节奏舞者：基于关键动作转换图和有条件姿态插值网络的3D舞蹈生成方法研究

贺亚运，彭俊清，王健宗，肖京
平安科技（深圳）有限公司，广东 深圳 518063

摘要

3D舞蹈是元宇宙中虚拟人的一种重要表现形式，它将音乐与舞蹈进行有机结合，大大增强了元宇宙中相关应用的趣味性。之前的工作通常把3D舞蹈生成简单视作一个序列生成任务，但是生成的舞蹈动作质量较差且与音乐的契合度较低。受人类学习舞蹈过程的启发，提出了一种新颖的3D舞蹈框架——“节奏舞者”来解决上述问题。该框架首先使用VQ-VAE-2对舞蹈进行分层编码量化，可有效改善舞蹈生成质量；然后使用节奏点上的关键动作编码建立关键动作转换图，既可保证生成的舞蹈动作与音乐节拍的契合度，又可增加舞蹈动作的多样性。为了确保关键动作之间平滑自然地连接，提出了一个姿态插值网络来学习关键动作之间的转换动作。通过大量实验证明，该框架避免了长序列生成的不稳定和不可控问题，实现了舞蹈动作与音乐节奏的高度契合，达到了当前最优效果。

关键词

3D舞蹈；元宇宙；舞蹈生成；深度学习

中图分类号：TP399 文献标志码：A doi: 10.11959/j.issn.2096-0271.2023004

Rhythm dancer: 3D dance generation by key-motion transition graph and pose-interpolation network

HE Yayun, PENG Junqing, WANG Jianzong, XIAO Jing
Ping An Technology (Shenzhen) Co., Ltd., Shenzhen 518063, China

Abstract

3D dance is an indispensable form of virtual humans in the metaverse. It organically combines music and dance art, which greatly increases the interest in the metaverse. Previous work usually treats it as a simple sequence generation task, but it is difficult to match the dance movements with the music beat perfectly and the quality of long sequence dance generation is

difficult to be guaranteed. Inspired by the process by which humans learn to dance, a novel 3D dance framework "Rhythm Dancer" to solve the above problems was proposed. The framework first uses VQ-VAE-2 to encode and quantify the dances in a hierarchical way, which effectively improves the quality of dance generation. Then, a key movement transition map was created using the core dance movements on the rhythm points, which not only ensures that the generated dance movements fit with the music beat, but also increases the diversity of dance movements. To ensure smooth and natural connections between the core dance moves, a pose-interpolation network was proposed to learn the transition movements between key moves. Extensive experiments demonstrate that the framework not only avoids the instability and uncontrollability problems of long sequence generation, but also achieves a higher match between dance movements and music rhythms, reaching state-of-the-art results.

Key words
3D dance, metaverse, dance generation, deep learning

0 引言

如何更好地由音乐生成舞蹈是一项有意义的研究。随着元宇宙变得火热，其中的虚拟角色需要有更多的技能，跳舞有助于虚拟角色在元宇宙或者其他线上多媒体平台更好地表现自己。然而，生成令人满意的舞蹈动作并非易事。

当前国内外对3D舞蹈生成的研究相对较少。目前主流的方法是将3D舞蹈生成看作一个序列生成任务，例如参考文献[1]提出的AI编舞框架，采用全注意力机的模型交叉Transformer网络，将一段音乐和2 s的初始舞蹈序列作为输入，生成后续的舞蹈动作；参考文献[2]提出了一种先使用VQ-VAE对舞蹈动作进行编码和量化，再使用GPT网络生成舞蹈动作序列的方法，有效地对舞蹈动作空间进行降维，使生成的舞蹈动作更加自然合理。

目前舞蹈生成过程中主要存在两个问题：①生成的舞蹈动作很难与音乐节拍保持一致；②生成长序列的舞蹈动作会出现一些问题，如随着时间推移，生成的动作质量越来越差，或者动作一直简单地重复。

许多研究者试图设计一个模型，这个模型将从音乐中提取的特征作为输入，进而生成对应的舞蹈动作序列[1,3-7]，但是这样的模型通常是不稳定的，有时生成的舞蹈动作会很奇怪。这是因为舞蹈动作空间是一个高维度的空间，而在模型训练时并没有对动作空间的范围进行严格限制，而且舞蹈动作与音乐节拍是否一致最影响人们对生成舞蹈的直观感受。但是直接让模型从数据中学习音乐与动作的这种一致性是很难的。

为了解决以上问题，本文提出了一种新的3D舞蹈生成方法，此方法受到人类学习舞蹈的过程的启发。当学习一个舞蹈时，舞蹈老师通常会先告诉学生这个舞蹈包含哪些关键动作，学会了关键动作之后，再学习如何衔接它们。非常重要的一点是，舞蹈的关键动作通常会出现在音乐节拍点上，因此通过分析舞蹈配乐的节拍，可以找出舞蹈的关键动作。但是，找到关键动作之后，它们之间是彼此孤立的，接下来需要重新扫描舞蹈数据集，建立关键动作的转换图。该图是一个有向有权图，通过该图可以获取关键动作的转换关系和转换概率。

为了避免舞蹈动作空间维度太高导致生成的舞蹈动作不稳定，本文先使用VQ-VAE-2[8]这种无监督的学习方式，对所有的舞蹈动作进行量化编码，最终生成

一个舞蹈动作的编码簿。编码簿的大小通常只有几百个编码向量，其中每一个编码都表示一种唯一的舞蹈动作。通过这种方式可以对舞蹈动作空间进行降维，这保证了生成的舞蹈动作更加稳定，并且不会有奇怪的舞蹈动作。关键动作的衔接也很重要，本文使用有条件姿态插值网络（以下简称姿态插值网络），以两个关键动作和舞蹈风格特征为输入，生成中间的舞蹈动作序列。

完成以上步骤，模型就具备了为一首音乐生成对应舞蹈动作的能力。首先，通过分析节拍找出音乐的节拍点，这些节拍点就是需要放置关键动作的位置；然后，在关键动作转换图中进行采样，生成一串关键动作，并将其放置到节拍点的位置；最后使用训练好的姿态插值网络，生成关键动作之间的衔接动作序列。

综上所述，本文的主要贡献如下：

- 使用VQ-VAE-2对舞蹈动作进行编码量化，有效降低了舞蹈动作空间维度，并且这种分层的模型结构使生成的舞蹈动作更加流畅和稳定；
- 提出通过建立关键动作转换图的方式来生成关键动作；
- 提出先生成关键动作，再使用姿态插值网络生成衔接舞蹈动作的方式来生成3D舞蹈；
- 提出基于联合因果注意力的姿态插值网络，并在模型训练中引入了新的损失函数。

1 相关工作介绍

1.1 人物动作合成与3D舞蹈生成

如何生成更加真实的人体动作序列一直是学者研究的一个重要方向。早些时候，人们使用动作图[9-10]生成人体动作。动作图是在大量人体动作捕捉数据[11]上建立的有向图，图上的节点表示一个人体动作，边表示不同动作之间的转换关系。将所有的动作归类并放到一个图上，然后在图上采样，就可以得到一串连续动作。然而这种动作生成的方式可控性很差，不能针对特定的场景生成对应的动作。之后随着深度学习的发展，一些研究者试图从一个大的舞蹈数据集中训练出一个深度模型，希望这个深度模型可以自动学习出音乐与舞蹈动作之间的关系。他们尝试了很多网络架构，如CNN[12-13]、GAN[14]、Transformer等，但是生成效果却不尽如人意。生成过程中经常会出现动作的简单重复，或者几乎静止不动的情况，而且会生成一些奇怪的动作，这说明模型没有很好地学习到人体正常的姿态范围。

1.2 舞蹈动作的编码和量化

目前已经有比较成熟的对人体进行建模[15]的方法，但是理论上人体动作空间是很大的。受人体结构的限制，每个关节只能在一定范围内移动，因此真正的人体动作空间实际上是大的动作空间中的一个子空间，如果不对这个动作空间进行空间限制，模型最终可能生成奇怪的人体动作。但是如果人为给各种可能的舞蹈动作做划分和编码，工作量是巨大的，几乎不可能实现。为了解决这个问题，参考文献[2]使用VQ-VAE[16]对舞蹈动作进行了编码和量化，此方法是一种无监督的方式，简洁且高效。与参考文献[2]不同的是，本文使用了更先进的VQ-VAE-2[8]的分层结构，更好地利用了局部信息和全局信息。而且本文提出的关键舞蹈动作是上半身动作和下半身动作结合的一个整体动作，因此本文

会对整体动作进行编码量化。在编码簿的基础上进行后序舞蹈动作的生成时，本文在步骤和方法上均与参考文献[2]不同。

1.3 分阶段的舞蹈生成框架

之前的研究表明，音乐节拍和运动的空间回折点在时间上具有强相关性[1,17]，因此保证生成舞蹈的运动节拍和音乐节拍的契合度尤为重要。人们在使用动画制作工具制作动画时，通常并不会制作动画的每一帧，而是制作动画的关键帧，关键帧制作完成后，就可以由此生成流畅的动画。与此过程类似，参考文献[18]提出了一种分两个步骤来生成3D舞蹈的方法：第一步，分析音乐节拍信息，找出音乐的节拍点，然后将节拍点处的音乐片段截取出来，并提取频谱特征，使用提取的频谱特征训练一个深度模型，使之能够通过音乐片段生成对应的关键动作；第二步，预测相邻关键动作之间的运动曲线参数，使用多结Kochanek-Bartels样条（multi-knots Kochanek-Bartels splines）方法对每个运动曲线进行建模。此种3D舞蹈生成方式逻辑上合理，但是音乐片段与关键动作并没有很强的相关性，生成的关键动作之间几乎没有关联。而且对于没有在训练集中的音乐，模型可能生成比较奇怪的动作。在预测相邻关键动作之间的运动曲线时，参考文献[18]使用了比较传统的建模方法，流程复杂且模型准确率较低。本文也采用了相似的分阶段生成3D舞蹈动作的框架，但是在生成关键动作时，本文使用了更合理的关键动作转化图，并且在生成关键动作之间的衔接动作时使用了更先进的深度模型，保证了更好的生成效果。

2 节奏舞者生成模型

节奏舞者3D舞蹈生成模型的工作流程如**图1**所示。与其他舞蹈生成模型[1,3,4,7,19-21]不同，本文没有一次性让模型生成所有的舞蹈动作序列，而是采用了“两步走”的方式，先生成关键动作，再生成关键动作之间的衔接动作。关键动作在整个舞蹈中起着至关重要的作用，关键动作与音乐节拍是否一致非常影响观众的直观感受。而且，本文没有直接使用原始的动作数据，

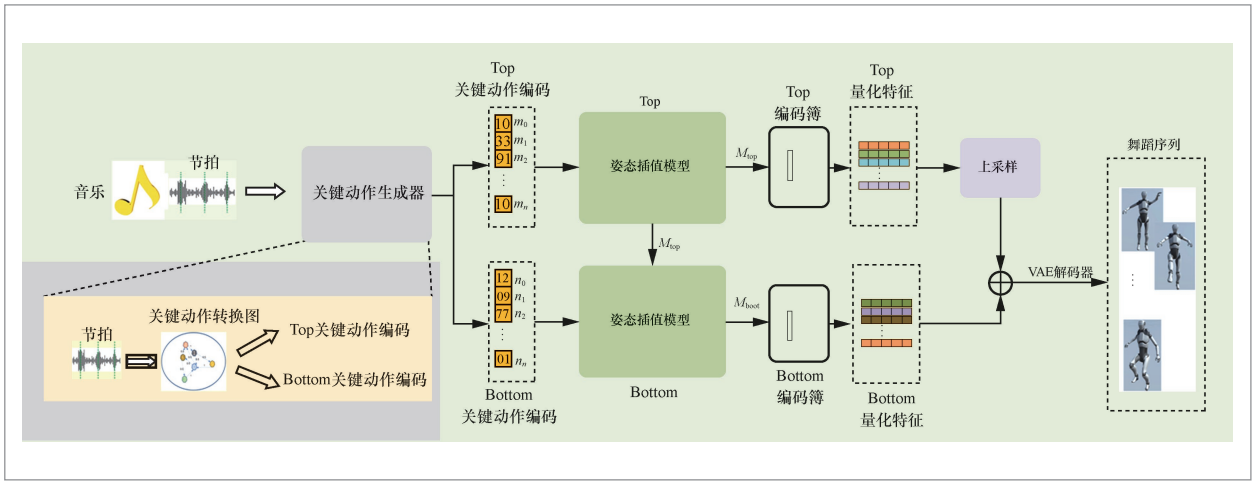

图1 3D 舞蹈生成过程总览

而是先对数据集中的舞蹈动作做了编码和量化，在编码量化过程中使用VQ-VAE-2[8]的包含Top层和Bottom层的双层结构，使模型可以同时考虑到局部信息和全局信息，这个过程会在第2.1节进行介绍。然后在第2.2节会介绍如何从数据集中提取所有的关键动作，并生成关键动作转换图[22]。

2.1 基于VQ-VAE-2的舞蹈动作编码簿生成

人体的运动涉及几十个关节[15]，用来表示人体动作的向量的维度很高，而且人体各个关节都有各自的运动范围，如果不加限制就可能生成很多奇怪的动作。

一种好的解决方案是先对高维的动作向量进行降维，把数据集中所有人体的动作压缩到一个有限的空间中，而且最好以一种无监督的方式进行。而这正是VQ-VAE[16,23]所擅长的。为了在舞蹈生成时可以让模型同时考虑局部信息和全局信息，从而使生成的舞蹈动作更加稳定和流畅，本文采用了VQ-VAE-2[8]的分层结构。

如图2所示，一段舞蹈动作 $M \in R^{T\times(J\times 3)}$，$T$ 是时长，J 是关节数量，可以用一段量化特征序列 $e^q \in R^{T'\times 2C}$ 表示，$T'=T/d$，d 是Bottom部分的下采样率，$2C$ 是量化特征的通道数。本文使用一个一维时域卷积 E_1 将动作序列 M 编码成向量 e_{bottom}，e_{bottom} 可以继续使用一维时域卷积 E_2 编码为 e_{top}。在训练时，共分两步，先训练Top部分，再训练Bottom部分，Top部分和Bottom部分分别包含一个编码簿。对 e_{top} 和 e_{bottom} 使用的量化方式相同。以 e_{top} 为例，对 e_{top} 中的每个向量 $e_{\text{top},i}$ 选取Top编码簿中与之最近的元素作为量化后的向量 $e_{\text{tq},i}$。

$$e_{\text{tq},i}=\arg\min_{z_j\in\mathcal{Z}}\left\|e_{\text{top},i}-z_j\right\| \tag{1}$$

最后，将 e_{tq} 和 e_{bottom} 拼接成 e_{q}，e_{q} 可以通过动作解码器重新解码为舞蹈动作序列 M。

为了避免关节整体位移对动作编码的

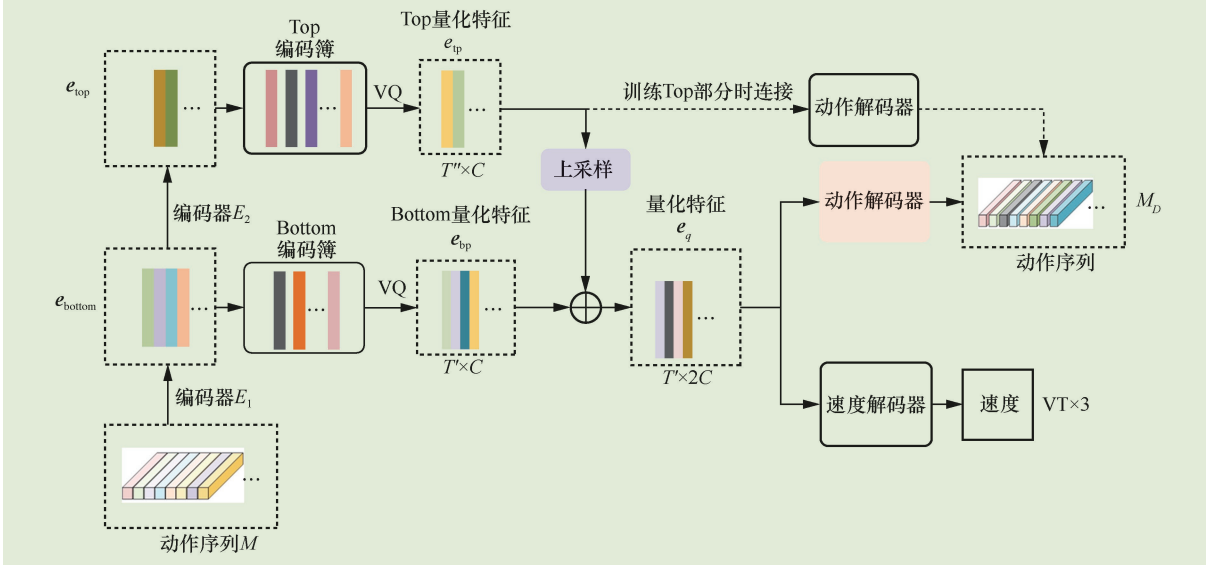

图 2 3D 舞蹈动作生成使用的 VQ-VAE-2 模型结构

影响(出现在不同位置的同一个动作应该被编码为同一个向量),本文先将输入的动作序列M进行归一化,即把根关节点的位置置零。为了表示做某个舞蹈动作时身体的整体移动速度,使用速度解码器获取整体移动速度V。V是一个矢量,包含速度的大小和方向。

训练时,先训练Top部分,再将Top部分固定,训练Bottom部分。两者训练的方法和使用的损失函数基本一致,只是在训练Bottom部分时,需要将e_{tq}进行上采样后与e_{tq}进行拼接,生成最终的量化特征e_q。下面以训练Top部分为例进行介绍,其损失函数如下:

$$L_M = L_{rec}(M, M_D) + \left\| sg\left[e_{top}\right] - e_{tq} \right\| + \omega \left\| e_{top} - sg\left[e_{tq}\right] \right\| \quad (2)$$

其中,L_{rec}为重建损失。在这个重建损失中,不仅包含3D关节点的相对位置损失,而且包含关节点运动的速度和加速度损失:

$$L_{rec}(M, M_D) = \left\| M - M_D \right\| + \alpha \left\| M' - M'_D \right\| + \beta \left\| M'' - M''_D \right\| \quad (3)$$

其中,M'和M''分别是M在时间上的一阶偏导和二阶偏导,α和β是可学习的参数。L_M第二项为“编码簿损失”,sg是停止计算梯度(stop gradient)[24]的缩写,sg[]表示不对方括号内的变量计算梯度。此部分计算编码器得到特征向量e_{top}和其对应的量化特征e_{tq}之间的距离,并将其作为辅助误差项。此误差项只向字典向量e_{tq}传递,通过对误差惩罚来学习e_{tq}向量,不更新编码器和解码器。第三项与第二项一致,也是计算e_{top}与e_{tq}的距离。不过这里对量化特征向量e_{tq}使用了stop gradient约束,使得此误差项只向编码器反向传递,训练速度解码器时的损失函数为$L_{rec}(V, V_t)$,意为速度预测值与真实值的差值,其中V_t为速度的真实值。

VQ-VAE-2训练完成后,Top编码簿和Bottom编码簿中分别包含了代表各种舞蹈动作的量化特征。因为Top部分使用了更大的下采样率,所以Top部分相较Bottom部分信息更加浓缩。Top编码簿中的量化特征包含更多的全局信息,而Bottom编码簿中的量化特征包含更详细的局部信息。在进行舞蹈动作生成时,同时考虑全局信息和局部信息,可以使生成的动作更加稳定流畅。

2.2 基于有向图的关键动作转换图生成

音乐节拍使用音频处理工具Librosa[25]进行提取,即用Librosa获取一段音乐中音乐节拍出现的时间点,此时间点处对应的舞蹈动作即关键动作。找出的关键动作可以用第2.1节中的Top编码簿和Bottom编码簿中的量化特征编号表示,分别记作T_code和B_code,通过分析整个舞蹈数据集,最终可以得到任何两个关键动作的转换关系和转换概率(数据集中相邻的两个关键动作视为具有转换关系),并且建立一个关键动作的转换图[22]。

关键动作转换图生成后,可以为一段新的音乐生成所需的关键动作。方法如下:先使用节拍分析工具获取音乐的节拍信息,即获取此段音乐中需要插入关键动作的时间点和关键动作数量;然后在关键动作转换图上进行随机游走采样,采样所得的动作序列即此段音乐所需的关键动作序列。

2.3 基于联合因果注意力[26]的姿态插值网络

生成关键舞蹈动作后,可以使用姿态插值网络生成两个关键动作之间的衔接动

作，如**图1**所示。先使用姿态插值模型生成Top部分的舞蹈动作编码 M_{top}，再将 M_{top} 作为Bottom部分姿态插值模型的条件输入来生成Bottom部分的舞蹈动作编码 M_{bott}。姿态插值模型如**图3**所示，相比Top部分的姿态插值模型，Bottom部分的姿态插值模型只是增加了 M_{top} 作为条件输入，其他结构相同。下面仅介绍Bottom部分舞蹈动作编码的生成过程。

图4所示是如何使用Bottom部分的姿态插值模型估计舞蹈动作的概率值。假设有一段Bottom部分的动作编码序列 M_{bott}，M_{bott} 需要符合特定规则，M_{bott} 序列的第一个元素为起始动作编码 p_{start}，第二个元素为终止动作编码 p_{end}，意为模型生成的整个序列要以 p_{start} 为开始，并以 p_{end} 为终止。假如现在已经生成了Top部分的舞蹈动作编码 M_{top}。接下来，首先将 M_{bott} 和 M_{top} 分别转化为可学习的特征向量$\boldsymbol{B}$和$\boldsymbol{T}$，并将它们与提取的舞蹈风格特征向量$\boldsymbol{S}$沿时间维度进行拼接[27]；然后将拼接的向量送入12层（该值可以调整，本文选择层数为12）的Transformer层；最后经过全连接和Softmax层输出表示动作编码概率的向量$\boldsymbol{R}$。$\boldsymbol{R}$的长度为Bottom编码簿的大小。对于时刻t，计算编码簿中每一个编码 z_j 在$\boldsymbol{R}$中对应的概率值，并选取概率最大值对应的编码作为t时刻的预测动作编码值 $\hat{m}_t$：

$$\hat{m}_t = \arg\max_k P\left(z_k \mid \boldsymbol{S}, M_{\text{top}}, m_{0\cdots t-1}\right) \tag{4}$$

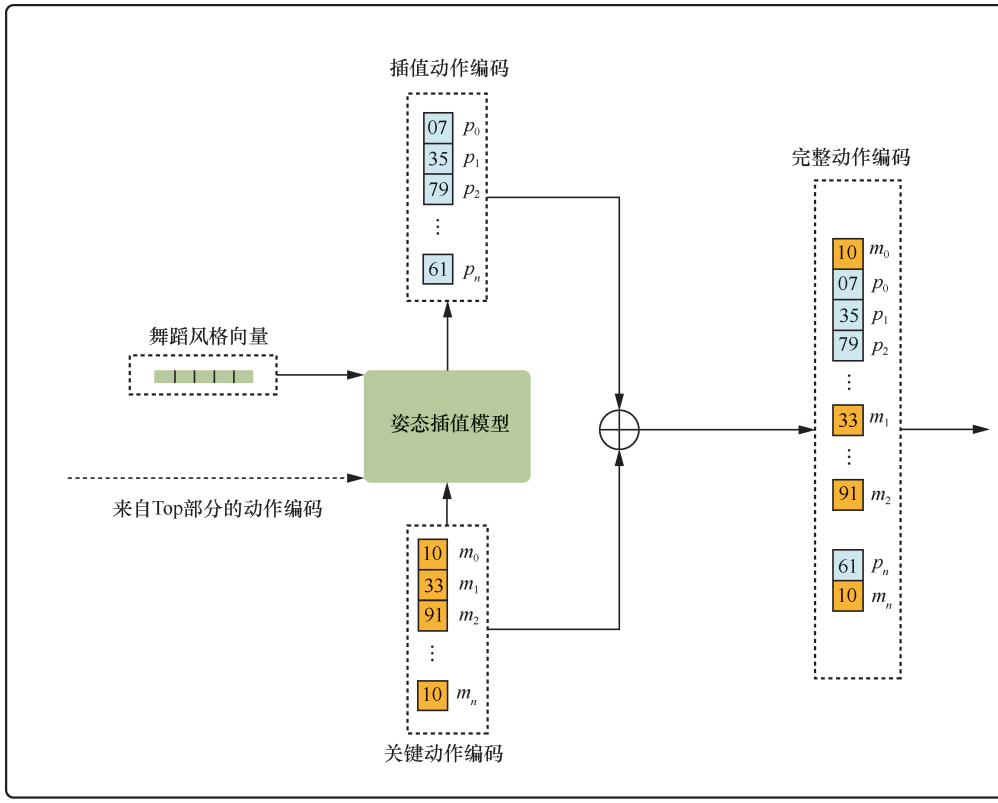

图 3 姿态插值模型

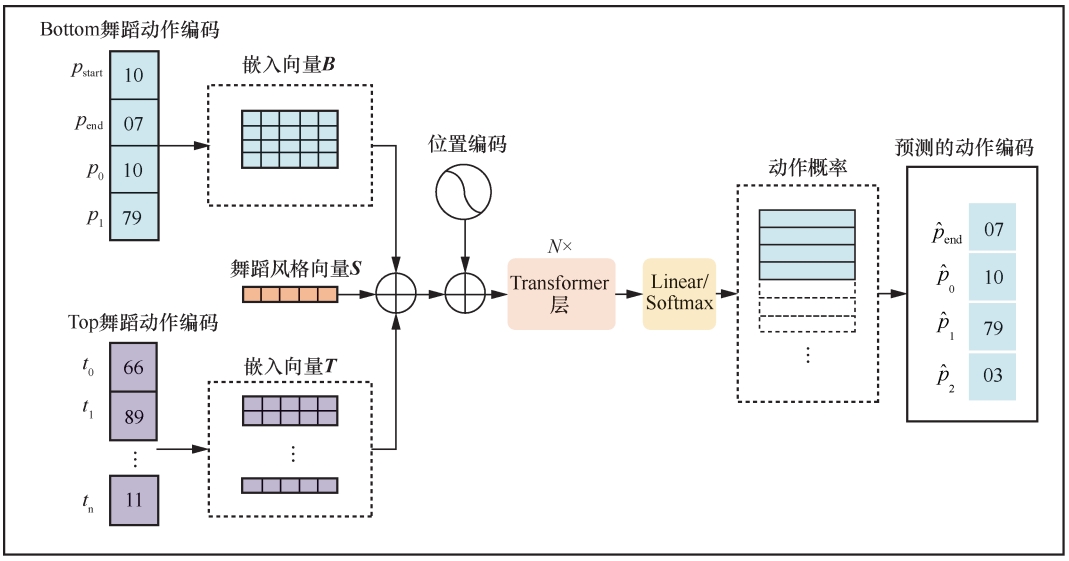

图 4 有条件姿态插值网络结构

2.3.1 Transformer层与联合因果注意力

Transformer[26]是一个基于注意力机制的网络，被广泛应用在自然语言处理和图像识别等领域。Transformer中最重要的是多头注意力机制的使用，输入$\boldsymbol{X}$经过多头注意力机制层后被转化为新的向量$\boldsymbol{U}$。计算过程如下所示：

$$\boldsymbol{U}=\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V},\boldsymbol{M})=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}+\boldsymbol{M}}{\sqrt{C}}\right)\boldsymbol{V} \quad (5)$$

$$\boldsymbol{Q}=\boldsymbol{X}\boldsymbol{W}^{Q},\boldsymbol{K}=\boldsymbol{X}\boldsymbol{W}^{K},\boldsymbol{V}=\boldsymbol{X}\boldsymbol{W}^{V}$$

其中，$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$由输入矩阵$\boldsymbol{X}$计算而来，$\boldsymbol{M}$是掩码矩阵，针对不同类型的注意力机制有不同的取值。本文在预测下一个Bottom动作编码时，应该防止模型使用未来信息，因此此处应该使用因果注意力，也就是模型只能用当前时刻之前的信息。但是本文中模型的输入较复杂，输入中包含由Bottom舞蹈动作编码转化而来的嵌入向量$\boldsymbol{B}$、由Top舞蹈动作编码转化而来的嵌入向量$\boldsymbol{T}$，以及舞蹈风格向量$\boldsymbol{S}$。针对特殊的输入结构，本文提出了一种新的注意力层，称之为联合因果注意力层，其结构如**图5**所示。因为要预测下一时刻Bottom舞蹈的动作编码，为了避免模型使用未来信息，所以对Bottom部分使用因果注意力，而对Top部分和舞蹈风格向量部分使用全注意力，并且只允许Top部分和舞蹈风格向量部分向Bottom部分单向传递信息。

联合因果注意力层既保证了不同种类输入之间进行充分的信息交换，同时保证了在预测下一时刻的Bottom动作编码时，模型不会使用未来信息。

2.3.2 姿态插值模型的训练

模型采用有监督学习的方式进行训练。但是因为姿态插值模型需要生成两个关键动作之间的衔接动作，所以笔者希望模型可以很好地学习到如何从前一个关键动作开始，顺滑地进行中间动作的过渡，并保证最后生成的动作一定落在后一

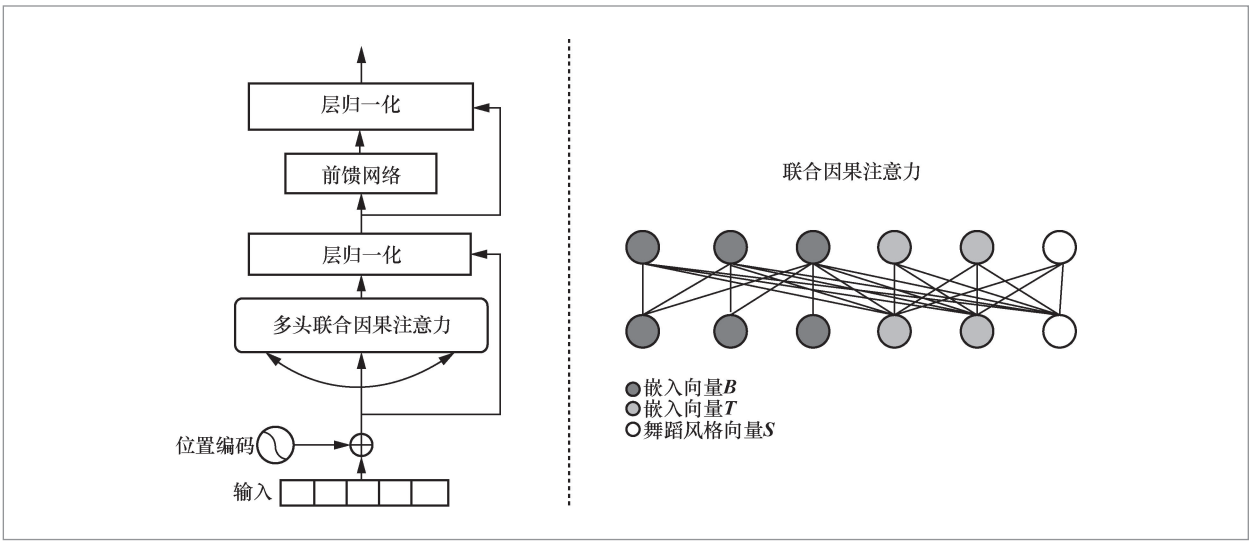

图 5 Transformer 层结构

个关键动作上。对于模型生成的序列，头部序列和尾部序列的生成质量非常重要，因此在训练时，本文提出了一种新的损失函数：

$$L=\frac{1}{N}\sum_{n=0}^{N-1}\left[\text{CrossEntropy}(\hat{p}_n,p_n)\times f\left(\frac{n}{N-1}\right)\right]$$
$$f(x)=\alpha x^2-\alpha x+1 \quad (0<\alpha<4) \tag{6}$$

其中，N为模型预测的两个关键动作之间衔接动作序列的长度，p_n为动作编码的真实值，$\hat{p}_n$为模型输出的动作编码的预测值。$f(x)=\alpha x^2-\alpha x+1\,(\alpha>0)$为二次函数，当$x\in[0,1]$时，$1-\frac{\alpha}{4}\leqslant f(x)\leqslant 1$，并且在$x=0$和$x=1$处，$f(x)=1$，在$x=0.5$处，$f(x)=1-\frac{\alpha}{4}$，即$x\in[0,1]$，函数呈现两端大、中间小的形态。本文中$\alpha$=2。损失函数中加入此函数后，会对预测动作序列的两端给予更严厉的惩罚，使模型更好地保障两端动作编码的生成质量。

训练完成后，姿态插值模型可以学习到如何生成p_{start}和p_{end}之间的动作序列，并根据输入的舞蹈风格调整模型的输出。

先生成Top部分的动作编码，再生成Bottom部分的动作编码，这是一个由粗略到精细的生成过程，可以同时考虑到全局信息和局部信息。

最后，可以使用生成的Top部分的动作编码和Bottom部分的动作编码按**图2**所示的方式生成最终的舞蹈动作序列。

3 实验结果分析

3.1 实验准备

本文在目前公开的最大的舞蹈数据集AIST++[1]上进行模型训练和测试。此数据集一共包含992段高质量的3D舞蹈片段。数据被划分为训练集和测试集两个集合，其中训练数据952条，测试数据40条。在训练集上进行模型训练，且只在测试集上进行模型测试。在进行数据集划分时，需要严格保证训练集和测试集没有音乐和舞蹈动作片段的重合。

本文中，当舞蹈动作序列编码为 e_{bottom} 特征向量时，下采样率为4，当将 e_{bottom} 编码为 e_{top} 时，下采样率也为4，也就是说原始的舞蹈动作序列编码为 e_{top} 向量的下采样率为16。Bottom编码簿和Top编码簿大小分别为512和256，编码后的特征向量通道数 C 都为512。训练VQ-VAE-2网络时，舞蹈数据被切割为长度为5 s（300帧）的片段。送入模型进行训练时，batchsize设为64，损失函数 L_M 中第三项的系数 ω 为0.15，重建损失函数 L_{rec} 中的系数 α 和 β 分别为0.8和0.9。采用Adam优化器以 1×10^{-5} 的学习率训练VE-VAE-2网络。训练姿态插值网络时，舞蹈动作编码转化为嵌入向量后的向量通道数为512，Transformer层的层数为12，注意力层的头数为12。舞蹈风格特征提取采用了参考文献[27]的方法，通过风格嵌入向量生成网络，将输入的一段舞蹈片段进行舞蹈风格提取，并生成一个维度为512的舞蹈风格向量，引入舞蹈风格向量的目的是希望模型在生成舞蹈时可以统一风格，增加舞蹈的观赏性。训练姿态插值网络时采用的是Adam优化器[28]，学习率设置为 3×10^{-5}，训练轮数为300。VQ-VAE-2和姿态插值模型使用一台Tesla V100 GPU进行训练，共耗时2天。

训练完成后，模型根据给定的音乐可以输出对应的舞蹈序列，舞蹈序列的每一帧为人体24个关节的坐标值，输出的舞蹈序列可以导入虚拟3D引擎Unity中进行虚拟人物驱动，Unity中需要加载FinalIK资源包。本文中演示使用的UNITy版本为2019.03.11f1，FinalIK版本为V2.1。

3.2 实验结果量化评估

本文使用3个量化指标来评估生成的舞蹈动作。这3个指标分别为舞蹈动作质量、舞蹈动作多样性、音乐与舞蹈节拍一致性。

- 舞蹈动作质量：舞蹈动作质量的获取方法是，计算生成的舞蹈动作和数据集中所有的舞蹈动作在动力学和几何学两种特征上的FID [29]值，即弗雷歇距离。舞蹈动作的动力学特征和几何学特征分别使用fairmotion[30]中实现的两个运动特征提取器[31]进行提取，分别使用 FID_{k} 和 FID_{g} 表示动力学特征对应的FID值和几何学特征对应的FID值。好的舞蹈动作质量可以保证生成的舞蹈动作更加真实合理。

- 舞蹈动作多样性：当给定不同的音乐时，人们希望模型可以生成更加多样性的舞蹈。计算在AIST++测试集上生成的40段舞蹈在特征空间中的平均欧氏距离，以此衡量生成舞蹈的多样性。在这里特征空间也分为两种，分别为动力学特征空间和几何学特征空间，记为 Dist_{k} 和 Dist_{g}。

- 音乐与舞蹈节拍一致性：节拍一致性是最直接影响人们对舞蹈观看体验的指标，这个指标可以衡量音乐和舞蹈动作的相关性。本文按照其他论文的做法，计算舞蹈动作和背景音乐的节拍一致性得分，这个值可以评估舞蹈动作的节拍和音乐节拍的相关程度，计算式为：

$$\frac{1}{\left|B^m\right|}\sum_{t^m\in B^m}\exp\left\{-\frac{\min_{t^d\in B^d}\left\|t^d-t^m\right\|^2}{2\sigma^2}\right\} \quad (7)$$

其中，$B^d=\{t^d\}$ 是舞蹈动作节拍，$B^m=\{t^m\}$ 是音乐节拍，σ 是针对不同FPS（每秒帧数）的归一化参数。本文中FPS值为60，对应的 σ 值为3。音乐节拍可以使用音频处理工具Librosa[25]进行提取，即通过工具分析得到音乐节拍出现的时间点，舞蹈动作节拍通过计算运动速度的局部最小值来获得。

本文利用以上3个指标对比了节奏

舞者模型与目前几个较好的舞蹈生成模型。参与对比的舞蹈生成模型包括Li J M等人所提模型[32]、DanceNet[33]、DanceRevolution[34]、FACT[1]、Bailando[2]。在AIST++测试集上，每种方法都生成40段对应的舞蹈序列，然后计算上述3个指标。对比结果见**表1**。通过**表1**可以看出，节奏舞者模型在各个指标上的结果优于绝大部分现有模型：在FID_k和FID_g上相较之前表现最好的模型Bailando有更好的效果；与Bailando相比在节拍一致性得分上提升了5%，说明节奏舞者模型生成的舞蹈与音乐更加契合。这说明先建立关键动作转化图，再从图中采样获取关键动作的方式可以更好地保证音乐节拍与舞蹈节拍的一致性。这验证了本文方法的有效性。

舞蹈生成的最终目的是供人欣赏。为了在人的主观感受上更好地评估生成舞蹈的效果，本文进行了用户体验调查。首先从每种舞蹈生成方法生成的舞蹈中选出30个生成样本，然后将每种方法生成的样本与节奏舞者模型生成的样本随机两两组合，之后选择10位观众，让每位观众观看每个组合的视频文件，并标记哪个文件中生成的舞蹈效果更好，最后将结果进行汇总。详细结果可见**表1**。相比之前最好的舞蹈生成模型Bailando，节奏舞者模型胜出率高达74.6%。节奏舞者模型生成的舞蹈样例如**图6**所示。

3.3 消融实验

本文进行了消融实验，以验证VQ-VAE-2和姿态插值网络中联合因果注意力层的有效性。有效性的评估使用了动作质量（FID_k和FID_g）和节拍一致性得分两种量化指标。除量化指标外，为了衡量不同生成结果对人产生的直观感受，笔者请观众对生成的样本进行了主观感受打分。

本文比较了VQ-VAE和VQ-VAE-2对舞蹈动作进行编码量化的效果差异，并且研究了在训练VQ-VAE-2网络时，式（3）中节点运动的速度和加速度损失的加入对结果的影响，结果见**表2**。从**表2**可以看出，相比VQ-VAE，VQ-VAE-2在FID_k、FID_g和节拍一致性得分3个指标上均有更好的表现，说明VQ-VAE-2这种分层的结构可以在舞蹈生成过程中兼顾局部信息和全局信息，有助于提升舞蹈生成质量，符合实验预期。损失函数中去除速度/加速度损失项后，FID_k和FID_g明显升高，分别上升了8.17%和6.39%；损失函数中去掉速度和加速度损失项后，节拍一致性得分变低，说明速度和加速度对于舞蹈动作序列来说是非常重要的信息，丢失这部分信息会影响模

表1 各舞蹈生成方法的结果

方法	动作质量		动作多样性		节拍一致性得分↑	用户体验调查本模型胜出率
	FID_k ↓	FID_g ↓	$Dist_k$ ↑	$Dist_g$ ↑		
Li J M等人所提模型	86.43	20.58	6.85*	4.93	0.161	100.00%
DanceNet	69.18	17.76	2.86	2.72	0.143	95.10%
DanceRevolution	73.42	31.01	3.52	2.46	0.195	87.20%
FACT	35.35	12.40	5.94	5.30	0.221	94.30%
Bailando	29.26	10.05	7.55	6.30	0.233	74.60%
节奏舞者	28.25	9.38	7.53	6.34	0.245	—

图 6 节奏舞者生成的舞蹈样例

表 2 不同量化方式和速度 / 加速度损失对结果的影响实验分析

方法	FID_k ↓	FID_g ↓	节拍一致性得分↑
VQ-VAE	29.21	10.26	0.238
VQ-VAE-2	28.25	9.38	0.245
无速度/加速度损失	30.56	9.98	0.215

型的舞蹈生成质量。

对于姿态插值网络，首先，将联合因果注意力层换成最简单的因果注意力层，即图5中嵌入向量***B***和嵌入向量***T***之间没有进行特征交叉融合。结果发现 FID_k 和 FID_g 大幅上升，说明在生成Bottom部分动作编码的过程中，特征向量***T***起到了关键的指导作用。为了探究舞蹈风格特征向量***S***对舞蹈生成效果的影响，实验中将舞蹈风格特征向量***S***去掉，即图5中，舞蹈风格特征向量***S***在注意力层中不参与特征融合，让模型根据测试集的音乐进行舞蹈生成，最后选择10位观众对舞蹈风格特征向量参与训练的模型生成的样本和舞蹈风格特征向量未参与训练的模型生成的样本分别打分（分数范围为1~5分），并计算平均分。结果显示，舞蹈风格特征向量参与训练的模型生成的样本分数更高，说明舞蹈风格特征向量在舞蹈生成时可以有效控制生成舞蹈的风格，给观众更好的视觉感受。详细结果见表3。

消融实验结果表明，VQ-VAE-2对舞蹈的生成质量有提升作用。将联合因果注意力层替换为简单的因果注意力层后，舞蹈生成质量大幅下降，这有效验证了Top部分舞蹈动作编码和Bottom部分舞蹈动作编码在舞蹈生成过程中协同配合的重要性。最后，实验结果表明，舞蹈风格向量的加入可以有效地控制生成舞蹈的风格，统

表 3 联合因果注意力层和舞蹈风格特征向量对生成结果的影响

方法	FID_k ↓	FID_g ↓	观众评价打分↑
无联合因果注意力层	78.95	17.34	—
有联合因果注意力层	28.25	9.38	—
无舞蹈风格特征向量	—	—	3.8
有舞蹈风格特征向量	—	—	4.3

一的舞蹈风格让整个舞蹈更加和谐流畅，给人更好的视觉感受。

4 结束语

本文提出了一种新的3D舞蹈生成框架。针对生成的舞蹈动作和音乐节拍难以契合的问题，本文提出了先构建关键动作转换图，再生成中间衔接动作的舞蹈生成方法，有效保证了动作节拍与音乐节拍的高度一致性。针对舞蹈动作空间维度太大而导致生成的舞蹈动作怪异的问题，本文提出使用VQ-VAE-2对舞蹈动作进行编码和量化，对舞蹈动作空间进行了有效降维，而且采用VQ-VAE-2的分层结构，让模型在生成舞蹈时可以兼顾全局信息和局部信息，使生成的舞蹈动作更加稳定流畅。为了更好地生成关键动作之间的衔接动作，本文提出了基于联合因果注意力的姿态插值网络，同时在模型训练时，引入了新的损失函数来保证生成的中间动作和两端的关键动作能更好地衔接。实验结果表明，在各量化指标上，节奏舞者模型相较之前的舞蹈生成模型均有更好的表现。

参考文献:

[1] LI R L, YANG S, ROSS D A, et al. AI choreographer: music conditioned 3D dance generation with AIST++[C]// Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2021.

[2] LI S Y, YU W J, GU T P, et al. Bailando: 3D dance generation by actor-critic GPT with choreographic memory[C]// Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 11040-11049.

[3] AHN H, KIM J, KIM K, et al. Generative autoregressive networks for 3D dancing move synthesis from music[J]. IEEE Robotics and Automation Letters, 2020, 5(2): 3501-3508.

[4] ALEMI O, FRANÇOISE J, PASQUIER P. GrooveNet: real-time music-driven dance movement generation using artificial neural networks[C]//Proceedings of the 23rd ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2017.

[5] GINOSAR S, BAR A, KOHAVI G, et al. Learning individual styles of conversational gesture[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 3492-3501.

[6] KAO H K, SU L. Temporally guided music-to-body-movement generation[C]// Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM Press, 2020: 147-155.

[7] REN X C, LI H R, HUANG Z J, et al. Self-supervised dance video synthesis conditioned on music[C]//Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM Press, 2020: 46-54.

[8] RAZAVI A, OORD A V D, VINYALS O. Generating diverse high-fidelity images with VQ-VAE-2[J]. arXiv preprint, 2019, arXiv:1906.00446.

[9] KOVAR L, GLEICHER M, PIGHIN F. Motion graphs[C]//Proceedings of ACM SIGGRAPH 2008. New York: ACM Press, 2008: 1-10.

[10] LEE J, SHIN S Y. A hierarchical approach to interactive motion editing for human-like figures[C]//Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques. [S.l.:s.n.], 1999: 39-48.

[11] PAPADOURAKIS A G. Motion capture and analysis: US20170348561A1[P]. 2017-12-07.

[12] HOLDEN D, SAITO J, KOMURA T. A deep learning framework for character motion synthesis and editing[J]. ACM Transactions on Graphics, 2016, 35(4): 1–11.
[13] HOLDEN D, SAITO J, KOMURA T, et al. Learning motion manifolds with convolutional autoencoders[C]// Proceedings of SIGGRAPH Asia 2015 Technical Briefs. New York: ACM Press, 2015: 1–4.
[14] HERNANDEZ A, GALL J, MORENO F. Human motion prediction via spatio-temporal inpainting[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 7133–7142.
[15] LOPER M, MAHMOOD N, ROMERO J, et al. Smpl[J]. ACM Transactions on Graphics, 2015, 34(6): 1–16.
[16] OORD A V D, VINYALS O, Kavukcuoglu K. Neural discrete representation learning[J]. arXiv preprint, 2017, arXiv:1711.00937.
[17] LEE H Y, YANG X D, LIU M Y, et al. Dancing to music[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. New York: ACM Press, 2019: 3586–3596.
[18] LI B Y, ZHAO Y C, SHI Z L, et al. DanceFormer: music conditioned 3D dance generation with parametric motion transformer[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(2): 1272–1279.
[19] CHEN K, TAN Z P, LEI J, et al. Choreomaster: choreography-oriented music-driven dance synthesis[J]. ACM Transactions on Graphics, 2021, 40(4): 1–13.
[20] FERREIRA J P, COUTINHO T M, GOMES T L, et al. Learning to dance: a graph convolutional adversarial network to generate realistic dance motions from audio[J]. Computers & Graphics, 2021, 94(Feb.): 11–21.
[21] TANG T R, JIA J, MAO H Y. Dance with melody: an LSTM-autoencoder approach to music-oriented dance synthesis[C]// Proceedings of the 26th ACM international conference on Multimedia. New York: ACM Press, 2018: 1598–1606.
[22] WEST D B. Introduction to graph theory[M]//Discrete mathematics in statistical physics. Berlin: Springer, 2001.
[23] YAN W, ZHANG Y Z, ABBEEL P, et al. VideoGPT: video generation using VQ-VAE and transformers[J]. arXiv preprint, 2021, arXiv:2104.10157.
[24] CHEN X L, HE K M. Exploring simple Siamese representation learning[C]// Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021.
[25] JIN Y H, ZHANG J K, LI M J, et al. Towards the automatic anime characters creation with generative adversarial networks[J]. arXiv preprint, 2017, arXiv:1708.05509.
[26] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM Press, 2017: 6000–6010.
[27] ZHANG X J, XU Y, YANG S, et al. Dance generation with style embedding: learning and transferring latent representations of dance styles[J]. arXiv preprint, 2021, arXiv:2104.14802.
[28] KINGMA D P, BA J. Adam: a method for stochastic optimization[C]//Proceedings of the 3rd International Conference for Learning Representations. [S.l.:s.n.], 2015.
[29] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. GANs trained by a two time-scale update rule converge to a local Nash equilibrium[J]. arXiv preprint, 2017, arXiv:1706.08500.
[30] WON D G A J. fairmotion – tools to load, process and visualize motion capture data[Z]. 2020.
[31] MÜLLER M, RÖDER T, CLAUSEN M. Efficient content-based retrieval of motion capture data[C]//Proceedings of ACM SIGGRAPH 2005. New York: ACM Press, 2005: 677–685.

[32] LI J M, YIN Y H, CHU H, et al. Learning to generate diverse dance motions with transformer[J]. arXiv preprint, 2020, arXiv:2008.08171.
[33] ZHUANG W L, WANG C Y, CHAI J X, et al. Music2Dance: DanceNet for music-driven dance generation[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2022, 18(2): 1-21.
[34] HUANG Y H, ZHANG J J, LIU S Y, et al. Genre-conditioned long-term 3D dance generation driven by music[C]// Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2022: 4858-4862.

作者简介

贺亚运(1990-),男,平安科技(深圳)有限公司资深算法工程师,主要研究方向为人工智能、声纹识别、元宇宙虚拟人等。

彭俊清(1973-),男,国家认证计算机系统架构设计师,平安科技(深圳)有限公司资深经理,高级人工智能算法研究员,在IT行业耕耘多年,精通架构设计、云平台和AI系统建设,发表多篇论文,获得多项专利授权。

王健宗(1983-),男,博士,平安科技(深圳)有限公司副总工程师,美国佛罗里达大学人工智能博士后,中国计算机学会(CCF)杰出会员,深圳市计算机学会理事,深圳市地方级领军人才,《大数据》期刊编委,曾任美国莱斯大学电子与计算机工程系研究员,主要研究方向为隐私计算、元宇宙、边缘计算和量子计算。曾获得中国专利奖优秀奖、深圳市科技进步奖、CCF科学技术奖、《麻省理工科技评论》中国2022年隐私计算科技创新人物称号等。

肖京(1972-),男,博士,平安集团首席科学家,深圳市政协委员,深圳市决策咨询委员会委员,CCF深圳分部副主席,广东省人工智能与机器人学会副理事长,上海市科协人工智能专业委员会委员,深圳市人工智能行业协会会长。先后在爱普生美国研究院及美国微软公司担任高级研发管理职务。发表学术论文249篇,美国授权专利101项,中国授权专利155项,参与及承担国家级项目11项,获吴文俊人工智能科学技术进步奖一等奖、上海市科学技术进步奖一等奖、中国专利优秀奖、广东省专利优秀奖,以及吴文俊人工智能"杰出贡献奖"。

收稿日期: 2022-08-24
通信作者: 王健宗, jzwang@188.com
基金项目: 广东省重点领域研发计划"新一代人工智能"重大科技专项(No.2021B0101400003)
Foundation Item: The Key Research and Development Program of Guangdong Province (No.2021B0101400003)