计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200900104

基于 CQT 和梅尔频谱的带有人声的音乐风格转换方法

叶洪良　朱皖宁　洪　蕾
金陵科技学院软件工程学院　南京 211100
(yehongliangiao@163.com)

摘　要　近年来,生成对抗网络在图像风格迁移领域中表现优秀,然而其在音乐领域表现一般。现有的音乐风格迁移对带有人声的音乐的风格迁移效果不佳。为了解决这些问题,首先提取音乐的 CQT 特征和梅尔频谱特征,然后采用 CycleGAN 对 CQT 特征和梅尔频谱的联合特征做风格迁移,再通过 WaveNet 声码器来对迁移后的谱图进行解码,最终实现了带有人声的音乐的风格迁移。在公开数据集 FMA 上对所提模型进行评估,符合要求的音乐的平均风格迁移率达到了 94.07%。与其他算法相比,该方法所产生的音乐的风格迁移率和音频质量都优于其他算法。

关键词:生成对抗网络;风格迁移;音乐处理;表征学习
中图法分类号　TP183

Music Style Transfer Method with Human Voice Based on CQT and Mel-spectrum

YE Hong-liang,ZHU Wan-ning and HONG Lei
School of Software Engineering,Jinling Institute of Technology,Nanjing 211100,China

Abstract　In recent years,the generative confrontation network has performed well in the field of image style transfer,but its performance in the field of music is average. The existing music style transfer has poor effect on the style transfer of music with human voice. In order to solve these problems,the CQT feature and Mel spectrum feature of the music are extracted,and then CycleGAN is used to transfer the style of the combined feature of CQT feature and Mel spectrum. Finally,the WaveNet vocoder is used to decode the migrated spectrum. Finally,we realize the style transfer of music with vocals. The proposed model is evaluated on the public data set FMA,and the average style transfer rate of music that meets the requirements reaches 94.07%. Compared with other algorithms,the style transfer rate and audio quality of the music produced by this method are better than other algorithms.

Keywords　Generative adversarial networks,Style transfer,Music processing,Representation learning

1　绪论

1.1　课题背景

过去 3 年来,神经风格迁移已经持续成长为了一个蓬勃发展的研究领域[1]。这一研究领域内越来越多的活动受到了科学挑战和工业需求的推动。风格迁移在社交、辅助用户创作和娱乐应用中都有着广阔的应用前景。

音乐风格迁移是风格迁移算法在另一个领域的尝试[2]。由于音乐是基于时间的片段并且音乐的成分较多,故提取特征较为复杂,特征之间的连接较为复杂紧密。目前学术界大多数的处理方法是将应用在图像风格迁移的算法直接应用在音乐风格迁移之上,并且大部分音乐都是乐器演奏的纯乐曲。但是这些算法在带有人声的音乐上取得的效果却不尽人意。当前大量的歌曲被翻唱为各种不同风格的版本,但是歌手的翻唱数量远远不能达到人们对于不同风格翻唱歌曲的需求,故研究一个适用于带有人声的流行音乐风格迁移的模型对计算机音乐领域具有重要意义。

1.2　国内外研究进展

Gatys 等[3]首次将神经网络用于图像风格迁移,并且表现出了优异的效果。在此之后,有学者尝试将生成对抗网络应用在图像风格迁移领域。随后,风格迁移领域逐渐在计算机视觉领域发展起来。这与此同时也促使了 GAN 的蓬勃发展。近几年来,CycleGAN[4],DualGAN[5],DiscoGAN[6] 等无监督学习的 GAN 逐渐被提出。通常,图片、音乐这种数据很难找到配对的数据,而以上这几种无监督算法的提出解决了这些问题。

目前学术界在音乐风格迁移上已做的工作主要在常见乐器的纯音乐音色风格迁移上。在进行处理的音乐可以分为原始音频和非原始音频。例如,Gino 等[7]通过将 midi 格式的音频转换成钢琴滚动矩阵,然后将矩阵输入 CycleGAN 进行训

基金项目:金陵科技学院高层次人才科研启动基金(jit-b-201624);江苏省大学生创新训练计划项目(202013573045Y);江苏高校哲学社会科学基金项目(2019SJA0485)
This work was supported by the Jinling Institute of Technology High-level Talent Research Startup Fund Support(jit-b-201624),Jiangsu Province University Student Innovation Training Program Project(202013573045Y) and Jiangsu University Philosophy and Social Science Foundation Project(2019SJA0485).
通信作者:朱皖宁(zhuwanning@jit.edu.cn)

练，最终生成转换后 midi 音频。该方法的一个显著的优点是计算开销较小，且转换纯乐器演奏的音乐的风格较为多变。但其缺点也很明显，首先其无法处理原始音频，其次通过此种方法进行风格迁移仅仅只能从弹奏的维度上进行。Huang 等[8]提出的 Timbretron 通过提取音频的 CQT 特征，然后通过 CycleGAN 对其进行音色转换，随后通过提前训练好声码器将转换后的 CQT 特征转换成原始的音频。但是由于其是在单一音色域上进行的风格变换，当歌声中出现突兀的声音时（例如人声），会大幅削减该声音的强度。Noam 等[9]提出了一个通用的音乐翻译网络，该网络通过训练一个 WaveNet 音乐编码器和多个 WaveNet 解码器来实现音乐音色转换。该网络可以实现从一种音色域转换到多种音色域，但是由于该算法需要为不同的风格训练多种解码器，机器需要承受巨大的计算开销，且在处理一种新的音乐时，需要训练一个新的解码器，模型缺乏通用性。上述算法在其各自研究的方向上展现出了不错的效果，但在笔者所阅读的文献之中，很少有对带有人声和背景音乐的原始音频做风格转换的研究，直接将这些算法强行嫁接到此问题上并没有产生能令人信服的结果。故本文通过提取音频的 CQT 特征和梅尔频谱，然后利用 CycleGAN 对特征进行风格转换，最后通过 WaveNet 解码器将特征转换为高质量的音乐，最终实现了带有人声的音乐的风格转换。

2 知识储备

2.1 CycleGAN

生成式对抗网络是一种深度学习模型，是 Gosodfellow 等[10]提出的一类隐式生成模型。模型通过框架中两个模块（生成模型和判别模型）的互相博弈学习产生高质量的输出。生成模型尝试生成假的样本来愚弄判别模型。而判别模型则尝试区分真实的数据和假的样本。假定 G 是生成器，D 是判别器，$P_{data}(x)$ 是真实样本的分布且 x 从该分布中采样，$P_z(z)$ 是 x 的潜在码 z 的分布。则目标方程为：

$$G,D=\min_G\max_D \mathbb{E}_{x\sim P_{data}(x)}[\log D(x)]+\mathbb{E}_{z\sim P_z(z)}[\log(1-D(G(z)))] \tag{1}$$

CycleGAN 是一种无监督的生成对抗网络，可以在没有任何成对数据的情况下学习两个域之间的映射。CycleGAN 包含两个生成器和两个判别器，两个生成器分别需要学习该域到对应域的映射。两个判别器则需要通过对各自域真实数据的学习，判断对应域生成器所生成的数据是否为本域数据。CycleGAN 的损失函数除了包含两个对抗性损失之外，还需要加上一个循环一致性损失，用来保留其输入结构，如式(2)所示：

$$\mathcal{L}_{cyc}(G,F)=\mathbb{E}_{x\sim P_{real}(x)}[\|F(G(x))-x\|_1]+\mathbb{E}_{y\sim P_{real}(y)}[\|F(G(y))-y\|_1] \tag{2}$$

其中，G 表示 CycleGAN 的前向转换，F 表示 CycleGAN 的后向转换。

2.2 时频分析

2.2.1 常数 Q 变换

在数学和信号处理中，常数 Q 变换将数据序列变换到频域。与快速傅里叶变换相比，该变换更适用于音乐数据。由于变换的输出实际上是幅度和相位相对于对数频率的信号，因此需要较少的频率区间才能有效地覆盖给定范围。该变换反映了人类的听觉系统，变换在较低频率下频谱分辨率更高，而在较高频率下时间分辨率更高。另外，在该变换中，音符的和声形成了乐器的音色的图案特征。假设每个谐波的相对强度相同，则随着基频的变化，这些谐波的相对位置将保持恒定。这可以使模型对音色的识别更加容易。

2.2.2 梅尔频谱

由于经傅里叶变换后得到的声谱图较大，为了得到合适大小的声音特征，通常将它通过梅尔尺度滤波器组，变为梅尔频谱。频率的单位是 Hz，人耳能听到的频率范围是 20 Hz～20 000 Hz，并且人耳对频率的感知不是线性关系，而是对低频率的声音敏感，对高频的声音不敏感。若将这种频率关系转换为梅尔频率，则人耳对于频率的感知就变为线性关系，这使得对特征进行提取和处理时更加容易。图 1 和图 2 分别展示了用 librosa 提取的 CQT 特征和梅尔频谱特征的谱图。

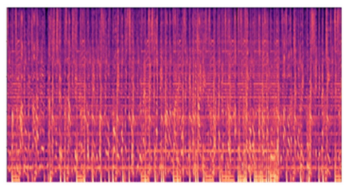


图 1 CQT 谱图

Fig. 1 CQT spectrum

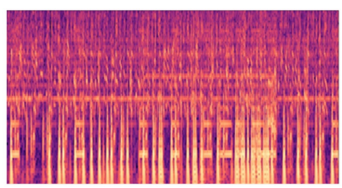


图 2 梅尔频谱

Fig. 2 Mel spectrum

2.3 WaveNet

WaveNet 由 Oord 等[11]提出，是一个用于生成高质量原始音频波形的自回归生成模型。该模型由具有残差连接和跳跃连接的带洞因果卷积层构成。WaveNet 采用了因果卷积和带洞卷积的思想，因果卷积保证了网络在预测的时候不会获取超前的信息，而带洞卷积可以使网络增加感受野，两者使产生的音频更加自然。

WaveNet 可以很容易地被修改为声码器。例如 Polyak 等[12]通过训练一个 WaveNet 自编码器和两个辅助网络，实现了人类语音的相互转换。Huang 等[8]用频谱图作为输入数据，训练了一个音乐解码器。该解码器通过输入频谱图可以使网络合成自然的、高质量的声音和音乐。

3 模型架构

本节介绍了模型的基本架构和模型的处理流程。模型基于 CycleGAN 和 WaveNet 解码器。模型处理流程如下：

1）提取出音频的梅尔频谱特征和 CQT 特征。

2）将提取的梅尔频谱特征和 CQT 特征合并为两层输入进 CycleGAN 模型，然后 CycleGAN 产生出风格迁移后的梅尔频谱特征和 CQT 特征。

3）将两层特征输入进提前训练好的 WaveNet 解码器产生音频。

模型结构图如图 3 所示。

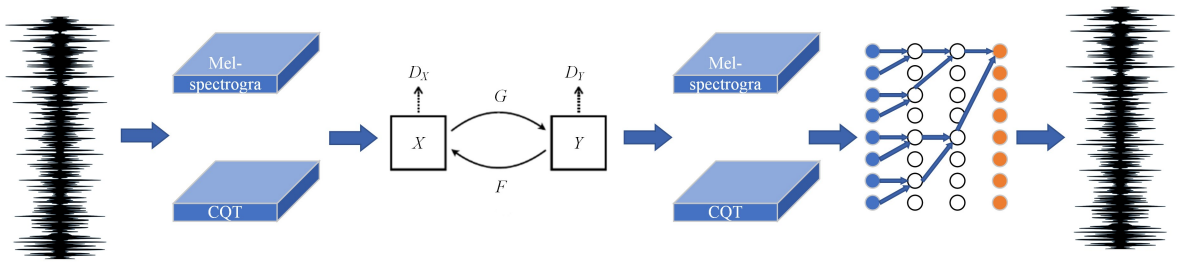


图 3 模型架构图

Fig. 3 Model architecture

3.1 问题定义以及数据预处理

给定音乐数据集 dataset，从 dataset 采样得到单个音乐样本。音乐样本表示为一个二维矩阵 $\boldsymbol{M}_i$，矩阵 $\boldsymbol{M}_i$ 的大小为 $m\times n$，其中 m 表示音乐片段在时间序列上的长度，n 表示音乐的通道数。对于训练集中的某一首音乐 $\boldsymbol{M}_i$，其中 $i\in\{1,2,3,\cdots,k\}$，k 表示训练集中音乐的数量。对于单个音乐序列 $\boldsymbol{M}_i$，对音乐序列在相同的窗口大小和步长进行变换得到谱图 $\boldsymbol{Q}$ 和梅尔频谱 $\boldsymbol{T}$，假设谱图 $\boldsymbol{Q}$ 的大小为 $i\times j$，谱图 $\boldsymbol{T}$ 的大小为 $i\times s$，那么矩阵 $\boldsymbol{T}$ 和矩阵 $\boldsymbol{Q}$ 可以表示为：

$$\boldsymbol{T}=\begin{bmatrix} t_1^1 & t_1^2 & \cdots & t_1^s \\ t_2^1 & t_2^2 & \cdots & t_2^s \\ \vdots & \vdots & \ddots & \vdots \\ t_i^1 & t_i^2 & \cdots & t_i^s \end{bmatrix} \tag{3}$$

$$\boldsymbol{Q}=\begin{bmatrix} q_1^1 & q_1^2 & \cdots & q_1^j \\ q_2^1 & q_2^2 & \cdots & q_2^j \\ \vdots & \vdots & \ddots & \vdots \\ q_i^1 & q_i^2 & \cdots & q_i^j \end{bmatrix} \tag{4}$$

因为在两种变换使用的窗口大小和步长相同，$j>s$，故矩阵 $\boldsymbol{T}$ 添加 $j-s$ 列向量，即：

$$\boldsymbol{Tp}=\begin{bmatrix} t_1^1 & t_1^2 & \cdots & t_1^s & 0_1^{s+1} & \cdots & 0_1^j \\ t_2^1 & t_2^2 & \cdots & t_2^s & 0_2^{s+1} & \cdots & 0_2^j \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ t_i^1 & t_i^2 & \cdots & t_i^s & 0_i^{s+1} & \cdots & 0_i^j \end{bmatrix} \tag{5}$$

将填补后的矩阵 $\boldsymbol{Tp}$ 和 $\boldsymbol{Q}$ 合并为两层，形成一个三维向量 $\boldsymbol{X}$，向量大小为 $i\times j\times 2$。由于 CycleGAN 是无监督算法，即没有相应的标签，则在实验中直接将两种风格域的向量 $\boldsymbol{X}$ 作为 CycleGAN 的输入和输出进行训练。

3.2 CycleGAN

3.2.1 损失函数

因为算法的目标是将音乐从一个域传输到另一个域，所以生成器实际上并没有将噪声作为输入，而是从源域获取真实的样本。在本文中，模型一次只处理两个域之间的翻译，因此将它们称为 $domain_X$ 和 $domain_Y$，这两个域对应于来自两种不同流派的音乐。由于传输应该是对称的，因此将样本从 $domain_X$ 传输到 $domain_Y$，同时从 $domain_Y$ 传输到 $domain_X$。CycleGAN 的 $X->Y$ 的基本损失函数如式(6)所示：

$$\mathcal{L}_{GAN}(G,D,X,Y)=\mathbb{E}_{x\sim P_{data}(x)}[\log D(x)]+\mathbb{E}_{y\sim P_{data}(y)}[\log(1-D(G(y)))] \tag{6}$$

除此之外，CycleGAN 还添加了 Zhu 等[4]提出的 identity loss，如式(7)所示。实验表明，当没有添加该损失时，生成的谱图会丢失其颜色成分，其表现为使最后生成的音频产生较大的杂音，故添加该损失有利于模型生成质量更高的谱图。

$$\mathcal{L}_{identity}(G,F)=\mathbb{E}_{y\sim P_{data}(y)}[\|G(y)-y\|_1]+\mathbb{E}_{x\sim P_{data}(x)}[\|F(x)-x\|_1] \tag{7}$$

添加式(2)表示的 cycleconsistency loss 和 identity loss，其中 λ_1 和 λ_2 分别表示 cycleconsistency loss 和 identity loss 所占的权重。则总的损失函数为：

$$\mathcal{L}(G,F,D_X,D_Y)=\mathcal{L}_{GAN}(G,D_Y,X,Y)+\mathcal{L}_{GAN}(F,D_X,Y,X)+\lambda_1\mathcal{L}_{cyc}(G,F)+\lambda_2\mathcal{L}_{identity}(G,F) \tag{8}$$

由于 cycleconsistency loss 和 identity loss 的权重 λ_1 和 λ_2 会显著影响生成的音频。就 λ_1 来说，当 λ_1 过大时，CycleGAN 会选择简单、低纬度的变换；而当 λ_1 过小时，CycleGAN 会寻求复杂度较高的变换，生成的结果的变换样式多样但不易控制。就 λ_2 来说，当 λ_2 过大时，CycleGAN 会对原风格造成较强的约束，从而会使生成的音频失去表现力；而当 λ_2 过小时，CycleGAN 会丢失谱图的颜色成分。由于随意变化 λ_1 的值会使算法无法正常收敛，故本文在实验时将 λ_1 设为一个固定的值。而对于 λ_2，研究表明随着算法迭代的次数增多，λ_2 的值应该减小。通常 λ_2 的衰减都为线性衰减，但是本文通过实验发现，λ_2 的曲线衰减比较符合音乐的风格转换算法。故本文就 λ_2 尝试提出一种非线性衰减函数，相比线性衰减函数，本文提出的非线性衰减函数使模型展现出了更好的鲁棒性。假设算法一共要迭代 t 步，那么在第 n 步，λ_2 为：

$$\lambda_2=\frac{\ln t+\sqrt{t}}{n+\sqrt{t}} \tag{9}$$

3.2.2 棋盘伪影

由于 CycleGAN 采用的是反卷积的操作，这会导致生成的谱图存在严重的棋盘伪影，对音频的影响表现为严重的间接性噪声(如图 4 红框标识处)。

为此本研究参考了 Odena 等[13]的工作，采用最近邻插值和正则卷积代替反卷积。文献[13]介绍了两种方法，第一种方法是确保使用的内核大小除以步幅，以免出现重叠问题。但是尽管该技术最近在图像超分辨率方面获得了成功，但反卷积仍然很容易出现带有棋盘伪影的结果。另一种方法是从卷积到计算特征中分离出较高分辨率的上采样。首先通过最近邻插值法调整输入大小，然后进入卷积层。与第一种方法相比，该方法不仅在图像超分辨率任务中能很好地工作，还能使生成的谱图不易出现棋盘伪影的结果(见图 5)。

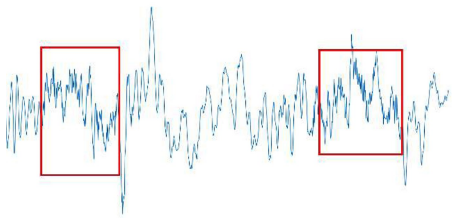


图 4 采用反卷积产生的音频(电子版为彩色)

Fig. 4 Audio generated by deconvolution

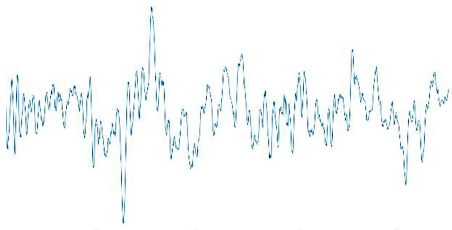


图 5 采用最近邻插值产生的音频

Fig. 5 Audio generated using nearest neighbor interpolation

3.3 WaveNet Decoder

3.3.1 训练数据预处理

由于 WaveNet 损失函数为 tanh 激活函数，而 tanh 函数所表示的波形处于[−1,1]之间，而通过音频生成的 CQT 谱图和梅尔频谱取自然对数的值符合 [−6,2]之间的正态分布。故在训练之前需要对输入数据进行全局标准化，使输入数据的分布符合 $N(0,1)$分布。

参考 Engel 等[14]的研究，在时频表示中很难由谱图直接预测出相位，因此需要丢弃梅尔频谱和 CQT 谱图的相位信息，然后对两层相位信息进行变形和补零填充操作，以便两层合并。

3.3.2 网络结构

本文的 WaveNet 解码器的网络结构和 WaveNet 原论文保持同样的结构，但是将其输入改为频谱图，网络扩张率为 2^k(k 表示网络处于第几层)。对于所有的带洞卷积和因果卷积层，都使用了大小为 3 的卷积核。对于所有的残差块来说，远跳连接和残差连接的长度都为 256。除此之外，每个残差层都包含一个 ReLu 非线性函数。

4 实验分析

由于音乐质量的好坏是一个高度主观的衡量标准，评价一个音乐生成系统的表现是困难的。评估风格则比较容易，因为可以根据已有的训练集训练一个分类器，然后根据分类器的结果来确定风格迁移的转换率。综合以上两点，本文采取了以下方案：

(1)由人们主观评价音乐品质是否合格，如是否有较大的杂音。

(2)根据已有的音乐数据训练一个分类器，计算音乐域的成功迁移率。

4.1 实验设备

4.1.1 硬件配置

实验配置为：CPU 为 Intel(R) Xeon(R) E5-2660 v4 @ 2.00GHz；GPU 为 NVIDIA GV100GL [Tesla V100 PCIe 32 GB]；GPU 大小为 32GB；内存为 48GB。

4.1.2 软件配置

操作系统为 64 位 Ubuntu 16.04.6 LTS，cuda 版本为 cuda10，python 版本为 3.6.4，深度学习框架 tensorflow-gpu 1.8.0，keras 2.2.4，音乐处理框架有 librosa0.72 等。

4.2 训练集

实验选择的数据集为 FMA[15]，它是一个开放且易于访问的数据集，适用音乐分类等研究分析。实验选择 FMA 训练集是 medium 版本。该版本一共包含了 25 000 个音乐片段，每个音乐片段的长度为 30 s，采样率为 22 050。训练模型所采用的音乐包含如下 6 种音乐流派：Pop，Blues，Folk，Jazz，Country，Classical。由于网络大小的限制，实验将一个 30s 的片段分成 6 个大约 5s 的片段输入进 CycleGAN 进行风格迁移。通过 librosa 提取 CQT 特征和梅尔频谱特征。

4.3 分类器训练

根据 Wu 等[16]在音乐流派分类的工作，本文在 FMA 数据集上训练了 DW-KNN 音乐流派分类器，其中所包含的音乐域及各种预测音乐流派的准确率如图 6 所示。

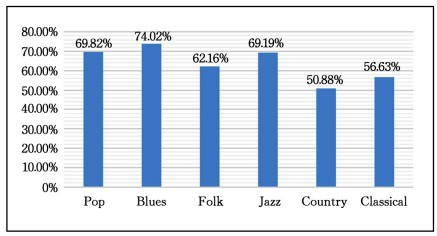


图 6 DW-KNN 的音乐分类准确率

Fig. 6 DW-KNN's music classification accuracy rate

4.4 评判标准

由于音乐流派种类多、边界模糊，且音乐具有时间特性，一首歌曲可能包含多个不同风格的片段。为了更好地评判此风格迁移算法的效果，我们摒弃了人为主观性。迁移后的音乐风格按照分类器的分类结果标注。在实验的多组数据中，选择了 6 组(见后文表 1、表 2 的第一列)和其逆过程进行评价。评价的指标有 AudioQualityRate(AQR)，Transfer Rate (TR)。

首先在每个转换后的音乐域的样本里随机抽取 30 首音乐片段，其中 AQR 由 5 位人员对音频的平均打分值决定。打分人员首先对每个域转换后的音乐样本进行甄别，然后对音频进行打分，分数区间为 0～5 分，打分值越接近 5，则表示打分人认为该音乐的品质越高。TR 表示成功进行风格迁移的音乐占总风格迁移音乐的比例。

4.5 超参数的确定

4.5.1 λ_1 值的确定

在实验开始时，实验将 λ_1 设为一个固定的值，来证明算法 λ_2 采用本文提出的衰减方案优于 Huang 等[8]提出的线性衰减方案。按照之前设定的评判标准，分别对两组方案进行实验，并将实验结果用于接下来超参数的一系列选值。

由表 1 和表 2 可知，在音频质量相差不大的情况下，CycleGAN 的风格转换率(TR)有了显著的提高，其中前向风格转换率(TR)平均提高了 2.52%，后向转换率(TR)平均提高了 1.85%。故 λ_2 采取本文提出的非线性衰减的方案相比线性衰减来看表现出更好的风格迁移能力。在某种程度上，算法的平均风格迁移率达到了 94.07%，表现出了不错的表征

学习能力，实现了对以上 6 种音乐流派的风格转换。除此之外，由于 FMA 包含大量的带有人声的音乐，并且算法所产生的平均音乐质量(AQR)达到了 85.44%，故在带有人声的音乐的处理上，本文算法也表现出了良好的效果，具有一定的鲁棒性。

表 1 λ_2 采用线性衰减所产生的音频质量合格率和风格迁移率

Table 1 AQT(mean±SD), TR scores for music style transfer task with lineardecline λ_2

(单位：%)

Origin domain−>Object domain	Forward transfer		Backward transfer	
	AQR	*TR*	*AQR*	*TR*
Pop−>Classical	86.23±5.65	93.66	80.17±4.64	92.34
Pop−>Blues	87.16±4.82	96.74	75.91±3.31	94.75
Blue−>Country	82.38±5.23	93.79	79.44±8.07	91.53
Folk−>Jazz	88.69±5.50	89.45	78.08±8.20	88.03
Jazz−>Classical	79.79±6.95	91.71	75.48±8.39	89.14
Pop−>Folk	84.23±4.68	94.42	80.77±6.12	91.41

表 2 λ_2 采用非线性衰减所产生的音频质量合格率和风格迁移率

Table 2 AQT(mean±SD), TR scores for music style transfer task with non-linear decline λ_2

(单位：%)

Origin domain−>Object domain	Forward transfer		Backward transfer	
	AQR	*TR*	*AQR*	*TR*
Pop−>Classical	85.72±6.78	96.42	82.27±5.34	95.12
Pop−>Blues	89.42±5.32	97.80	79.46±4.44	93.95
Blue−>Country	86.91±7.43	95.59	79.71±6.87	94.32
Folk−>Jazz	82.48±8.22	92.55	80.22±7.20	89.52
Jazz−>Classical	81.79±4.35	95.30	78.42±5.25	91.74
Pop−>Folk	86.32±6.62	94.73	81.64±6.86	91.82

4.5.2 迭代次数的确定

每经过 200 次迭代，记录一次损失函数的值，其中包括判别器的损失和生成器的损失，将判别器的损失值和生成器的损失值的波形绘制到一张图上，如图 7 所示。可以看出当算法迭代到 125000 步左右，生成器和判别器的损失都收敛到了一个局部的最小值。并且生成音乐风格和目标音乐风格基本接近，算法趋于稳定。由于计算限制，本文选择 125000 作为算法的迭代次数，并在此基础上探讨其他超参数的选值问题。

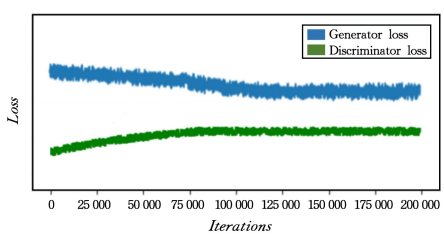


图 7 正向模型 loss 曲线

Fig. 7 Forward model loss curve

4.5.3 λ_2 值的确定

根据上述实验结果，最优的迭代次数大概是在 125000 次左右且 λ_2 的损失在非线性衰减的情况下要优于非线性衰减。故在本次实验中，λ_2 的衰减曲线大致为：

$$\lambda_2=\frac{\ln(125000)+\sqrt{125000}}{n+\sqrt{125000}} \tag{10}$$

故在此曲线约束之下，实验探讨 λ_1 的取值。在 λ_1 的设置上，首先在一个较大的尺度上进行实验，假设 $\lambda_1\in\{1,10,30,50,70,100\}$。在其他网络参数恒定的情况下进行训练，然后计算出平均的 AQR 和 TR，结果如表 3 所列。

表 3 λ_1 的取值对算法的影响($\lambda_1\in\{1,10,30,50,70,100\}$)

Table 3 Influence of value of λ_1 on algorithm($\lambda_1\in\{1,10,30,50,70,100\}$)

(单位：%)

Score	$\lambda_1=1$	$\lambda_1=10$	$\lambda_1=30$	$\lambda_1=50$	$\lambda_1=70$	$\lambda_1=100$
AQR	87.43	83.69	77.43	72.43	69.49	68.54
TR	90.45	94.68	95.03	95.22	95.02	95.43
AQR * *TR*	79.08	79.23	73.58	68.96	66.02	65.40

由表 3 可知，当 $\lambda_1=10$ 时，在以上 λ_1 所给范围内的音频质量和风格迁移效果最好。故本文在区间(5,15)继续探索 λ_1 的取值问题。故假设 $\lambda_1\in\{6,7,8,9,11,12,13,14\}$。在网络参数恒定的情况下进行训练，结果如表 4 所列。

表 4 λ_1 的取值对算法的影响($\lambda_1\in\{6,7,8,9,11,12,13,14\}$)

Table 4 Influence of value of λ_1 on algorithm($\lambda_1\in\{6,7,8,9,11,12,13,14\}$)

(单位：%)

Score	$\lambda_1=6$	$\lambda_1=7$	$\lambda_1=8$	$\lambda_1=9$	$\lambda_1=11$	$\lambda_1=12$	$\lambda_1=13$	$\lambda_1=14$
AQR	86.34	85.44	83.72	82.53	82.75	80.44	81.42	79.34
TR	92.12	94.07	94.34	94.64	94.69	94.22	94.83	94.66
AQT * *TR*	79.53	80.37	78.98	78.10	78.35	75.79	77.21	75.10

由表 4 可知，当 $\lambda_1=7$ 时，音频质量和风格迁移的效果都比较不错。当 λ_1 的取值比较大时，虽然风格迁移率有着一定的提升，但是音频质量在显著下降。由于两个指标都有着重要的意义，故计算出二者的乘积来综合评判。由表 4 可以看出，当 $\lambda_1=7$ 时，风格迁移率和音频质量的乘积最大。对此实验中 $\lambda_1=7$ 是一个较优的取值。

4.6 算法对比

本文选取了另外两个音乐风格迁移算法进行对比，分别是 Huang 等[8]在音乐风格转换的工作和 Noam 等[9]的工作。由于 Noam 等[9]的算法不具备双向对称性，故本文模型和 Huang 等的算法都采取前向转换来参与比较。

表 5 3 种算法的效果对比

Table 5 Comparison of effects of three algorithms

(单位：%)

Method	*AQT*	*TR*
Huang's	68.93	87.62
Noam's	72.34	82.91
Our's	85.44	94.07

由表 5 可以看出算法无论是在音频的生成质量还是风格迁移的效果上都比另外两种算法更加优秀，因为 Huang 等的算法仅仅提取了音频的 CQT 特征，较少地保留了人声和音乐的因素；而 Noam 等的算法仅仅在纯乐器上表现优秀，但是当输入带有人声的音乐时，音乐产生了较严重的变形。而对比上面两种算法，本文的算法提取了 CQT 和梅尔频谱两种声学特征，并且用 WaveNet 代替其他声码器，保留了较多的声学约束，并且两种特征共同对波形进行了约束，故所提算法能够在生成的音频质量和风格迁移效果上优于另外两种算法。

结束语 音乐风格迁移在近年来已经初步形成为一个新的研究领域。这一研究领域内越来越多的活动受到了科学挑战和工业需求的推动。音乐风格迁移在包括社交、辅助用户创作和娱乐应用都有着广阔的应用前景。本文先提取音乐的

(下转第 363 页)

Learning to Spot Artifacts[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 2733-2742.

[30] GIDARIS S, SINGH P, KOMODAKIS N. Unsupervised Representation learning by predicting image rotations[J]. arXiv: 1803.07728, 2018.

[31] PATHAK D, KRAHENBUHL P, DONAHUE J, et al. Context encoders: Feature learning by inpainting[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 2536-2544.

[32] GIDARIS S, SINGH P, KOMODAKIS N. Unsupervised representation learning by predicting image rotations[J]. arXiv: 1803.07728, 2018.

WU Lan, born in 1981, Ph.D, professor, master tutor. Her main research interests include information security, artificial intelligence, multisensor networked information fusion theory, fault diagnosis of complex systems and devices, intelligent information processing.

（上接第 330 页）

CQT 特征和梅尔频谱特征，然后采用 CycleGAN 对 CQT 特征和梅尔频谱特征映射的图片做风格迁移，最后通过 WaveNet 解码器生成音乐波形，最终实现了带有人声的音乐的风格迁移。在实验训练的分类器上，符合要求的音乐的平均风格迁移率达到了 94.07%。本文提取以上两种特征，既保留了音乐的特征，也保留了较多的人声的特征。与此同时，这两种特征的重叠部分共同约束了波形，使得 WaveNet 对波形的预测更加准确。

在以后的工作里，可以考虑如何将巨量的特征数据输入网络和如何保留前后音乐片段之间的关联性。除此之外，如何提高风格转换后音乐的质量也是需要着重解决的问题。

参 考 文 献

[1] JING Y, YANG Y, FENG Z, et al. Neural style transfer: A review[J]. IEEE Transactions on Visualization and Computer Graphics, 2019, 26(11): 3365-3385.

[2] DAI S, ZHANG Z, XIA G G. Music style transfer: A position paper[J]. arXiv: 1803.06841, 2018.

[3] GATYS L A, ECKER A S, BETHGE M. A neural algorithm of artistic style[J]. arXiv: 1508.06576, 2015.

[4] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 2223-2232.

[5] YI Z, ZHANG H, TAN P, et al. Dualgan: Unsupervised dual learning for image-to-image translation[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 2849-2857.

[6] KIM T, CHA M, KIM H, et al. Learning to discover cross-domain relations with generative adversarial networks[J]. arXiv: 1703.05192, 2017.

[7] BRUNNER G, WANG Y, WATTENHOFER R, et al. Symbolic music genre transfer with cyclegan[C]// 2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, 2018: 786-793.

[8] HUANG S, LI Q, ANIL C, et al. Timbretron: A wavenet (cyclegan (cqt (audio))) pipeline for musical timbre transfer[J]. arXiv: 1811.09620, 2018.

[9] MOR N, WOLF L, POLYAK A, et al. A universal music translation network[J]. arXiv: 1805.07848, 2018.

[10] GOSODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]// Advances in Neural Information Processing Systems. 2014: 2672-2680.

[11] OORD A, DIELEMAN S, ZEN H, et al. Wavenet: A generative model for raw audio[J]. arXiv: 1609.03499, 2016.

[12] POLYAK A, WOLF L. Attention-based wavenet autoencoder for universal voice conversion[C]// ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 6800-6804.

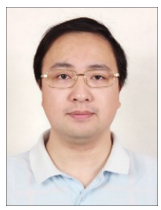
[13] ODENA A, DUMOULIN V, OLAH C. Deconvolution and checkerboard artifacts[J]. Distill, 2016, 1(10): e3.

[14] ENGEL J, RESNICK C, ROBERTS A, et al. Neural audio synthesis of musical notes with wavenet autoencoders[C]// International Conference on Machine Learning. PMLR, 2017: 1068-1077.

[15] DEFFERRARD M, BENZI K, VANDERGHEYNST P, et al. Fma: A dataset for music analysis[J]. arXiv: 1612.01840, 2016.

[16] WU M, LIU X. A Double Weighted KNN Algorithm and Its Application in the Music Genre Classification[C]// 2019 6th International Conference on Dependable Systems and Their Applications (DSA). IEEE, 2020: 335-340.

YE Hong-liang, born in 1999. His main research interests include deep learning and music processing.



ZHU Wan-ning, born in 1983, Ph.D. His main research interests include quantum information technology and quantum computing.