Machine Vision and Applications (2023) 34:48
https://doi.org/10.1007/s00138-023-01399-x

ORIGINAL PAPER

Multimodal dance style transfer

Wenjie Yin[1] · Hang Yin[2] · Kim Baraka[3] · Danica Kragic[1] · Mårten Björkman[1]

Received: 20 March 2023 / Revised: 20 March 2023 / Accepted: 11 April 2023

Abstract

This paper first presents CycleDance, a novel dance style transfer system that transforms an existing motion clip in one dance style into a motion clip in another dance style while attempting to preserve the motion context of the dance. CycleDance extends existing CycleGAN architectures with multimodal transformer encoders to account for the music context. We adopt a sequence length-based curriculum learning strategy 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. Building upon CycleDance, we further propose StarDance, which enables many-to-many mappings across different styles using a single generator network. Additionally, we introduce new metrics for gauging transfer strength and content preservation in the context of dance movements. To evaluate the performance of our approach, we perform an extensive ablation study and a human study with 30 participants, each with 5 or more years of dance experience. Our experimental results show that our approach can generate realistic movements with the target style, outperforming the baseline CycleGAN and its variants on naturalness, transfer strength, and content preservation. Our proposed approach has potential applications in choreography, gaming, animation, and tool development for artistic and scientific innovations in the field of dance.

Keywords Style transfer · Dance motion · Multimodal learning · Generative models

1 Introduction

Style transfer methods facilitate and streamline the art creation process for media such as images [1] and music [2]. Similar techniques in the field of dance show promise for enabling creators, such as choreographers and dancers, to generate variations across different movement styles, leveraging an existing dance sequence as a starting point. In a video game context, these style variations may e.g. be associated with different characters with unique attributes or personalities. In a choreographic context, style transfer may lead to hybrid human-artificial creative processes that combine human and artificial intelligence, where choreographers can use a tool to iterate over interesting, unexpected, or complementary variations of the initial choreographic material. For example, a choreographer could use style transfer to generate variations of a particular dance sequence and then select the most interesting variations to incorporate into the final choreography.

Existing studies on human movement style transfer primarily focus on simple locomotive or exercise motions [3, 4] and domain transfers between adults and children [5]. Technical methods including cycle-consistent adversarial networks (CycleGAN) [6] and adaptive instance normalization (AdaIN) [7] have been employed to transfer such sequential data. However, a research gap remains in applying these techniques to enable style transfer for more complex movements, particularly in the domain of dance. Dance movements usually have no explicit functional purpose and tend to exhibit a considerable richness in posture, rhythm, and their composition. Consequently, generating dance movements can be particularly challenging, as it requires a

✉ Wenjie Yin
yinw@kth.se

Hang Yin
hayi@di.ku.dk

Kim Baraka
k.baraka@vu.nl

Danica Kragic
dani@kth.se

Mårten Björkman
celle@kth.se

[1] KTH Royal Institute of Technology, Stockholm, Sweden

[2] University of Copenhagen, Copenhagen, Denmark

[3] Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

Published online: 09 May 2023

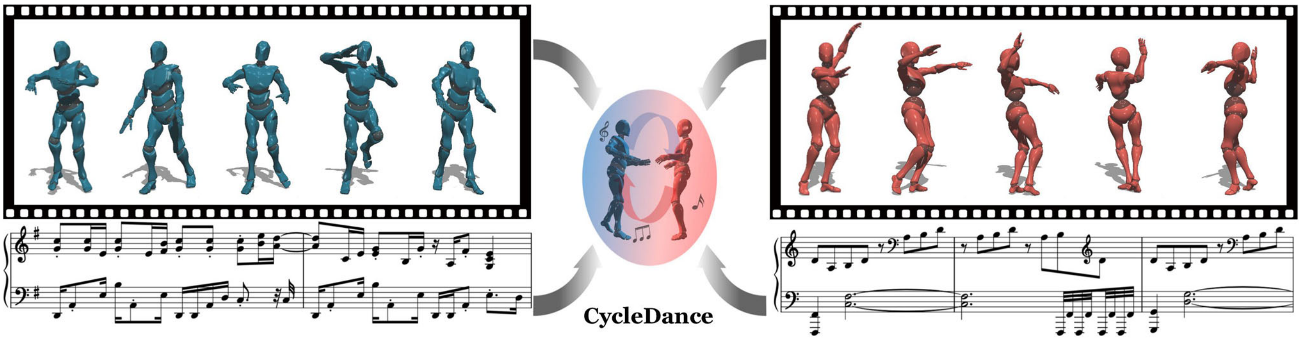

Fig. 1 Dance style transfer using CycleDance between two different dance styles: locking dance and ballet-jazz dance. The CycleDance framework is trained with unpaired dance motion and music context, enabling it to generate realistic dance movements faithful to the target style

multi-layer approach that captures the coordination of joint dynamics and the socio-cultural factors associated with the production and perception of the movement. At the same time, diverse characteristics can be found within different dance styles, stemming from distinct historical origins. Dance styles could be thought of more generally 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.

In addition, dance is commonly accompanied by music, which can provide tremendous clues for understanding and composing movement. Recent studies have investigated the effectiveness of music-conditioned dance synthesis, which can generate dance motion directly from musical context [8, 9]. However, it is unclear whether music context will also facilitate style transfer tasks. Further research is needed to determine how multi-modal input should be processed in the context of dance style transfer, and to identify the most effective methods for incorporating music.

In this paper, we present CycleDance, a multimodal system for dance style transfer (see Fig. 1). CycleDance adopts a generative approach by extending CycleGAN-VC2 [10] to work with unpaired data. To facilitate high-quality style transfer, we leverage a cross-modal transformer architecture [8] that effectively captures relevant features across different modalities. Specifically, we design a two-pathway transformer-based architecture to extract temporally aligned motion and music representations in the context of style transfer. A progressive curriculum learning scheme inspired by Fu et al. [11] is adopted to mitigate instability and premature convergence in training large adversarial models. To assess the quality of dance style transfer quantitatively, two new metrics based on probabilistic divergence and selected key pose frames are proposed. We also sought subjective evaluations and insights from a group of human participants with extensive experience in dance, providing quantitative analysis based on valuable experts' feedback. Our evaluations show that our proposed approaches outperform the baseline method and its ablative versions, achieving significant improvements in both the proposed metrics and subjective evaluations. The qualitative examples can be accessed at the following URL: https://youtu.be/kP4DBp8OUCk.

This paper is an extended version of our conference paper [12], where the notion of CycleDance was introduced for the first time. In the present paper, we give a more detailed presentation of the method design of CycleDance and related research. In addition, we propose a new model StarDance as the multi-domain extension of CycleDance for transfer between more than two styles. CycleGAN-based methods often require training of $k\,(k-1)$ generators to learn all mappings among k domains, resulting in slow training and limited generalization to new domains [13, 14]. To address these issues, StarDance is designed as a many-to-many dance style transfer method to effectively capture the complexities of dance movements and styles, following a similar idea in the image transfer domain [13, 14]. StarDance uses a single generator that is conditioned on both music context and style attribute, independent of the number of styles and hence alleviating the scalability issue of original CycleDance. New results on StarDance show that the new framework enables transfer styles among multiple dance styles without adding more models (Fig. 2).

In summary, our main contributions are as follows:

- To the best of our knowledge, CycleDance is the first approach 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.
- For evaluation, 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 also provide an extensive user study of the proposed models. The evaluations and insights from a group of experienced dance performers reveal essential aspects of designing future systems for dance style transfer.

2 Related work

In this section, we first provide an overview of prior works on general style transfer in Sect. 2.1, including image, audio, and text style transfer. Then, we focus on motion style transfer in Sect. 2.2. We also review motion synthesis from multi-modal data in Sect. 2.3.

2.1 Style transfer

In recent years, style transfer has achieved impressive progress across various fields, including computer vision, speech and music processing, natural language processing, motion animation, etc.

In the field of computer vision, the pioneering work of Gatys et al. [1] introduces the concept of style transfer and leverages the hierarchical layers in convolutional neural networks (CNNs) to extract both the underlying content structures and stylistic elements. They utilize an optimization-based technique to transfer styles between arbitrary images. Later, Li et al. [15] propose whitening and coloring transform (WCTs) to stylize images by analyzing second-order correlations of content and style features. More generally, Huang et al. [16] introduce an adaptive instance normalization (AdaIN) layer to solve the challenge of arbitrary target style application in image style transfer, which has been broadly adopted to fuse the style and content information for image generation and image-to-image translation [17–19]. Zhu et al. [6] introduce CycleGAN which uses a pair of generators and discriminators to learn the mapping between two unpaired image domains. This general idea has been further developed in StarGAN [13], which incorporates domain labels as additional input and enables image style transfer among multiple corresponding domains, such as facial appearances and expressions.

Voice conversion (VC) refers to a technique of converting non-linguistic or para-linguistic information from the original speech into the desired target speech while retaining the linguistic content unchanged. While some early VC frameworks have achieved success [20, 21], they rely on precisely aligned parallel data of source and target speech. To address this challenge, researchers have turned to non-parallel VC techniques. For example, Hsu et al. [22] construct a VC system from non-parallel speech with variational autoencoders and Wasserstein GANs. Kameoka et al. [23] build an auxiliary classifier VAE with information-theoretic regularization for the model training. Kaneko and Kameoka [24] propose CycleGAN-VC, which is a variation of the CycleGAN architecture using gated CNNs and an identity-mapping loss. This was later improved by CycleGAN-VC2 [10] with the addition of a 2-1-2D convolution structure and two-step adversarial losses to improve performance. This approach has also been extended to a StarGAN-based architecture to enable many-to-many mappings across different domains [14, 25]. Fu et al. [11] incorporated transformers and curriculum learning in voice conversion to facilitate training efficiency.

With the development of MIDI format parsing, research has also been carried out to transfer symbolic music styles, as demonstrated by studies such as Groove2Groove [26], which employs an encoder-decoder architecture and parallel data, and Malik et al. [27] that introduce StyleNet with a shared GenreNet, which aimed to learn various styles for music translation. Brunner et al. [2] use a CycleGAN-based approach for MIDI music. Ding et al. [28] design Steely-GAN, a symbolic-domain transfer approach that combines both pixel-level and latent-level features. Regarding style transfer in natural language processing (NLP), Mueller et al. [29] propose recurrent variational auto-encoders (VAE) to modify text sequences. Fu et al. [30] develop a multi-decoder and style-embedding model using adversarial networks to learn content and style representations. Dai et al. [31] propose a Style Transformer network with a tailored training scheme that integrates an attention mechanism and makes a latent representation-agnostic assumption. Finally, Xu et al. [32] introduce a cycled reinforcement learning approach focusing on unpaired sentiment-to-sentiment translation.

Our research focuses on transferring motion data, specifically dance movements. We use the CycleGAN-VC2 backbone, originally designed for voice conversion, as our foundation. To improve scalability, we extend the model to a StarGAN-based framework. We also augment an additional music modality in our approach to improve the training performance.

2.2 Motion style transfer

Motion style transfer has been a longstanding challenge in the field of computer animation, which involves transferring the motion style of a source animation to a target animation while preserving the key content, such as its structure, timing, spatial relationships, etc. Prior research in motion style transfer relied on handcrafted features [33–38]. Since style is a challenging attribute to define precisely, most modern studies advocate data-driven approaches for feature extractions [4, 39–45]. Commonly used models for style transfer include K nearest neighbors (KNNs) [46], convolutional auto-encoders [39], temporal invariant AdaIN layers [5], CycleGAN [45], spatial-temporal graph neural networks [44], and autoregressive flows [47]. Furthermore, certain studies focus on the issue of efficient real-time style transfer [41, 43, 46]. However, it should be noted that all these studies focus on relatively simple human movements, such as exercise and locomotion, where the stylistic variation is often limited. For example, the transfer between children and adult locomotion [45]. In contrast, our work deals with the transfer of dance movements that possess a significant level of complexity in terms of postures, transitions, rhythms, and artistic styles. Consequently, our research may have more empirical and practical value for video games or film industries. Given the intricacies involved, our method differs significantly from the reviewed research. We utilize transformer and curriculum learning on top of CycleGAN-VC2 to enable more effective training on more complex motion data.

Another important task that accompanies the transfer of motion styles is to evaluate the quality of the synthesized animation. While subjective surveys help estimate movement quality, such as recruiting a group of dance experts with defined requirements, relying on them for evaluation can be expensive, time-consuming, and have low reproducibility [38]. Utilizing objective metrics for quantitative evaluation eliminates the need for human involvement, avoiding the issues associated with subjective surveys. The Fréchet Inception Distance (FID) metric [48], which has proven effective in assessing synthesized images in computer vision, has become a standard in evaluating image generative models. Building on the success of FID, Wang et al. [49] extended the FID concept to motion data. Yoon et al. [50] defined the Fréchet Gesture Distance (FGD) as a metric to evaluate speech-driven gesture generation based on the distance between gesture feature distributions. Maiorca et al. [38] transform motions into image representations and introduced the Fréchet Motion Distance (FMD) to assess the quality and diversity of synthesized motion. Valle-Pérez et al. [8] evaluated the realism of music-based dance generation by measuring the Fréchet distance between the distributions of poses and movements. For the motion style transfer task, we propose a Fréchet Pose Distance (FPD) based on the distribution of key poses to assess the content preservation, as well as a Fréchet Motion Distance (FMD), the Fréchet distance between the distribution of the true dance motion and the generated dance motion to evaluate transfer strength.

2.3 Music-conditioned motion synthesis

Numerous studies have focused on human motion synthesis, utilizing various techniques such as deep feedforward networks [51], convolutional networks [52], recurrent models [53], graph neural networks [54], and autoencoders [55]. Dance and music are often intertwined, leading to an emerging research topic known as cross-modal motion generation. This field aims to understand the association between different modalities better and improve music-conditioned motion synthesis. Early works in cross-modal motion generation mostly focused on statistical models [56–58]. Specifically, these models typically generate motions by selecting preexisting dance moves that match particular music features, such as rhythm, intensity, and structure. With the recent advances in deep learning and the availability of large-scale datasets, learning-based methods have been developed to learn the patterns between music and motion. For example, ChoreoMaster [9] propose an embedding module to capture music-dance connections, while in DeepDance [59], a cross-modal association system is designed to correlate dance motion with music. Lee et al. [60] propose a decomposition-to-composition framework that leverages MM-GAN for music-based dance unit organization. The DanceNet model, as proposed in [60], uses a musical context-aware encoder to fuse music and motion features. In DanceFormer [61], kinematics-enhanced transformer-guided networks are utilized to perform motion curve regression. In a recent work by Valle-Pérez et al. [8], cross-modal transformers were successfully employed to model the relationship between music and motion distributions.

Music-conditioned dance synthesis refers to the task of generating dance motion sequences that are synchronized with a given musical context. In contrast, our work focuses on the dance style transfer task, which involves manipulating the style of existing dance movements while preserving contextual information. Although our style transfer model does not require music as an input for conditioning, incorporating it can enhance the quality of the generated movements.

3 Methodology

This paper aims to explore the problem of transferring dance styles. This section formulates the target problem and introduces notations used throughout the paper. A brief overview of CycleGAN, CycleGAN-VC2, and StarGAN, as well as relevant preliminaries, are provided to ensure self-containment and facilitate understanding of the proposed method. Building on these foundations, we introduce our novel technical frameworks CycleDance and StarDance. The StarDance part is not included in the conference version.

3.1 Problem formulation

Our study aims to develop mapping functions between two distinct domains, denoted as X and Y, without relying on paired data between these domains. In our scenario, we focus on the transfer of dance movements between two different styles, corresponding to domains X and Y given dance sample $x \sim P_X$ and $y \sim P_Y$, where x is a sample from

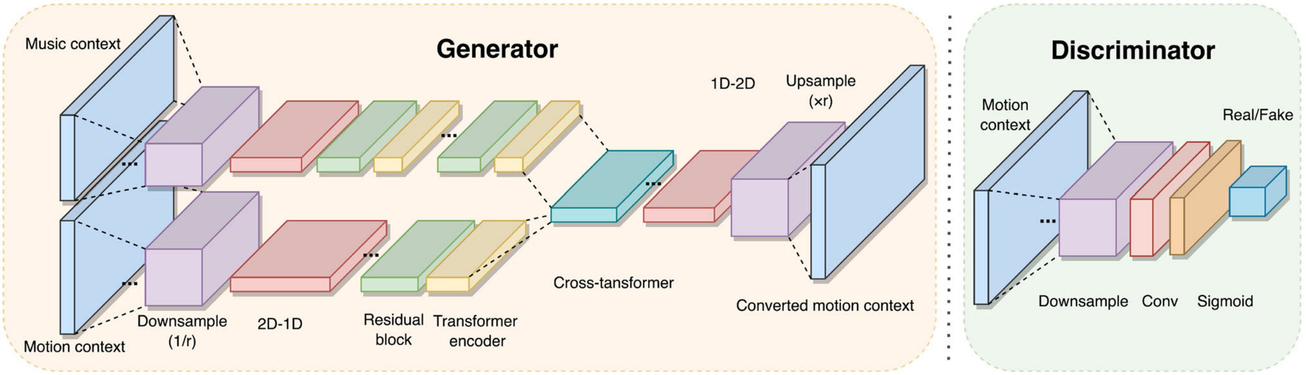

Fig. 2 The CycleDance architecture is composed of a generator and a discriminator. The generator consists of a motion pathway and a music pathway. Both pathways begin with downsampling blocks followed by a 2D–1D block. The motion, music, and cross-modal transformer blocks are standard full-attention transformer encoders. The fused pathway is followed by a 1D–2D block and upsampling blocks. The discriminator uses convolution in the last layer, following the approach of Kaneko et al. [10]

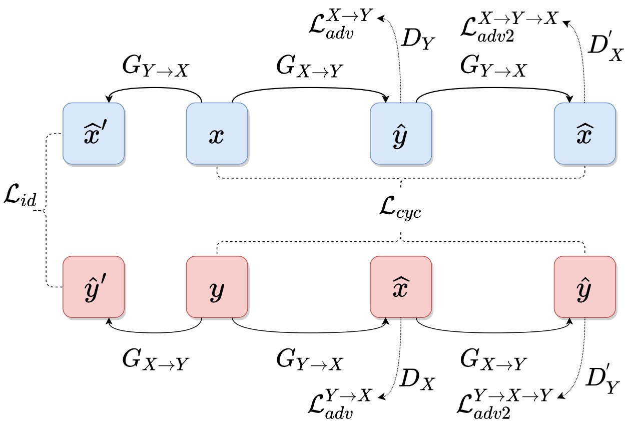

Fig. 3 The two-step adversarial generative training strategy, which involves four types of losses: adversarial loss $\mathcal{L}_{\text{adv}}$, cycle-consistency loss $\mathcal{L}_{\text{cyc}}$, identity-mapping loss $\mathcal{L}_{\text{id}}$, and second adversarial loss $\mathcal{L}_{\text{adv2}}$. See Sect. 3.2 for the definition of notations

domain X, drawn from the probability distribution P_X, and y is a sample from domain Y, drawn from the probability distribution P_Y. The dance samples may also be accompanied by music, with $m_x \in M_x$ and $m_y \in M_y$, respectively, with associated styles. The inclusion of music modality is optional in the style transfer task.

3.2 CycleDance training objective and strategy

To tackle the problem described above, we employ a strategy inspired by CycleGAN-like framework [6], which is depicted in Fig. 3. The CycleDance architecture we use incorporates two discriminators, D_X and D_Y, which differentiate between real and generated data. In addition, the architecture includes two mappings, $G_{X\to Y}$ and $G_{Y\to X}$, which are responsible for generating patterns in the target style. The mappings are cycled to enable the generated patterns to be converted back to their original domains. To achieve this goal, we adopt a strategy similar to that used in CycleGAN-VC2 [10] and utilize four types of losses, also see Fig. 3.

Adversarial loss $\mathcal{L}_{\text{adv}}^{X\to Y}$: this loss measures the difference between the transferred data $G_{X\to Y}(x, m_x)$ and the target y, where the discriminator D_Y distinguishes between the transferred data and the real data:

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

Similarly, we define the adversarial loss $\mathcal{L}_{\text{adv}}^{Y\to X}$ for $G_{Y\to X}$ and discriminator D_X.

Cycle-consistency loss $\mathcal{L}_{\text{cyc}}$: this loss helps to account for the loss of contextual information by recovering the original x and y from generated patterns, by $G_{X\to Y}(x, m_x)$ and $G_{Y\to X}(y, m_y)$, where 1-norm is adopted to minimize the absolute difference:

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

Identity-mapping loss $\mathcal{L}_{\text{id}}$: this loss further promotes content preservation by enforcing an identity transformation when applying $G_{X\to Y}$ and $G_{Y\to X}$ to the other domain:

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

Two-step adversarial loss $\mathcal{L}_{\text{adv2}}$: a second adversarial loss is used to mitigate the over-smoothing reconstruction statistics in the cycle-consistency loss [10]:

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

Note that this introduces an additional discriminator, denoted as D'_X. The loss $\mathcal{L}_{\text{adv2}}^{Y\to X\to Y}$ can be defined in a similar manner.

The final objective is expressed as a weighted sum of the above-mentioned loss terms

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

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

In addition, we utilize a curriculum learning algorithm as a training scheme. The rationale for this is that training can be made more efficient by initially handling simpler data and then gradually increasing the complexity of the training task. Similar strategies have been applied in various applications and scenarios, demonstrating its ability to improve the convergence rate, generalization capacity, and training stability [62]. We implement a curriculum learning strategy based on the length of the data by learning from short samples to longer ones as the training process progresses. The training data are truncated by following the method in [11].

3.3 StarDance training objectives and strategy

While CycleDance allows for generation of dance movement when a sufficient number of training examples are available, one limitation is that it only learns one-to-one mappings. New models need to be trained from scratch to transfer a new pair of dance styles. We thus extend CycleDance to a StarGAN-based backbone for a more scalable solution called StarDance. The StarDance architecture has one generator G and one discriminator D to accommodate for all styles with the help of an additional domain classifier C. The generator can take a dance sample x of style attribute c_x, with a target style attribute c_y, accompanied by music m, and generate a target dance sample $\hat{y} = G(x, m, c_y)$. The discriminator D produces a probability $D(x, c_x)$ that can be used to distinguish between real and transferred data, while the additional domain classifier C in StarDance is designed to predict which style an input dance sequence belongs to. Here we represent c as a one-hot vector, where each element is associated with a dance style. Similar to CycleDance, we define four types of losses in StarDance, as well as a domain classification loss for the classifier.

Adversarial loss $\mathcal{L}_{\text{adv}}$:

$$\begin{aligned}\mathcal{L}_{\text{adv}} &= \mathbb{E}_{x\sim P_X}[\log D(x, c_x)] \\ &+ \mathbb{E}_{x\sim P_X}[\log(1 - D(G(x, m, c_y), c_y))].\end{aligned} \tag{6}$$

Cycle-consistency loss $\mathcal{L}_{\text{cyc}}$:

$$\mathcal{L}_{\text{cyc}} = \mathbb{E}_{x\sim P_X}[\left\| G(G(x, m, c_y), m, c_x) - x \right\|_1]. \tag{7}$$

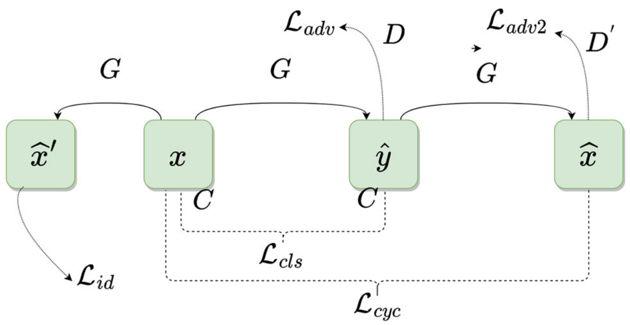

Fig. 4 The training strategy of StarDance, which involves five types of losses: adversarial loss $\mathcal{L}_{\text{adv}}$, cycle-consistency loss $\mathcal{L}_{\text{cyc}}$, identity-mapping loss $\mathcal{L}_{\text{id}}$, second adversarial loss $\mathcal{L}_{\text{adv2}}$, and domain classification loss $\mathcal{L}_{\text{cls}}$. See Sect. 3.3 for the definition of notations

Identity-mapping loss $\mathcal{L}_{\text{id}}$:

$$\mathcal{L}_{\text{id}} = \mathbb{E}_{x\sim P_X}[\|G(x, m, c_x) - x\|_1] \tag{8}$$

Two-step adversarial loss $\mathcal{L}_{\text{adv2}}$:

$$\begin{aligned}\mathcal{L}_{\text{adv2}} &= \mathbb{E}_{x\sim P_X}[\log D'_X(x)] \\ &+ \mathbb{E}_{x\sim P_X}[\log(1 - D'_X(G(G(x, m, c_y), m, c_x), c_x)]\end{aligned} \tag{9}$$

Domain classification loss $\mathcal{L}_{\text{cls}}$:

$$\begin{aligned}\mathcal{L}_{\text{cls}} &= \mathbb{E}_{x\sim P_X}[-\log C(c_x \mid x)] \\ &+ \mathbb{E}_{x\sim P_X}[-\log C(c \mid G(x, m, c_y))].\end{aligned} \tag{10}$$

To summarize, the full objective of StarDance to be minimized with respect to G, D, and C is given by

$$\begin{aligned}\mathcal{L}_{\text{full}} &= \mathcal{L}_{\text{adv}} + \lambda_{\text{cyc}}\mathcal{L}_{\text{cyc}} + \lambda_{\text{id}}\mathcal{L}_{\text{id}} \\ &+ \mathcal{L}_{\text{adv2}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}},\end{aligned} \tag{11}$$

where λ_{cyc}, λ_{id}, and λ_{cls} are regularization parameters to weight the losses. The training strategy of StarDance is illustrated in Fig. 4. With the StarDance architecture, there are fewer coefficients to tune, given a more compact generator.

3.4 Network architecture

Our Cycle/StarDance framework builds upon a cross-modal transformer, as illustrated in Fig. 2. In the CycleDance framework, the cross-modal transformer is employed to concatenate motion and music encodings, both of which are obtained through a sequence of layers including 2D convolution (purple blocks in Fig. 2), 2D–1D reshaping (red), residual convolution (green), and modality-specific transformers (yellow). The 2D convolutional layers are utilized for downsampling while keeping the sequential structure of the input data. The resulting downsampled features are then reshaped and passed through the residual blocks of 1D CNNs.

The reshaped 1D sequences are further processed by transformers, which use a position embedding to output encodings that capture temporal relationships among timesteps. Finally, the generator takes the concatenated encodings and feeds them into a 1D–2D reshape block (red) and an upsampling block (purple) to synthesize transferred dance motions. Within these blocks, we employ gated linear units (GLUs) [63], a tunable activation function, to learn hierarchical and sequential structures. In our StarDance framework, the transformer is adapted to incorporate style labels. Before feeding the concatenated encodings into the 1D–2D reshape block, we concatenate the encoded features and style attribute along the channel dimension.

For the discriminator, CycleDance first downsamples the motion data with a 2D CNN. We adopt the same approach as proposed in [10] and use convolution only at the last layer of the discriminator to alleviate training instability. The output layer of the discriminator employs a sigmoid activation function to make the final prediction on the motion clip. Similarly, the StarDance discriminator first concatenates the motion data and the style attribute along the channel dimension before the downsampling block.

We also devise a domain classifier in StarDance, which downsamples a motion data input with 2D CNNs and produces a sequence of class probability distributions at the last layer that predicts how likely the motion sequence is to belong to the respective style attributes.

3.5 Training details

We implement and train the Cycle/StarDance model with the PyTorch-Lightning framework, with Adam optimizers used for training. The batch size is fixed to 1, given that $batch\_size = 1$ with instance normalization achieves invariance to mean and variance of features, which is beneficial for the style transfer task. We set the initial learning rate of 0.0002 for the generator and 0.0001 for the discriminator and domain classifier, with momentum terms of 0.5 for both optimizers. As for loss weights, we set $\lambda_{\text{cyc}} = 10$, $\lambda_{\text{id}} = 5$, and $\lambda_{\text{cls}} = 5$. The identifying-mapping loss $\mathcal{L}_{\text{id}}$ is only active for the first 10^4 iterations for regulating consistency in the initial training stage. Our framework adopts a curriculum learning strategy to alleviate the challenges of training models for extremely long sequential data. Specifically, we gradually increase the sample length from 32 frames to 800 frames, adding 8 frames to data fragments every 500 epochs. In total, we train the entire framework for 5×10^5 iterations.

4 Experiments and evaluations

To demonstrate the capabilities of our proposed approach, we compare it to baseline models and ablations on a large dance dataset. This section begins with a description of the dataset used in our experiments, as well as the processing steps and the experimental setup (Sect. 4.1). We then provide details about our objective (Sect. 4.3) and subjective (Sect. 4.4) evaluations, which aim to benchmark different dance style transfer methods and ablations. We present the results of our evaluations and discuss their implications. Demonstrations of our novel framework are available at https://youtu.be/kP4DBp8OUCk. Extending the conference version, we append the details of the evaluation metrics, the results of StarDance and the accessibility to our preprocessed data, as well as the user study questionnaire.

4.1 Dataset

We utilize the AIST+ Dance Database [42] as our source of data to generate 3D dance motion samples with paired music. The AIST++ database reconstructs 3D motions with SMPL parameters from multi-view videos in the AIST Dance Database [64], a large-scale dance video database containing various dance genres recorded by professional dancers. To extract motion features, we first downsample all motion data to 30 frames per second (fps) and retarget it to a 21-body-joint skeleton using Autodesk MotionBuilder. We adopt exponential map parametrization to represent the 3D rotation of all joints. The root joint (hip) is characterized by four additional features representing changes in the vertical root position, ground-projected position, and 2D facing angle. Consequently, a 67-dimensional vector is employed to represent the motion features of each frame. The music features are extracted with the Librosa toolbox in a similar way to [64]. 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. Six dance genres were selected for analysis based on suggestions from professional dancers we recruited, specifically ballet-jazz, locking, hip-hop, pop, waacking, and house dance. Each dance set consists of 141 motion sequences and six songs, spanning approximately 2000 s in total. We make the preprocessed data publicly available at https://urlis.net/aist to further facilitate research in the community.

4.2 Baseline models and ablations

To evaluate the impact of design choices such as the cross-modal transformer and curriculum learning strategy, we compare our proposed CycleDance and StarDance models to the CycleGAN-VC2 baseline. Additionally, we implement three alternative architectures for an ablation study. In the first ablation configuration, CycleTransGAN, we eliminate the music pathway, cross-modal transformer, and curriculum learning strategy. Our aim is to highlight the effectiveness of the transformer architecture we introduced. In the Cycle-

TransGAN+CL ablation, we apply curriculum learning to the CycleTransGAN model. We aim to gauge the performance gains achieved by meticulously modulating the complexity of the samples used to train the model. The final ablation, CycleCrossTransGAN, employs cross-modal transformers for motion and music information in the encoder without using the curriculum learning strategy. Through this ablation, we aim to evaluate the influence of the cross-modal transformers by comparing the discrepancies between CycleTransGAN and CycleCrossTransGAN.

4.3 Objective evaluation

The primary objective of all these models is to transfer the dance style from a specified source to a particular target dance style. We conduct assessments from both objective and subjective perspectives to ensure a comprehensive evaluation of the complex motion patterns common in dance. For the objective evaluation, we analyze 17 dance sequences for each style. Style transfer was performed on each ablated model, and two metrics were used to evaluate the performance on transfer strength and content preservation. The two metrics are designed based on Fréchet distance, which is computed by Eq. 12:

$$\mathrm{FID} = \|\mu_r - \mu_g\|_2^2 + \mathrm{Tr}(\Sigma_r + \Sigma_g - 2\sqrt{\Sigma_r \Sigma_g}), \qquad (12)$$

where (μ_r, Σ_r) and (μ_g, Σ_g) are, respectively, the mean and the covariance matrix of the real and generated dance movement distribution.

Transfer strength The transfer strength, which measures the extent that the source style is converted to the target style, is the most crucial aspect of style transfer. Specifically, we adopt the Fréchet distance between the true and generated dance motion distributions. To capture features that are more style-correlated such as intensity, we utilize joint velocity and acceleration. To be specific, we use pairs of consecutive raw poses without normalization (x_{i-1}, x_i) to convert the pose representations to joint velocity v_i. Similarly, we use triples of consecutive poses (x_{i-1}, x_i, x_{i+1}) to estimate the joint acceleration a_i. We refer to this metric as the Fréchet motion distance (FMD), which is used to quantify the transfer strength between the generated motion and target style.

Content preservation We use the Fréchet pose distance (FPD) as a measure of content preservation, which evaluates how well the salient poses of the original dance movement are preserved after the transfer. For this dimension, we use the Fréchet distance between distributions of key poses x_k for a given dance movement. We detect frames presenting local maxima in joint acceleration as the key frames that segment the full dance movement. To ensure comparability across frames, key poses in these frames are normalized with respect to the hip-centric origin.

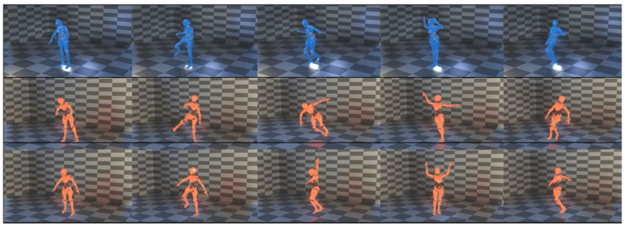

Fig. 5 An example of transferring locking dance sequences (top, blue y-bot) to ballet-jazz dance using CycleGAN-VC2 (middle, red x-bot) and CycleDance (bottom, red x-bot)

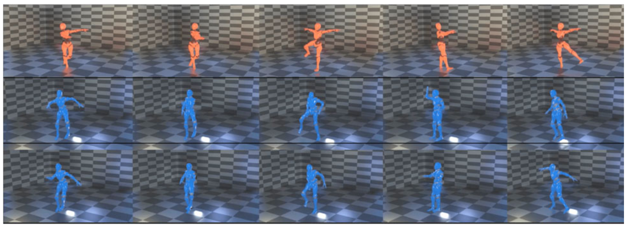

Fig. 6 An example of transferring ballet-jazz dance sequences (top, red x-bot) to locking dance using CycleGAN-VC2 (middle, blue y-bot) and CycleDance (bottom, blue y-bot)

In Table 1, we present the quantitative results of the proposed model and ablations for style transfer across three pairs of dance genres in both directions. These pairs include ‘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).

The baseline model, CycleGAN-VC2, appears to struggle with the style transfer task for locking dance to ballet-jazz, as evidenced by the significantly higher MFD compared to all other ablation methods. The complete framework, CycleDance, outperforms all other ablation methods and achieves the best performance on both metrics and almost all transfer pairs. This emphasizes the necessity of all introduced techniques in this task.

Figure 5 presents an example of a synthesized motion clip that demonstrates the transfer of dance style from locking to ballet-jazz. An example of style transfer from ballet-jazz to locking dance is shown in Fig. 6. The top keyframe sequence shows the original locking dance. The middle sequence is generated by the baseline model, CycleGAN-VC2, while the bottom sequence is generated by the proposed CycleDance. By comparing the poses of each column in these two figures, it can be observed that the extracted key gestures are representative of the pose sequences. Notably, CycleDance achieves a higher similarity to the source gestures, preserving more content while aligning better with the target dance style.

Table 1 Quantitative objective evaluation

Method	FMD						FPD					
	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
StarDance	3.7153	1.4823	1.2818	1.0893	1.6145	1.5801	0.5217	0.4378	0.4707	0.4267	0.4410	0.4138

Fréchet Motion distance (FMD) and Fréchet Pose distance (FPD) are employed to evaluate the performance of the baseline model, our proposed CycleDance model and StarDance model, and the three ablations. The acronym BJ2LC refers to the transfer of dance style from ballet-jazz to locking dance. Correspondingly, LC2BJ denotes the reverse transfer from locking dance to ballet-jazz. Similarly, WK, HP, PO, and HO are acronyms used to represent waacking, hip-hop, pop, and house dance, respectively

In addition, the ablation study revealed that CycleTransGAN (the combination of CycleGAN-VC2 and transformer) achieved lower FMD scores, suggesting that the model benefited from capturing richer intra-relations among frames with the help of the transformer. By comparing the results of 'CycleTransGAN and CycleCrossTransGAN', we can observe an improvement in both FMD and FPD metrics. This suggests that the music information aids in generating accurate target-style movements and that this contextual information is effectively encoded by the cross-modal transformer. The comparison between 'CycleTransGAN and CycleTransGAN+CL' as well as 'CycleCrossTransGAN and CycleDance' indicates that curriculum learning greatly enhances transfer strength. This demonstrates the effectiveness of gradually increasing the level of difficulty by training with longer dance sequences.

In Table 1, we also present results from StarDance, which demonstrates an ability to solve the task of transferring among multiple dance styles using a more scalable model. The performance of StarDance exhibits a similar trend to that of CycleDance. It performs reasonably well on certain pairs, such as transferring from 'house to pop dance,' but shows poor performance when transferring from 'ballet-jazz to locking dance'. The poorer performance in general is natural, given that CycleDance is specifically trained for particular pairs of dance styles, whereas StarDance is expected to cover them all. It is worthwhile to investigate the factors that influence style transfer in different dance pairs in future studies.

4.4 Subjective evaluation

To obtain a more comprehensive evaluation of our model and the baseline, we conducted a user study in addition to the objective assessment, asking participants to rate three aspects: motion naturalness, transfer strength, and content preservation. This study also includes open-ended questions to gather additional feedback and suggestions for future work. We make the user study questionnaire publicly available at https://urlis.net/ques to further support relevant studies.

Our subjective analysis primarily focuses on the style transfer between ballet-jazz and locking dance, as these dance styles are well-established and well-understood by dance professionals, providing a diverse representation of the challenges in style transfer. An online survey was performed to evaluate the transfer tasks for both 'ballet-jazz to locking dance' and 'locking dance to ballet-jazz', gathering participant feedback. To generate the videos for each source and target dance sequence, we utilized the Blender software to create 8-second clips with an x-bot character (representing ballet-jazz) and a y-bot character (representing locking dance). During the survey, participants were presented with a source dance video clip followed by a generated target dance clip. Prior to this, an introduction phase allowed participants to become familiar with the animated dance through video clips that could be played. To avoid potential order effects, the order of target dance clips was randomly selected and balanced. Each target dance clip was generated either from CycleDance or from the baseline. The participants were allowed to view the clips multiple times before answering three questions:

- **Motion naturalness**: *To what extent do you agree with the following statement?—The generated motion clip looks natural after the 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 resulting 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 [65].

During the study, 30 participants equipped with at least 5 years of dance experience, including training, choreographing, teaching, or performing, were recruited. Participants in the study were aged between 20 and 41 years (median 30), with 37.9% identifying as male, 58.6% as female, and 3.4% as other. According to the demographic questions, the participants reported their familiarity with ballet-jazz dance and locking dance as M = 3.93 (SD = 1.05) and M = 3.03 (SD = 1.18), respectively, on a scale from 1 (not at all familiar) to 5 (very familiar). As the generated motions were presented using virtual characters, we also assessed the frequency at which participants played video games. Results indicated that 34.5% of participants played video games weekly, 13.8% played monthly, 13.8% yearly, and 37.9% rarely.

We performed a statistical analysis of the subjective responses from the user study to support our findings, and evaluated whether our proposed method could be further enhanced. Based on Fig. 7, the results indicate that CycleDance received higher ratings from experts than the baseline model on both motion naturalness and transfer strength. The Wilcoxon signed-rank test was used to assess the statistical significance of the subjective responses. The results showed that 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. Therefore, from the perspective of dance experts, CycleDance is preferred for its improved naturalness and similarity to the target dance style, which is consistent with what we observe from the objective quantitative results (Sect. 4.3). For the responses on content preservation, Fig. 8 presents the overall statistics for the four queried aspects. The experts chose CycleDance more often than the baseline model on all four aspects, indicating that they believed that CycleDance better preserved specific features of the source dance style. We conducted a McNemar test to assess the statistical significance of the differences between the baseline model and CycleDance. The test showed no significant differences between the two models on 'Orientation through space' ($p = 0.1724$) and 'Shapes of the limbs' ($p = 0.1573$). On the other

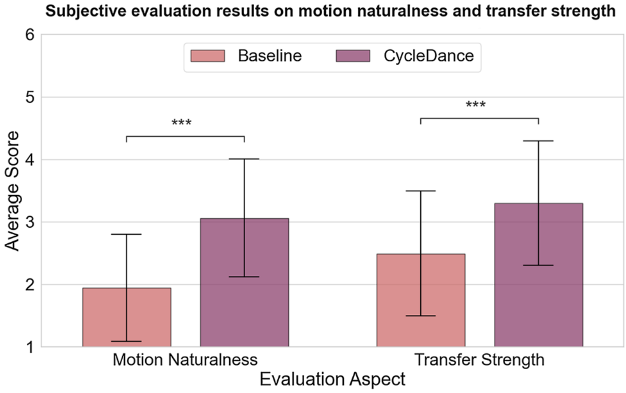

Fig. 7 The subjective evaluation results on motion naturalness and transfer strength, where the error bars represent the standard errors of the averages. Statistical significance was determined using the Wilcoxon signed-rank test, which compares the medians (*** means $p < 0.0001$)

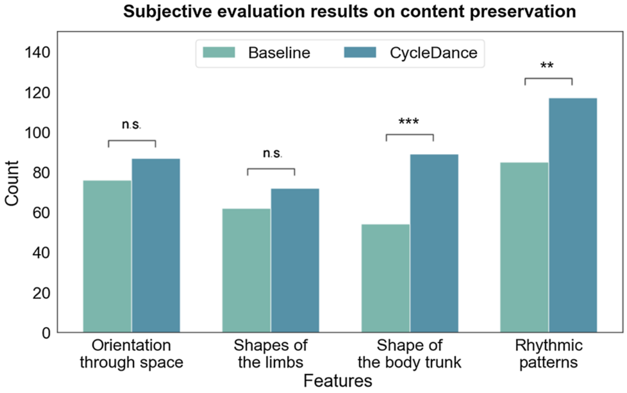

Fig. 8 The subjective evaluation results on content preservation. The results show that CycleDance outperforms the baseline model in several aspects, including orientation through space, shape of the limbs, shape of the body trunk, and rhythmic patterns. Statistical significance was determined using the Wilcoxon signed-rank test, which compares the medians (*** means $p < 0.00001$, ** means $p < 0.0001$, and n.s. means $p > 0.05$)

hand, there was a strong statistical significance supporting the proposed CycleDance for ‘Shapes of the body trunk’ ($p = 0.000002$) and ’Rhythmic patterns’ ($p = 0.00004$), based on the median value. Both CycleDance and the baseline model received higher scores on rhythmic patterns and orientation through space among the four aspects evaluated. This suggests preserving dance orientation, and rhythm is relatively easier when performing dance style transfer. Preserving the shape of the limbs is found to be more challenging, where no significant differences were found between CycleDance and the baseline. However, CycleDance outperforms the baseline in preserving the shape of the body trunk.

In response to open-ended questions, dance experts provided feedback that, for the task from ‘ballet jazz to lock dance,’ both methods produced a jerky style that mimicked pop and lock dance. The example illustrated in Fig. 9 is frequently mentioned as a significant instance of successful "transfer" with a noticeable locking dance style from the perspective of experienced dancers. Regarding the CycleDance samples of transferring ‘locking dance to ballet-jazz’, the dance experts noted that the character arms clearly exhibit jazz or ballet characteristics and effectively maintain traditional shapes. The dance experts also provided feedback on some limitations of the proposed method. One common observation is that some generated motions appear wobbly, indicating a need for applying smoothing filters to improve the overall quality of the results. Experts have also pointed out that in jazz ballet, dancers typically point their toes, whereas the generated movements always show ankle flexion. This demonstrates the limitation that the available data do not adequately represent the nuances of foot movements required in ballet-jazz, highlighting the need for more comprehensive data collection in future studies. This limitation also leads to some physically unrealistic effects, such as the character appearing to float when their body does not have any contact with the floor.

5 Discussion of societal impact

We present a style transfer framework that offers both artistic and scientific contributions to the field of dance. We anticipate several potential impacts on industries and society in the near future. This work has the potential to positively impact the field of choreography and dance research by unlocking new possibilities for the hybrid human-artificial co-creation of dance material. This work could also benefit industries such as video games and animation. For example, the framework can be used to create group dances where each character has a unique motion style. Such effects could lead to a transformation of the job market, with a shift towards jobs that rely more on a combination of creativity and automation. Additionally, it may lead to the development of new user-friendly interfaces and tools for various industries. However, a potential negative impact of the technology is that it may blur the ownership in creative processes, i.e., determining who should be credited as the creator(s) of the generated dance movements. Transfer models 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. This raises concerns about the potential perpetuation of existing societal biases and discrimination in the application of these models.

6 Conclusion and future work

This study tackles the challenging task of style transfer for sequential data with intricate variations and complex frame dependencies, specifically in the domain of dance movements. To address these challenges, we first propose CycleDance, which leverages expressive data encoders, cross-modal contexts, and a curriculum-based training scheme. We also propose StarDance, which extends the backbone from the CycleGAN-based model to the StarGAN-based model to handle more than two dance genres. The effectiveness of our proposed frameworks is confirmed by

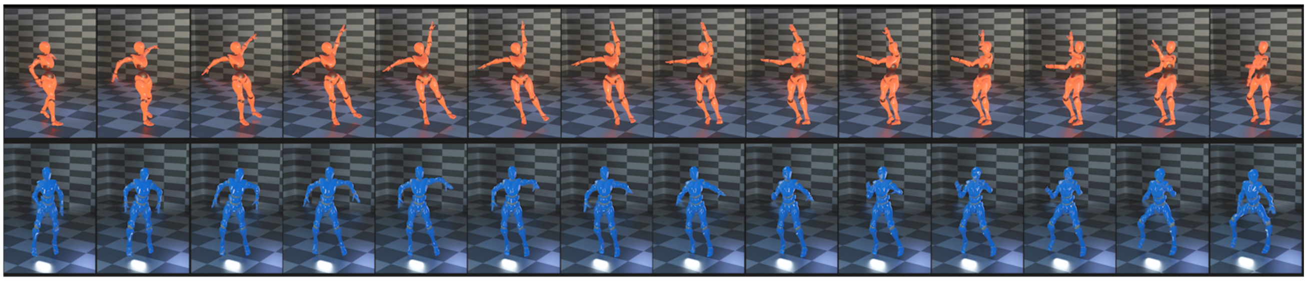

Fig. 9 An illustrative example generated by CycleDance, where ballet-jazz dance (red x-bot) is transferred to locking dance (blue y-bot). The top panel shows the original ballet-jazz dance motion, while the bottom panel shows the transferred locking dance motion

quantitative results and human expert evaluations on similarity aspects and transfer aspects. To the best of our knowledge, this is the first work to use music context for dance or general motion style transfer. Recently, diffusion models have emerged as a promising approach for generative modeling, demonstrating state-of-the-art results in image and video generation tasks. In the future, we plan to extend our GAN-based style transfer framework to incorporate diffusion-based architecture. Research is also needed to address identified limitations on preserving limb shapes. We also will explore motion style transfer in video settings. Based on these techniques, we envision new tools in dance motion design for choreography, the film industry, and video games.

Acknowledgements This study has received funding from the European Commission Horizon 2020 research and innovation program under Grant Agreement Number 824160 (EnTimeMent). This work benefited from access to the HPC resources provided by the Swedish National Infrastructure for Computing (SNIC), partially funded by the Swedish Research Council through Grant Agreement No. 2018-05973. We thank the dancers from Stockholm University of the Arts, KTH Dance Club, and others. We are also grateful to the reviewers for their thoughtful comments.

Funding Open access funding provided by Royal Institute of Technology.

References

1. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2414–2423 (2016)
2. Brunner, G., Wang, Y., Wattenhofer, R., Zhao, S.: Symbolic music genre transfer with cyclegan. In: 2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI), pp. 786–793. IEEE (2018)
3. Mason, I., Starke, S., Zhang, H., Bilen, H., Komura, T.: Few-shot learning of homogeneous human locomotion styles. In: Computer Graphics Forum, vol. 37, pp. 143–153. Wiley Online Library (2018)
4. Du, H., Herrmann, E., Sprenger, J., Cheema, N., Hosseini, S., Fischer, K., Slusallek, P.: Stylistic locomotion modeling with conditional variational autoencoder. In: Eurographics (Short Papers), pp. 9–12 (2019)
5. Aberman, K., Weng, Y., Lischinski, D., Cohen-Or, D., Chen, B.: Unpaired motion style transfer from video to animation. ACM Trans. Graph. **39**(4), 1–64 (2020)
6. Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)
7. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4401–4410 (2019)
8. Valle-Pérez, G., Henter, G.E., Beskow, J., Holzapfel, A., Oudeyer, P.-Y., Alexanderson, S.: Transflower: probabilistic autoregressive dance generation with multimodal attention. ACM Trans. Graph. **40**(6), 1–14 (2021)
9. Chen, K., Tan, Z., Lei, J., Zhang, S.-H., Guo, Y.-C., Zhang, W., Hu, S.-M.: Choreomaster: choreography-oriented music-driven dance synthesis. ACM Trans. Graph. **40**(4), 1–13 (2021)
10. Kaneko, T., Kameoka, H., Tanaka, K., Hojo, N.: Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion. In: ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6820–6824. IEEE (2019)
11. Fu, C., Liu, C., Ishi, C.T., Ishiguro, H.: Cycletransgan-evc: a cyclegan-based emotional voice conversion model with transformer. arXiv preprint arXiv:2111.15159 (2021)
12. Yin, W., Yin, H., Baraka, K., Kragic, D., Björkman, M.: Dance style transfer with cross-modal transformer. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 5058–5067 (2023)
13. Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., Choo, J.: Stargan: unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8789–8797 (2018)
14. Kameoka, H., Kaneko, T., Tanaka, K., Hojo, N.: Stargan-vc: non-parallel many-to-many voice conversion using star generative adversarial networks. In: 2018 IEEE Spoken Language Technology Workshop (SLT), pp. 266–273. IEEE (2018)
15. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.-H.: Universal style transfer via feature transforms. Adv. Neural Inf. Process. Syst. **30**, 66 (2017)
16. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1501–1510 (2017)
17. An, J., Li, T., Huang, H., Shen, L., Wang, X., Tang, Y., Ma, J., Liu, W., Luo, J.: Real-time universal style transfer on high-resolution images via zero-channel pruning. arXiv preprint arXiv:2006.09029 (2020)
18. Huang, X., Liu, M.-Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 172–189 (2018)
19. Wang, H., Li, Y., Wang, Y., Hu, H., Yang, M.-H.: Collaborative distillation for ultra-resolution universal style transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1860–1869 (2020)
20. Chen, L.-H., Ling, Z.-H., Liu, L.-J., Dai, L.-R.: Voice conversion using deep neural networks with layer-wise generative training. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(12), 1859–1872 (2014)
21. Saito, Y., Takamichi, S., Saruwatari, H.: Voice conversion using input-to-output highway networks. IEICE Trans. Inf. Syst. **100**(8), 1925–1928 (2017)
22. Hsu, C.-C., Hwang, H.-T., Wu, Y.-C., Tsao, Y., Wang, H.-M.: Voice conversion from unaligned corpora using variational autoencoding Wasserstein generative adversarial networks. arXiv preprint arXiv:1704.00849 (2017)

23. Kameoka, H., Kaneko, T., Tanaka, K., Hojo, N.: Acvae-vc: non-parallel voice conversion with auxiliary classifier variational autoencoder. IEEE/ACM Trans. Audio Speech Lang. Process. **27**(9), 1432–1443 (2019)
24. Kaneko, T., Kameoka, H.: Parallel-data-free voice conversion using cycle-consistent adversarial networks. arXiv preprint arXiv:1711.11293 (2017)
25. Kaneko, T., Kameoka, H., Tanaka, K., Hojo, N.: Stargan-vc2: Rethinking conditional methods for Stargan-based voice conversion. arXiv preprint arXiv:1907.12279 (2019)
26. Cífka, O., Şimşekli, U., Richard, G.: Groove2groove: one-shot music style transfer with supervision from synthetic data. IEEE/ACM Trans. Audio Speech Lang. Process. **28**, 2638–2650 (2020)
27. Malik, I., Ek, C.H.: Neural translation of musical style. arXiv preprint arXiv:1708.03535 (2017)
28. Ding, Z., Liu, X., Zhong, G., Wang, D.: Steelygan: semantic unsupervised symbolic music genre transfer. In: Pattern Recognition and Computer Vision: 5th Chinese Conference, PRCV 2022, Shenzhen, China, November 4–7, 2022, Proceedings, Part I, pp. 305–317 (2022). Springer
29. Mueller, J., Gifford, D., Jaakkola, T.: Sequence to better sequence: continuous revision of combinatorial structures. In: International Conference on Machine Learning PMLR, pp. 2536–2544 (2017)
30. Fu, Z., Tan, X., Peng, N., Zhao, D., Yan, R.: Style transfer in text: exploration and evaluation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
31. Dai, N., Liang, J., Qiu, X., Huang, X.: Style transformer: unpaired text style transfer without disentangled latent representation. arXiv preprint arXiv:1905.05621 (2019)
32. Xu, J., Sun, X., Zeng, Q., Ren, X., Zhang, X., Wang, H., Li, W.: Unpaired sentiment-to-sentiment translation: a cycled reinforcement learning approach. arXiv preprint arXiv:1805.05181 (2018)
33. Amaya, K., Bruderlin, A., Calvert, T.: Emotion from motion. In: Graphics Interface, vol. 96, pp. 222–229 (1996). Toronto, Canada
34. Unuma, M., Anjyo, K., Takeuchi, R.: Fourier principles for emotion-based human figure animation. In: Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques, pp. 91–96 (1995)
35. Witkin, A., Popovic, Z.: Motion warping. In: Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques, pp. 105–108 (1995)
36. Aristidou, A., Zeng, Q., Stavrakis, E., Yin, K., Cohen-Or, D., Chrysanthou, Y., Chen, B.: Emotion control of unstructured dance movements. In: Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation, pp. 1–10 (2017)
37. Hsu, E., Pulli, K., Popović, J.: Style translation for human motion. In: ACM SIGGRAPH 2005 Papers, pp. 1082–1089 (2005)
38. Maiorca, A., Yoon, Y., Dutoit, T.: Evaluating the quality of a synthesized motion with the fréchet motion distance. In: ACM SIGGRAPH 2022 Posters, pp. 1–2 (2022)
39. Holden, D., Saito, J., Komura, T.: A deep learning framework for character motion synthesis and editing. ACM Trans. Graph. **35**(4), 1–11 (2016)
40. Holden, D., Habibie, I., Kusajima, I., Komura, T.: Fast neural style transfer for motion data. IEEE Comput. Graph. Appl. **37**(4), 42–49 (2017)
41. Smith, H.J., Cao, C., Neff, M., Wang, Y.: Efficient neural networks for real-time motion style transfer. Proc. ACM Comput. Graph. Interact. Tech. **2**(2), 1–17 (2019)
42. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Ai choreographer: music conditioned 3d dance generation with aist++. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 13401–13412 (2021)
43. Mason, I., Starke, S., Komura, T.: Real-time style modelling of human locomotion via feature-wise transformations and local motion phases. arXiv preprint arXiv:2201.04439 (2022)
44. Park, S., Jang, D.-K., Lee, S.-H.: Diverse motion stylization for multiple style domains via spatial-temporal graph-based generative model. Proc. ACM Comput. Graph. Interact. Tech. **4**(3), 1–17 (2021)
45. Dong, Y., Aristidou, A., Shamir, A., Mahler, M., Jain, E.: Adult2child: motion style transfer using cyclegans. In: Motion, Interaction and Games, pp. 1–11 (2020)
46. Xia, S., Wang, C., Chai, J., Hodgins, J.: Realtime style transfer for unlabeled heterogeneous human motion. ACM Trans. Graph. **34**(4), 1–10 (2015)
47. Wen, Y.-H., Yang, Z., Fu, H., Gao, L., Sun, Y., Liu, Y.-J.: Autoregressive stylized motion synthesis with generative flow. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13612–13621 (2021)
48. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local Nash equilibrium. Adv. Neural Inf. Process. Syst. **30**, 66 (2017)
49. Xi, W., Devineau, G., Moutarde, F., Yang, J.: Generative model for skeletal human movements based on conditional dc-gan applied to pseudo-images. Algorithms **13**(12), 319 (2020)
50. Yoon, Y., Cha, B., Lee, J.-H., Jang, M., Lee, J., Kim, J., Lee, G.: Speech gesture generation from the trimodal context of text, audio, and speaker identity. ACM Trans. Graph. **39**(6), 1–16 (2020)
51. Butepage, J., Black, M.J., Kragic, D., Kjellstrom, H.: Deep representation learning for human motion prediction and classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6158–6166 (2017)
52. Yan, S., Li, Z., Xiong, Y., Yan, H., Lin, D.: Convolutional sequence generation for skeleton-based action synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4394–4402 (2019)
53. Habibie, I., Holden, D., Schwarz, J., Yearsley, J., Komura, T.: A recurrent variational autoencoder for human motion synthesis. In: 28th British Machine Vision Conference (2017)
54. Yin, W., Yin, H., Kragic, D., Björkman, M.: Graph-based normalizing flow for human motion generation and reconstruction. In: 2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN), pp. 641–648 (2021). IEEE
55. Li, Z., Zhou, Y., Xiao, S., He, C., Huang, Z., Li, H.: Auto-conditioned recurrent networks for extended complex human motion synthesis. arXiv preprint arXiv:1707.05363 (2017)
56. Shiratori, T., Nakazawa, A., Ikeuchi, K.: Dancing-to-music character animation. In: Computer Graphics Forum, vol. 25, pp. 449–458 (2006). Wiley Online Library
57. Fan, R., Xu, S., Geng, W.: Example-based automatic music-driven conventional dance motion synthesis. IEEE Trans. Vis. Comput. Graph. **18**(3), 501–515 (2011)
58. Lee, M., Lee, K., Park, J.: Music similarity-based approach to generating dance motion sequence. Multimedia Tools Appl. **62**(3), 895–912 (2013)
59. Sun, G., Wong, Y., Cheng, Z., Kankanhalli, M.S., Geng, W., Li, X.: Deepdance: music-to-dance motion choreography with adversarial learning. IEEE Trans. Multimedia **23**, 497–509 (2020)
60. Zhuang, W., Wang, C., Xia, S., Chai, J., Wang, Y.: Music2dance: Dancenet for music-driven dance generation. arXiv preprint arXiv:2002.03761 (2020)
61. Li, B., Zhao, Y., Sheng, L.: Dancenet3d: music based dance generation with parametric motion transformer. arXiv preprint arXiv:2103.10206 (2021)
62. Wang, X., Chen, Y., Zhu, W.: A survey on curriculum learning. IEEE Trans. Pattern Anal. Mach. Intell. **6**, 66 (2021)

63. Dauphin, Y.N., Fan, A., Auli, M., Grangier, D.: Language modeling with gated convolutional networks. In: International Conference on Machine Learning, PMLR, pp. 933–941 (2017)
64. 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: ISMIR, vol. 1, p. 6 (2019)
65. Newlove, J., Dalby, J.: Laban for all (2004)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.