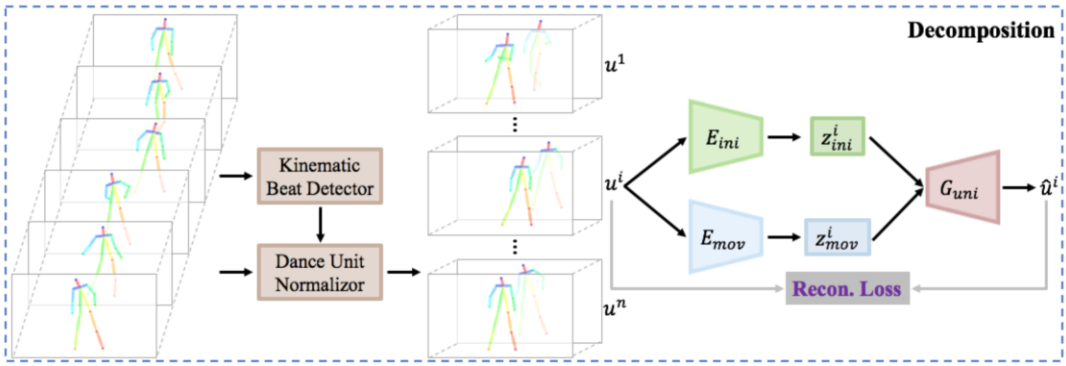
本文围绕着 music-to-dance 这一生成任务，以 google 今年 nips 的文章 Dance to Music 为主，结合其他相关的技术进行介绍。

利用一段音乐生成舞蹈主要有以下几个难点，1.舞蹈动作除了要逼真之外，还要和音乐的**风格**、**节拍**保持一致。2. 舞蹈动作是复杂的，某一时刻的姿态（pose）可能会跟随后续不同的动作，如果把一个舞蹈动作（movement）分解为初始时刻的姿态和后续动作，则需要后续动作即具有随机性，同时又可以和这个姿态自然衔接。3. 在具体设计模型时的困难在于，舞蹈是一个 long term、具有时空结构的 movement 的序列，同样面临文本生成中 exposure bias 的问题，而在连续空间的生成会进一步加剧。

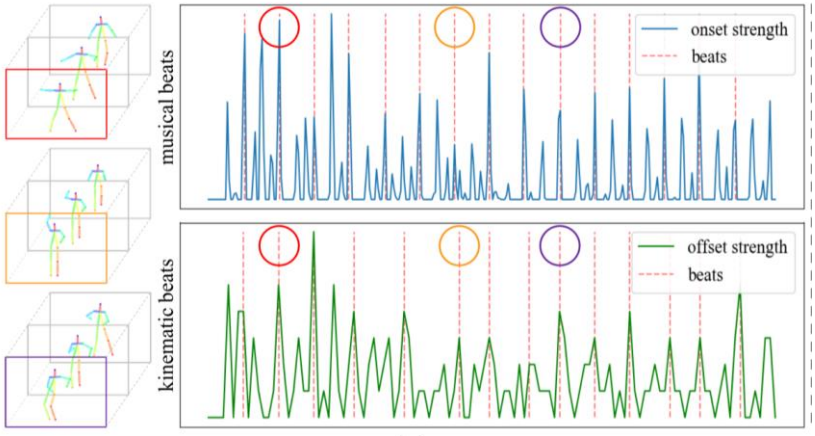
尽管利用完整音乐生成具有重复动作的、风格统一的舞蹈是一个有趣的课题，但是由于高质量数据的匮乏、以及 exposure bias 会随着序列的增长加剧，所以这篇文章作为开坑之作，简化了这个任务。文章提出了先分解、再组合的结构，将舞蹈切割成一些 movements，然后将 movements 作为生成的基本单位，然后再组装成连贯的舞蹈。

1. 分解阶段

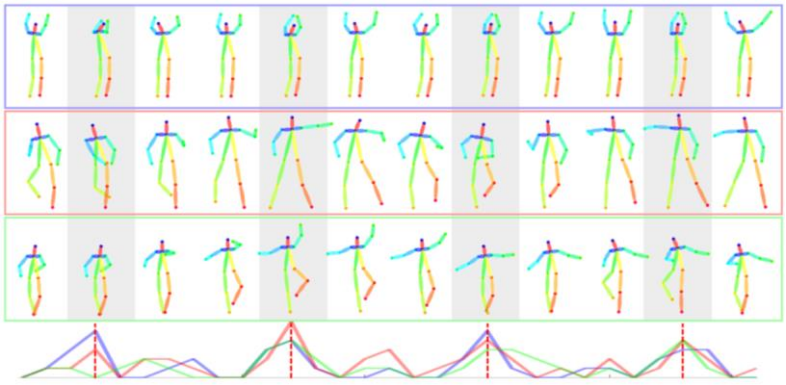


分解阶段分为两步，首先是将数据集中的舞蹈分割成 movement，然后学习一个模型来生成 movement。

对于音乐节拍进行分割的技术已经很成熟了（beat tracking），即根据所谓 onset strength 来判断一个小节音乐的起始位置。而对于舞蹈动作来说，运动节拍的分割与音频有所不同，文章提出用动作的突然减速来作为动作节拍的切分点。文章追踪了人体各个关键点的运动幅度和角度，检测幅度忽然减小或者角度有剧烈的变化的情况。此外，考虑舞蹈动作和音乐的相关性，常见的情况是舞蹈的节拍一定会和音乐节拍对齐，但不一定要符合每个音乐节拍。如下图所示：



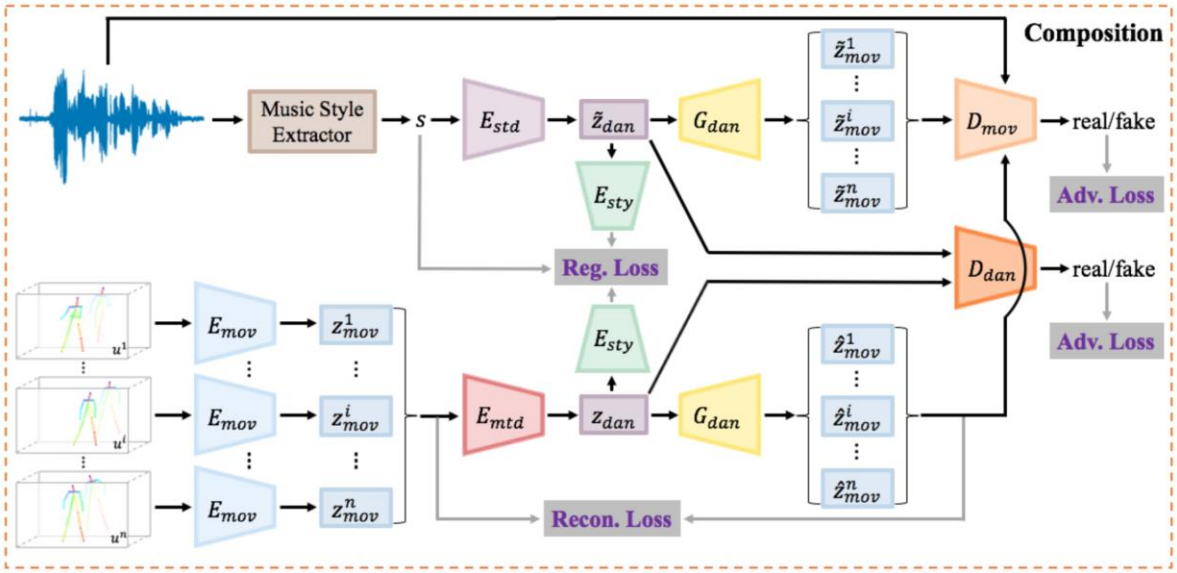
将舞蹈切割成小的节拍之后，在每个节拍等时间间隔的取某些时刻的姿态，这些姿态即为构成一个 movement 的数据，被储存的每个 movement 的姿态的数量相同：



下一步，需要学习一个生成模型来生成 movement，文章提出了一种 Dance Unit VAE 模型。具体来说，将 movement 分解到两个隐空间：初始姿态空间和动作空间，这样进行分解是为了方便进行 movement 序列化生成：上一个 movement 的最终姿态可以用来作为下一 movement 的起始姿态。该 VAE 包含上述两个 encoder 和一个 decoder，loss 为：

$$L^u_{\text{recon}} = \mathbb{E}[\|G_{uni}(z_{ini}, z_{mov}) - u\|_1],$$
$$L^u_{\text{KL}} = \mathbb{E}[\text{KL}(\mathcal{Z}_{ini}\|N(0,\text{I}))] + \mathbb{E}[\text{KL}(\mathcal{Z}_{mov}\|N(0,\text{I}))]$$

2. 组合阶段



文章提出 music-to-move GAN 来基于音乐生成 movement 的序列。数据集是 music 和 dance 的 pair，对于 dance，先通过上节训练的模型切分成 movement 序列，再压缩到 movement 隐空间，然后经过一个从该空间到 dance 隐空间的 encoder，得到 dance 的隐变量，然后用一个 generator 序列化的生成一组 movement；对于 music，文章主要利用的信息就是音乐的风格，收集的数据集包括三种类型的舞蹈：ballet、zumba、hip-hop，图示的音乐风格提取器实际上就是一个分类器，将得到的类别 label/概率向量（参考源码）加上一个噪音送入一个 encoder，也映射到 dance 隐空间（如果没有其他约束，最终该空间可能是一个简单的混合高斯分布），然后经过同样的 generator 产生 movement 序列。

首先注意到，比起让音乐生成的 movement 与真实的 movement 去一一 match（相当于 AE），只是用一个判别器去判断 real 还是 fake 是一个宽松得多的目标，但是为了保证生成的舞蹈和音乐还是匹配的，而不是对任意音乐都会生成同一组 real 的 movenment，

文章将对抗 loss 定义如下：

$$L_{\text{adv}}^{m} = \mathbb{E}[\log D_{mov}(\{\hat{z}_{mov}^{i}\}, a) + \log(1 - D_{mov}(\{\tilde{z}_{mov}^{i}\}, a))]$$

其中 a 是提取的音频特征，要判断的是在某个音乐下得到这样的舞蹈是不是 real 的。然后再看 dance 部分的 AE，如果去掉这个网络结构，将真实的 movement 放到 AE 生成的 movement 的位置也是可以训练的。但是这样设计有两个好处，一个是利用重建损失，增强 G_{dan} 的能力（z_{dan} 接近一个简单的分布，所以主要的信息量都压缩在生成器里）。

$$L_{\text{recon}}^{m} = \mathbb{E}[\|\{\hat{z}_{mov}^{i}\} - \{z_{mov}^{i}\}\|_1].$$

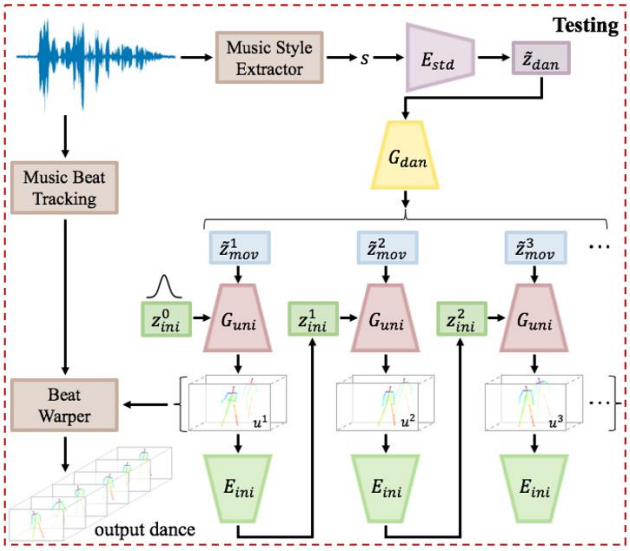
第二点是可以在 z_{dan} 空间中就可以约束来自 music 和 dance 的隐变量分布对齐：

$$L_{\text{adv}}^{d} = \mathbb{E}[\log D_{dan}(z_{dan}) + \log(1 - D_{dan}(\tilde{z}_{dan}))],$$
$$L_{\text{KL}}^{d} = \mathbb{E}[\text{KL}(\mathcal{Z}_{dan} \| N(0, \text{I}))].$$

最后再加上对风格一致性的约束，就构成了 MM-GAN 的训练目标：

$$L_{\text{recon}}^{s} = \mathbb{E}[\|E_{sty}(z_{dan}) - s\|_1 + \|E_{sty}(\tilde{z}_{dan}) - s\|_1].$$

3. 测试阶段



当训练完成后，输入一段音乐，我们可以按图示进行生成，这时我们将舞蹈节拍的时长设定为与音乐的节拍一致，同时利用提取到的音乐风格，自回归的生成舞蹈动作。