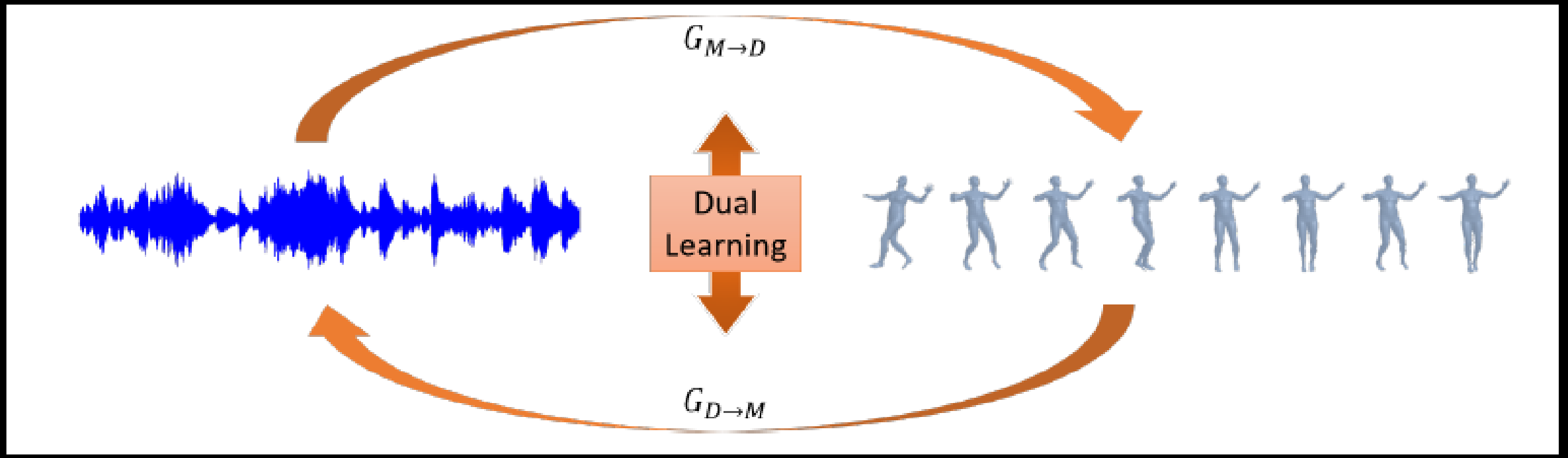
SEMINAR OF TECHNOLOGY & ARTS 20211214

授課教師:許素朱教授
報告學生: 翁政弘 iPhD109003815

DUAL LEARNING MUSIC COMPOSITION AND DANCE CHOREOGRAPHY

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). Association for Computing Machinery, New York, NY, USA, 3746–3754. DOI:https://doi.org/10.1145/3474085.3475180



ABSTRACT

- 人工智慧輔助，為特定舞蹈創作音樂的雙重任務在很大程度上被忽視了。
- 在本文中，出了一個新的擴展，在雙重學習方法中聯合對這兩個任務進行建模。
- 為了利用兩種模式的二重性，引入了一個最佳傳輸目標來對齊特徵嵌入，以及一個循環一致性損失來促進整體一致性。
- 實驗結果表明，我們的雙重學習框架提高了個人任務的表現，提供了真實且忠實於條件輸入的生成的音樂作品和舞蹈編排。

KEYWORDS

cross-modal generation,
dual learning,
optimal transport

CCS CONCEPTS

- Computing methodologies → Neural networks;
  Multi-task learning;

- Applied computing → Media arts.

* ACM Computing Classification System (CCS)

INTRODUCTION

- 從進化的角度來看，音樂和舞蹈起到了對人類社會功能的重要作用。
  他們在人類活動中無處不在，從個人娛樂到社交活動和禮儀活動。

- 在之上形成了人類文化& 現代文明不可或缺的一部分，並為人類社會做出巨大貢獻對個人和社會的意識感知等不一樣的影響。

- 在過去的幾個世紀裡，音樂和舞蹈逐漸系統化為兩種獨立的藝術形式但他們親密而深厚的聯繫是明確無誤的。

- 兩者都需要將我們的內部情感表達為外部運動。

- 為了舞蹈，表達的媒介是身體動作的視覺形式，而對於音樂，動作是通過聽覺表現出來的旋律和節奏

Dance

RELATED WORKS

早期的作品根據音樂源的相似性匹配從固定模板生成舞蹈序列。

自從合成的編舞只是重新安排了舞蹈動作從嚴格的訓練數據來看，有一個缺點不自然的過渡和缺乏創造力。

隨著深度學習方法的出現，序列到序列已經提出了用於生成舞蹈序列的模型編碼的音樂功能。一項開創性工作 使用 L2 損失比較有以下傾向的舞蹈序列動作凍結。為了緩解這種情況，Ruozi Huang 提出了學習加上 L1 損失，而 Zijie Ye採用測Baseline Loss。到能夠生成不同的舞蹈序列，介紹與編碼音樂特徵一起的隨機種子向量，而另一種方法 [Jan Kautz. 2019] 使用 GAN。一個關鍵的評論是 [Daxin Jiang. 2021] 專注於 2D 運動。我們的 3D 表示工作重要線索，例如相對位置或不變性在骨骼長度上要清楚地透視，從而出現更逼真、更有吸引力和幾何豐富。

另一研究方向[Hao Li. 2020] 採用跨模式架構用於生成以音樂和音樂為條件的舞蹈序列舞步。

在這個研究工作中，Music to dance 生成的主要任務僅取決於單個音樂和不需要任何額外的舞蹈。

計算音樂生成有兩種通用方法。第一個側重於符號表示 [5, 14, 16]。這 second 在音頻 [7, 9, 10] 中將音樂作為原始波形處

Music

RELATED WORKS

計算音樂生成有兩種通用方法。第一個側重於符號表示 [François-David Pachet. 2017]。另一種則在音頻 [Kristina Toutanova. 2018] 中將音樂作為原始波形處理或頻域 。理想情況下，波形表示實現更豐富的特徵波動感受，例如允許在音樂表演中產生人聲或細微差別在樂譜的頂部附加一層解釋。然而，這帶來了極高的計算成本。

把聲音特徵放在上下樂譜當中，以5 秒的波形表示用 48kHz 採樣的音樂序列將產生 240,000 的長度聲譜圖序列。儘管最近取得了一些進展，例如稀疏表示 [Ilya Sutskever. 2019] 或離散表示 [Karen Simonyan. 2018]，學習表現出多層次的音樂結構和層次在不同的尺度上，計算成本仍然非常高。在鑑於此，本研究求助於輕量級的符號表示用於音樂生成的部分。

MOTIVATION

- Music and dance are intimately connected.
  - Themes and emotions manifest auditorily as melodies and rhythms for music and visually as body movements for dance
- Recent work have explored AI generation of dance choreographs from music.
- This Paper propose to concurrently tackle the dual task, i.e. composing music for dance.
- Leverage dual learning to enhance the modeling of each task

CHALLENGES

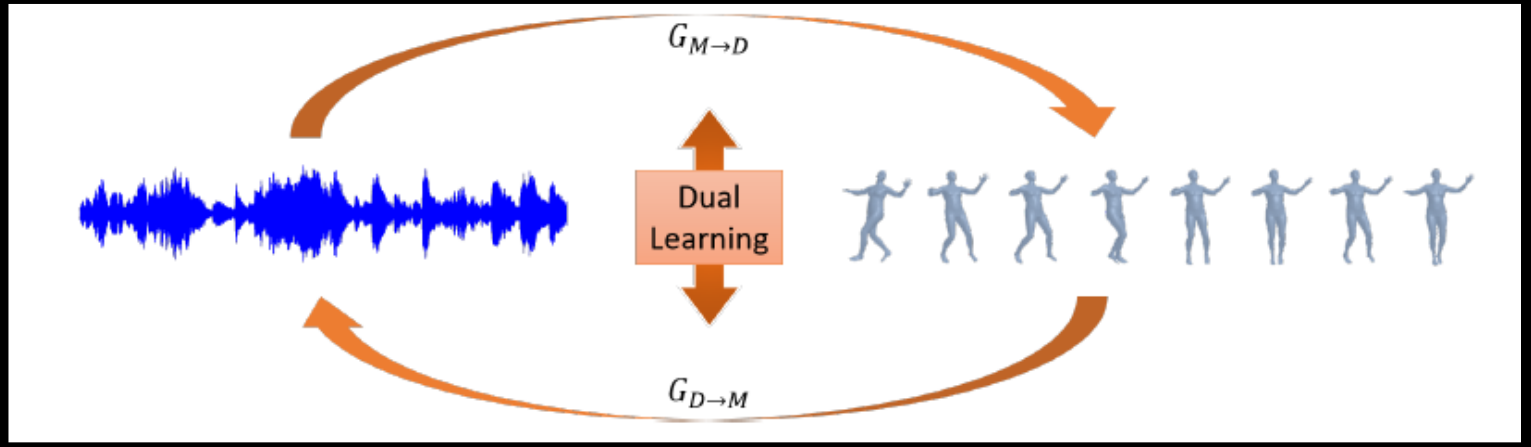
- Cross-domain generation
- Creativity and diversity in generated music and dance sequences
- Consistency between music and dance

KEY CONCEPT

- 本文提出了一個新問題跟如何解決：即同時學習舞蹈編排和音樂創作。

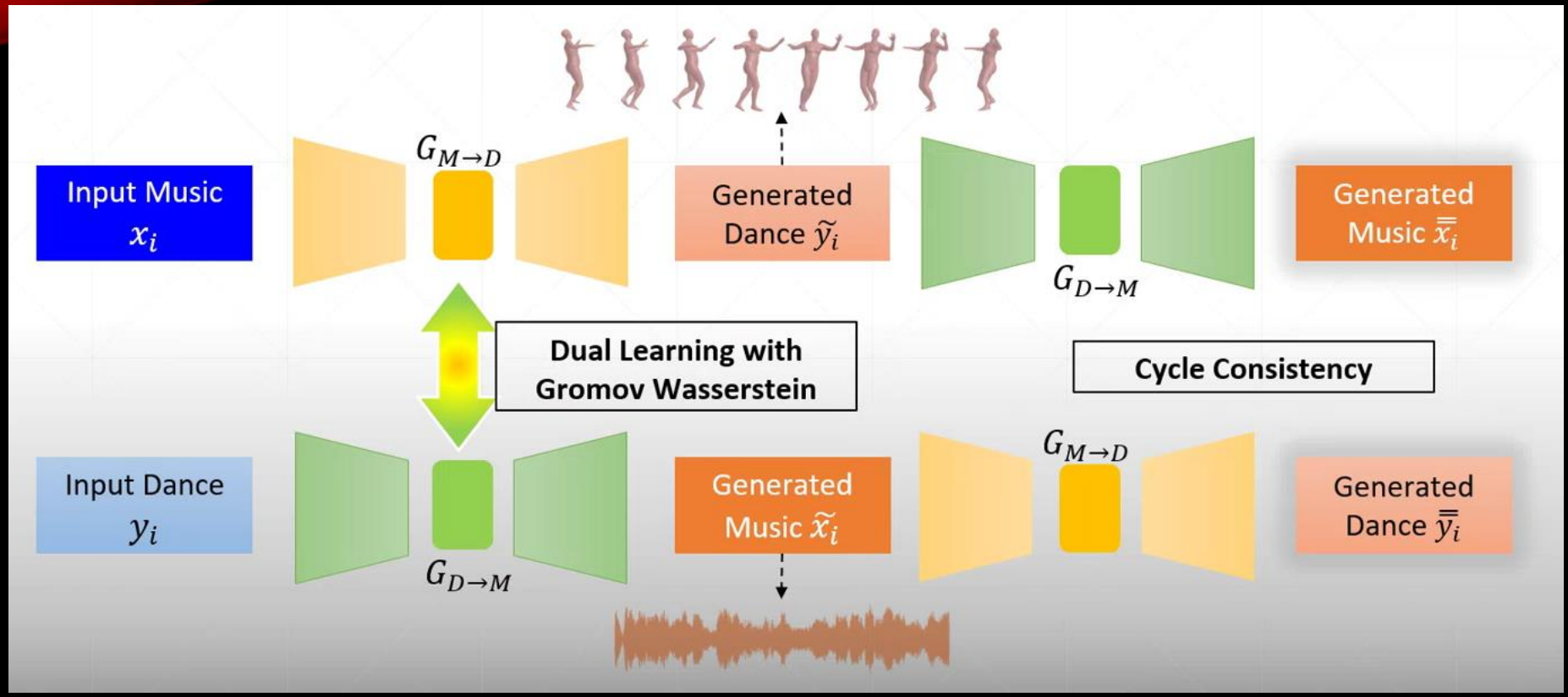
- 具體來說，本文的框架由兩個網絡組成:
  $GM \rightarrow D$ 用於從輸入音樂生成 3D 舞蹈編排，以及
  $GD \rightarrow M$ 用於合成給定舞蹈序列的音樂作品。
  作者利用這些任務的二元性來提取共同的底層

- 主題並確保生成的輸出和條件輸入之間的一致性。



APPROACH

A high-level overview of the pipeline.



The two generative networks:

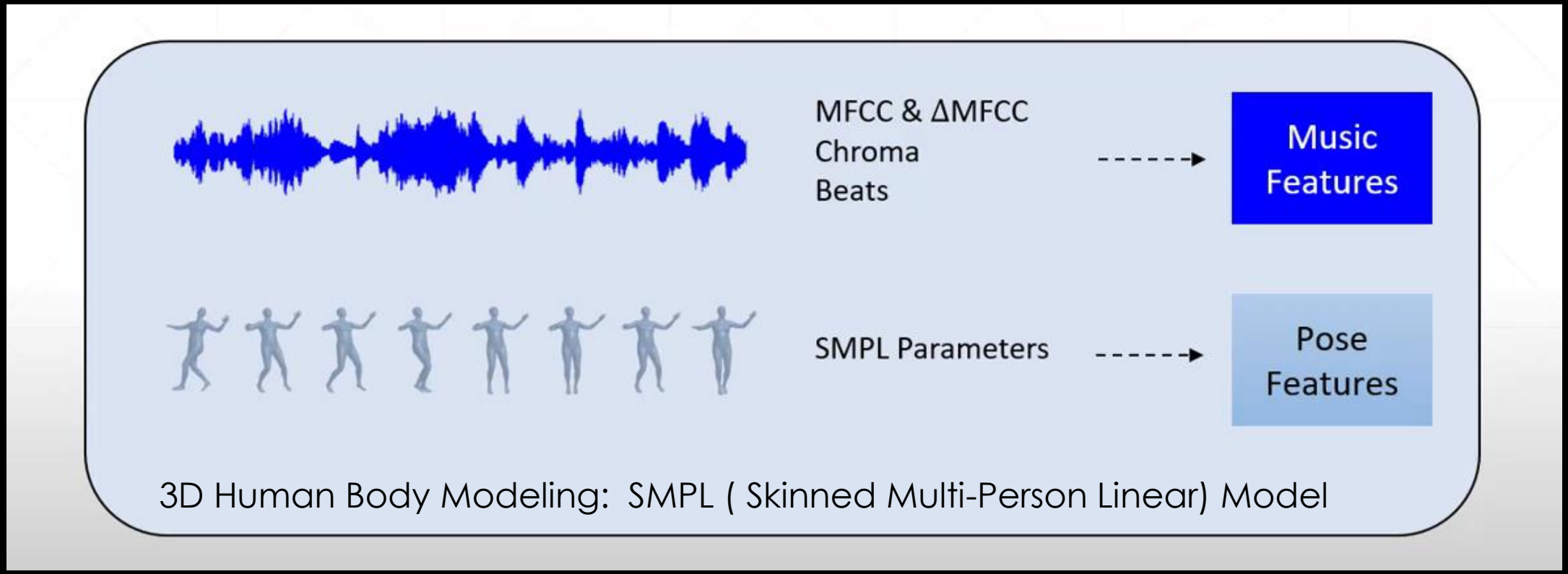
$GM\to D$ (music-to-dance choreography)

$GD\to M$ (dance-to-music composition)

in this framework comprise a sequence-to-sequence architecture where the encoder and decoder are both transformer networks

INPUT DATA/SIGNAL PREPROCESSING

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



MFC 常用於音樂片段訊號中 可以得到一組足以代表此音樂訊號之 倒頻譜(Cepstral)
MFCC: Mel-Frequency Cepstral Coefficients 梅爾頻率 倒譜係數

梅爾頻率倒譜(MFC)

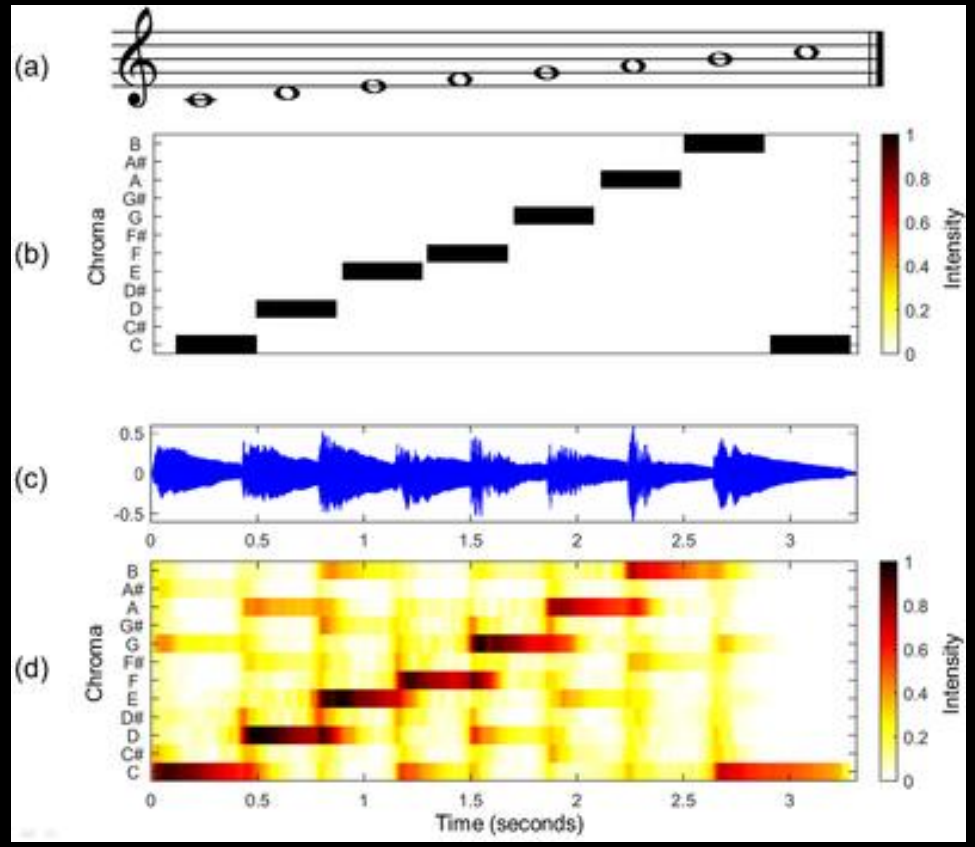
- **梅爾頻率倒譜系數 (Mel-Frequency Cepstral Coefficients，MFCCs)**
- 由Davis和Mermelstein在1980年代提出，廣泛被應用於語音識別的功能

- 在MFCC之前，線性預測係數（LPCS）和線性預測倒譜系數（LPCCs）是自動語音識別的的主流方法。

- MFCC通常有以下之過程:[1]
  - 將一段語音訊號分解為多個訊框。
  - 將語音訊號預強化，通過一個高通濾波器。
  - 進行傅立葉轉換，將訊號轉換至頻域。
  - 將每個訊框獲得的頻譜通過梅爾濾波器(三角重疊窗口)，得到梅爾刻度。
  - 在每個梅爾刻度上提取對數能量。
  - 對上面獲得的結果進行離散餘弦轉換，轉換到倒頻譜域。
  - MFCC就是這個倒頻譜圖的幅度(amplitudes)。一般使用12個係數，與訊框能量疊加得13維的係數。

CHROMA & BEATS

Chroma



(a) Musical score of a C-major scale.
(b) Chroma 色度 obtained from the score.
(c) Audio recording of the C-major scale played on a piano.
(d) Chromagram 色度圖譜 obtained from the audio recording.

Beats

BEATS 節奏分析與設計：

意涵與類別**Context and Genres**

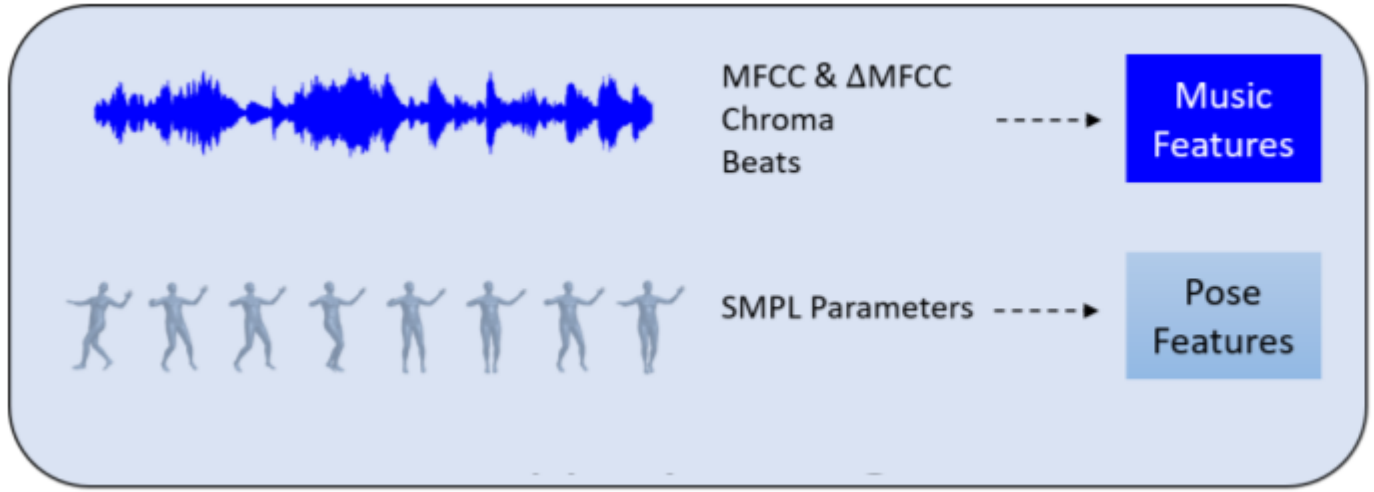
3 Beats(三拍子):華爾滋家族　Waltz Family

2 Beats(二拍子):進行曲家族　March Family

DATASET

Dataset Summary

Dance Genre	# of Sequences	# Frames	
Rumba	10	20950	From Tang et al. (2018). Dance sequences are longer at around 150 sec. Reparameterized in SMPL pose parameters.
Cha Cha	8	20425	
Tango	9	49165	
Waltz	34	43298	
Break Dance	141	46526	From Li et al. (2021). Dance sequences generally range from 8 to 12 sec.
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	



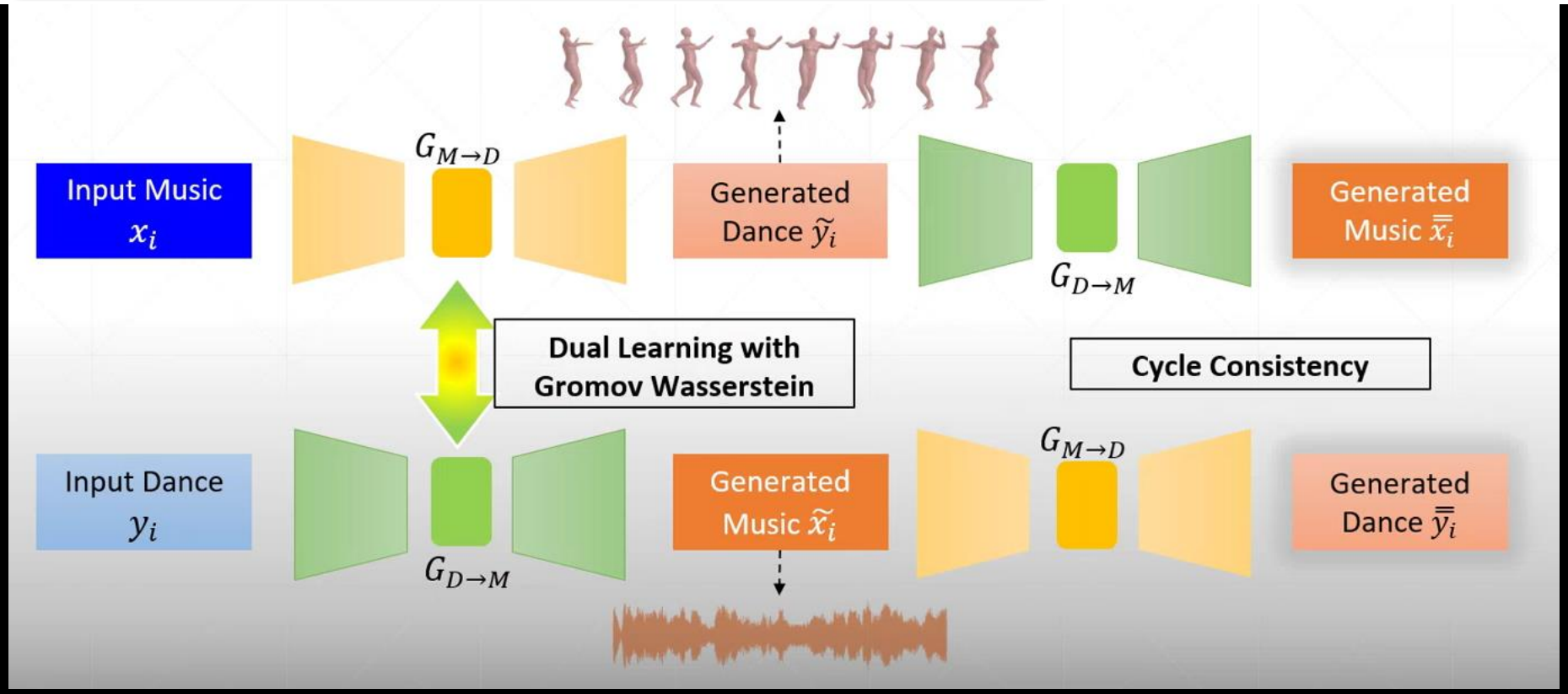
Gromov Wasserstein

Reconstruction Loss

$$\mathcal{L}_x(x_i, \tilde{x}_i) + \mathcal{L}_y(y_i, \tilde{y}_i)$$

Cycle Consistency Loss

$$\mathcal{L}_x(x_i, \bar{\bar{x}}_i) + \mathcal{L}_y(y_i, \bar{\bar{y}}_i)$$



RELATED FUNCTIONS

- **Initial Stage: Gromov Wasserstein Loss**

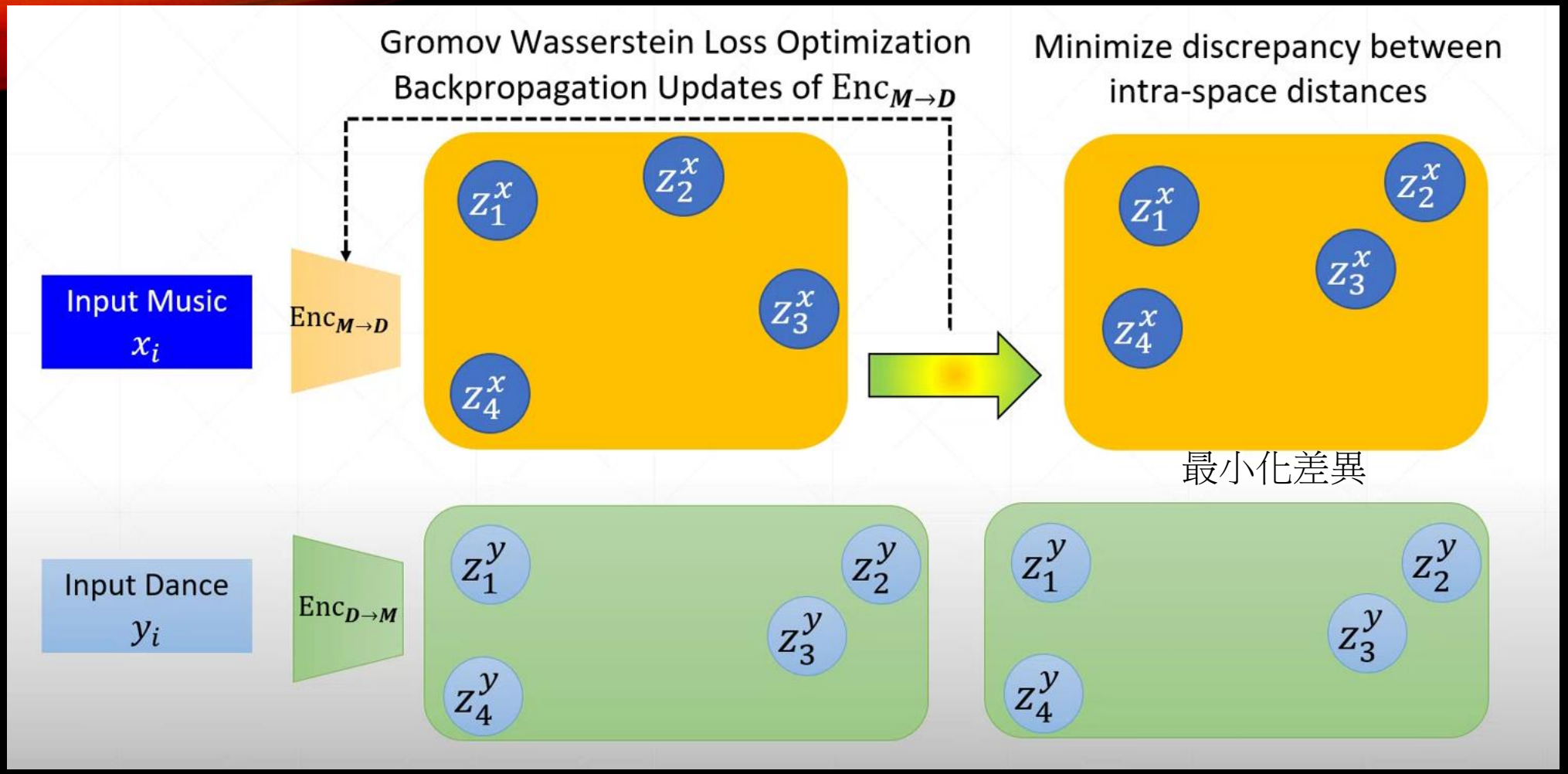
- **Reconstruction Loss**

$$\mathcal{L}_X(x_i, \tilde{x}_i) + \mathcal{L}_Y(y_i, \tilde{y}_i)$$

- **Posterior Stage: Cycle Consistency Loss**

$$\mathcal{L}_X(x_i, \bar{\bar{x}}_i) + \mathcal{L}_Y(y_i, \bar{\bar{y}}_i)$$

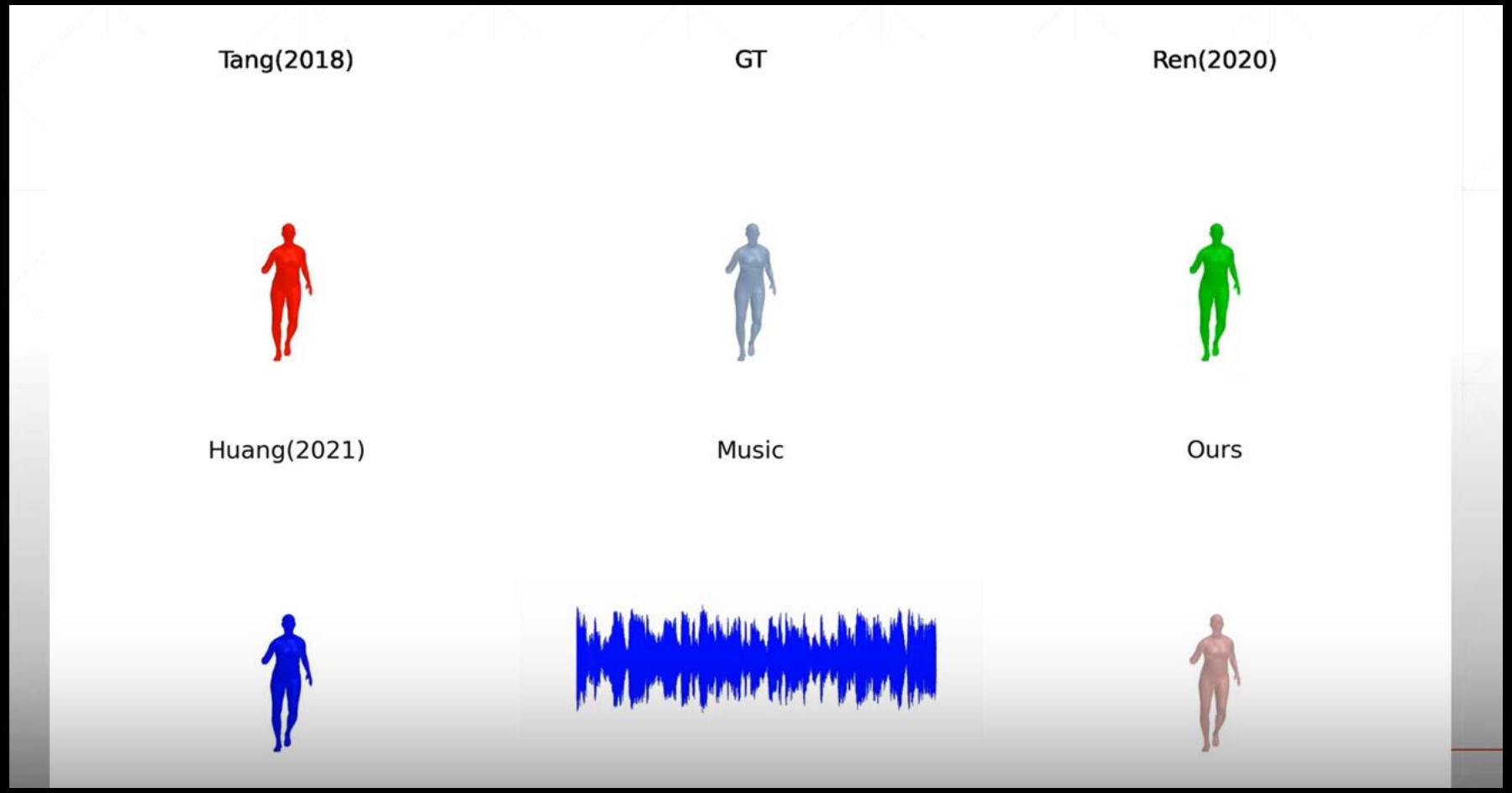
DUAL ENCOER/DECODER FRAMEWORK



In optimizing for the Gromov-Wasserstein loss,
the Enc$M\to D$ network parameters are updated,
re-positioning the music embedding vectors {z x i } 4 i=1
such that the discrepancy between intra-space distances is minimized

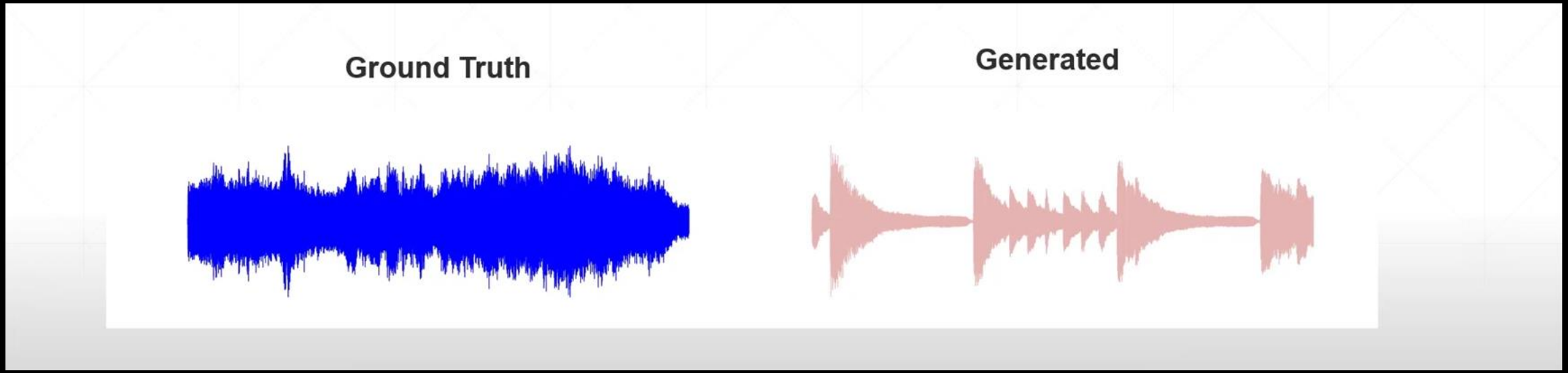
RESULT

- Comparsion of Dance Generated by Different Methods
- Same Music Input and Same Initial Pose for Cha Cha



SAMPLE GENERATED MUSIC

Ground True



RESULTS

- Quantitative Results
- Comparing Different Method for Dance Generation

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

NOTE:Fréchet 距離是曲線之間相似性的度量

CONCULSION

- Simultaneously learn music-to-dance and dance-to-music generation
- Cross domain sequence-to-sequence learning setting
- Gromov Wasserstein objective for aligning music and dance embeddings

WHAT I LEARN/GET?

- Art Model learning simulate human with Multiple-Sense
- Music vs Movement
- Music vs Emotion
- ….

CONNECTION

- Gromov-Wasserstein Alignment of Word Embedding Space, David Alvarez-Melis CSAIL, MIT dalvmel@mit.edu https://aclanthology.org/D18-1214.pdf

- The Gromov–Wasserstein distance between networks and stable network invariants, Titouan Vayer Univ. Bretagne-Sud, CNRS, IRISA F-56000 Vannes titouan.vayer@irisa.fr, https://proceedings.neurips.cc/paper/2019/file/a9cc6694dc40736d7a2ec018ea566113-Paper.pdf

- **Cross-Modal Dual Learning for Sentence-to-Video Generation** Yue Liu, MM '19: Proceedings of the 27th ACM International Conference on MultimediaOctober 2019 Pages 1239–1247, https://doi.org/10.1145/3343031.3350986