Music-driven Motion Editing: Local Motion Transformations Guided by Music Analysis. In *Proceedings 20th Eurographics UK Conference*, 2002, 38–44.

[44] Laichuthai A, Kanongchaiyo P. Synchronization Between Motion and Music Using Motion Graph. In *The 8th Electrical Engineering/ Electronics, Computer, Telecommunications and Information Technology (ECTI) Conference*, 2011, 496–499.

[45] Davis A, Agrawala M. Visual Rhythm and Beat. *ACM Transactions on Graphics*, 2018, 37(4).

[46] Bellini R, Kleiman Y, Cohen-Or D. Dance to the Beat: Synchronizing Motion to Audio. *Computational Visual Media*, 2018, 4(3).

[47] Chung JS, Zisserman A. Out of Time: Automated Lip Sync in the Wild. In *Asian Conference on Computer Vision*, 2016, 251–263.

[48] Halperin T, Ephrat A, Peleg S. Dynamic Temporal Alignment of Speech to Lips. In *IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, 3980–3984.

[49] Wang J, Fang Z, Zhao H. AlignNet: A Unifying Approach to Audio-Visual Alignment. In *Proc, of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020, 3309–3317.

[50] Phillips GM. Interpolation and Approximation by Polynomials, 2003, springer Science & Business Media.

[51] Holden D, Komura T, Saito J. Phase-Functioned Neural Networks for Character Control. *ACM Transactions on Graphics*, 2017, 36(4).

[52] Aristidou A, Lasenby J, Chrysanthou Y, Shamir A. Inverse Kinematics Techniques in Computer Graphics: a Survey. *Computer Graphics Forum*, 2018, 37(6): 35–58.

[53] McFee B, Raffel C, Liang D, Ellis DP, McVicar M, Battenberg E, Nieto O. librosa: Audio and Music Signal Analysis in Python. In *Proceedings of the 14th Python in Science Conference*, volume 8, 2015, 18–24.

[54] Sakoe H, Chiba S. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1978, 26(1): 43–49.

[55] Rabiner L, Juang BH. Fundamentals of Speech Recognition, 1993, prentice-Hall, Inc.

[56] Daugman JG. Uncertainty Relation for Resolution in Space, Spatial Frequency, and Orientation Optimized by Two-dimensional Visual Cortical Filters. *Journal of the Optical Society of America A*, 1985, 2(7): 1160–1169.

[57] Dowson D, Landau B. The Fréchet Distance between Multivariate Normal Distributions. *Journal of Multivariate Analysis*, 1982, 12(3): 450–455.

[58] Aristidou A, Cohen-Or D, Hodgins JK, Chrysanthou Y, Shamir A. Deep Motifs and Motion Signatures. *ACM Transactions on Graphics*, 2018, 37(6): 187:1–187:13.

[59] Zhou Y, Barnes C, Lu J, Yang J, Li H. On the Continuity of Rotation Representations in Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, 5745–5753.

[60] Andreou N, Lazarou A, Aristidou A, Chrysanthou Y. A Hierarchy-Aware Pose Representation for Deep Character Animation. *CoRR*, 2021, abs/2111.13907.

Author biography

Qiu Zhou is a postgraduate in the School of Computer Science and Technology at Shandong University. She received BSc from Shandong University in 2019. Her main interests are motion analysis and synthesis. You can contact her at: zhouqiulv@gmail.com

Manyi Li is an Associate Researcher in the School of Software at Shandong University. She received BSc and PhD degrees from Shandong University in 2013 and 2018 respectively and was a postdoc fellow at the GrUVi Lab, Simon Fraser University during 2019-2021. Her main interests are 3D content creation and understanding. You can contact her at: manyili@sdu.edu.cn.

Qiong Zeng is an Associate Researcher in the School of Computer Science and Technology at Shandong University. She received BSc and PhD degrees from Nanchang University and Shandong University in 2010 and 2015 respectively. Her main interests are focused on motion analysis and visualization. You can contact her at: qiong.zn@sdu.edu.cn.

Andreas Aristidou is an Assistant Professor at the Department of Computer Science, University of Cyprus. He had been a Cambridge European Trust fellow at the University of Cambridge, where he obtained his PhD. Andreas received his BSc from the National and Kapodistrian University of Athens and he is an MSc honor graduate of Kings College London. His main research interests are focused in the areas of computer graphics and character animation. You can contact him at: a.aristidou@ieee.org.

Xiaojing Zhang is an undergraduate student at Taishan College of Shandong University. She entered the university in 2019. Her main interests are focused on computer graphics and visualization. You can contact her at: xj.zhang@mail.sdu.edu.cn.

Lin Chen is an Associate Professor of the Qingdao Institute of Humanities and Social Sciences, Shandong University. She received her doctorate at the Freie Universität Berlin. Her research interests include the aesthetic ideas of Baumgarten and their far-reaching influence, theatre and dance research, and cultural studies.You can contact her at: chenlin.spring@gmail.com.

Changhe Tu is a Professor in the School of Computer Science and Technology, Shandong University, China. He received his BSc, MEng, and PhD degrees from Shandong University, Jinan, P.R. China, in 1990, 1993, and 2003, respectively. His research interests are in the areas of computer graphics and robotics. You can contact him at: chtu@sdu.edu.cn.