arXiv:2304.12704v1 [cs.SD] 25 Apr 2023

GTN-Bailando：基于预训练风格令牌网络的风格一致的长期三维舞蹈生成

Haolin Zhuang[1, *], *Shun Lei*[1], *Long Xiao*[1], *Weiqin Li*[1], *Liyang Chen*[1],
Sicheng Yang[1], *Zhiyong Wu*[1, 3, †], *Shiyin Kang*[2, †], *Helen Meng*[3]

[1] 清华大学深圳国际研究生院，中国深圳[2] XVerse 公司，中国深圳[3] 香港中文大学，中国香港特别行政区{zhuanghl21, leis21}@mails.tsinghua.edu.cn, zywu@sz.tsinghua.edu.cn, kangshiyin@xverse.cn,

摘要

近年来，音乐驱动的三维舞蹈生成已成为备受关注的研究课题，在现实应用中具有巨大潜力。大多数现有方法未考虑舞蹈风格，导致生成的舞蹈动作风格不一致。此外，舞蹈风格与音乐之间的关联尚未得到研究。为解决这些问题，我们提出了一种风格一致的舞蹈生成框架 GTN-Bailando。首先，我们提出流派令牌网络（GTN），从音乐中推断舞蹈风格，以增强长期舞蹈生成的流派一致性。其次，为提升模型的泛化能力，我们采用预训练与微调策略。在 AIST++ 数据集上的实验结果表明，所提出的舞蹈生成框架在动作质量和流派一致性方面均优于当前最先进的方法[1]。

索引词— 三维舞蹈生成、风格令牌、多模态、音乐驱动

1. 引言

舞蹈流派 [1]是舞蹈发展过程中形成的表现形式，每种流派都有其独特的特征。多样化的舞蹈流派提升了舞蹈的丰富性，使其成为全球最广为人知的艺术表达形式之一。随着人工智能（AI）的发展，人们可以生成舞蹈姿态序列，用于编舞辅助和虚拟偶像表演等具有发展前景的应用。然而，由于难以保持流派一致性，利用人工智能编舞生成令人满意的舞蹈姿态序列仍然面临挑战。在某一特定舞蹈流派中，并非所有物理上可行的舞蹈姿态都合适；编排出的舞蹈姿态不仅受到更严格的位置限制，还与给定音乐存在关联。

现有的大多数人工智能编舞都侧重于舞蹈与音乐旋律之间的对齐。早期方法 [2–11]将音乐和舞蹈直接输入单一网络，以自回归方式生成舞蹈序列。然而，这类方法会在自回归生成过程中累积误差，并且很可能退化为生成非舞蹈动作。与此同时，一些研究 [12–17] 采用检索方法，将舞蹈划分为固定长度的单元，并根据音乐的旋律对这些单元进行拼接编舞。尽管这些方法能够确保生成舞蹈的质量，但无法兼容不同的拍号和每分钟节拍数（BPM）。最近，一种出色的舞蹈生成框架 Bailando [18], 结合向量量化变分自编码器（VQ-VAE） [19] 与演员-评论家 [20]生成式预训练变换器（GPT） [21] ，解决了自回归舞蹈生成和检索式舞蹈生成方法中的问题。

然而，上述所有研究在舞蹈生成过程中都忽略了流派。缺乏流派信息时，生成的舞蹈会在同一音乐片段中呈现多种流派的舞蹈动作，导致表演与音乐不相称（例如在嘻哈音乐中加入芭蕾舞动作）。因此，近期有一些舞蹈生成框架开始致力于处理流派问题。GCDG [22]使用独热向量表示流派，并将流派嵌入变换器解码器中，以生成具有流派一致性的舞蹈。MNET [23] 使用映射网络将潜在编码转换为适用于多流派的风格编码，同时生成舞蹈动作。这些方法能够生成具有特定流派的舞蹈，但在推理时需要辅助输入，例如人工确定的流派标签。此外，舞蹈流派与其背景音乐之间存在特定的关联，使编舞者能够根据音乐确定舞蹈的流派。因此，理想情况下，人工智能编舞应根据音乐推断流派，并使用推断出的流派生成舞蹈。

本文提出了一种基于流派令牌网络（GTN）的流派一致长期舞蹈生成框架。首先，我们介绍 GTN，通过学习音乐与流派之间的相关性来推断流派。对于给定的一段音乐，可以通过 GTN 推断其流派，并将其用作舞蹈生成的条件，确保生成的每个舞蹈姿态都符合某一流派的位置限制。其次，我们提出了一种 GTN 预训练策略，以提升其泛化能力。由于现有舞蹈-音乐对齐数据集中的音乐数据不足，GTN 很难从音乐中准确推断流派。因此，为了加强各流派与其对应音乐之间的关联，我们收集了大规模带有流派标签的舞蹈背景音乐，用于预训练 GTN。之后，我们使用舞蹈-音乐对齐数据集训练舞蹈生成框架，并加载预训练 GTN 的权重进行微调。这样，GTN 就能更有效地推断音乐的流派，从而进一步提升流派一致性。实验结果表明，我们提出的舞蹈生成框架和预训练策略在评估指标和可视化评判方面均显著优于现有方法。

---

* 第一作者在 XVerse 公司实习期间完成的工作。
† 通讯作者。[1]生成的演示：
https://im1eon.github.io/ICASSP23-GTNB-DG/

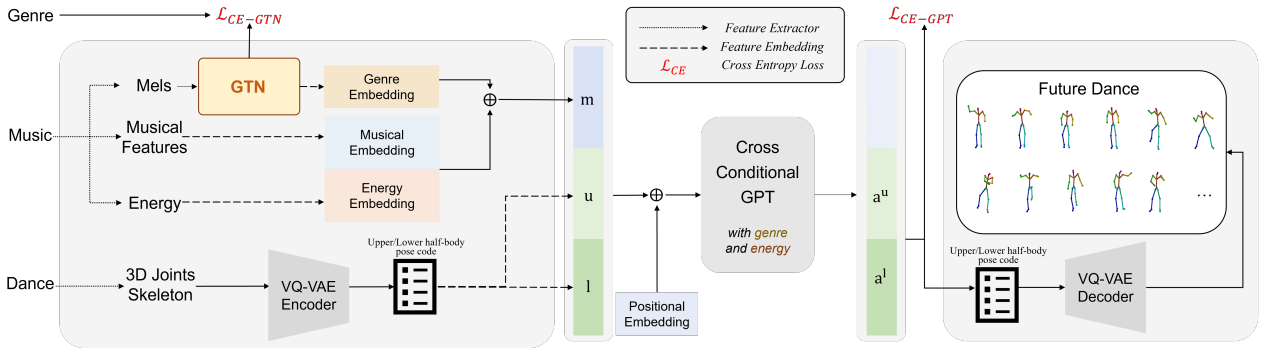

图 1。GTN-Bailando。音乐的梅尔频谱图被输入流派令牌网络（GTN）。生成的流派嵌入被用作 GPT 的条件，以确保每个生成的舞蹈都与其流派保持一致。通过交叉熵损失学习参数：$\mathcal{L}_{CE-GTN}$ 使用流派标签，$\mathcal{L}_{CE-GPT}$ 使用真实舞蹈姿态编码。

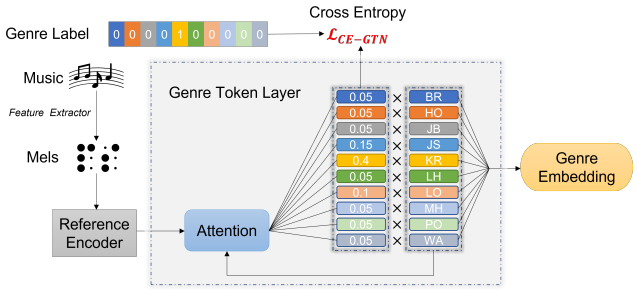

图 2。所提出的流派令牌网络（GTN）的架构。

2. 所提出的方法

问题定义。我们的目标是从音乐中推断流派，并根据推断出的流派生成舞蹈。形式化地说，给定一段音乐 $X = \{x_t\}_{t=1}^{N}$，其中 N 是音乐长度，以及一个初始舞蹈姿态$y_{t=1}$，我们的目标是：首先，从 X推断流派$\hat{g} \in G$，其中 G 是预定义流派类别的集合；其次，将$\hat{g}$ 用作条件，并 X 来生成舞蹈 $\hat{Y} = \{\hat{y}_t\}_{t=2}^{N}$。

我们的目标是生成具有流派一致性的舞蹈。在本节中，我们首先介绍 GTN，并展示如何从音乐中推断流派。其次，我们介绍流派一致性的舞蹈生成框架。之后，我们将解释 GTN 的预训练和微调策略。所提出的 GTN-Bailando 的架构如图 1 所示。

2.1. 流派令牌网络

受基于情感风格令牌的语音合成工作的启发 [24,25],，我们提出对流派令牌进行监督训练，以实现流派与音乐之间的关联。如图 2 所示，所提出的 GTN 架构由三个模块组成：参考编码器、流派令牌层和流派嵌入。

参考编码器采用自 [24],，用于将音频信号压缩为固定长度的向量。在本研究中，将音乐片段的梅尔频谱图输入参考编码器，并压缩为可学习的参考嵌入。

流派令牌层包含一组流派令牌嵌入和一个注意力模块，该模块使用参考嵌入作为查询向量。注意力模块学习参考嵌入与相似度一组随机初始化嵌入中的每个令牌之间的度量。这组嵌入也称为流派令牌，在所有音乐片段之间共享。

流派令牌层的输出是输入音乐属于各个流派类别的概率。为提高 GTN 的鲁棒性，我们采用软嵌入方法 [26] 来表示流派。这样，便可根据概率对令牌进行加权求和，形成流派嵌入。

为增强音乐与流派之间的相关性，令牌数量设为与流派类别数量一致。同时，将流派标签转换为独热嵌入，并输入流派令牌层，作为令牌权重的目标。因此，GTN 通过流派标签与流派令牌权重之间的交叉熵损失进行监督训练，优化方式如下：

$$\mathcal{L}_{CE-GTN} = \frac{1}{T}\sum_{t=0}^{T-1} \mathrm{CE}\left(g_t, \hat{g}_t\right), \tag{1}$$

其中，g_t 和$\hat{g}_t$ 分别表示第 t 个音乐片段的流派向量和流派 token 权重向量。CE（·）表示交叉熵损失函数，T 表示音乐片段的数量。

2.2. 舞蹈生成框架

受 Bailando 启发，[18], 我们基于 VQ-VAE 和交叉条件 GPT 生成舞蹈。由于舞蹈动作的速度与音乐能量之间存在相关性，我们进一步考虑能量特征，以提升生成舞蹈的动作质量。

具体而言，给定一段输入音乐，我们首先提取能量和音乐特征，然后分别将它们嵌入可学习向量$\boldsymbol{Z}_e$ 和 $\boldsymbol{Z}_m$中。同时，我们提取输入音乐片段的梅尔频谱图，再将其输入 GTN，以生成流派嵌入 $\boldsymbol{Z}_g$。然后，我们沿时间维度拼接 $\boldsymbol{Z}_m$ 和 $\boldsymbol{Z}_e$，并加上 $\boldsymbol{Z}_g$，形成$\mathbf{m}$。对于舞蹈，我们首先将舞蹈片段的骨架关节位置输入 VQ-VAE 编码器，生成上半身和下半身姿态编码，然后分别将它们嵌入可学习向量$\mathbf{u}$ 和$\mathbf{l}$中。

得到$\mathbf{m}$、$\mathbf{u}$、$\mathbf{l}$后，我们在时间维度上将它们拼接起来，并添加一个可学习的位置嵌入，然后将其输入交叉条件 GPT。最后，我们得到 GPT 的输出，即上半身和下半身姿态编码的概率。我们根据这些概率得到预测的上半身和下半身姿态编码对，并将其

输入 VQ-VAE 解码器，以生成未来舞蹈。

请注意，在训练期间，我们对 GTN 采用教师强制方案，以提高舞蹈生成框架整体的流派一致性。通过监督训练，利用预测的动作概率 $\mathbf{a}$ 与真实姿态编码 p 之间的交叉熵损失，对交叉条件 GPT 进行优化，如下所示：

$$\mathcal{L}_{CE-GPT} = \frac{1}{T'} \sum_{t=0}^{T'-1} \sum_{h=u,l} \mathrm{CE}\left(\mathbf{a}_t^h, p_{t+1}^h\right). \quad (2)$$

最后，舞蹈生成框架 $\mathcal{L}_{DG}$ 的损失可计算为：

$$\mathcal{L}_{DG} = \alpha \cdot \mathcal{L}_{CE-GTN} + \beta \cdot \mathcal{L}_{CE-GPT}. \quad (3)$$

2.3. 预训练与微调

尽管 GTN 能够推断流派，但其泛化能力仍然较差，难以建立流派与音乐之间的关联，也无法确保生成舞蹈的流派一致性。这是因为舞蹈-音乐对齐数据集中的音乐训练数据规模不足（十种流派的音乐总时长不到 1 小时）。

因此，我们提出了一种预训练方法，以增强 GTN 的泛化能力，并提高生成舞蹈的流派一致性。首先，我们从互联网收集大量舞蹈背景音乐数据，标注相应的舞蹈类型标签，然后使用这些数据训练 GTN。之后，在训练舞蹈生成框架的交叉条件 GPT 时，我们加载预训练 GTN 的参数。请注意，为防止 GTN 在微调过程中出现过拟合，当达到阈值轮次时，GTN 将被冻结。

3. 实验

3.1. 数据集

3.1.1. 用于预训练 *GTN* 的数据集

在预训练 GTN 时，我们从互联网上收集与不同流派舞蹈相对应的背景音乐。每种流派收集 1 小时的音乐数据，构成一个涵盖十种流派、总时长为 10 小时的音乐数据集。这些流派与舞蹈-音乐对齐数据集 AIST++ [9] 中的十种流派相同，分别为霹雳舞（BR）、芭蕾爵士（JB）、浩室舞（HO）、街头爵士（JS）、克朗普舞（KR）、洛杉矶风格嘻哈（LH）、锁舞（LO）、中部风格嘻哈（MH）、流行舞（PO）和甩手舞（WA）。

3.1.2. 舞蹈生成数据集

在使用微调后的 GTN 训练和评估舞蹈生成框架时，我们使用 AIST++ [9] 数据集。据我们所知，这是目前公开可用的、规模最大的带有流派标签的 3D舞蹈-音乐对齐数据集。AIST++ 数据集包含 992 段高性能的 60 帧每秒 3D 姿态序列，采用 SMPL格式，分别划分为 952 段训练集和 40 段验证集。

3.2. 实验设置

对于 GTN，我们使用 Librosa 从音乐中提取 80 维梅尔频谱图。对于舞蹈生成框架，使用 Librosa 提取音乐特征，包括梅尔频率倒谱系数（MFCC）、MFCC差分、恒定 Q 色度图、节拍图和起始强度，最终得到 438 维音乐特征。同时，我们参考 Fastspeech2 [27], 从音乐中提取能量，得到 1 维能量特征。

舞蹈生成框架和 GTN 的设置遵循以往研究，对于公式（3），我们设定 $\alpha = 1, \beta = 0.001$。GTN 中的流派令牌层包含 10 个令牌。框架中的 VQ-VAE 采用 Bailando [18] 中的编舞记忆。所有舞蹈和音乐数据均裁剪为 4 秒。我们首先对 GTN 进行 250 个周期的预训练，并对 VQ-VAE 进行 500 个周期的预训练，然后加载预训练的 VQ-VAE，生成与舞蹈动作对应的姿态编码。之后，我们加载预训练的 GTN，并对舞蹈生成框架训练 400 个周期；训练 90 个周期后，冻结 GTN。

我们将所提出的方法与 FACT [9] 和 Bailando [18], 进行比较，这两种方法均属于最先进的舞蹈生成方法。对于每种方法，我们以十首不同的 AIST++ 测试音乐以及两种不同的起始姿态编码或序列为条件生成舞蹈，共得到 20 个舞蹈片段。我们将生成的舞蹈剪辑为 20 秒，以用于后续实验。

3.3. 实验结果与分析

3.3.1. 流派嵌入可视化

在图3中，我们使用t-SNE方法对不同流派的流派嵌入进行可视化。我们使用AIST++ 测试集验证舞蹈生成框架中的GTN，结果表明，不同的流派嵌入彼此之间分离良好，证明GTN能够从音乐中正确推断流派。

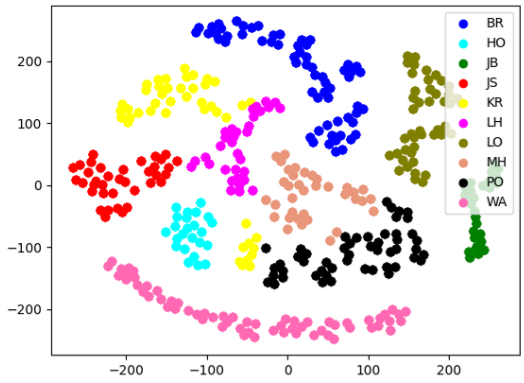

图3。流派嵌入的t-SNE可视化。不同颜色表示不同的舞蹈流派。

3.3.2. 主观评估

由于缺乏用于评估流派一致性的客观量化指标，我们参考情感语音合成 [25, 26], 中的方法进行主观评估，以进一步评估所提出的舞蹈生成框架生成的舞蹈的视觉表现。测试由24名受试者参与完成，他们均参加过舞蹈训练，并且对所包含的舞蹈流派有一定了解。受试者需要评估舞蹈质量和流派一致性，并以1至5分的量表对舞蹈进行评分（最低分为1分，最高分为5分），分值间隔为1分。

表 1 的最后两列报告了舞蹈质量和流派一致性的平均主观评分（MOS）。所提出的方法优于所有基线模型，表明 GTN 能够建立音乐与流派之间的关联。此外，以推断出的流派作为条件，所提出的舞蹈生成框架能够生成质量更高且流派一致的舞蹈。

表 1。不同舞蹈生成框架的评估结果。舞蹈质量和流派一致性为 MOS 结果，置信区间为 95%。“*”表示所提出的方法。“w/o”是消融研究中“without”（不含）的缩写。

	动作质量		动作多样性			用户研究	
	$FID_k \downarrow$	$FID_g \downarrow$	$DIV_k \uparrow$	$DIV_g \uparrow$	$BAS \uparrow$	质量↑	一致性↑
真实值	17.10	10.60	8.19	7.45	0.2484	-	-
FACT	37.31	34.87	5.75	5.47	0.2175	2.15±0.10	2.19±0.11
Bailando	32.11	9.95	6.09	5.70	0.2299	3.19±0.11	3.11±0.11
所提出的方法*	**29.51**	**8.57**	**6.15**	**6.65**	**0.2352**	**3.65±0.10**	**3.60±0.10**
不含 $\mathcal{L}_{CE-GTN}$	34.35	10.92	5.51	5.95	0.2292	3.35±0.13	3.25±0.13
无教师强制	41.77	14.13	5.96	3.54	0.2298	3.49±0.13	3.29±0.14
无预训练 GTN	30.55	7.03	7.03	6.60	0.2466	3.10±0.14	3.06±0.15
无能量	32.66	6.23	6.23	6.94	0.2366	3.40±0.15	3.33±0.15

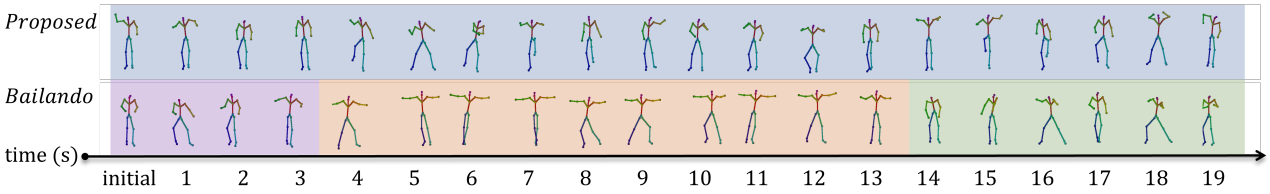

图 4。流派一致性可视化。不同颜色表示不同流派的舞蹈 的一致性可视化。不同颜色表示不同流派的舞蹈

3.3.3. 客观评估

对于客观评估，依据 [18], ，我们评估生成的舞蹈动作的质量与多样性，以及生成动作与音乐节拍之间的对齐程度。具体而言，对于动作质量，我们通过 Fr´echet Inception Distance（FID） [28]，分别在运动学特征（记为“FID_k”）和几何特征（记为“FID_g”）上计算 AIST++ 中所有动作序列与生成的舞蹈动作之间的分布距离。 FID越低，生成的舞蹈就越接近真实值。对于动作多样性，我们分别在运动学特征（记为“DIV_k”）和几何特征（记为“DIV_g”）上计算平均欧几里得距离。 DIV越高，距离越大，意味着生成的舞蹈包含更多样的动作。对于对齐程度，我们计算音乐节拍与动作节拍之间的节拍对齐分数（BAS）。 BAS越高，舞蹈越符合节拍。不同方法的结果如表 1 所示。

可以看出，我们所提出的框架在各个方面均优于基线框架。具体而言，我们所提出的舞蹈生成框架取得了最低的 FID_k 和 FID_g，以及最高的 DIV_k、 DIV_g 和 BAS。所提出的方法通过考虑流派和能量，生成了质量更高、更多样化的舞蹈，同时改善了舞蹈动作与音乐旋律之间的匹配度。

3.3.4. 消融研究

此外，我们进行了消融研究，以探究模型生成逼真舞蹈的能力。结果如表 1 所示，消融研究的可视化对比也可在演示网页中查看。

如表 1 所示，当优化框架时不使用 $\mathcal{L}_{CE-GTN}$，且训练过程中不采用教师强制方案时，生成的舞蹈与真实值中的舞蹈差异会很大，且多样性较低。如果不考虑舞蹈动作的速度与音乐能量之间的相关性，舞蹈的表现力就会降低。如果不采用预训练和微调策略，尽管该框架能够生成与真实值相似的舞蹈，但由于 GTN 的泛化能力较差，其流派一致性仍会受到限制。

3.3.5. 风格一致性可视化

为进一步评估生成舞蹈的流派一致性，我们对所提出的框架和 Bailando [18]生成的舞蹈关节骨架进行了定性结果可视化。我们从各框架生成的“LO”（Locking）流派舞蹈中随机选取一个 20 秒片段，并以 1 帧每秒的频率采样结果。

如图 4 所示，对于给定的音乐片段，Bailando 生成的舞蹈包含多种舞蹈流派，与音乐并不协调。相比之下，我们的框架能够推断流派，并生成与音乐旋律相匹配且符合“LO”流派的舞蹈。更多流派一致性可视化结果请参见演示页面。

4. 结论

本文针对流派一致性这一挑战，提出了一种舞蹈生成框架 GTN-Bailando。该框架能够建立音乐与流派之间的关联，并基于能量、音乐以及通过 GTN 推断出的流派，生成高质量且流派一致的舞蹈。同时，所提出的在大规模数据集上对 GTN 进行预训练的策略，有效提升了其泛化能力。在 AIST++ 上的实验结果表明，我们的方法能够根据音乐生成高质量且流派一致的舞蹈。

致谢： 本工作得到国家自然科学基金委员会（62076144）、深圳市下一代互动媒体创新技术重点实验室（ZDSYS 20210623092001004）和深圳市科技计划（ WDZC20220816140515001、JCYJ20220818101014030）的资助。

5. 参考文献

[1] Shuhei Tsuchida、Satoru Fukayama、Masahiro Hamasaki 和 Masataka Goto，“Aist舞蹈视频数据库：用于舞蹈信息处理的多流派、多舞者和多摄像机数据库”，载于 第*20*届国际音乐信息检索学会会议论文集，*ISMIR 2019*，荷兰代尔夫特，2019年11月，第501–510页。

[2]Omid Alemi、Jules Franc¸oise 和 Philippe Pasquier，“GrooveNet：使用人工神经网络进行实时音乐驱动的舞蹈动作生成”，网络，第 8 卷，第 17 期，第 26 页，2017 年。

[3] Taoran Tang、Jia Jia 和 Hanyang Mao，“Dance with Melody：一种面向音乐的舞蹈合成 LSTM自编码器方法”，*ACM Multimedia*，2018 年。[4] Nelson Yalta、Shinji Watanabe、Kazuhiro Nakadai 和 Tetsuya Ogata，“用于基本舞步生成的弱监督深度循环神经网络”，国际神经网络联合会议，2019 年。[5] Hsuan-Kai Kao 和 Li Su，“时间引导的音乐到身体动作生成”，*ACM Multimedia*，2020 年。

[6] Hyemin Ahn、Jaehun Kim、Kihyun Kim 和 Songhwai Oh，“基于音乐的 3D 舞蹈动作合成生成式自回归网络”，*IEEE*机器人与自动化快报，第 5 卷，第 2 期，第 3501–3508 页，2020 年。

[7] Jiaman Li、Yihang Yin、Hang Chu、Yi Zhou、Tingwu Wang、Sanja Fidler 和 Hao Li，《利用 Transformer 学习生成多样化舞蹈动作》，*arXiv* 预印本 *arXiv:2008 08171*，2020。。[8] Ruozi Huang、Huang Hu、Wei Wu、Kei Sawada、Mi Zhang 和 Daxin Jiang，《舞蹈革命：通过课程学习结合音乐进行长期舞蹈生成》，学习，2020。[9] Ruilong Li、Shan Yang、David A. Ross 和 Angjoo Kanazawa，《AI 编舞师：基于 AIST++ 的音乐条件三维舞蹈生成》，载于*IEEE* 国际计算机视觉会议（*ICCV*），2021。[10] Guillermo Valle-P´érez、Gustav Eje Henter、Jonas Beskow、An-dre Holzapfel、Pierre-Yves Oudeyer 和 Simon Alexander-son，《Transflower：利用多模态注意力进行概率自回归舞蹈生成》，*ACM Transactions on Graphics*（*TOG*），第40卷，第6期，第1–14页，2021。[11] Mingao Zhang、Changhong Liu、Yong Chen、Zhenchun Lei 和 Mingwen Wang，《使用多个 Conformer 进行音乐到舞蹈生成》，载于*2022*年国际多媒体检索会议论文集，2022年，第34–38页。[12] Rukun Fan、Songhua Xu 和 Weidong Geng，《基于示例的自动音乐驱动传统舞蹈动作合成》，*IEEE*可视化与计算机图形学汇刊，第18卷，第3期，第501–515页，2011。[13] Minho Lee、Kyogu Lee 和 Jaeheung Park，《基于音乐相似度生成舞蹈动作序列的方法》，多媒体工具与应用，第62卷，912，2013 3，895期，第–页。[14] Hsin-Ying Lee、Xiaodong Yang、Ming-Yu Liu、Ting-Chun Wang、Yu-Ding Lu、Ming-Hsuan Yang 和 Jan Kautz，《随乐起舞》，神经信息处理系统进展，第32卷，2019。

[15] Kang Chen、Tan Zhipeng、Lei Jin、Song-Hai Zhang、Yuan-Chen Guo、Weidong Zhang 和 Shi-Min Hu，《Choreomaster：面向编舞的音乐驱动舞蹈合成》，*ACM Transactions onGraphics*，2021。[16] Zijie Ye、Haozhe Wu、Jia Jia、Yaohua Bu、Wei Chen、Fanbo Meng 和 Yanfeng Wang，《Choreonet：迈向基于编舞动作单元的音乐到舞蹈合成》，*ACM* 多媒体，2020。[17] Ferda Ofli、Engin Erzin、Y¨ucel Yemez 和 A Murat Tekalp，《Learn2dance：学习用于编舞合成的统计音乐到舞蹈映射》，*IEEE* 多媒体汇刊，第14卷，第3期，第747–759页，2011。[18] Li Siyao、Weijiang Yu、Tianpei Gu、Chunze Lin、Quan Wang、Chen Qian、Chen Change Loy 和 Ziwei Liu，《Bailando：通过带有编舞记忆的演员-评论家GPT进行三维舞蹈生成》，载于*CVPR*，2022。[19] Aaron van den Oord、Oriol Vinyals 和 Koray Kavukcuoglu，《神经离散表示学习》，神经信息处理系统，2017。

[20] Vijay R Konda、John N Tsitsiklis，《演员-评论家算法》，神经信息处理系统，2002。

[21] Alec Radford、Jeffrey Wu、Rewon Child、David Luan、Dario Amodei、Ilya Sutskever 等，《语言模型是无监督多任务学习者》，*OpenAI* 博客，第 1 卷，第 9 期，2019 年。[22] Yuhang Huang、Junjie Zhang、Shuyan Liu、Qian Bao、Dan Zeng、Zhineng Chen 和 Wu Liu，《音乐驱动的流派条件长期三维舞蹈生成》，载于*ICASSP 2022——2022*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2022 年，第4858–4862 页。[23] Jinwoo Kim、Heeseok Oh、Seongjean Kim、Hoseok Tong 和 Sanghoon Lee，《全新的舞伴：由多种舞蹈流派控制的音乐条件多样化舞蹈》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2022 年，第 3490–3500 页。[24] Yuxuan Wang、Daisy Stanton、Yu Zhang、RJ Skerry-Ryan、Eric Battenberg、Joel Shor、Ying Xiao、Fei Ren、Ye Jia 和 Rif A. Saurous，《风格标记：端到端语音合成中的无监督风格建模、控制与迁移》，国际机器学习会议，2018 年。[25]Wu Pengfei、Ling Zhen-Hua、Liu Lijuan、Jiang Yuan、Wu Hongchuan 和 Dai Lirong，《使用风格标记和半监督训练的端到端情感语音合成》，*IEEE*会议论文集，2019 年。[26] Yi Lei、Shan Yang、Xinsheng Wang 和 Lei Xie，《Msemotts：情感语音合成的多尺度情感迁移、预测与控制》，*IEEE/ACM*音频、语音与语言处理汇刊，第 30 卷，第 853–864 页，2022 年。[27] Yi Ren、Chenxu Hu、Xu Tan、Tao Qin、Sheng Zhao、Zhou Zhao 和 Tie-Yan Liu，《FastSpeech 2：快速、高质量的端到端文本转语音》，载于国际学习表示会议，2020 年。[28] Martin Heusel、Hubert Ramsauer、Thomas Unterthiner、Bernhard Nessler 和 Sepp Hochreiter，《采用双时间尺度更新规则训练的生成对抗网络收敛于局部纳什均衡》，亚洲应用科学与工程期刊，2017 年。