Music-to-Dance Generation with Optimal Transport

Shuang Wu[1,2] , **Shijian Lu**[2] and **Li Cheng**[3]

1Black Sesame Technologies
2Nanyang Technological University
3University of Alberta
wushuang@outlook.sg, shijian.lu@ntu.edu.sg, lcheng5@ualberta.ca

Abstract

Dance choreography for a piece of music is a challenging task, having to be creative in presenting distinctive stylistic dance elements while taking into account the musical theme and rhythm. It has been tackled by different approaches such as similarity retrieval, sequence-to-sequence modeling and generative adversarial networks, but their generated dance sequences are often short of motion realism, diversity and music consistency. In this paper, we propose a Music-to-Dance with Optimal Transport Network (MDOT-Net) for learning to generate 3D dance choreographies from music. We introduce an optimal transport distance for evaluating the authenticity of the generated dance distribution and a Gromov-Wasserstein distance to measure the correspondence between the dance distribution and the input music. This gives a well defined and non-divergent training objective that mitigates the limitation of standard GAN training which is frequently plagued with instability and divergent generator loss issues. Extensive experiments demonstrate that our MDOT-Net can synthesize realistic and diverse dances which achieve an organic unity with the input music, reflecting the shared intentionality and matching the rhythmic articulation. Sample results are found at https://www.youtube.com/watch?v=dErfBkrlUO8.

1 Introduction

Dance and music are intimately related. They share a movement form wherein for dance, movements are articulated visually as body motion whereas for music, movements manifest themselves via an auditory and allegorical nature. These movements, whether visual in dance or auditory in music, evoke emotions and feelings of intentionality. Looking at this intimate connection between dance and music from an artificial intelligence perspective, an intriguing question would be whether a computational model can generate a coherent and meaningful dance sequence for a given piece of music. This is an ambitious task which is useful for simulation, behavior understanding and ultimately benefit the vast community of dancers and musicians.
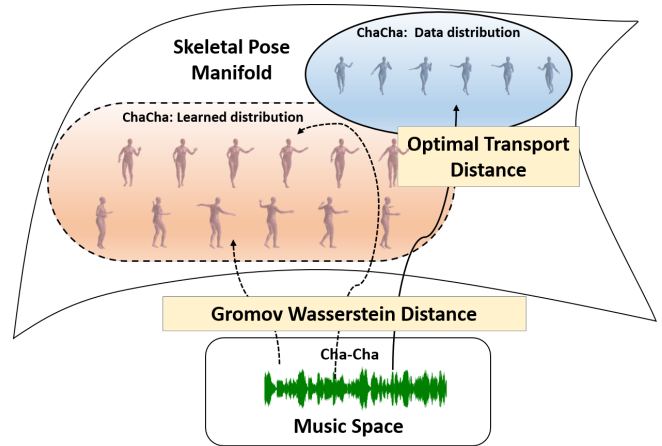

Figure 1: The proposed MDOT-Net generates a matching dance sequence given a music input. We model different genres of music and dance styles (e.g. cha cha and tango) by distributions lying within *Music Spaces* and an *Articulated Pose Manifold*, where the authenticity of the generated dance distributions is evaluated by an *Optimal Transport Distance* and the harmony between the music and the generated dance is measured by a *Gromov-Wasserstein Distance*.

In this paper, we present a framework generating 3D dance choreographies from music sequences. There are multiple challenges to be addressed for this task. 1) The generated dance motions have to be realistic and adhere to the idiosyncratic distinctions of the dance style. For example, a generated waltz sequence should reflect stylistic elements that are recognizable (even to the non-expert observer). 2) Dance choreography is inherently diverse and multiple choreographic interpretations for the same musical piece are ubiquitous. An adequate computational model would have to generate diverse and multimodal dance kinematics. 3) Choreographing for a piece of music has to take into consideration the rhythmic articulation, melody, theme and variation to achieve an organic unity, a challenging feat even for a professional human expert. The generated dance should be intricately bonded with the music input through a shared intentionality of the movement form, and reflect the melodic styles and rhythmic articulation.

Earlier works typically adopted a similarity retrieval ap-

proach [Fan *et al.*, 2011; Lee *et al.*, 2013] which lacks creativity. The sequence to sequence modeling approach in [Tang *et al.*, 2018] is also limited to a single output and unable to generate diverse dances for any piece of input music. A recent class of works turned to Generative Adversarial Networks (GANs) [Lee *et al.*, 2019; Ren *et al.*, 2020] for multimodal generation and enhance diversity. However, both [Lee *et al.*, 2019; Ren *et al.*, 2020] focused on 2D choreographies which lack in the dynamic richness and pose realism of 3D choreographies. Furthermore, tuning the adversarial training schemes in these GAN approaches is an arduous task [Artetxe *et al.*, 2018]. Discriminating the music and dance correspondence by mapping to a common embedding space also tends to be inadequate, resulting in lack of coherence between the dance and music.

To address these challenges and overcome the shortcomings of existing methods, we leverage optimal transport (OT) theory [Villani, 2008] and propose a Music-to-Dance with Optimal Transport Network (MDOT-Net). As illustrated in Figure 1, for a given music input, MDOT-Net generates diverse dance sequences that correspond to trajectories over a articulated pose manifold. Directly working with dance distributions supported on the articulated pose manifold is advantageous for realism of the dance poses, allowing subtle stylistic distinctions and nuances to be reflected. We evaluate the optimal transport distance between the generated and data distribution on this manifold. This offers several advantages such as the capability to handle non-overlapping distributions (a major issue for Jensen-Shannon divergence) [Arjovsky *et al.*, 2017] and non-divergent generator loss by reframing the adversarial training as an optimization problem [Genevay *et al.*, 2018; Salimans *et al.*, 2018].

Since dance and music exist in different domains, it is difficult to quantitatively gauge their differences directly. Therefore, we propose a Gromov-Wasserstein distance [Mémoli, 2011; Bunne *et al.*, 2019] to compare distributions over different domains (music space and articulated pose manifold). The Gromov-Wasserstein distance compares distributions in *relational* terms. The intuition is that for a matching pair of dance and music (D, M), a generated dance sequence D' is likely a good match for music sequence M' if D' is close to D and M' is close to M. Similar to the optimal transport distance in facilitating adversarial training, the Gromov-Wasserstein distance enables a more efficient approach in assessing the music and dance correspondence.

Our contribution are summarized as follows. 1) We develop a novel optimal transport framework for music to dance generation. The authenticity of the generated dance is measured via an optimal transport distance on the manifold of articulated poses. 2) A Gromov-Wasserstein distance is incorporated to facilitate learning cross modal generations from the music space to the articulated pose manifold through a *relational* rather than absolute measure of the music and dance similarity. 3) Our MDOT-Net can generate realistic and diverse 3D dance sequences faithful to the rhythm and melody of a given music input.

2 Related Works

Optimal Transport for Generative modeling. Optimal transport [Villani, 2008] defines a metric distance for probability distributions over arbitrary spaces. The generative modeling problem is reframed as finding an optimized transport for aligning the model distribution and data distribution. However, solving the optimal transport problem is expensive, and this computational burden presented major hurdles for employing optimal transport for generative modeling. The Wasserstein GAN [Arjovsky *et al.*, 2017] turned to the dual optimal transport problem and proposed a discriminator approximating 1-Lipschitz functions for GAN training. An alternative line of work was pursued in [Cuturi, 2013; Genevay *et al.*, 2016], in which the introduction of an entropic regularization term reduces the computational cost. The regularized primal optimal transport problem is amenable to backpropagation training [Genevay *et al.*, 2018; Salimans *et al.*, 2018].

[Mémoli, 2011] generalizes optimal transport for comparing distributions supported on different spaces, introducing the Gromov-Wasserstein distance as a notion of distance between intra-domain distances. The Gromov-Wasserstein distance is a promising metric for learning cross-domain correspondences [Bunne *et al.*, 2019] such as unsupervised language translation [Alvarez-Melis and Jaakkola, 2018] or graph matching [Xu *et al.*, 2019].

Dance Generation. Earlier works generally utilised similarity retrieval [Fan *et al.*, 2011; Lee *et al.*, 2013]. A major drawback is that the synthesized choreography appears rigid and lacks creativity, simply arranging the dance moves in the training data with unnatural transitions. [Tang *et al.*, 2018] employs Long Short Term Memories (LSTM) in a sequence to sequence modeling framework that generates motion features from encoded musical features. However, in using a L2 loss for the dance sequences, the synthesized motion are unrealistic and tends to incur motion freezing for longer sequences. Another shortcoming is the inability to generate diverse dance sequences.

[Huang *et al.*, 2021] proposes curriculum learning with L1 loss on the dance sequences to alleviate the motion freezing issue. It also introduces a noise vector on top of the encoded musical feature to enable multimodal generation. An alternative approach utilises GANs to synthesize multimodal dances. [Ren *et al.*, 2020] adopts a GAN framework with both a local and global discriminator to measure discrepancies between dance sequences. [Lee *et al.*, 2019] proposes a two-phase framework by first learning a decomposition of dance sequences into basic dance motion units and subsequently composing these dance units into a dance sequence with a GAN. These works focused on 2D poses, losing the geometric richness and realism of 3D motion. Important cues such as the dancer's position cannot be clearly put into perspective and invariance of bone lengths across frames is not enforced. More recently, [Valle-Pérez *et al.*, 2021; Li *et al.*, 2021; Wu *et al.*, 2021] explored transformer-based models for music-to-dance generation with [Wu *et al.*, 2021] proposing a dual learning framework of concurrently learning music composition conditional on dance inputs.
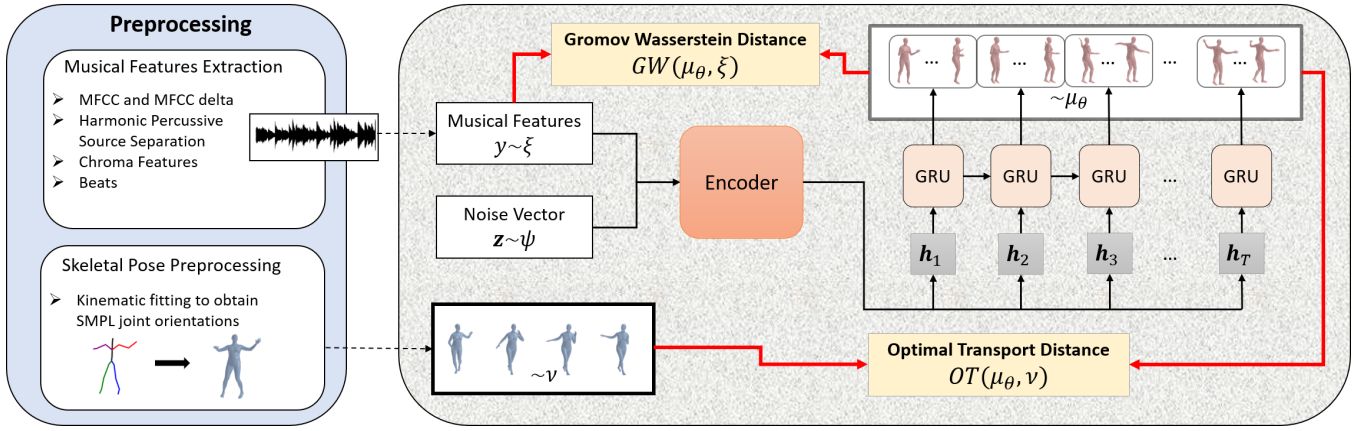
Figure 2: Overview: An *Encoder* maps the extracted *Musical Features* and *Noise Vector* into latent vectors $\mathbf{h}_1, \cdots, \mathbf{h}_N$. Each $\mathbf{h}_i$ corresponds to an inter-beat sequence, comprising a hierarchical representation of global musical feature and local beat-level features. A *GRU decoder* generates a dance sequence from $\mathbf{h}_i$ for each beat event. The authenticity of the generated sequences is measured by an *Optimal Transport Distance*, and the music-to-dance matching is measured by a *Gromov-Wasserstein Distance*.

3 Our Approach

Music Input Preprocessing. The music waveform is sampled at 48kHz. We do not use the raw waveform as input as it would be too computationally expensive with too much redundancies. Following existing works [Tang *et al.*, 2018; Ren *et al.*, 2020], we adopt a similar procedure of extracting Mel-frequency cepstral coefficients (MFCC) and MFCC delta features which constitute low level sound features [Müller, 2015]. To incorporate additional high level musical information, we further extract chroma features which correspond to pitch and melody as well as beats which relate to rhythm.

Pose Preprocessing. We first perform an inverse kinematics fitting to obtain the joint orientation parameters for each dance pose in the Skinned Multi-Person Linear (SMPL) model [Loper *et al.*, 2015]. Key advantages of this over 3D positions representation include: 1) bone length invariance and rotational degrees of freedom are inherently inbuilt in this manifold; 2) We can easily normalize the bone lengths across performers and the global orientation of a dance sequence.

3.1 Formal Problem Statement

We denote the dance distribution as ν and the music distribution as ξ. As illustrated in Figure 2, the generator with parameters θ learns a parametric mapping g_θ that maps an input music sequence y (sampled from ξ) and noise vector z (sampled from a Gaussian distribution ψ) to a generated dance sequence $\tilde{x}$. This gives the model distribution of generated dances $\mu_\theta = g_{\theta_\#}(\xi, \psi)$. A crucial aspect of our MDOT-Net is in introducing an optimal transport distance $OT(\mu_\theta, \nu)$ and a Gromov-Wasserstein distance $GW(\mu_\theta, \xi)$ as objective functions to facilitate learning of the generator parameters θ

$$\arg\min_\theta \ OT(\mu_\theta, \nu) + GW(\mu_\theta, \xi). \tag{1}$$

In what follows, we present the definitions for $OT(\mu_\theta, \nu)$ and $GW(\mu_\theta, \xi)$ and algorithms for computing them.

3.2 Optimal Transport Distance

For the generated dance distribution μ_θ and the data dance distribution ν, we consider each as a discrete distribution with m samples. We have $\mu_\theta = \frac{1}{m} \sum_{i=1}^m \delta_{\tilde{x}_i}$, $\nu = \frac{1}{m} \sum_{j=1}^m \delta_{x_j}$ where δ denotes the Dirac delta distribution. The optimal transport distance between μ_θ and ν is obtained via the following optimal transport problem [Kantorovitch, 1958]:

$$OT_c(\mu_\theta, \nu) = \min_{\gamma \in \Gamma} \sum_{i,j} \gamma_{ij} c(\tilde{x}_i, x_j). \tag{2}$$

Intuitively, the optimal transport distance is the minimum total cost of matching μ_θ with ν with $c(\tilde{x}, x)$ denoting the unit cost of moving generated sequence $\tilde{x}$ to data sequence x. This is optimized over the set Γ of all possible transport plans:

$$\Gamma = \left\{ \gamma \in \mathbb{R}_+^{m \times m} \mid \forall i \sum_j \gamma_{ij} = 1, \forall j \sum_i \gamma_{ij} = 1 \right\}. \tag{3}$$

Cost function c. A T-frames dance sequence is given by $x = (\mathbf{p}_1, \cdots, \mathbf{p}_T) \in \mathcal{X}^T$ where each pose comprises of joint rotations, *i.e.* $\mathbf{p}_i = (R_{i,1}, \cdots, R_{i,J})$ [1]. We propose a cost function measuring the squared geodesic distance on the $SO(3)$ rotation manifold:

$$
\begin{aligned}
c_R(\tilde{x}, x) &= \sum_{i=1}^T \sum_{j=1}^J \text{geodesic}_{SO(3)}(\tilde{R}_{i,j}, R_{i,j})^2 \\
&= \sum_{i=1}^T \sum_{j=1}^J \left| \arccos\left[\frac{\text{Tr}(\tilde{R}_{i,j}^T R_{i,j}) - 1}{2} \right] \right|^2.
\end{aligned} \tag{4}
$$

Solving for Equation 2. For distributions supported on an Euclidean space $\mathbb{R}^d$ with L1 cost function $c(\tilde{x}, x) = |\tilde{x} - x|$, the optimal transport distance is known as the Wasserstein distance. The can be solved through 1-Lipschitz functions in

---

[1] We disregard global translation of the pose.

the dual formulation as proposed in Wasserstein GAN [Arjovsky *et al.*, 2017]. Our case is more complicated in that the distributions are over a manifold with a squared geodesic cost function. To solve for this, we follow [Genevay *et al.*, 2016] in introducing a regularization term

$$OT_{c,\epsilon}(\mu_\theta, \nu) = \min_{\gamma \in \Gamma} \sum_{i=1}^n \sum_{j=1}^n \gamma_{ij} c(\tilde{x}_i, x_j) + \epsilon I(\gamma) \quad (5)$$

where $I(\gamma) = \sum_{i,j} \gamma_{ij} \log_2 \gamma_{ij}$ is the mutual information of μ_θ, ν. This regularization transforms the primal Kantorovich problem into a convex optimization problem and Eqn 5 admits a solution of the form $\gamma* = \text{diag}(\mathbf{a}) K \text{diag}(\mathbf{b})$ where $K_{ij} \stackrel{\text{def}}{=} -\exp(c(\tilde{x}_i, x_j)/\epsilon)$. We compute the optimal transport distance with Algorithm 1. This algorithm has the crucial advantage of being amenable to backpropagation [Genevay *et al.*, 2018] and faster convergence.

Algorithm 1 OT Distance for batch of m samples with Sinkhorn-Knopp algorithm

---

Input: gen. dance sequences $\widetilde{\mathbf{X}} = \{\tilde{x}_i\}_{i=1}^m$
Input: data dance sequences $\mathbf{X} = \{x_j\}_{j=1}^m$
Hyperparameters: regularization ϵ, Sinkhorn iterations L
  Dance Cost Matrix $C_{ij} = c_R(\tilde{x}_i, x_j)$ from Eqn 4,
  $K = \exp(-C/\epsilon)$
  $\mathbf{b}^{(0)} = \mathbb{1}_m$ where $\mathbb{1}_m = (1, \cdots, 1)^T \in \mathbb{R}^m$
  for $\ell = 1 : L$ **do**
    $\mathbf{a}^{(\ell)} = \mathbb{1}_m \oslash K\mathbf{b}^{(\ell-1)}, \mathbf{b}^{(\ell)} = \mathbb{1}_m \oslash K^T\mathbf{a}^{(\ell)}$
    $\oslash$ denotes component-wise division
  end for
Output: $OT_\epsilon(\widetilde{\mathbf{X}}, \mathbf{X}) = \sum_{i,j} C_{ij} a_i^{(L)} K_{ij} b_j^{(L)}$

Following [Salimans *et al.*, 2018], we sample two independent mini-batches of data and generated sequence pairs $(\widetilde{\mathbf{X}}, \mathbf{X}), (\widetilde{\mathbf{X}}', \mathbf{X}')$ in order to compute the following unbiased optimal transport distance $\overline{OT}_\epsilon$

$$\overline{OT}_\epsilon = OT_\epsilon(\widetilde{\mathbf{X}}, \mathbf{X}) + OT_\epsilon(\widetilde{\mathbf{X}}', \mathbf{X}) + OT_\epsilon(\widetilde{\mathbf{X}}, \mathbf{X}')$$
$$+ OT_\epsilon(\widetilde{\mathbf{X}}', \mathbf{X}') - 2OT_\epsilon(\widetilde{\mathbf{X}}, \widetilde{\mathbf{X}}') - 2OT_\epsilon(\mathbf{X}, \mathbf{X}'). \quad (6)$$

3.3 Gromov-Wasserstein Distance

Comparing the similarity of dance distribution μ_θ and music distribution ξ invokes a cross domain learning problem. The Gromov-Wasserstein distance is defined as a *relational* distance between the respective costs within each distribution. The cost function for the dance pose manifold is defined in Eqn 4. For music distributions, we learn an embedding f and define the cost as the L1 distance in this embedding space

$$d(y_i, y_j) = \|f(y_i) - f(y_j)\|_1. \quad (7)$$

The Gromov-Wasserstein distance for our task is given by

$$\Pi = \left\{ \pi \in \mathbb{R}_+^{m \times m} \mid \forall i \sum_j \pi_{ij} = 1, \forall j \sum_i \pi_{ij} = 1 \right\} \quad (8)$$

$$GW(\mu_\theta, \xi) = \min_{\pi \in \Pi} \sum_{i,j,k,l} |c_R(\tilde{x}_i, \tilde{x}_k) - d(y_j, y_l)|^2 \pi_{ij} \pi_{kl}.$$

Solving for Eqn 8. This may be solved via entropic regularization and projected gradient descent as in Algorithm 2.

Algorithm 2 GW Distance for 2 independent batches of m samples

---

Input: gen. dance sequences $\widetilde{\mathbf{X}} = \{\tilde{x}\}_{i=1}^m, \widetilde{\mathbf{X}}' = \{\tilde{x}'\}_{i=1}^m$
Input: music sequences $\mathbf{Y} = \{y\}_{i=1}^m, \mathbf{Y}' = \{y'\}_{i=1}^m$
Hyperparameters: regularization ε, projection iterations M, Sinkhorn iterations L
Initialize: $\pi_{ij}^{(0)} = \frac{1}{n} \forall i, j$
  Dance Cost Matrix $C_{ij} = c_R(\tilde{x}_i, \tilde{x}'_j)$ from Eqn 4
  Music Cost Matrix $D_{ij} = d(y_i, y'_j)$ from Eqn 7
  for $l = 1 : M$ **do**
    $E = \frac{1}{m} D^2 \mathbb{1}_m \mathbb{1}_m^T + \frac{1}{m} \mathbb{1}_m \mathbb{1}_m^T C^2 - 2D\pi^{(l-1)} C^T$
    $K = \exp(-E/\varepsilon)$
    $\mathbf{b}^{(0)} = \mathbb{1}_m$
    for $\ell = 1 : L$ **do**
      $\mathbf{a}^{(\ell)} = \mathbb{1}_m \oslash K\mathbf{b}^{(\ell-1)}, \mathbf{b}^{(\ell)} = \mathbb{1}_m \oslash K^T\mathbf{a}^{(\ell)}$
    end for
    $\pi^{(l)} = \text{diag}(\mathbf{a}^{(L)}) K \text{diag}(\mathbf{b}^{(L)})$
  end for
Output: $GW_\varepsilon(\widetilde{\mathbf{X}}, \widetilde{\mathbf{X}}', \mathbf{Y}, \mathbf{Y}') = \sum_{i,j,k,l} |C_{ik} - D_{jl}|^2 \pi_{ij}^{(M)} \pi_{kl}^{(M)}$

3.4 Algorithmic Pipeline

An overview of our MDOT-Net is presented in Figure 2. We generate an inter-beat dance sequence directly instead of a frame-by-frame synthesis since this improves the temporal smoothness and also facilitates modeling nuances in dance motions. Our encoder network, as illustrated in Figure 3 processes the musical feature y and a noise vector $\mathbf{z}$ into latent vectors $\mathbf{h}_1, \cdots, \mathbf{h}_T$ where T denotes the total beat events in the music input y. Each $\mathbf{h}_i$ consists of a hierarchical representation of the global musical feature and local beat level features. Each $\mathbf{h}_i$ is decoded via a GRU network into a dance sequence $\tilde{x}_i$ consisting of 10 articulated poses. For post-processing, we employ spherical interpolation to fix the frames per second to 25. The training procedure is summarized in Algorithm 3.
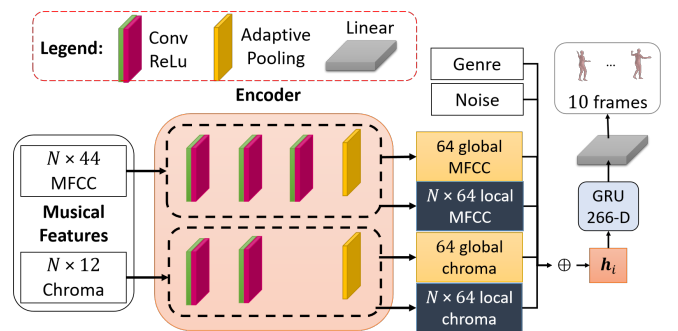

Figure 3: Architecture details of encoder and decoder

Algorithm 3 Overall Algorithmic Pipeline

Input: music dataset with data distribution ξ
Input: dance dataset with data distribution ν
Hyperparameters: regularization parameters ϵ, ε, batch size m, learning rate α, training epochs T, generator to critic update ratio n_{gen}, noise vector distribution ψ
Initialize: generator parameters θ_0, music embedding f_0.
  for $l = 1 : T$ **do**
      Sample 2 independent mini-batches of music-dance pairs from data and noise vectors from ψ, $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}), (\mathbf{X'}, \mathbf{Y'}, \mathbf{Z'})$
      Generate dance sequences as $\widetilde{\mathbf{X}} = g_\theta(\mathbf{Y}, \mathbf{Z})$ and $\widetilde{\mathbf{X}}' = g_\theta(\mathbf{Y'}, \mathbf{Z'})$.
      Compute $\overline{OT}_\epsilon$ with Algorithm 1 and Equation 6.
      Compute GW_ε with Algorithm 2.
      if $l \mod (n_{\text{gen}}) > 0$ **then**
         $\theta \leftarrow \theta - \alpha\nabla_\theta \overline{OT}_\epsilon - \alpha\nabla_\theta GW_\varepsilon$
      else
         $f \leftarrow f + \alpha\nabla_f GW_\varepsilon$
      end if
  end for
Output: Generator network parameters θ

---

4 Experimental Results

4.1 Dataset and Implementation Details

We adopt the public dataset of [Tang *et al.*, 2018] comprising four dance styles, waltz, tango, cha-cha, and rumba. The dances were performed by professional dancers and comprises 3D motion capture data obtained via Vicon MoCap devices. We perform an inverse kinematics fitting to reparameterize the dance sequence as SMPL [Loper *et al.*, 2015] parameters, consisting of 3D orientations for 24 joints.

We implement MDOT-Net in PyTorch. Additional libraries included Librosa [McFee *et al.*, 2015] for music processing and the Python Optimal Transport Toolbox [Flamary *et al.*, 2021]. The hyperparameters are as follows: regularization parameters are set to $\epsilon = 0.1$, $\varepsilon = 0.5$; mini-batch size is set to $m = 128$; learning rate is set to $\alpha = 0.001$; number of Sinkhorn iterations is set to $L = 30$; projection iterations is set to $M = 20$; generator to critic update ratio is set to $n_{\text{gen}} = 10$. The RMSprop optimizer is used. Convergence occurs around 150 epochs.

4.2 Baselines and Ablation Studies

Baselines. As a relatively novel task, few methods have been developed for generating 3D dance sequences from music. Existing methods [Ren *et al.*, 2020; Huang *et al.*, 2021] focus on 2D generation, making it difficult to visualize the full dynamic richness of dance motions. As such, we adapted them for 3D generation for fair comparison.

Ablation Studies. We further perform two ablation studies for validating the efficacy of the optimal transport distance and Gromov-Wasserstein distance. **1) WGAN**: Here we replace the optimal transport and Gromov-Wasserstein objectives with WGAN discriminators. The discriminator for dance sequences is adapted from AGCN [Shi *et al.*, 2019]

(SOTA for modeling skeletal based motion). A second discriminator serves to determine if the dance sequence matches input music. **2) Remove GW**: We remove the Gromov-Wasserstein objective to investigate its effectiveness in establishing music and dance correspondence.

4.3 Evaluation: Realism and Consistency

Sample generated dance sequences for tango are illustrated in Figure 4. We observe that [Tang *et al.*, 2018] tends to converge to mean pose and is lacking in dynamical variations. Comparatively, GAN methods including [Ren *et al.*, 2020] and our ablation experiment with WGAN demonstrates more variation in the range of dance motions, but often lacks in naturalness and appears unrealistic, especially in turning dance motions. Our MDOT-Net generates the most realistic looking dance motions which are also consistent with the tango style.

User Study. A single blind user study involving 8 dancers is engaged to judge the authenticity of the generated dance sequences and tabulated in Table 1. For each dance style, 5 sample generated sequences of 20 seconds duration are ranked according to two criteria, namely 1) the naturalness and realism of the motion and 2) adherence to the specific dance style. The rankings echo our above qualitative observations, with the samples generated by our proposed approach being preferred over previous methods. This demonstrates the effectiveness of the optimal transport objective in generating realistic and consistent dance sequences.

Method	Realism	Consistency	FID
[Tang *et al.*, 2018]	5.9	5.8	105.5 ±17.2
[Ren *et al.*, 2020]	4.6	4.8	67.0 ±9.3
[Huang *et al.*, 2021]	3.3	3.1	43.4 ±7.5
Ablation: WGAN	3.8	4.0	49.3 ±5.2
Ablation: Remove GW	2.3	2.1	32.4 ±4.8
MDOT-Net (Ours)	**1.2**	**1.3**	**25.6** ±3.3

Table 1: Results for Realism and Consistency

FID. We randomly sample data dance sequences ranging from 100 frames to 250 frames and employ a pre-trained AGCN [Shi *et al.*, 2019] for feature extraction to evaluate the Fréchet Inception Distance [Heusel *et al.*, 2017]. The better FID result suggests that the optimal transport objective is more effective than existing methods and the WGAN in matching the model distribution with data distribution. This could be explained by the optimal transport objective not being prone to instability issues of adversarial training.

Method	Diversity	Multimodality				
		Rumba	Cha Cha	Tango	Waltz	Average
Ground Truth	63.5	-	-	-	-	-
[Tang *et al.*, 2018]	18.2	13.2	16.7	14.8	10.7	13.9
[Ren *et al.*, 2020]	35.4	32.1	30.5	27.4	25.9	29.0
[Huang *et al.*, 2021]	34.7	28.9	28.5	31.9	34.4	30.9
Ablation: WGAN	41.5	38.9	46.5	33.5	35.4	38.6
Ablation: Remove GW	58.7	52.3	55.4	46.8	**50.7**	51.3
MDOT-Net (Ours)	**60.1**	**55.6**	**59.7**	**49.4**	48.8	**53.4**

Table 2: Results for Diversity (dance generated via different music) and Multimodality (dance generated on the same music)

Method	Click ↓	Tango dance sequences shown at 1 second intervals
Ground Truth		
[Tang *et al.*, 2018]		
[Ren *et al.*, 2020]		
[Huang *et al.*, 2021]		
Ablation: WGAN		
Ablation: Remove GW		
MDOT-Net (Ours)		

Figure 4: Visualization of sample generated tango dance sequences. Dance animations will be played in Adobe Acrobat Reader upon clicking.

Method	Beats Matching (%)				Music-Dance Matching Rank	
	Rumba	Cha Cha	Tango	Waltz	Dataset Music	New Music
Ground Truth	65.1	68.4	62.4	72.3	1.2	-
[Tang *et al.*, 2018]	13.7	11.2	14.2	16.9	6.9	5.9
[Ren *et al.*, 2020]	44.7	40.2	39.9	50.7	5.6	4.7
[Huang *et al.*, 2021]	52.3	46.5	52.6	58.2	3.6	1.9
Ablation: WGAN	49.8	49.4	55.7	58.8	3.0	3.8
Ablation: Remove GW	54.3	60.2	60.5	64.3	3.8	2.8
MDOT-Net (Ours)	**63.2**	**65.4**	**64.7**	**73.1**	**1.9**	**1.2**

Table 3: Results for beats matching and user preference

4.4 Evaluation: Diversity and Multimodality

We adopt the terminology of [Lee *et al.*, 2019]: *diversity* refers to variations over the entire ensemble of dances generated from different music inputs whereas *multimodality* pertains to dances generated from the same music. For quantitative evaluation, we employ a perceptual similarity metric [Zhang *et al.*, 2018]. For diversity, we generate 20 dance sequences conditioned on different music inputs. For multimodality, 5 dance sequences are conditioned on the same music. The pairwise feature distance is evaluated for each collection and experiments are averaged over 20 independent trials. The results reported in Table 2 demonstrate significant improved diversity and multimodality over existing methods and the ablation setting of WGAN, indicating that MDOT-Net is more effective in preventing mode collapse.

4.5 Evaluation: Cohesion and Unity with Music

Beats Matching. We evaluate the consistency of rhythmic articulation through beats matching. A dance beat is a local minimum in the mean joint speeds and it matches a music beat if the two events are within ± 2 frames. The beats matching ratio measures matching beats against total dance beats.

Music-Dance Harmony. Evaluating the music and dance coherence is a rather subjective task and we again engage an user study. 10 dance sequences (20 sec duration) are ranked by 8 users according to the perceived harmony with music. The consistent preference for MDOT-Net justify the effectiveness of the Gromov-Wasserstein distance.

5 Conclusion

In this work, we propose a MDOT-Net framework for generating 3D dance sequences conditioned on music. Through an optimal transport objective for matching the model and data dance distributions as well as a Gromov-Wasserstein objective for aligning the music and dance, MDOT-Net proves capable of generating realistic dances that are consistent with the dance style, display diversity and match the music. Extensive experiments demonstrate the effectiveness of the optimal transport and Gromov-Wasserstein objectives. For future work, we will generalize this cross-domain sequence-to-sequence generation framework to more applications.

Acknowledgements

References

[Alvarez-Melis and Jaakkola, 2018] David Alvarez-Melis and Tommi Jaakkola. Gromov-wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890, 2018.

[Arjovsky *et al.*, 2017] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 214–223, 2017.

[Artetxe *et al.*, 2018] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, 2018.

[Bunne *et al.*, 2019] Charlotte Bunne, David Alvarez-Melis, Andreas Krause, and Stefanie Jegelka. Learning generative models across incomparable spaces. In *International Conference on Machine Learning*, pages 851–861, 2019.

[Cuturi, 2013] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

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

[Flamary *et al.*, 2021] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

[Genevay *et al.*, 2016] Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Advances in neural information processing systems*, pages 3440–3448, 2016.

[Genevay *et al.*, 2018] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617, 2018.

[Heusel *et al.*, 2017] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pages 6626–6637, 2017.

[Huang *et al.*, 2021] Ruozi Huang, Huang Hu, Wei Wu, Kei Sawada, Mi Zhang, and Daxin Jiang. Dance revolution: Long-term dance generation with music via curriculum learning. In *International Conference on Learning Representations*, 2021.

[Kantorovitch, 1958] Leonid Kantorovitch. On the translocation of masses. *Management science*, 5(1):1–4, 1958.

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

[Lee *et al.*, 2019] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. Dancing to music. In *Advances in Neural Information Processing Systems*, pages 3586–3596, 2019.

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

[Loper *et al.*, 2015] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.

[McFee *et al.*, 2015] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, pages 18–25, 2015.

[Mémoli, 2011] Facundo Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487, 2011.

[Müller, 2015] Meinard Müller. *Fundamentals of music processing: Audio, analysis, algorithms, applications*. Springer, 2015.

[Ren *et al.*, 2020] Xuanchi Ren, Haoran Li, Zijian Huang, and Qifeng Chen. Self-supervised dance video synthesis conditioned on music. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 46–54, 2020.

[Salimans *et al.*, 2018] Tim Salimans, Han Zhang, Alec Radford, and Dimitris Metaxas. Improving gans using optimal transport. In *International Conference on Learning Representations*, 2018.

[Shi *et al.*, 2019] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12026–12035, 2019.

[Tang *et al.*, 2018] Taoran Tang, Jia Jia, and Hanyang Mao. Dance with melody: An lstm-autoencoder approach to music-oriented dance synthesis. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1598–1606, 2018.

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

[Villani, 2008] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[Wu *et al.*, 2021] Shuang Wu, Zhenguang Liu, Shijian Lu, and Li Cheng. Dual learning music composition and dance choreography. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3746–3754, 2021.

[Xu *et al.*, 2019] Hongteng Xu, Dixin Luo, Hongyuan Zha, and Lawrence Carin Duke. Gromov-wasserstein learning for graph matching and node embedding. In *International conference on machine learning*, pages 6932–6941. PMLR, 2019.

[Zhang *et al.*, 2018] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.