

Dual Learning Music Composition and Dance Choreography

Shuang Wu

wushuang@outlook.sg

Nanyang Technological University

Shijian Lu

Shijian.Lu@ntu.edu.sg

Nanyang Technological University

Zhenguang Liu*

liuzhenguang2008@gmail.com

Zhejiang Gongshang University

Li Cheng*

lcheng5@ualberta.ca

University of Alberta

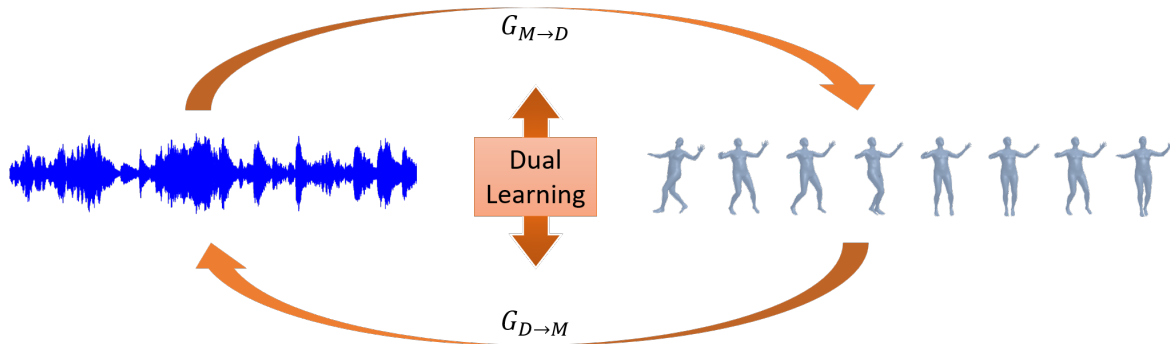


Figure 1: We propose a novel problem where we concurrently learn dance choreography and music composition. Specifically, our framework consists of two networks, $G_{M \rightarrow D}$ for generating 3D dance choreographs from input music, and $G_{D \rightarrow M}$ that synthesizes music compositions given dance sequences. We leverage the duality of these tasks to extract the common underlying themes and ensure consistency between the generated output and the conditional input.

ABSTRACT

Music and dance have always co-existed as pillars of human activities, contributing immensely to the cultural, social, and entertainment functions in virtually all societies. Notwithstanding the gradual systematization of music and dance into two independent disciplines, their intimate connection is undeniable and one art-form often appears incomplete without the other. Recent research works have studied generative models for dance sequences conditioned on music. The dual task of composing music for given dances, however, has been largely overlooked. In this paper, we propose a novel extension, where we jointly model both tasks in a dual learning approach. To leverage the duality of the two modalities, we introduce an optimal transport objective to align feature embeddings, as well as a cycle consistency loss to foster overall consistency. Experimental results demonstrate that our dual learning framework improves individual task performance, delivering generated music compositions and dance choreographs that are realistic and faithful to the conditioned inputs.

*Corresponding Authors

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '21, October 20–24, 2021, Virtual Event, China

© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-8651-7/21/10...\$15.00

<https://doi.org/10.1145/3474085.3475180>

CCS CONCEPTS

• Computing methodologies → Neural networks; Multi-task learning; • Applied computing → Media arts.

KEYWORDS

cross-modal generation, dual learning, optimal transport

ACM Reference Format:

Shuang Wu, Zhenguang Liu, Shijian Lu, and Li Cheng. 2021. Dual Learning Music Composition and Dance Choreography. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21)*, October 20–24, 2021, Virtual Event, China. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3474085.3475180>

1 INTRODUCTION

From an evolutionary perspective, music and dance have played a vital role for the social function of the human species [35]. They are ubiquitous in human activities, ranging from personal entertainment to social functions and ceremonial activities. On top of forming an indispensable tapestry in human culture, they are also integral to modern civilization and contribute immensely to our individual and social well-being. Over the past centuries, music and dance are gradually systematized into two separate art-forms but their intimate and deep connection is unmistakable. Both entail expressing our internal emotions as external movements. For dance, the medium of expression is visual in the form of body movements whereas for music, movements manifest auditorily through melodies and rhythms.

It is widely acknowledged that music evokes motions and feelings of intentionality [30]. We may find ourselves moving along to the beats and dancing to music, perhaps even unaware of our movements. Neuroscience research elucidates how music and dance activities may involve similar stimuli in our brain [6, 20]. Recently, the artificial intelligence community has also taken an interest. Several works [11, 15, 18, 19, 29, 31] have investigated the task of generating dance choreographs from music.

Reciprocally, dance would appear incomplete and unadorned if there were no accompanying music. Since generating music from dance remains largely overlooked in the literature, we propose to concurrently tackle this dual task. As summarized in Fig. 1, we develop a bi-directional generative model for synthesizing realistic and matching dance from music *and* music from dance. This dual learning has the additional advantage of enhancing the modeling of each modality. Given the popularity of music and dance videos in entertainment and multimedia, our ultimate goal is to enable effective engagement of the public, contribute to the user experience, and benefit the vast community of musicians and dancers.

There are several challenges in this task.

Cross-domain generation Translating between music and dance constitutes a cross-domain sequence-to-sequence generative modeling task. Most source-to-target domain learning tasks entail image-to-image [17, 23, 40, 42] or language translation [13], for which the data lies in topologically identical spaces. However, our task is further complicated by the fact that the ambient spaces for our data have totally distinct topological properties. Specifically, music is represented as waveforms, whereas a dance sequence is represented as 3D motion trajectories on a pose manifold. This increases the difficulty of learning a network mapping between music and dance with realistic outputs.

Creativity and Diversity Earlier works in dance choreography [11, 19] adopted a similarity retrieval approach which simply glues together dance moves from a learned template and is lacking in both innovation and diversity. [31] utilized a sequence-to-sequence model with Long Short Term Memory (LSTM) units which suffered from the limitation of a single output. However, multiple interpretations for music composition or dance choreography is commonplace and diversity should also be reflected in the model. One viable approach [18, 29] is to employ Generative Adversarial Networks (GANs) that enables a distribution of plausible outputs instead of a single deterministic one.

Consistency between music and dance On top of securing the realism of the generated music pieces or dance sequences, we need to ensure harmony between the generated output and the conditioned input. In other words, we need to extract the shared themes and intentionality between the two media as domain-independent features and ensure that such domain-independent abstractions are reflected in the target output. Furthermore, the kinematic beats in dance and the acoustic beats in music should be aligned.

To address these challenges, our proposed approach employs transformer network architectures [34] as encoders and decoders in a sequence-to-sequence framework. We incorporate a full attention mechanism [8] for feature learning in both the music and dance domain. This has the key advantage of a global level understanding of underlying themes. We instill diversity into our model by

concatenating a random vector with the encoded feature. Furthermore, to effectively leverage the dual structure of the problem, we also propose an optimal transport inspired alignment that serves to match cross-domain features. Specifically, we define a Gromov-Wasserstein distance [26] which measures the *relational* distance between intra-domain distances while preserving the domain topological structures. Optimizing the Gromov-Wasserstein objective facilitates feature learning in each encoder network. It therefore promotes the proximity and similarity of the learnt embeddings despite inherent differences in domain topology.

Our contributions may be summarized as follows:

- (1) We consider a new problem of both music-to-dance generation and its dual dance-to-music generation.
- (2) A novel dual learning strategy is proposed, which incorporates a Gromov-Wasserstein distance to facilitate feature learning of each task, as well as promote coherence between input and output.
- (3) Empirical experiments demonstrate the applicability of our approach in delivering realistic, diverse generations of dance choreographs and music compositions faithful to the input. Moreover, superior performance is observed when comparing to the state-of-the-arts in music-to-dance generation.

2 RELATED WORK

2.1 Dance generation

Earlier works generate dance sequences from fixed templates following a similarity matching for the music source [11, 19]. Since the synthesized choreographs simply re-arrange the dance moves from the training data in rigid fashions, there is the drawback of unnatural transitions and a lack of creativity.

With the advent of deep learning methods, sequence-to-sequence models have been proposed for generating dance sequences from encoded music features. A pioneering work [31] uses a L2 loss for comparing the dance sequences, which suffers from a tendency of motion freezing. To alleviate this, [15] proposes curriculum learning coupled with a L1 loss whereas [39] employs a geodesic loss. To enable the generation of diverse dance sequences, [15, 39] introduce a random seed vector alongside the encoded music features, while an alternative approach [18, 29] utilises GANs. One key remark is that [15, 18, 29] focus on 2D motion. The 3D representation in our work allows important cues such as relative positions or invariance on bone lengths to be clearly put into perspective, thus appearing more realistic, appealing and geometrically rich.

Another line of work [21, 22] employs a cross-modal architecture for generating dance sequences conditional on both music *and* previous dance moves. In our work, the primary task of music-to-dance generation is only conditional on a single music stream and does not require any additional dance stream.

2.2 Music generation

There are two general approaches in computational music generation. The first focuses on symbolic representations [5, 14, 16]. The second handles music as raw waveforms either in the audio [7, 9, 10] or frequency domain [33]. Ideally, the waveform representation

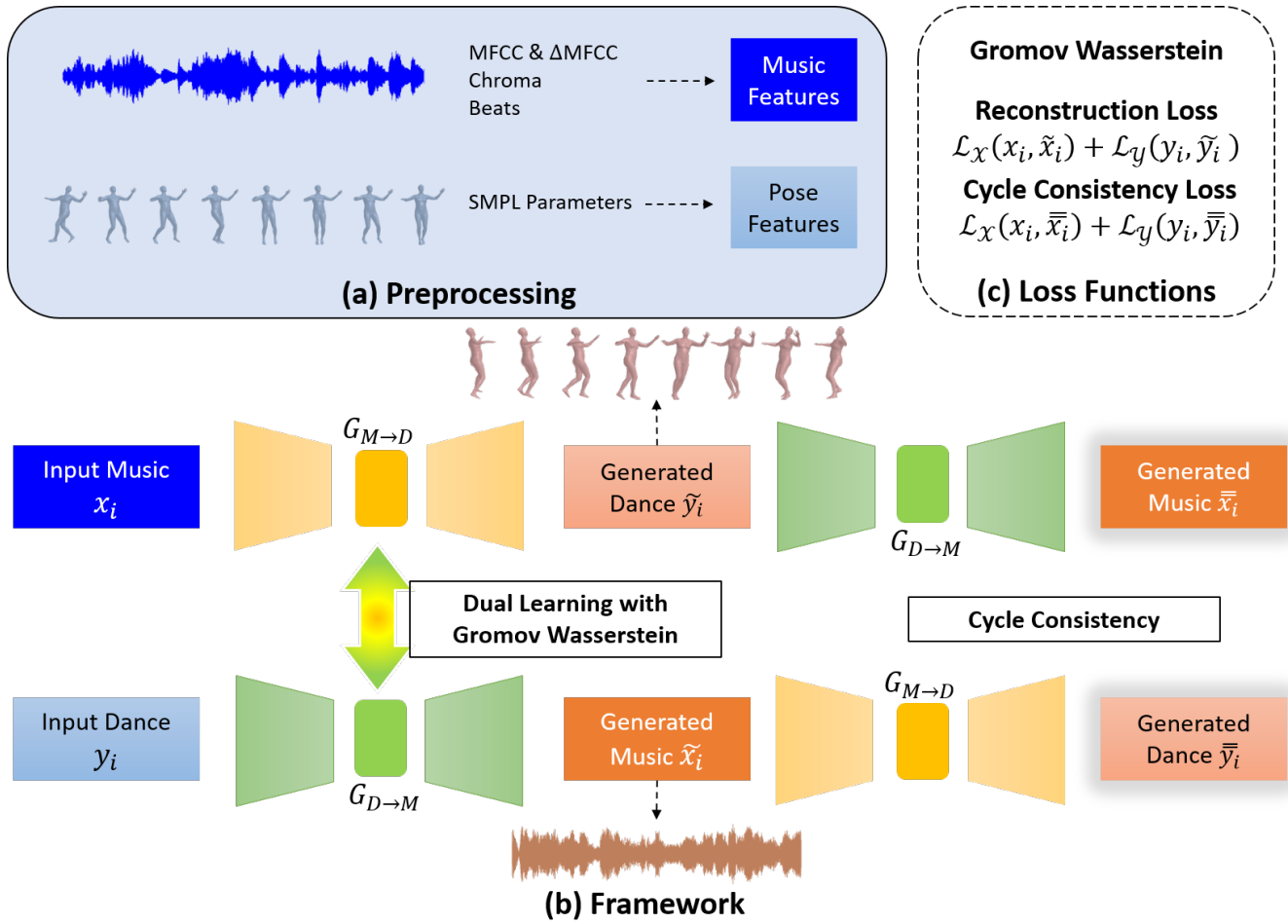


Figure 2: A high-level overview of our pipeline. (a) For data preprocessing, we extract MFCC, chroma and beats features from the music waveform raw data, and represent the dance sequence as pose and translation parameters in the SMPL model [24]. (b) The two generative networks $G_{M \rightarrow D}$ (music-to-dance choreography) and $G_{D \rightarrow M}$ (dance-to-music composition) in our framework comprise a sequence-to-sequence architecture where the encoder and decoder are both transformer networks. (c) We train our network through a Gromov-Wasserstein loss that facilitates dual learning in this cross-domain situation, on top of the reconstruction loss and cycle consistency loss.

enables a much richer generative landscape, such as allowing generation of human voices or nuances in musical performances which attach an additional layer of interpretation on top of the sheet music. However, this comes at an extremely high computational cost. To put this in context, the waveform representation for a 5 seconds music sequence sampled at 48kHz would incur a 240,000 length spectrogram sequence. Despite recent advancements such as sparse attentions [7] or discretized representations [9, 10], learning the multiple levels of musical structure and hierarchy which manifest at different scales remains extremely computationally expensive. In light of this, we resort to the lightweight symbolic representation for our music generation task.

2.3 Dual Learning

Dual learning [36] is a paradigm which jointly trains a primary task and its dual task. The symmetry and duality at the data or model level [37] can be leveraged, ideally improving the performance of both tasks compared to training for each independently. Dual learning can be in a supervised setting such as in image-to-image translations for paired data [17] or machines language translation [13]. It may likewise be extended to the unsupervised setting [40, 42] whereby the notion of a cycle consistency loss is introduced, compelling the primary task and dual task to learn inverse maps of each other. In our dual learning scheme, on top of the cycle consistency loss, we also introduce a novel Gromov-Wasserstein loss to facilitate feature space alignment.

3 OUR APPROACH

An overview of our pipeline is outlined in Fig. 2. In the following, we present the details of each component.

3.1 Data Preprocessing

Music The raw music input are sampled at 48kHz, thereby obtaining a waveform representation in the form of a time-frequency spectrogram. However, such high sampling rates may well lead to high computational costs and lots of redundancy. To this end, we employ the Librosa toolbox [25] to perform feature engineering and extraction. These extracted features will be used in our networks instead of the raw waveform.

Following [18], we first extract Mel-frequency cepstral coefficients (MFCC) [38] and MFCC delta from the spectrogram. However, these low level features (typically used for speech recognition) may be inadequate for conveying high level musical information. Thereafter, we perform a harmonic percussive source separation [27] which decomposes the spectrogram into harmonic components and percussive components. The harmonic components correspond to the pitch and melody of the music from which we extract chroma features. The percussive components provide the rhythmic information from which we extract the beats and onsets. Specifically, for each frame, our music features is a 53 dimensional vector, comprising 20-dim MFCC, 20-dim MFCC delta, 12-dim chroma features, and 1-dim one-hot encoding for beats.

Dance We work with two 3D dance sequence datasets from [22, 31]. [22] performed a 3D annotation of the AIST Dance database [32] and each 3D pose is parameterized in the Skinned Multi-Person Linear Model [24]. For consistency, we performed a fitting with SMPLify [3] on the dataset in [31] to obtain SMPL pose parameters.

The SMPL model has 24 skeletal joints and a pose is represented as 24×3 axis-angle parameters, which characterizes the 3D orientation or rotation of each joint. However the axis-angle (and likewise the quaternion) parameterization is not globally continuous over the 3D rotational group $SO(3)$ [2]. Furthermore, it is cumbersome to define a geometrically meaningful loss for axis-angle parameters and we adopt a Stiefel manifold representation [41]:

$$\hat{R} = \begin{pmatrix} | & | \\ \mathbf{R}_1 & \mathbf{R}_2 \\ | & | \end{pmatrix}, \quad (1)$$

which essentially amounts to discarding the last column for a rotation matrix. Such a representation is smooth over $SO(3)$, thus offering empirical advantages for backpropagation. Overall, the skeletal pose is parameterized as a $24 \times 6 + 3 = 147$ -dim vector.

3.2 Problem Formulation

We denote the music space as \mathcal{X} and the dance space as \mathcal{Y} . We sample fixed duration (T frames) music and dance sequence pairs $(x_i, y_i)_{i=1}^N$ i.i.d from $\mathcal{X} \times \mathcal{Y}$. Formally, the problem statement can be formulated as follows.

- The primary task is to learn a mapping $G_{M \rightarrow D} : \mathcal{X} \rightarrow \mathcal{Y}$ such that $\sum_{i=1}^N \mathcal{L}_{\mathcal{Y}}(G_{M \rightarrow D}(x_i), y_i)$ is minimized. Here $\mathcal{L}_{\mathcal{Y}}$ denotes a metric in dance space \mathcal{Y} .

- The dual task is to learn a mapping $G_{D \rightarrow M} : \mathcal{Y} \rightarrow \mathcal{X}$ such that $\sum_{i=1}^N \mathcal{L}_{\mathcal{X}}(G_{D \rightarrow M}(y_i), x_i)$ is minimized. Here $\mathcal{L}_{\mathcal{X}}$ denotes a metric in music space \mathcal{X} .

In the ideal case, $G_{M \rightarrow D}$ and $G_{D \rightarrow M}$ would be inverse of each other, i.e. $G_{D \rightarrow M}(G_{M \rightarrow D}(x_i)) = x_i$ and $G_{M \rightarrow D}(G_{D \rightarrow M}(y_i)) = y_i$. The discrepancy from this ideal case may be leveraged in the form of a cycle discrepancy loss [42] in our dual learning context.

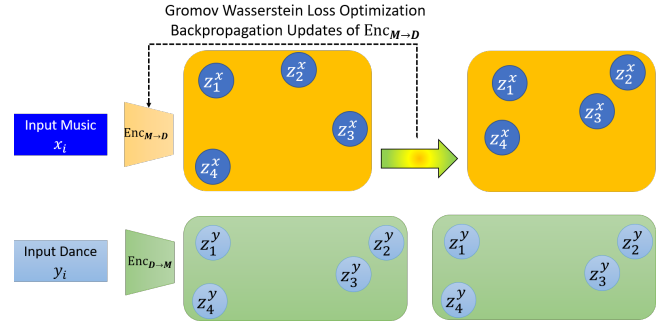


Figure 3: In optimizing for the Gromov-Wasserstein loss, the $Enc_{M \rightarrow D}$ network parameters are updated, re-positioning the music embedding vectors $\{z_i^x\}_{i=1}^4$ such that the discrepancy between intra-space distances is minimized.

3.3 Overall Framework

Transformer and attention Both the $G_{M \rightarrow D}$ and $G_{D \rightarrow M}$ consist of transformer networks that serve as encoders and decoders as illustrated in Fig. 2. We adopt a full attention mechanism [8] throughout for all transformers. This allows full access to the entire context, without masking of the future contexts. Explicitly, for a given sequence input S , an attention layer learns the context Z :

$$Z = \text{Softmax} \left(\frac{\langle \mathbf{S} \mathbf{W}_q, \mathbf{S} \mathbf{W}_k \rangle}{\sqrt{D}} \right) \mathbf{S} \mathbf{W}_v, \quad (2)$$

where D is the number of channels in the attention layer and $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v$ denote the query, key and value weights respectively.

Intuitively, the full attention mechanism aligns with how humans go about composing music for dance or choreographing dance from music. We make a global consideration of the structure and themes in our design. In allowing access to all contextual information, we expect that the structural constructs and hierarchical abstractions can be better modeled in the transformer networks.

Gromov-Wasserstein Loss to Facilitate Dual Learning in Encoders Each encoder network learns an embedding map from its respective input space into a latent space. Explicitly, we have:

$$\begin{aligned} Enc_{M \rightarrow D} : \mathcal{X} &\rightarrow \mathcal{Z}_{\mathcal{X}} \\ x &\mapsto z^x, \\ Enc_{D \rightarrow M} : \mathcal{Y} &\rightarrow \mathcal{Z}_{\mathcal{Y}} \\ y &\mapsto z^y. \end{aligned} \quad (3)$$

The music space \mathcal{X} and dance space \mathcal{Y} have different topological structures. As such, it would be mathematically impossible to embed them into the same space since embeddings are topology-preserving maps. This means that we cannot directly gauge the similarity

between a piece of music and a dance sequence by defining a metric on their embedding vectors.

On the other hand, embeddings do introduce metric distortions [4]. This motivates us to propose a Gromov-Wasserstein loss [1, 26, 28] which measures the discrepancy between the metrics on the two embedding spaces instead of directly comparing between cross-domain samples. By defining an *inter-space* distance between the respective *intra-space* distances for \mathcal{Z}_X and \mathcal{Z}_Y , the Gromov-Wasserstein loss gives a well-defined notion of distance between music and dance pairs. Heuristically, we present in Fig. 3 how optimizing this loss amounts to aligning the feature embeddings.

Give two point sets of m embedded vectors $\{z_i^x\}_{i=1}^m, \{z_i^y\}_{i=1}^m$, we may view them as two discrete empirical distributions μ, ν as:

$$\mu = \sum_{i=1}^m \frac{1}{m} \delta_{z_i^x}, \quad \nu = \sum_{i=1}^m \frac{1}{m} \delta_{z_i^y}, \quad (4)$$

where δ denotes the Dirac delta distribution. Formally, the Gromov-Wasserstein distance for our task is given by

$$GW(\mu, \nu) = \min_{\pi \in \Pi} \sum_{i,j,k,l} \left| \|z_i^x - z_k^x\|_1 - \|z_j^y - z_l^y\|_1 \right|^2 \pi_{ij} \pi_{kl}. \quad (5)$$

Here Π defines the set of all joint distributions with marginals μ and ν . The goal is to find the optimal transport matrix π that minimizes the squared distance between intra-space L1 costs.

Following [28], we introduce an entropic regularization term which allows much more efficient solving of Eq. (5). The entropy regularized Gromov-Wasserstein distance can then be solved via the Sinkhorn algorithm and projected gradient descent [28]. We outline the steps in Algorithm 1.

Algorithm 1 GW Distance for 2 batches of m samples

Input: music embeddings $Z^x = \{z_i^x\}_{i=1}^m, Z'^x = \{z_i'^x\}_{i=1}^m$

Input: dance embeddings $Z^y = \{z_i^y\}_{i=1}^m, Z'^y = \{z_i'^y\}_{i=1}^m$

Hyperparameters: regularization $\varepsilon > 0$, projection iterations M , Sinkhorn iterations L

Initialize: $\pi_{ij}^{(0)} = \frac{1}{n}, \forall i, j$

Cost Matrix for music embeddings $C_{ij} = \|z_i^x - z_j^x\|_1$

Cost Matrix for dance embeddings $D_{ij} = \|z_i^y - z_j^y\|_1$

for $l = 1 : M$ **do**

$$E = \frac{1}{m} D^2 \mathbb{1}_m \mathbb{1}_m^\top + \frac{1}{m} \mathbb{1}_m \mathbb{1}_m^\top C^2 - 2D\pi^{(l-1)}C^\top$$

$$K = \exp(-E/\varepsilon)$$

$$\mathbf{b}^{(0)} = \mathbb{1}_m$$

for $\ell = 1 : L$ **do**

$$\mathbf{a}^{(\ell)} = \mathbb{1}_m \oslash K \mathbf{b}^{(\ell-1)}, \quad \mathbf{b}^{(\ell)} = \mathbb{1}_m \oslash K^\top \mathbf{a}^{(\ell)}$$

where \oslash denotes component-wise division

end for

$$\pi^{(l)} = \text{diag}(\mathbf{a}^{(L)}) K \text{diag}(\mathbf{b}^{(L)})$$

end for

Output: $GW_\varepsilon(Z^x, Z'^x, Z^y, Z'^y) = \sum_{i,j,k,l} \|D_{ik} - C_{jl}\|^2 \pi_{ij}^{(M)} \pi_{kl}^{(M)}$

Optimizing the Gromov-Wasserstein distance through Algorithm 1 updates the music encoder network parameters $\text{Enc}_{M \rightarrow D}$ through backpropagation. This matching of the music embeddings with the dance embeddings facilitates learning the duality of our two tasks.

Decoding and Generation During training phase, the start token for the decoder network is taken from the training data. Specifically, for the dance generation task, the start token is the initial dance pose. During inference phase, we sample a random seed vector for the initial pose instead. For the music generation task, during inference, we randomly sample a chord root from $\{C, C\#, D, D\#, E, F\#, G, G\#, A, A\#, B\}$ and a chord quality among {major, minor}. These random seeds during the inference phase introduce diversity in the generated dance and music.

3.4 Loss Functions

As illustrated in Fig. 2, three loss functions are defined. The first is our Gromov-Wasserstein loss, which has been discussed in detail in the previous subsection. Essentially it serves as an auxiliary regularization loss that promotes a better learning of the correspondence between the music and dance embeddings. This facilitates extracting the commonalities and shared structural similarities between the two media, which would enhance the consistency between inputs and generated outputs in our tasks.

The reconstruction loss has the general form:

$$\begin{aligned} \mathcal{L}_{\text{dance}}^{\text{reconstruction}} &= \sum_i \mathcal{L}_Y(y_i, G_{M \rightarrow D}(x_i)), \\ \mathcal{L}_{\text{music}}^{\text{reconstruction}} &= \sum_i \mathcal{L}_X(x_i, G_{D \rightarrow M}(y_i)). \end{aligned} \quad (6)$$

\mathcal{L}_Y denotes the metric over the dance space \mathcal{Y} . Specifically, we have $\mathcal{Y} = (\mathbb{R}^3 \times SO(3) \times \dots \times SO(3))^T$ where T is the total number

of frames in our model. A point $y \in \mathcal{Y}$ may then be written as $y = [y_t^{\text{trans}}, y_t^{\text{rot}_1}, \dots, y_t^{\text{rot}_{24}}]_{t=1}^T$. We define $\mathcal{L}_Y : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ as follows:

$$\mathcal{L}_Y(y, \tilde{y}) = \underbrace{\sum_{t=1}^T \|y_t^{\text{trans}} - \tilde{y}_t^{\text{trans}}\|_1}_{\text{L1 loss for translation}} + \sum_{t=1}^T \sum_{j=1}^{24} \text{geodesic}(y_t^{\text{rot}_j}, \tilde{y}_t^{\text{rot}_j})^2. \quad (7)$$

Here, $\text{geodesic} : SO(3) \times SO(3) \rightarrow \mathbb{R}_+$ defines the shortest distance between two 3D rotations. Recall that $y_t^{\text{rot}_j}$ constitutes the first two columns of its associated rotation matrix R_t^j as defined in Eq. (1), we can recover its third column through the cross product. The geodesic distance between rotation matrices R, \tilde{R} is given by:

$$\text{geodesic}(R, \tilde{R}) = \left| \arccos \left[\frac{\text{Tr}(R\tilde{R}^\top) - 1}{2} \right] \right|. \quad (8)$$

For the music space \mathcal{X} , we compare only the chroma and beats. The metric $\mathcal{L}_X : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$ is defined as:

$$\mathcal{L}_X(x, \tilde{x}) = \underbrace{\sum_{t=1}^T \|x_t^{\text{chroma}} - \tilde{x}_t^{\text{chroma}}\|_1}_{\text{L1 loss for chroma}} + \underbrace{\sum_{t=1}^T \|x_t^{\text{beats}} - \tilde{x}_t^{\text{beats}}\|_1}_{\text{L1 loss for beats}}. \quad (9)$$

The cycle consistency loss is another regularization term that we incorporate to facilitate dual learning. Whereas the Gromov-Wasserstein loss is applied at an initial stage to update the music embedding network parameters, the cycle consistency loss is applied at the posterior stage to gauge the generated rendition against the input.

With the same metrics defined in Eqs. (7) and (9), the cycle consistency loss measures the discrepancy of the two dual networks $G_{M \rightarrow D}$ and $G_{D \rightarrow M}$ from being inverse to each other:

$$\begin{aligned}\mathcal{L}_{\text{dance}}^{\text{cycle}} &= \sum_i \mathcal{L}_{\mathcal{Y}}(y_i, G_{M \rightarrow D}(G_{D \rightarrow M}(y_i))), \\ \mathcal{L}_{\text{music}}^{\text{cycle}} &= \sum_i \mathcal{L}_{\mathcal{X}}(x_i, G_{D \rightarrow M}(G_{M \rightarrow D}(x_i))).\end{aligned}\quad (10)$$

Our transformer networks $G_{M \rightarrow D}$ and $G_{D \rightarrow M}$ are trained concurrently and updated according to the following prescriptions:

$$\begin{aligned}\mathcal{L}^{\text{GW}} + \mathcal{L}_{\text{dance}}^{\text{reconstruction}} + \mathcal{L}_{\text{dance}}^{\text{cycle}} &\xrightarrow{\text{backpropagation}} G_{M \rightarrow D}, \\ \mathcal{L}_{\text{music}}^{\text{reconstruction}} + \mathcal{L}_{\text{music}}^{\text{cycle}} &\xrightarrow{\text{backpropagation}} G_{D \rightarrow M}.\end{aligned}$$

4 EXPERIMENTS

4.1 Dataset Description

As mentioned in subsection 3.1, we employ two public datasets from [31] and [22, 32], which consists of 3D dance sequences with accompanying music. Since these two datasets do not overlap in their dance genres, we combine them into a single entity. We first re-parameterized the skeletal pose representation in [31] in the SMPL model [24] pose parameters (3D positional representation of 17 keypoints \rightarrow rotational pose representation of 24 keypoints) through an inverse kinematics fitting. For consistency in frame rate, we also perform a downsampling (via spherical interpolation) of the frames per second (FPS) in [22] from 60 to 25. The statistics for the aggregated dataset is summarized in the following table.

Dance Genre	# of Sequences	Frames	Remarks
Rumba	10	20950	From [31]. Dance sequences are typically much longer at around 150 seconds.
Cha Cha	8	20425	
Tango	9	49165	
Waltz	34	43298	
Break Dance	141	46526	From [22]. Dance sequences generally range from 8 to 12 seconds.
House	141	40050	
Ballet Jazz	141	47727	
Street Jazz	141	47920	
Krump	141	47534	
LA Style Hip Hop	141	48323	
Lock	141	47388	
Middle Hip Hop	141	48276	
Pop	140	46749	
Waack	140	47355	

Table 1: Summary of dataset, aggregated from [31] and [22, 32]. For our tasks, the frame rate is set to 25 FPS. During training, the input and output sequences are fixed at 75 frames (or 3 seconds).

4.2 Implementation Details

Our framework is implemented with PyTorch. We set the input and output sequence lengths to $T = 75$ frames (equivalent to 3 seconds). Both $G_{M \rightarrow D}$ and $G_{D \rightarrow M}$ have 6 transformer layers and 8 attention heads. The hidden units dimensions for the transformer layers is set to 512 for $G_{M \rightarrow D}$ and 256 for $G_{D \rightarrow M}$. For our Gromov-Wasserstein loss, the hyperparameters for Algorithm 1 are as follows: the entropic regularization parameter is set to $\varepsilon = 0.2$ while the number

of Sinkhorn iterations and projection iterations are respectively set to $L = 30$ and $M = 20$. During training, we use the Adam optimizer with a batch size of 16 and an initial learning rate of $1e-4$ (decays to $\{1e-5, 5e-6\}$ after $\{20k, 40k\}$ iterations).

During inference, the decoder start token may either be sampled randomly from a pool (for diverse generation setting) or provided as an auxiliary input. For inference over 75 frames, we simply feed the last generated frame as the decoder start token.

4.3 Dance Generation

For generation of music-to-dance choreography, we compare our work against [15, 29, 31]. For [31], we re-implement it since the source code is not available. We also adapt [15, 29] (both implemented for 2D dance sequences) for 3D dance generation. We employ three quantitative metrics, namely the Fréchet Distance, Diversity and Beats Alignment. The details of these metrics are discussed below and the quantitative results are reported in Tab. 2.

Furthermore, we engage a qualitative user study to rate the realism and the genre consistency of the generated dances. The results are tabulated in Tab. 3.

We showcase sample dance animations in Fig. 4 and 5. These sequences spanning 10 seconds are animated at 5 FPS (requires Adobe Acrobat Reader). More sample sequences are available in the supplementary video.

Method	Fréchet Distance	Diversity	Beats Alignment (%)
Ground Truth	-	-	68.7
Tang et al. (2018) [31]	986.4	10.3	31.2
Ren et al. (2020) [29]	1526.3	48.2	49.5
Huang et al. (2021) [15]	384.2	37.2	62.3
Ours	140.5	49.8	64.5

Table 2: Quantitative results for dance generation. We adopt three metrics: 1) *Fréchet distance* measures the difference from the ground truth, 2) *Diversity* measures the variation in dance moves, and 3) *Beats Alignment* evaluates the percentage matching of kinematic beats and music beats. Best performance highlighted in bold.

Fréchet Distance We define the Fréchet distance as the average 3D joint distance of the generated dance sequence from the ground truth. In our experiments, we evaluate the Fréchet distance for 280 sequences (20 per genre, each spanning 9 seconds) in our pre-allocated test set. Our framework achieves significantly better performance than compared methods. This superior matching of the generated dances with the ground truth suggests that the music-to-dance correspondence is better modeled in our approach.

Sample generated Cha Cha sequences are shown in Fig. 4 for the same initial pose and input music. We observe that our network generates sequences more consistent with the ground truth. [31] has a inclination to gravitate towards motionless states, showing very limited range of motions. On the other hand, while not suffering from lack of motions, [29] tends to deviate from the ground truth.

Diversity Recall that we incorporate randomness into our decoder through a randomly seeded start token. This allows our framework to generate diverse dance moves from the same input music but a different initial pose. We define our diversity metric as the variation in 3D joints position, evaluated over 5 generated dance sequences conditioned on the same music and different initial






Method	Click ↓	Cha Cha dance sequences displayed at 0.2 second intervals
Ground Truth		
Tang et al. (2018) [31]		
Ren et al. (2020) [29]		
Huang et al. (2021) [15]		
Ours		

Figure 4: Visual comparison of sample generated Cha Cha dance sequences. Best viewed via Adobe Acrobat Reader. Click the figures under ‘Click ↓’ to show dance animations (4 seconds at 5 FPS).









Genre	Click ↓	Dance sequences displayed at 0.2 second intervals
Cha Cha 		
		
		
Break Dance 		
		
		

Figure 5: Our framework can generate diverse dance sequences given the same input music but a different starting pose. Best viewed via Adobe Acrobat Reader where dance animations will be played upon clicking the figures under ‘Click ↓’.

poses. This is averaged over 20 independent trials for each genre. In Fig. 5, We showcase the dance choreographs obtained by our model given the same input music. The diverse generation is indicative that our model is adept at learning the abstract movements in music and translating them to kinematic movements in dance.

Beats Alignment Beats alignment measures the consistency of rhythmic articulation. For simplicity, we follow [22] in defining dance kinematics beats as local minima in average joints 3D speed. We then define a matching beat if the occurrence of the kinematics beat occurs within 5 frames or 0.2 seconds of a music beat. The beats alignment ratio can then be defined as the ratio of matching beats

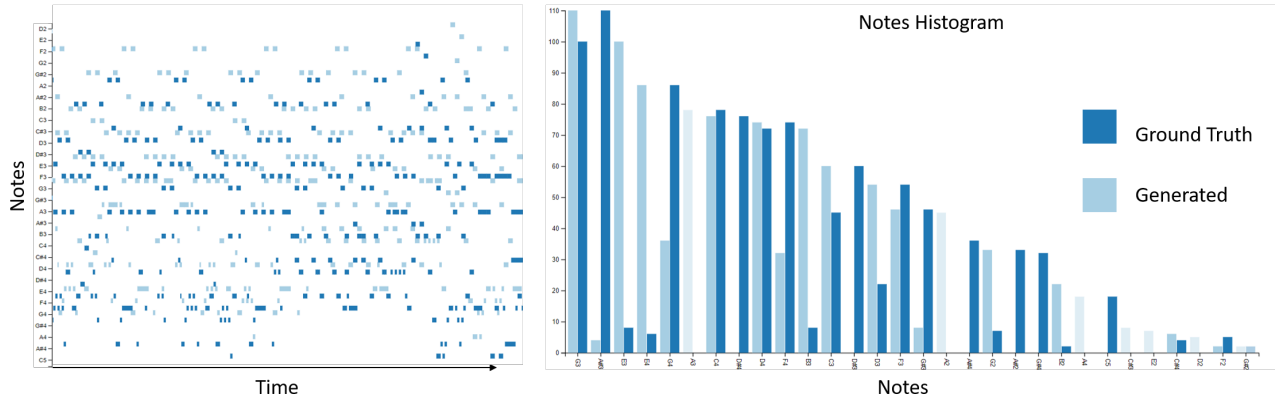


Figure 6: Visualization of sample generated music (light blue) and original music (dark blue). Left: Visualization of music notes played out in time. Right: Histogram of music notes.

Method	Realism (Ranking)	Genre Consistency (Ranking)
Tang et al. (2018) [31]	3.9	3.80
Ren et al. (2020) [29]	2.98	2.58
Huang et al. (2021) [15]	2.05	2.35
Ours	1.07	1.27

Table 3: User study for dance generation. We report the average ranking of each method for motion realism and dance genre consistency.

to the total dance beats. Our framework again achieves compelling performance for beats matching, demonstrating its effectiveness in learning the rhythmic articulation of the input music.

User Study We engage 5 users (amateur dancers) to assess the generated dances in a single blind study. For each dance genre, we generate 3 dance sequences of 12 seconds duration via different methods. Each user then rank the generated dances according to two criteria: 1) the naturalness or realism of the dance poses; 2) the consistency with the dance genre. As shown in Tab. 3, the dance sequences generated from our framework is overwhelmingly ranked most preferred for both realism and genre consistency.

4.4 Music Generation

To our knowledge, we are the first to investigate the task of composing music for given dance sequences. We propose to quantify the performance of our model via a notes accuracy metric. We first transpose both compared music pieces to either the C major or A minor scale (depending on its original chord quality), before computing the average notes accuracy. Our model achieves an accuracy of 72%, as tabulated in Tab. 4.

However, this metric is not ideal since music composition is rather subjective task and is best assessed via human listening tests [16]. To this end, we invite the reader to look over and listen to the sample music compositions in our supplementary file. A sample generated music piece, is shown in tandem with the ground truth in Fig. 6. The visualization is done with JuxtaMidi [12].

4.5 Ablation Studies

We perform ablation studies focused on investigating the effectiveness of our dual learning scheme. The first key component

GW	CC	Dance			Music
		Fréchet Distance	Diversity	Beats Alignment (%)	Notes Accuracy
	✓	335.4	46.3	58.5	0.51
✓		196.2	50.1	60.3	0.59
		376.5	46.2	56.4	0.37
✓	✓	140.5	49.8	64.5	0.72

Table 4: Ablation experiments performed on two components of our framework. GW refers to the Gromov-Wasserstein loss and CC refers to the cycle consistency loss.

in our framework is the Gromov-Wasserstein loss, applied at the early stage of a training iteration to align the music embeddings with dance embeddings. The second component is the cycle consistency loss, optimized at the final stage of a training iteration as an additional regularization to enforce overall consistency.

The ablation experiments are reported in Tab. 4. We observe significant performance drops for the Fréchet distance, beats alignment ratio and music notes accuracy upon removal of the Gromov-Wasserstein objective. This provides strong justification of its potency for handling the duality of the tasks. The cycle consistency loss, is also effective, albeit to a lesser extent. Overall, compared to independently training separate task-specific networks, the conjoined learning in our framework delivers significant improvements through a balanced set-up of well-tailored components.

5 CONCLUSION

In this work, we propose a novel problem of simultaneously learning music-to-dance choreography and dance-to-music composition. The crucial ingredient is how to effectively leverage the duality of the tasks and integrate the information from both domains. To overcome the challenge of our cross-domain setting, we design a Gromov-Wasserstein objective for aligning the music embeddings vis-à-vis the dance embeddings, coupled to a cycle consistency loss. These auxiliary losses and our dual learning scheme prove capable in boosting the performance of individual tasks. Our framework delivers realistic and genre-consistent dance generations, as well as viable music compositions. For future work, we seek to extend our framework to raw waveform based music composition and explore multi-persons dance choreography.

REFERENCES

- [1] David Alvarez-Melis and Tommi Jaakkola. 2018. Gromov-Wasserstein Alignment of Word Embedding Spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 1881–1890.
- [2] Luis Astey et al. 1987. A cobordism obstruction to embedding manifolds. *Illinois Journal of Mathematics* 31, 2 (1987), 344–350.
- [3] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. 2016. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European conference on computer vision*. Springer, 561–578.
- [4] Jean Bourgain. 1985. On Lipschitz embedding of finite metric spaces in Hilbert space. *Israel Journal of Mathematics* 52, 1-2 (1985), 46–52.
- [5] Jean-Pierre Briot, Gaëtan Hadjeres, and François-David Pachet. 2017. Deep learning techniques for music generation—a survey. *arXiv preprint arXiv:1709.01620* (2017).
- [6] Steven Brown and Lawrence M Parsons. 2008. The neuroscience of dance. *Scientific American* 299, 1 (2008), 78–83.
- [7] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509* (2019).
- [8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [9] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. 2020. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341* (2020).
- [10] Sander Dieleman, Aäron van den Oord, and Karen Simonyan. 2018. The challenge of realistic music generation: modelling raw audio at scale. *arXiv preprint arXiv:1806.10474* (2018).
- [11] Rukun Fan, Songhua Xu, and Weidong Geng. 2011. Example-based automatic music-driven conventional dance motion synthesis. *IEEE transactions on visualization and computer graphics* 18, 3 (2011), 501–515.
- [12] Jeremy Grifski. [n.d.]. JuxtaMidi. <https://therenegadecoder.com/code/juxtamidi-a-midi-file-visualization-dashboard/>. Accessed: 2021-04-17.
- [13] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. *Advances in neural information processing systems* 29 (2016), 820–828.
- [14] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M Dai, Matthew D Hoffman, Monica Dinulescu, and Douglas Eck. 2018. Music transformer. *arXiv preprint arXiv:1809.04281* (2018).
- [15] Ruozhi Huang, Huang Hu, Wei Wu, Kei Sawada, Mi Zhang, and Daxin Jiang. 2021. Dance Revolution: Long-Term Dance Generation with Music via Curriculum Learning. In *International Conference on Learning Representations*.
- [16] Yu-Siang Huang and Yi-Hsuan Yang. 2020. Pop Music Transformer: Beat-based modeling and generation of expressive Pop piano compositions. In *Proceedings of the 28th ACM International Conference on Multimedia*. 1180–1188.
- [17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1125–1134.
- [18] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. 2019. Dancing to music. In *Advances in Neural Information Processing Systems*. 3586–3596.
- [19] Minho Lee, Kyogu Lee, and Jaeheung Park. 2013. Music similarity-based approach to generating dance motion sequence. *Multimedia tools and applications* 62, 3 (2013), 895–912.
- [20] Daniel J Levitin and Anna K Tirovolas. 2009. Current advances in the cognitive neuroscience of music. *Annals of the New York Academy of Sciences* 1156, 1 (2009), 211–231.
- [21] Jiaman Li, Yihang Yin, Hang Chu, Yi Zhou, Tingwu Wang, Sanja Fidler, and Hao Li. 2020. Learning to Generate Diverse Dance Motions with Transformer. *arXiv preprint arXiv:2008.08171* (2020).
- [22] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. 2021. Learn to Dance with AIST++: Music Conditioned 3D Dance Generation. *arXiv preprint arXiv:2101.08779* (2021).
- [23] Jianxin Lin, Yingce Xia, Tao Qin, Zhibo Chen, and Tie-Yan Liu. 2018. Conditional image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5524–5532.
- [24] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2015. SMPL: A skinned multi-person linear model. *ACM transactions on graphics (TOG)* 34, 6 (2015), 1–16.
- [25] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, Vol. 8. 18–25.
- [26] Facundo Mémoli. 2011. Gromov-Wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics* 11, 4 (2011), 417–487.
- [27] Meinard Müller. 2015. *Fundamentals of music processing: Audio, analysis, algorithms, applications*. Springer.
- [28] Gabriel Peyré, Marco Cuturi, and Justin Solomon. 2016. Gromov-wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*. 2664–2672.
- [29] Xuanchi Ren, Haoran Li, Zijian Huang, and Qifeng Chen. 2020. Self-supervised Dance Video Synthesis Conditioned on Music. In *Proceedings of the 28th ACM International Conference on Multimedia*. 46–54.
- [30] Christopher Small. 1998. *Musicking: The meanings of performing and listening*. Wesleyan University Press.
- [31] Taoran Tang, Jia Jia, and Hanyang Mao. 2018. Dance with melody: An LSTM-autoencoder approach to music-oriented dance synthesis. In *Proceedings of the 26th ACM international conference on Multimedia*. 1598–1606.
- [32] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. 2019. AIST Dance Video Database: Multi-Genre, Multi-Dancer, and Multi-Camera Database for Dance Information Processing. In *ISMIR*. 501–510.
- [33] Sean Vazquez and Mike Lewis. 2019. Melnet: A generative model for audio in the frequency domain. *arXiv preprint arXiv:1906.01083* (2019).
- [34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762* (2017).
- [35] Tianyan Wang. 2015. A hypothesis on the biological origins and social evolution of music and dance. *Frontiers in neuroscience* 9 (2015), 30.
- [36] Yingce Xia, Tao Qin, Wei Chen, Jiang Bian, Nenghai Yu, and Tie-Yan Liu. 2017. Dual supervised learning. In *International Conference on Machine Learning*. PMLR, 3789–3798.
- [37] Yingce Xia, Xu Tan, Fei Tian, Tao Qin, Nenghai Yu, and Tie-Yan Liu. 2018. Model-level dual learning. In *International Conference on Machine Learning*. PMLR, 5383–5392.
- [38] Min Xu, Ling-Yu Duan, Jianfei Cai, Liang-Tien Chia, Changsheng Xu, and Qi Tian. 2004. HMM-based audio keyword generation. In *Pacific-Rim Conference on Multimedia*. Springer, 566–574.
- [39] Zijie Ye, Haozhe Wu, Jia Jia, Yaohua Bu, Wei Chen, Fanbo Meng, and Yanfeng Wang. 2020. ChoreoNet: Towards Music to Dance Synthesis with Choreographic Action Unit. In *Proceedings of the 28th ACM International Conference on Multimedia*. 744–752.
- [40] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. 2017. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*. 2849–2857.
- [41] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. 2019. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5745–5753.
- [42] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2223–2232.