Dance Style Transfer with Cross-modal Transformer

Wenjie Yin*, Hang Yin*, Kim Baraka†, Danica Kragic*, and Mårten Björkman*

*KTH Royal Institute of Technology, Stockholm, Sweden
†Vrije Universiteit Amsterdam, Amsterdam, Netherlands
yinw@kth.se, hyin@kth.se, k.baraka@vu.nl, dani@kth.se, celle@kth.se

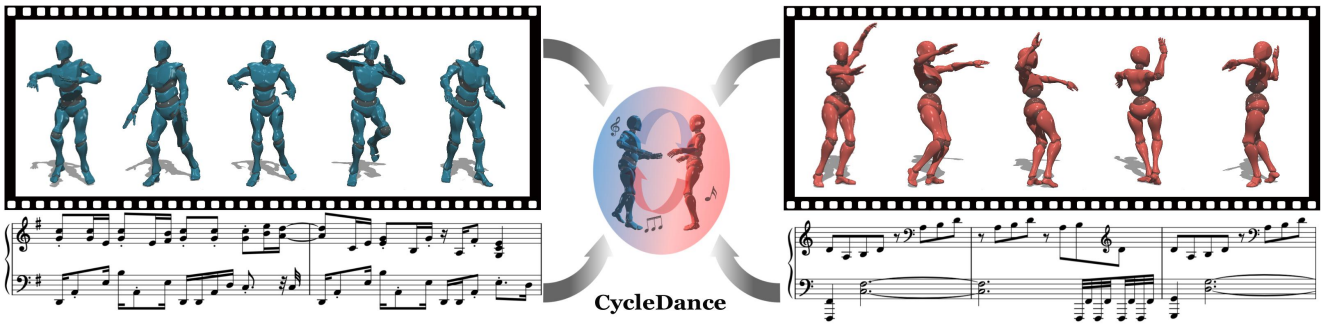

Figure 1: Dance style transferred by CycleDance between two dance styles: left) locking dance and right) ballet-jazz dance. The CycleDance framework is trained with unpaired dance motion together with music context.

Abstract

We present CycleDance, a dance style transfer system to transform an existing motion clip in one dance style to a motion clip in another dance style while attempting to preserve motion context of the dance. Our method extends an existing CycleGAN architecture for modeling audio sequences and integrates multimodal transformer encoders to account for music context. We adopt sequence length-based curriculum learning to stabilize training. Our approach captures rich and long-term intra-relations between motion frames, which is a common challenge in motion transfer and synthesis work. We further introduce new metrics for gauging transfer strength and content preservation in the context of dance movements. We perform an extensive ablation study as well as a human study including 30 participants with 5 or more years of dance experience. The results demonstrate that CycleDance generates realistic movements with the target style, significantly outperforming the baseline CycleGAN on naturalness, transfer strength, and content preservation. [1]

[1] Demo at https://youtu.be/kP4DBp8OUCk.

1. Introduction

Style transfer methods facilitate art creation of a target style for media such as images [16] and music [4]. Similar methods are promising for creators to use an existing dance sequence as a starting point to generate variations across different movement styles. In a video game context, these style variations may -e.g. be associated to different characters with different attributes or personalities. In a choreographic context, such a tool may lead to hybrid human-artificial creative processes, where style transfers are used to iterate on interesting, unexpected, or complementary variations of an initial choreographic material.

Existing research on transferring human movement styles largely focuses on simple locomotive or exercise motions [30, 12, 1] and domain transfers between adults and children. Technical methods for transferring such sequential data include cycle-consistent adversarial networks (CycleGAN) [46] and adaptive instance normalization (AdaIN) [24]. However, a research gap remains for applying similar techniques to enable style transfer of more complex movements, such as dance movements. Dance movements usually have no explicit functional purpose and tend to exhibit considerable richness in posture, rhythm and their compo-

sition. Generation of dance movements can be particularly challenging since it demands a multi-layer approach that captures motion qualities such as the coordination of joint dynamics and socio-cultural factors associated with the production and perception of the movement. Meanwhile, there exist a variety of such characteristics within different dance styles, originating from different historical backgrounds. Dance styles could be thought more generally of as styles of performing certain dance movements rather than strictly dance genres. This adds another layer of complexity to the generation of high-quality dance movements of a specific target style. All these challenges call for computational models that can capture both high-frequency features and long-term dependencies over time, and as such generate realistic dance with aesthetic and coherence.

Moreover, dance is commonly accompanied by music which can provide tremendous clues for understanding and composing movement. Recent works have shown the effectiveness of music-conditioned dance synthesis [39, 6], which can directly generate dance motion given music context. However, it is unclear whether music context will also facilitate style transfer tasks and how such a multi-modal input should be processed in this context.

In this paper, we propose CycleDance, a multimodal system (see Figure 1) for dance style transfer. CycleDance adopts a generative scheme by extending CycleGAN-VC2 [23] to work with unpaired data. To tackle the challenges identified above, we exploit a cross-modal transformer architecture [39] that aims to effectively capture relevant features among different modalities so as to enhance style transfer quality. Specifically, we design a two pathways transformer-based architecture to extract temporally aligned motion and music representations in the context of style transfer. We further propose to train CycleDance progressively with a curriculum learning scheme inspired by Fu et al. [14]. This alleviates issues of instability in training large adversarial models and premature convergence that can lead to inferior performance. We evaluate our framework on the AIST++ [37] dance database, with the analysis focused on transfer between various dance genres. Two new metrics based on probabilistic divergence and selected key pose frames are proposed to quantitatively assess the quality of dance style transfer. Moreoever, we invite a group of human participants with extensive dancing experience to provide a subjective evaluation and insights from an expert perspective. These evaluations show that CycleDance greatly outperforms a baseline method and its ablative versions. As an illustration, a video with generated examples can be found at https://youtu.be/kP4DBp8OUCk.

In summary, our contributions are mainly as follows:

- Our approach is, to the best of our knowledge, the first to combine complex dance motion and music context in the style transfer task, unlocking potential applications in choreography, gaming, and animation, as well as in tool development for artistic and scientific innovations in the field of dance.
- We introduce new metrics based on probabilistic divergence and selected key pose frames for gauging transfer strength and content preservation in the context of dance movements.
- We provide an extensive user study of the proposed model. The evaluations and insights from a group of experienced dance performers reveal essential aspects of designing such systems.

2. Related Work

In this section, we first provide an overview of prior works on general style transfer in Section 2.1 and focus on motion style transfer in Section 2.2. As another relevant topic, motion synthesis from multi-modal data will be briefly reviewed in Section 2.3.

2.1. Style Transfer

In recent years, style transfer has achieved impressive progress in computer vision, speech processing, music processing, natural language processing, motion animation, etc. In computer vision, the pioneering work of Gatys et al. [16] introduces the concept of style transfer and uses convolutional neural networks (CNNs) to merge the style and content between arbitrary images. Huang et al. [21] propose an adaptive instance normalization (AdaIN) layer to directly apply arbitrary target styles to an image. Zhu et al. [46] propose CycleGAN, using a pair of generators and discriminators to translate image style. The general idea of CycleGAN has been further developed and improved in StarGAN [7], with domain labels as additional input, so that image styles can be transformed to multiple corresponding domains, such as facial appearances and expressions.

In research on voice conversion (VC), Kaneko and Kameoka [22] develop CycleGAN-VC based on CycleGAN, but with gated CNNs and an identity-mapping loss. This is further improved by CycleGAN-VC2 [23] which adopts two-step adversarial losses, a 2-1-2D convolution structure, and PatchGAN. Fu et al. [14] further incorporate transformers and curriculum learning in voice conversion. Research has also been conducted to transfer symbolic music styles, with examples such as Groove2Groove [8], which employs an encoder-decoder structure and parallel data, and [4] for MIDI music with a CycleGAN-based approach.

For style transfer in natural language processing (NLP), Mueller et al. [31] propose recurrent variational autoencoders (VAE) to revise text sequences. Fu et al. [15] construct a multi-decoder and a style-embedding model to learn independent content and style representations with adversarial networks. Dai et al. [9] propose a Style Transformer

network with a special training scheme, which employs an attention mechanism and makes no assumption about the latent representation.

Our work focuses on transferring motion data, in particular, dance movements. We adopt CycleGAN-VC2, previously used for voice conversion, as the basis of our framework and augment training with an extra music modality.

2.2. Motion Style Transfer

Early works on motion style transfer rely on handcrafted features [2, 38, 42, 3, 20], while most modern studies advocate learning by extracting features from data [19, 12, 18, 35, 27, 29, 33, 11]. Typical models in use include convolutional auto-encoders [19], CycleGAN [11], temporal invariant AdaIN layers [1], autoregressive flows [41] and spatial-temporal graph neural networks [33]. Some research also concerns efficient generation for real-time style transfer [43, 35, 29]. All these works target relatively simple human movements, such as locomotion and exercise, for which the variation in style is often limited e.g. transferring between adult and children locomotion [11].

Our work handles transfer of dance movements that exhibit substantial richness in terms of postures, rhythms, transitions and artistic styles and as such may be of greater empirical value for e.g. video game or film industries. To handle such complexities, our method significantly differs from the reviewed work, with transformer and curriculum learning leveraged on top of CycleGAN-VC2 for more effective training on more complex movement data.

2.3. Music-conditioned Motion Synthesis

A plethora of research works have focused on human motion synthesis [5, 44, 17, 45, 28]. Since dance is often combined with music, cross-modal motion generation, an emerging research topic that explores the association between different modalities, is often explored for better understanding of human motion and music-conditioned motion synthesis. Most early works focus on statistical models [34, 13, 25] and typically generate motions by selection. To be specific, this means synthesizing motion by selecting the motion segments in a database whose features (such as rhythm, structure, and intensity) match each music segment. With the development of deep learning, learning-based methods have also been explored.

For example, in ChoreoMaster [6], an embedding module is designed to capture music-dance connections. Sun et al. [36] propose DeepDance, a cross-modal association system, which correlates dance motion with music, and Lee et al. [47] a decomposition-to-composition framework that leverages MM-GAN for music based dance unit organization. In DanceNet [47], a musical context aware encoder fuses music and motion features, while in DanceFormer [26] kinematics enhanced transformer guided networks are used for motion curve regression. More recently, cross-modal transformers have been successfully applied to model the distribution between music and motion [39].

Music-conditioned dance synthesis aims to generate dance motion sequences associated to a given music context. Our work explores the dance style transfer task, focusing on manipulating the style of the existing dance movements while keeping the contextual information. The music modality is not mandatory for our style transfer model to work, but can be incorporated to benefit the generation quality when the data is available.

3. Methodology

This section formulates our target problem and establishes notations used throughout the paper. Preliminaries about CycleGAN and CycleGAN-VC2 are also given for self-containment. On the basis of these, we present the contributed technical framework CycleDance.

3.1. Problem Formulation

Our goal is to learn mapping functions between two domains X and Y without relying on paired data between these domains. In our scenario, we transfer dance between two style domains X and Y given dance sample $x \sim P_X$ and $y \sim P_Y$. The dance samples may be paired with music $m_x \in M_x$ and $m_y \in M_y$ with associated styles, although the music modality is only optional in the transfer task.

3.2. Adversarial Training Loss and Strategy

We address the formulated problem with a CycleGAN-like architecture [46], as illustrated in Figure 3. The architecture includes two discriminators D_X and D_Y which are used to distinguish the real and generated data, as well as two mappings $G_{X\to Y}$ and $G_{Y\to X}$ for generating patterns of the target style. The mappings are also cycled such that the generated patterns can be converted back to the original domains. To this end, we follow CycleGAN-VC2 [23] and incorporate four types of losses, also see Figure 3.

Adversarial loss $\mathcal{L}_{adv}^{X\to Y}$: this loss measures the discrepancy between the transferred data $G_{X\to Y}(x, m_x)$ and the target y, with the discriminator D_Y attempts to distinguish the transferred data from real data:

$$\begin{aligned}\mathcal{L}_{adv}^{X\to Y} &= \mathbb{E}_{y\sim P_Y}[\log D_Y(y)] \\ &\quad + \mathbb{E}_{x\sim P_X}[\log(1 - D_Y(G_{X\to Y}(x, m_x)))].\end{aligned} \tag{1}$$

Correspondingly, the adversarial loss $\mathcal{L}_{adv}^{Y\to X}$ can be defined for $G_{Y\to X}$ and discriminator D_X.

Cycle-consistency loss $\mathcal{L}_{cyc}$: this accounts for the loss of contextual information by recovering the original x and y from generated patterns through $G_{X\to Y}(x, m_x)$ and

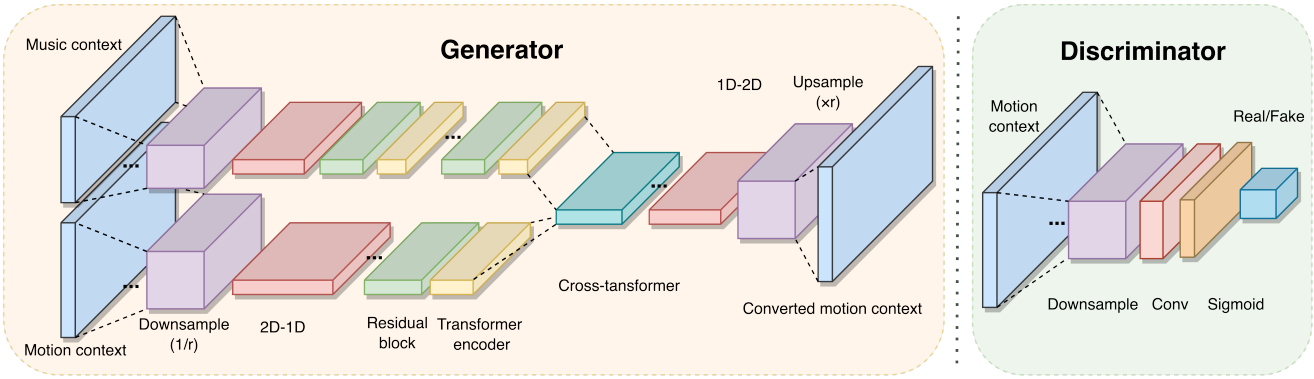

Figure 2: The CycleDance architecture. In the generator, there is a motion pathway and a music pathway. Each pathway starts with downsampling blocks, followed by a 2D-1D block. The motion, music and cross-modal transformer blocks are standard full-attention transformer encoders. Then the fused path is followed by a 1D-2D block and upsampling blocks. In the discriminator, like in Kaneko et al. [23], convolution at used in the last layer.

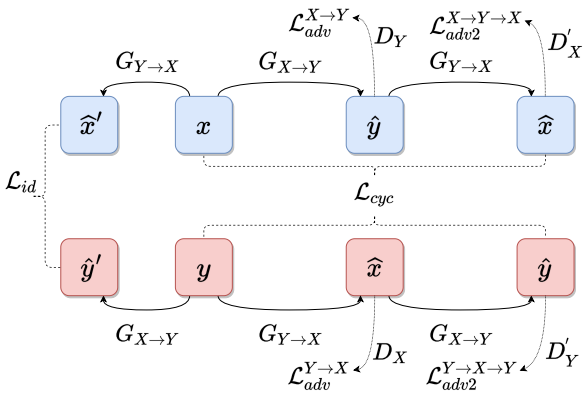

Figure 3: The two step adversarial generative training strategy. The full objective includes four types of losses; adversarial loss $\mathcal{L}_{adv}$, cycle-consistency loss $\mathcal{L}_{cyc}$, identity-mapping loss $\mathcal{L}_{id}$, and second adversarial loss $\mathcal{L}_{adv2}$. See Section 3.2 for the definition of notations.

$G_{Y\to X}(y, m_y)$:

$$\begin{aligned}\mathcal{L}_{cyc} &= \mathbb{E}_{x\sim P_X}[\|G_{Y\to X}(G_{X\to Y}(x, m_x)) - x\|_1] \\ &+ \mathbb{E}_{y\sim P_Y}[\|G_{X\to Y}(G_{Y\to X}(y, m_y)) - y\|_1].\end{aligned} \tag{2}$$

Identity-mapping loss $\mathcal{L}_{id}$: this further encourages input preservation by enforcing an identity transformation when $G_{X\to Y}$ and $G_{Y\to X}$ are applied to the other domain:

$$\begin{aligned}\mathcal{L}_{id} &= \mathbb{E}_{x\sim P_X}[\|G_{Y\to X}(x, m_x) - x\|_1] \\ &+ \mathbb{E}_{y\sim P_Y}[\|G_{X\to Y}(y, m_y) - y\|_1]\end{aligned} \tag{3}$$

Two-step adversarial loss $\mathcal{L}_{adv2}$: this is a second adversarial loss to alleviate the over-smoothing reconstruction statistics in the cycle-consistency loss [23]. Note this introduces an additional discriminator D'_X and $\mathcal{L}_{adv2}^{Y\to X\to Y}$ can be defined in a similar way:

$$\begin{aligned}\mathcal{L}_{adv2}^{X\to Y\to X} &= \mathbb{E}_{x\sim P_X}[\log D'_X(x)] \\ &+ \mathbb{E}_{x\sim P_X}[\log(1 - D'_X(G_{Y\to X}(G_{X\to Y}(x, m_x))))]\end{aligned} \tag{4}$$

The overall objective is finally written as a weighted sum of the above terms

$$\begin{aligned}\mathcal{L}_{full} &= \mathcal{L}_{adv}^{X\to Y} + \mathcal{L}_{adv}^{Y\to X} + \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_{id}\mathcal{L}_{id} \\ &+ \mathcal{L}_{adv2}^{X\to Y\to X} + \mathcal{L}_{adv2}^{Y\to X\to Y},\end{aligned} \tag{5}$$

where λ_{cyc} and λ_{id} trade off the consistency and identity loss terms.

Besides, we adopt a curriculum learning algorithm as the training scheme. The intuition is that the training can be more effective by starting with simpler data and progressively handle more complex data. Such a strategy has been applied to various applications and scenarios, showing an ability to improve the convergence rate and generalization capacity, and providing better training stability [40]. We adopt a length-based curriculum learning strategy by training data truncation, similar to [14]. The length of input sequences is increased gradually to allow the model to learn from short samples to long samples.

3.3. Network Architecture

Our CycleDance framework adopts CycleGAN-VC2 as the backbone and extends it with a cross-modal transformer, as depicted in Figure 2. The cross-modal transformer concatenates two pathways of motion and music encodings, both of which are obtained through a sequence of layers including 2D convolution (purple blocks in Figure 2), 2D-1D reshaping (red), residual convolution (green) and modality-specific transformers (yellow). The 2D CNN layers are used to perform downsampling while preserve the original sequential structure. The downsampled features are reshaped

and pass through the residual blocks of 1D CNNs. The reshaped 1D sequences are processed by transformers which adopt a position embedding and output encodings capturing temporal relations among timesteps. Finally, the generator takes the concatenated encodings and feed them into a 1D-2D reshape block (red) and a upsampling block (purple) for synthesizing transferred dance motions. In these blocks, we adopt gated linear units (GLUs) [10] as a tunable activation function to learn a sequential and hierarchical structure.

For the discriminator, CycleDance also first downsamples motion data with a 2D CNN. We only use convolution at the last layer to alleviate training instability, as is suggested in [23]. The output layer uses sigmoid activation to predict whether the motion clip is real or generated.

4. Experiments and Evaluations

In this section, we first describe the used dataset (Section 4.1), how they are processed and the concrete experimental setup. We then detail our assessments, including both objective (Section 4.3) and subjective (Section 4.4) evaluations, and report their results on benchmarking different dance style transfer methods and ablations.

4.1. Dataset

We generate 3D dance motion samples with paired music from an existing database called AIST++ Dance Database [27]. AIST++ reconstructs 3D motions from multi-view videos in the AIST Dance Database in terms of SMPL parameters [37]. To obtain motion features, we downsample all the motion data to 30 frames per second (fps) and retarget the motion to a skeleton with 21 body joints using Autodesk MotionBuilder. Similarly to [39], we use exponential map parametrization of the 3D rotation to represent all the joints (non-root). The root joint (hip) has four extra features representing the vertical root position, the ground-projected position changes, and the 2D facing angle changes. In total, each motion frame of dance sequences is represented by a 63-dim vector. The music features are extracted with the Librosa toolbox in a similar way to [37]. We combine 20-dim MFCC, 12-dim chroma, 1-dim one-hot peaks, and 1-dim one-hot beats, resulting in a total 35-dim audio feature.

The selected dance styles are ballet-jazz, locking, waacking, hip-hop, pop and house dance. As the example of data statistics, the ballet-jazz set and the locking dance set both contain 141 motion sequences with 6 songs, lasting 1910.8 seconds and 1898.5 seconds respectively.

4.2. Baseline Models and Ablations

We implement the proposed CycleDance model and the CycleGAN-VC2 baseline with PyTorch and train both models on the preprocessed dataset. In order to assess the significance of contributed design choices, such as the cross-modal transformer and curriculum learning strategy, three alternative architectures are also implemented for an ablation study. The first ablated configuration is CycleTransGAN, which removes the music pathway and the cross-modal transformer and disables the curriculum learning strategy. We expect to use this comparison to highlight the utility of the introduced transformer architecture. The second ablation, CycleTransGAN+CL, applies curriculum learning to CycleTransGAN. We aim to assess the performance gains by meticulously modulating the complexities of samples that the model is exposed to during training. The final ablation, CycleCrossTransGAN, also uses cross-modal transformers for motion and music information as the encoder. Curriculum learning is, however, not adopted in this configuration. We aim to see the impact of having cross-modal transformers by analyzing the differences between CycleTransGAN and CycleCrossTransGAN.

4.3. Objective Evaluation

The main task of all these models is to transfer dance style from a source to a target dance style. To allow for thorough quality assessment of complex motion patterns, common in dance, we perform evaluations from both objective and subjective perspectives. In the objective evaluation, we use 17 dance sequences per style. We transferred the style for each ablated model and evaluated two metrics that capture how well the style is transferred and how well the content is preserved. To this end, we design metrics based on the Fréchet distance, similar to [39].

Transfer strength. The most important aspect of style transfer is transfer strength, which measures the degree of conversion from the source style to the target style. To assess the transfer strength, for one dance style, we use the Fréchet distance between the true dance motion and the generated dance motion. Specifically, we use two consecutive raw poses (x_{i-1}, x_i) to convert the representations of both true and generated motions to joint velocity v_i, without normalization. Similarly, we use three consecutive poses (x_{i-1}, x_i, x_{i+1}) to calculate the joint acceleration a_i. We call this measure the motion Fréchet distance (MFD) and use it to measure how close the generated motion is to the true motion of a target style.

Content preservation. Another indispensable evaluation metric for style transfer is content preservation. For this dimension, for one dance movements, we use the Fréchet distance between distributions of key poses x_k. Keyframes containing such poses are extracted by detecting local maxima in joint acceleration. To make frames comparable, skeleton poses in keyframes are normalized to a hip-centric origin. We call this measure the pose Fréchet distance (PFD) and evaluate to what extent these salient poses are kept after the transfer.

Table 1 summarises the quantitative results of the pro-

Table 1: **Quantitative objective evaluation:** Motion Fréchet distance (MFD) and pose Fréchet distance (PFD) for the five competing models, includes the baseline model, our proposed CycleDance, as well as the ablations. BJ2LC denotes transferring from ballet-jazz to locking dance. Correspondingly, LC2BJ denotes transferring from locking dance to ballet-jazz. Similarly, WK, HP, PO, HO denotes waacking, hip-hop, pop, and house dance.

Method	MFD						PFD					
	BJ2LC	LC2BJ	WK2HP	HP2WK	PO2HO	HO2PO	BJ2LC	LC2BJ	WK2HP	HP2WK	PO2HO	HO2PO
CycleGAN-VC2	9.9430	3.4063	1.4354	1.2645	2.2841	1.9515	0.4897	0.3499	0.4847	0.3313	0.5212	0.5625
CycleTransGAN	3.5643	0.7886	1.0564	0.9464	1.5515	1.5354	0.4749	0.2501	0.4754	0.2834	0.4048	0.5215
CycleTransGAN+CL	2.9188	0.5848	1.0847	0.9847	1.4852	1.5521	0.4897	0.2543	0.4644	0.2882	0.4125	0.4185
CycleCrossTransGAN	2.7446	0.5819	0.9872	1.0782	1.4254	1.5251	0.4419	0.2244	0.4490	0.2880	0.3841	0.4126
CycleDance	2.6109	0.5755	0.8752	0.9501	1.3452	1.4855	0.4216	0.2230	0.4485	0.2960	0.3954	0.3827

posed model and ablations. We transfer the dance style between three pairs of dance genres in both directions, including 'ballet-jazz to locking dance' (BJ2LC) and 'locking dance to ballet-jazz' (LC2BJ), 'waacking to hip-hop dance' (WK2HP) and 'hip-hop to waacking dance' (HP2WK), as well as 'pop to house dance' (PO2HO) and 'house to pop dance' (HO2PO).

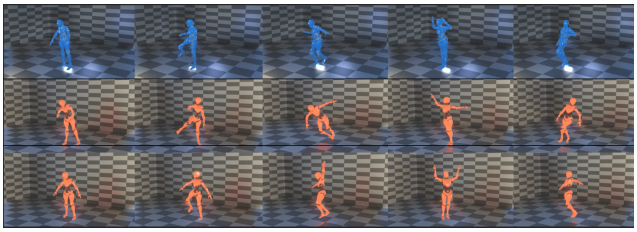

Figure 4: Example locking dance sequences (top, blue y-bot) transferred to ballet-jazz dance by CycleGAN-VC2 (mid, red x-bot) and CycleDance (bottom, red x-bot).

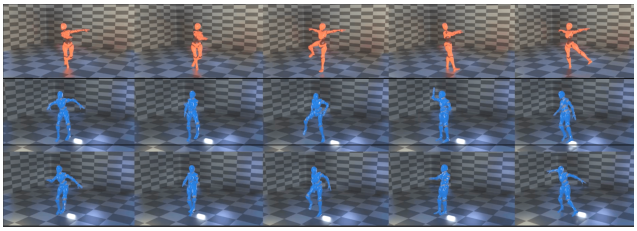

Figure 5: Example ballet-jazz dance sequences (top, red x-bot) transferred to locking dance by CycleGAN-VC2 (mid, blue y-bot) and CycleDance (bottom, blue y-bot).

We observe that the baseline model CycleGAN-VC2 struggles in this style transfer task, evident from the much higher MFD for the baseline model than for all other ablation methods. The complete framework, CycleDance, achieves the best performance on both metrics and almost all transfer pairs. This highlights the necessity of all introduced techniques in this task.

An example of synthesized motion clip is presented in Figure 4, which illustrates dance style transfer from locking to ballet-jazz dance. The top keyframe sequence shows the original locking dance. The sequence in the middle is generated by CycleGAN-VC2 and the bottom one by the proposed CycleDance. Another example in Figure 5 shows dance style transfer from ballet-jazz to locking dance. By comparing the poses of each column, it can be observed that the extracted key gestures are representative to the pose sequences. CycleDance has a higher similarity to the source gestures and can thus preserve more content while having better alignment to the target dance style.

In addition, through the ablation study, we observe that CycleTransGAN (CycleGAN-VC2 and transformer combined) achieves lower MFD, which can be seen as, with the help of the transformer, the model benefits from capturing richer intra-relations among frames. By comparing CycleTransGAN and CycleCrosTransGAN, both MFD and PFD are improved. We take this as evidence that the music information facilitates accurate generation of the target style and that this context information is successfully encoded by the cross-modal transformer. The comparison between of CycleTransGAN and CycleTransGAN+CL reveals that curriculum learning greatly improves transfer strength, showing the effectiveness of gradually increasing the difficulty by training with longer clips.

4.4. Subjective Evaluation

In addition to the objective evaluation, we conducted a user study to evaluate our model and the baseline by scoring three aspects: motion naturalness, transfer strength, and content preservation. We also ask some open-ended questions to gather opinions that may not be covered by the above aspects, to provide suggestions for future work.

Our analysis mainly focuses on ballet-jazz and locking dance, since the characterics of these are well understood by dance professionals. The user study was performed through an online survey covering transfer tasks for both 'locking dance to ballet-jazz' and 'ballet-jazz to locking dance'. We used Blender to render 8-second video clips with an x-bot character (for ballet-jazz) and a y-bot character (for lock-

ing dance) for each source and target dance sequence. The participants could play video clips and get acquainted with the animated dance in an introduction phase. In the actual survey, the participants were asked to watch a source dance video clip and a corresponding generated target dance clip. The target dance video clip was generated either from CycleDance or from the baseline, and the order of target dance clips was randomly selected and balanced to relieve potential order effects. The participants could repeatedly play the clips before answering three questions:

- **Motion naturalness**: *To what extent do you agree with the following statement? — The generated motion clip looks natural after style transfer.* (Likert item ranging from 1 (strongly disagree) to 5 (strongly agree)).

- **Transfer strength**: *To what extent do you agree with the following statement? — The generated motion clip looks like the target dance style.* (Likert item ranging from 1 (strongly disagree) to 5 (strongly agree)).

- **Content preservation**: *Which feature(s) do you think is (are) the most preserved between the original and the result video? — Orientation through space; — Shapes of the limbs; — Shape of the body trunk; — Rhythmic patterns — Other: ___.* (One or more of these four aspects could be selected). This list was based on the most salient features that dance analysts look at when analyzing expressive movement [32].

In the study, 30 participants with at least 5 (cumulative) years of dance experience (including training, performing, choreographing, or teaching) were recruited. Participants were between 20 and 41 years of age (median 30), 37.9% male, 58.6% female, and 3.4% others. According to the demographic questions, the participants' familiarity with the ballet-jazz dance and locking dance were M=3.93 (SD=1.05) and M=3.03 (SD=1.18) respectively, where 1 meant not at all and 5 meant very familiar. Since the generated motions were shown using virtual characters, we also counted the frequency at which participants played video games, which were 34.5% weekly, 13.8% monthly, 13.8% yearly, and 37.9% rarely.

We analyzed the subjective responses to provide statistical support for the results of the user study, and assessed whether the proposed method could be further improved. Figure 6 demonstrates the responses of the motion naturalness and transfer strength. On both aspects, the experts rated CycleDance higher on average compared to the baseline model. The subjective responses were compared through a Wilcoxon signed-rank test statistical significance. Both the median value of motion naturalness ($Z = -9.2262, p < 0.0001$) and transfer strength ($Z = -8.7677, p < 0.0001$) were significantly higher for CycleDance compared to the baseline model. Thus from the view of dance experts, CycleDance is favoured for improved naturalness as well as similarity to the target dance style, consistent with what we observe from the objective quantitative results (Section 4.3). As for the responses of the content preservation, Figure 7 summarizes the total statistics on the four aspects queried. On all four aspects, the experts chose CycleDance more often than the baseline model, when asked which specific features they believe are preserved. We ran a McNemar test for assessing the statistical significance of these gaps. The test revealed no significant

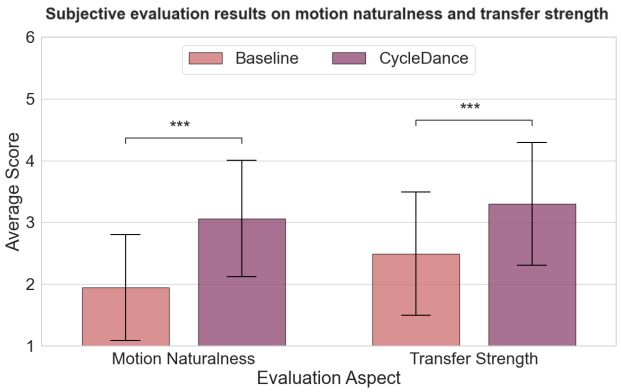

Figure 6: Subjective evaluation results on motion naturalness and transfer strength. Error bars represent standard errors of the averages. Statistical significance is the result of the Wilcoxon signed-rank test that compares the medians ($***$ means $p < 0.0001$).

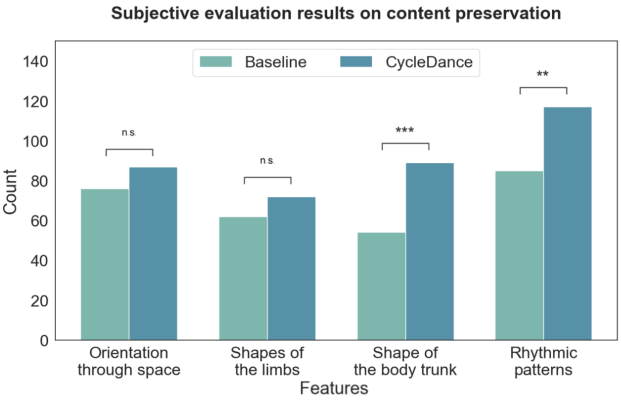

Figure 7: Subjective evaluation results on content preservation. The CycleDance outperforms the baseline model on orientation through space, shape of the limbs, shape of the body trunk, and rhythmic patterns. Statistical significance represents the results of the Wilcoxon signed-rank test that compares the medians ($***$ means $p < 0.00001$, $**$ means $p < 0.0001$, and $n.s.$ means $p > 0.05$).

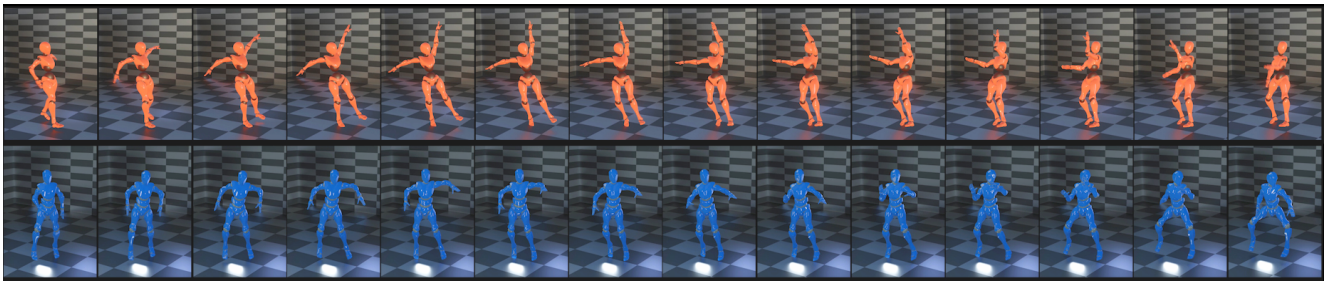

Figure 8: Example ballet-jazz dance (top, red x-bot) transferred to locking dance (bottom, blue y-bot) by CycleDance.

statistical differences between the Baseline model and CycleDance on 'Orientation through space' ($p = 0.1724$) and 'Shapes of the limbs' ($p = 0.1573$). In terms of median value of the 'Shapes of the body trunk' ($p = 0.000002$) and 'Rhythmic patterns' ($p = 0.00004$), the McNemar test showed a strong significance in support of the proposed CycleDance model. Among the four aspects, both CycleDance and the baseline received higher scores on rhythmic patterns and orientation through space. This implies that it is comparatively easier to keep dance orientation and rhythm while performing dance style transfer. CycleDance outperforms the baseline on preserving the shape of the body trunk. Preserving the shape of the limbs, on the other hand, appears to be more challenging.

Responding to open-ended questions, the dance experts commented that for 'ballet-jazz to locking dance', both methods have a jerky style that emulates pop and lock dance. The example shown in Figure 8 is frequently mentioned as a major indication of 'transfer' with a visible locking dance style from the view of experienced dancers. For CycleDance samples of transferring 'locking dance to ballet-jazz', the dance experts responded that the character arms are clearly jazz or ballet and are really good at holding 'traditional' shapes. The dance experts also commented on some limitations. One commonly mentioned point is that some motions look wobbly, which may indicate the need for applying some filters to smooth the generated results. The experts also pointed out that ballet-jazz usually requires dancers to point their feet while the generated motions always show flexed ankle joints. This shows the limitation of the considered data which currently do not capture fine foot movements. This caveat also causes some physically unrealistic effects such as the character appears floating when sometimes its body does not have a contact point on the floor.

5. Discussion of Societal Impact

This work contributes a framework for style transfer that aims to offer artistic and scientific innovations to the field of dance. In the short term, we could foresee several impacts on industries and society. The positive effect would be the progress in choreographic practice and dance research, which unlock new possibilities in terms of hybrid human-artificial co-creation of dance material. Certain industry sectors could benefit as well, such as video game and animation industries (e.g., group dances where each character has a different motion style). Such effects could lead to displacement of jobs and a shift towards jobs that relies more on a combination of creativity and automation, as well as development of new user-friendly interfaces and tools. We also foresee potentially negative impacts or misuses. This technology could blur the lines of ownership in creative processes, i.e., who is/are the creator(s). For movement styles beyond dance, such transfer models, if trained on non-representative datasets, could reinforce movement stereotypes of certain societal groups by learning a biased association between group membership and movement styles, e.g. elderly people or people with disabilities.

6. Conclusion and Future work

This work explores challenging style transfer for sequential data with rich variations and complex frame dependencies such as dance movements. The proposed CycleDance manages to alleviate these challenges by exploiting expressive data encoders, cross-modal contexts and a curriculum based training scheme. Quantitative results from similarity metrics and human expert evaluations confirm the effectiveness of CycleDance. To the authors' knowledge, this is the first work where music context is used for dance or general motion style transfer. In the future, we plan to extend the backbone from the CycleGAN-based model to StarGAN or an AdaIN-based model to handle more than two dance genres. Research is also needed to address identified limitations on preserving limb shapes. Based on these techniques, we envision new tools in dance motion design for choreography, film industry, and video games.

Acknowledgements

This research has received funding from the EC Horizon 2020 research and innovation program under grant agreement n. 824160 (EnTimeMent).

References

[1] Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Unpaired motion style transfer from video to animation. *ACM Transactions on Graphics (TOG)*, 39(4):64–1, 2020.

[2] Kenji Amaya, Armin Bruderlin, and Tom Calvert. Emotion from motion. In *Graphics interface*, volume 96, pages 222–229. Toronto, Canada, 1996.

[3] Andreas Aristidou, Qiong Zeng, Efstathios Stavrakis, KangKang Yin, Daniel Cohen-Or, Yiorgos Chrysanthou, and Baoquan Chen. Emotion control of unstructured dance movements. In *Proceedings of the ACM SIGGRAPH/Eurographics symposium on computer animation*, pages 1–10, 2017.

[4] Gino Brunner, Yuyi Wang, Roger Wattenhofer, and Sumu Zhao. Symbolic music genre transfer with cyclegan. In *2018 ieee 30th international conference on tools with artificial intelligence (ictai)*, pages 786–793. IEEE, 2018.

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

[6] Kang Chen, Zhipeng Tan, Jin Lei, Song-Hai Zhang, Yuan-Chen Guo, Weidong Zhang, and Shi-Min Hu. Choreomaster: choreography-oriented music-driven dance synthesis. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021.

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

[8] Ondřej Cífka, Umut Şimşekli, and Gaël Richard. Groove2groove: One-shot music style transfer with supervision from synthetic data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2638–2650, 2020.

[9] Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. Style transformer: Unpaired text style transfer without disentangled latent representation. *arXiv preprint arXiv:1905.05621*, 2019.

[10] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pages 933–941. PMLR, 2017.

[11] Yuzhu Dong, Andreas Aristidou, Ariel Shamir, Moshe Mahler, and Eakta Jain. Adult2child: Motion style transfer using cyclegans. In *Motion, Interaction and Games*, pages 1–11. 2020.

[12] Han Du, Erik Herrmann, Janis Sprenger, Noshaba Cheema, Somayeh Hosseini, Klaus Fischer, and Philipp Slusallek. Stylistic locomotion modeling with conditional variational autoencoder. In *Eurographics (Short Papers)*, pages 9–12, 2019.

[13] Rukun Fan, Songhua Xu, and Weidong Geng. Example-based automatic music-driven conventional dance motion synthesis. *IEEE transactions on visualization and computer graphics*, 18(3):501–515, 2011.

[14] Changzeng Fu, Chaoran Liu, Carlos Toshinori Ishi, and Hiroshi Ishiguro. Cycletransgan-evc: A cyclegan-based emotional voice conversion model with transformer. *arXiv preprint arXiv:2111.15159*, 2021.

[15] Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[16] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

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

[18] Daniel Holden, Ikhsanul Habibie, Ikuo Kusajima, and Taku Komura. Fast neural style transfer for motion data. *IEEE computer graphics and applications*, 37(4):42–49, 2017.

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

[20] Eugene Hsu, Kari Pulli, and Jovan Popović. Style translation for human motion. In *ACM SIGGRAPH 2005 Papers*, pages 1082–1089. 2005.

[21] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.

[22] Takuhiro Kaneko and Hirokazu Kameoka. Parallel-data-free voice conversion using cycle-consistent adversarial networks. *arXiv preprint arXiv:1711.11293*, 2017.

[23] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6820–6824. IEEE, 2019.

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

[25] Minho Lee, Kyogu Lee, and Jaeheung Park. Music similarity-based approach to generating dance motion sequence. *Multimedia tools and applications*, 62(3):895–912, 2013.

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

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

[28] Zimo Li, Yi Zhou, Shuangjiu Xiao, Chong He, Zeng Huang, and Hao Li. Auto-conditioned recurrent networks for ex-

tended complex human motion synthesis. *arXiv preprint arXiv:1707.05363*, 2017.

[29] Ian Mason, Sebastian Starke, and Taku Komura. Real-time style modelling of human locomotion via feature-wise transformations and local motion phases. *arXiv preprint arXiv:2201.04439*, 2022.

[30] Ian Mason, Sebastian Starke, He Zhang, Hakan Bilen, and Taku Komura. Few-shot learning of homogeneous human locomotion styles. In *Computer Graphics Forum*, volume 37, pages 143–153. Wiley Online Library, 2018.

[31] Jonas Mueller, David Gifford, and Tommi Jaakkola. Sequence to better sequence: continuous revision of combinatorial structures. In *International Conference on Machine Learning*, pages 2536–2544. PMLR, 2017.

[32] Jean Newlove and John Dalby. *Laban for all*. Taylor & Francis US, 2004.

[33] Soomin Park, Deok-Kyeong Jang, and Sung-Hee Lee. Diverse motion stylization for multiple style domains via spatial-temporal graph-based generative model. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 4(3):1–17, 2021.

[34] Takaaki Shiratori, Atsushi Nakazawa, and Katsushi Ikeuchi. Dancing-to-music character animation. In *Computer Graphics Forum*, volume 25, pages 449–458. Wiley Online Library, 2006.

[35] Harrison Jesse Smith, Chen Cao, Michael Neff, and Yingying Wang. Efficient neural networks for real-time motion style transfer. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 2(2):1–17, 2019.

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

[37] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *ISMIR*, volume 1, page 6, 2019.

[38] Munetoshi Unuma, Ken Anjyo, and Ryozo Takeuchi. Fourier principles for emotion-based human figure animation. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 91–96, 1995.

[39] Guillermo Valle-Pérez, Gustav Eje Henter, Jonas Beskow, Andre Holzapfel, Pierre-Yves Oudeyer, and Simon Alexanderson. Transflower: probabilistic autoregressive dance generation with multimodal attention. *ACM Transactions on Graphics (TOG)*, 40(6):1–14, 2021.

[40] Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[41] Yu-Hui Wen, Zhipeng Yang, Hongbo Fu, Lin Gao, Yanan Sun, and Yong-Jin Liu. Autoregressive stylized motion synthesis with generative flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13612–13621, 2021.

[42] Andrew Witkin and Zoran Popovic. Motion warping. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 105–108, 1995.

[43] Shihong Xia, Congyi Wang, Jinxiang Chai, and Jessica Hodgins. Realtime style transfer for unlabeled heterogeneous human motion. *ACM Transactions on Graphics (TOG)*, 34(4):1–10, 2015.

[44] Sijie Yan, Zhizhong Li, Yuanjun Xiong, Huahan Yan, and Dahua Lin. Convolutional sequence generation for skeleton-based action synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4394–4402, 2019.

[45] Wenjie Yin, Hang Yin, Danica Kragic, and Mårten Björkman. Graph-based normalizing flow for human motion generation and reconstruction. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pages 641–648. IEEE, 2021.

[46] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

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