参赛队员姓名：刘衍东

中学：北京市第一 0 一中学

省份：北京

国家/地区：中国

指导教师姓名：董豪，周宇辰

指导教师单位：北京大学,北京市第一 0 一中学

论文题目：基于深度神经网络的 MP3 有损压缩还原算法研究

摘要：

音频文件经过 MP3 有损压缩后会造成音质损失，从而使听觉感受下降，无法满足广泛的高质量音乐播放需求。针对这个问题，通过分析音乐中人声和不同种类的乐器在高频部分和低频部分的特征及相关性，经过长期探索,本文提出了一种新的基于 CNN 和 GAN 的时域和频域带宽扩展方法,支持高质量的 MP3 有损压缩音乐复原。对于频域的带宽扩展，我们使用了类似图片补全的方法；对于时域的带宽扩展，使用了超分辨率的方法。我们将所提出的方法与 RNN、BPCNN 等方法进行比较实验。实验结果证明本文所提出的方法具有最低的频谱损失和最好的还原效果。人耳区分实验结果进一步证明了该音频增强算法的有效性。

关键词：MP3 有损音乐压缩，超分辨率，带宽扩展，CNN，GAN

Deep Neural Network Based Recovery of MP3 Lossy Compressed Music

Yandong Liu
Beijing No. 101 Middle School

Abstract

The lossy compression of music audio files through MP3 cause loss of sound quality, which results in the decline of auditory experience and cannot meet the requirements of high-quality music playback in a wide variety of occasions. To solve this problem, after long-term exploration, we proposes an approach of time-domain and frequency-domain bandwidth expansion based on CNN and GAN to support high-quality MP3 lossy compressed music recovery, by analyzing the characteristics and correlation of human voice, and different kinds of musical instruments in high-frequency part and low-frequency part in music,, For bandwidth expansion in the frequency domain, a method similar to image inpainting is designed; for bandwidth expansion in the time domain, a super-resolution method is designed. We compare the proposed method with RNN, BPCNN, and other methods. The experimental results prove that the method proposed in this paper has the lowest spectral loss and the best reconstruction quality. The experimental results of human ear discrimination further prove the effectiveness of the audio enhancement algorithm.

Keywords: MP3 lossy music compression, Super-resolution, Bandwidth extension, CNN, GAN

目录

1. 引言

当前音频数据大多以数字方式进行存储。人们希望获得高质量无损音频，但无损音频占用空间大，存储困难，很难通过网络进行实时传输。人们不得不对其先进行压缩，以节省存储空间和增加传输速度。因此，一种音质好、占用空间小的音频压缩格式显得尤为重要。现有常见的无损压缩格式有 FLAC、APE 等，不过压缩效果均不理想，压缩率只能到达 60%左右。而有损压缩格式如 MPEG-1 Layer3(MP3)等虽然能够将音频信号压缩至其原有大小的 20%，比有损压缩节省了更多的存储空间，但会损失大量高频信息，使音频听觉效果下降。

针对这个问题，本文提出了一种创新性的基于深度神经网络带宽扩展方法支持高质量的音乐复原方法，将有带宽限制的音频信号增强至 CD 质量（44.1kHz 采样率），以还原原始数据的听觉感受。通过分析音乐中人声和不同种类乐器在高频和低频的特征和相关性，经过长期探索，我们设计了一套基于 CNN 和 GAN 方法，从频域和时域进行带宽扩展。对于频域的带宽扩展，我们使用了类似图片补全的方法；对于时域的带宽扩展，我们使用了超分辨率的方法。我们将所提出的方法与 RNN、BPCNN、SRCNN 等方法进行比较。实验证明本文所提出的方法具有最低的频谱损失和最好的还原效果。人耳区分实验结果进一步证明了音频增强算法的有效性。

2. 研究背景

2.1 音乐的特点

人耳对声波频率的感受不是线性的，而是以 2 为底的对数关系。声音频率每提高一倍，人耳会认为音调升高了一个八度。恒定 Q 变换(CQT)是一种生成以对数方式划分频段的频谱的方法，能够将音频以更接近人耳感受的方式呈现出来。但 CQT 在高频区域频段数量较少，从而降低了高频部分的分辨率，限制了还原的频谱质量。短时傅里叶变换(STFT)的频段是线性分布的，其逆变换能够快速地生成高质量的音频信号，因此我们使用 STFT 来构建音乐的频谱。

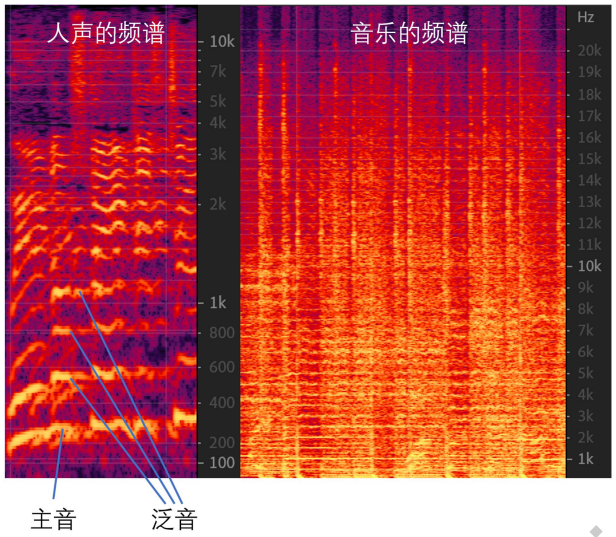
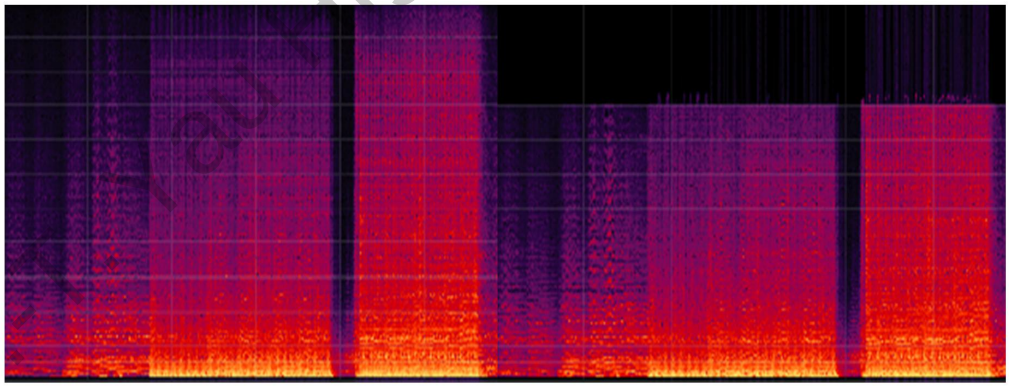


图1 人声和音乐的频谱图

音乐是由各种乐器、人声或者乐器和人声混合而成的，如图 1 所示。这些声音都是由主音（频率最低）和数个泛音（频率更高）构成的。对于某个乐器发出的声音或人声而言，频谱中所有的泛音形状都和主音形状非常相似，因此可以通过主音和频率较低的泛音来推测出频率较高的泛音。在 MP3 压缩之前，所有主音和泛音都存在，压缩之后，所有主音和部分泛音被保留，频率较高的泛音丢失，如图 2 所示，因此可以根据压缩后所保留的信息识别音乐的特征来还原丢失的高频部分。



压缩前　　压缩后

图2 MP3压缩前后的频谱信息对照

音乐信号有和声部分和非和声部分。和声部分包括各种非打击乐器，比如管弦乐器、电子合成器。非和声部分包括打击乐器，比如鼓或一些声音效果。人声是音乐中比较特殊的一部分，因为人声在较低频率处是和声的，在较高频率处是

非和声的。另外，不同的乐器有着不同的高频延伸范围，当高频延伸超过 MP3 压缩的截止频率时，这些乐器的高频泛音就会丢失，因此在还原高频信息时，应该补全这些乐器的高频泛音，而对高频延伸较低、泛音未丢失的乐器声音不予处理。

2.2 人耳听觉原理及特点

人耳的耳蜗中有很多共振频率不同的纤毛，它们与传入的声波进行共振。声波传入耳蜗后，对应频率的纤毛受到刺激，人耳就会认为声波中有这些频率成分。负责接收低频信号的纤毛可以感受自己的震动方向，因此人耳可以识别低频声音的相位并据此判断声音的来向。当声音频率升高时，耳蜗里的纤毛不再能感受自己震动的方向，即人耳对高频的相位不敏感，因此高频声音出现时，人耳仅能根据双耳声音强度的差别来判断声音的来向。综上，低频部分的相位对听觉感受有着巨大的影响，而高频部分的相位没有明显的作用，所以在 MP3 无损还原时仅需还原振幅谱，可以忽略相位谱，再用 Griffin-lim 算法[12]估算高频部分的相位即可。Griffin-lim 算法如下：

$$X^{[m+1]} = P_C(P_A(X^{[m]}))$$

$$P_c(X) = gg^{\dagger}X$$

$$P_A(x) = A \odot X \oslash |X|$$

由于人耳对高频相位的变化不敏感，当我们使用 Griffin-lim 算法来估算高频部分的相位时，人耳无法感知相位变化带来的波形变化，因此我们就可以不处理相位信息，从而降低神经网络的拟合任务量。

2.3 相关研究工作

目前补全 MP3 音乐高频损失的算法有非机器学习法和机器学习法。非机器学习法有“刺激”效果器和 Somesh Ganesh 研发的全波形矫正法(FWR)[3]。这些方法都通过将低频的部分信息复制到缺失的高频部分来补全丢失的高频信息，只能较好地处理某些乐器独奏的声音，不能有效地处理复杂的音乐，特别是交响乐和一些加有滤波器效果的电子音乐。机器学习法有 Marius Miron 发表的 CNN 方法

[1]、Binqiang Si 发表的金字塔小波卷积神经网络（PWCNN）方法[2]、Sung Kim 和 Visvesh Sathe 的 MU-GAN 方法[11]等，在处理频谱复杂的音乐上有所提升。Marius 的 CNN 和 Sung & Visvesh 的 MU-GAN 的主要问题是生成的高频部分频谱过于模糊，没有泛音应有的条纹，这是由于普通的损失函数无法在神经网络输出接近正确结果时给出有效的指导。PWCNN 的主要问题是生成的高频频谱基本上是对低频的部分频谱的对称复制，并没有起到真正的还原作用，原因是其网络层数较少，虽然有着高达 512 的通道数，但仍然无法学习到足够的特征图以应对各种不同乐器的泛音特点。其他利用机器学习来实现高频还原的研究主要针对人声增强，如 Yunpeng Li 等的 SEANnet[14],不适用于处理频谱更加复杂的音乐。

机器学习在图片超分辨率领域的应用已经十分成熟。图片超分辨率任务中最常用的卷积神经网络（CNN）能够高效地提取信号的特征，被证实十分适合对各种信号进行分辨率提升[6,17,18,19,20]。音频波形可以看作是二维图片的一维展开，那么 CNN 可以像处理图片一样处理音频。。

3.基于神经网络的有损压缩音乐复原

3.1 超分卷积网络结构(SRCNN)

音频信号是连续而有规律的，其中低频部分表现了波形的大致趋势，高频部分表现了波形的微小细节。在 MP3 压缩过程中，波形的大致趋势被保留，微小的细节被抛弃以节省空间。微小细节的丢失可以被看作是音频的采样率降低。

使用卷积-反卷积的图片超分辨率方法可以补全低分辨率的二维图片的微小细节。因此我们将这种方法从二维改为一维，应用到音频超分辨率即可补全音频波形的微小细节，从而还原有损压缩丢失的高频信息。

卷积的操作通过降采样和提升通道数来学习音频波形的特征，反卷积则利用这些特征重建高分辨率的波形，以实现对损失的高频信息的预测。此网络输入为将连续 128 个无损音频信号降采样一倍得到的 64 个音频信号，以模拟有损压缩的极端情况。输出为降采样之前的连续 128 个无损音频信号，即该网络可把音频信号的分辨率提升 1 倍，将网络输出的波形重采样至原来的采样率即可得到频谱完整的音乐。如图 3 所示，此网络由四个部分组成：

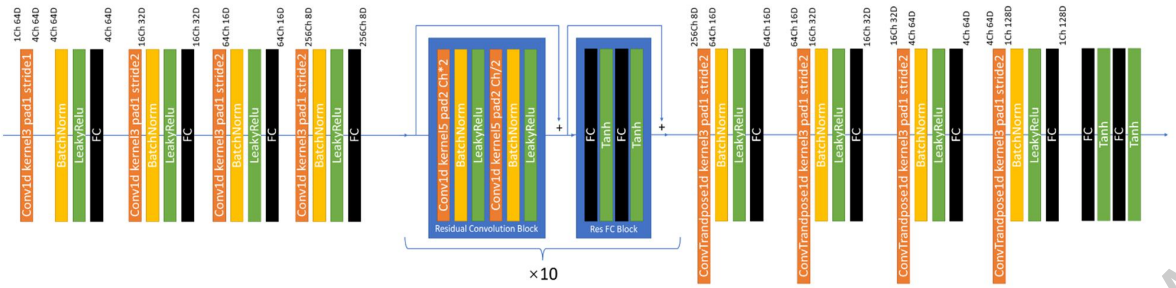


图3 SRCNN网络结构图

- 第一部分为一个卷积编码器。此部分使用卷积层来将音频信号降采样并提取特征。第一个卷积层步长为 1，其他的卷积层步长均为 2，每经过一个步长为 2 的卷积层，波形长度降低 1 倍，通道数提升 4 倍。

- 第二部分是残差卷积块结构和残差全连接块结构，可以防止梯度爆炸和梯度消失。残差卷积块通过将通道数先乘 2 再除 2 进行特征提取和特征选择。残差全连接块使用 tanh 激活函数，可以将数值变得更符合音频信号的范围，因为音频波形在预处理时被标准化到了(-0.5,0.5)的范围，而 Leaky ReLU 激活函数会把大部分数值限制在正数范围。

- 第三部分是一个解码器，它使用与前部对应的反卷积结构来利用学习到的特征生成高分辨率的音频信号。每个反卷积层步长均为 2，通道数降低 4 倍。卷积层之间设有全连接层，可以消除 0 填充带来的伪影并增强网络的拟合能力，使其能应对更复杂的音频信号。

- 第四部分的全连接层起到消除 0 补全带来边缘伪影的作用。

SRCNN 设立了 4 层卷积/反卷积结构，能够逐级提取音频信号的特征，从而提高了卷积核的利用效率。经过 4 层卷积后，音频信号的长度已经下降至 8 帧，此时如果继续卷积降采样并提升通道数，卷积时的边缘填充会带来高达 20%的冗余数据，对信号造成过大的影响，而且过多的通道数会给后续的特征选择带来巨大的运算量，使网络难以实时运行，得不偿失。如果卷积层小于 4 层，则会欠拟合，出现类似 PWCNN 的频谱对称复制问题。

SRCNN 的卷积残差块能利用通道数的变化进行特征提取和选择，其拟合能力优于通道数不变的普通卷积残差块。卷积层之间的全连接层和网络尾部的全连接

层起到了消除边缘填充伪影的作用，这也是 4.实验分析部分的对比实验中其他网络所不具备的。

3.2 基于对抗学习的频谱细节增强(SRGAN)

普通的损失函数在神经网络输出结果接近正确值时 loss 非常小，无法提供有效的指导。这导致还原的波形缺乏细节，无法起到补全高频信息的作用。在图片超分辨率领域，使用对抗生成网络(GAN)训练的生成器往往能还原出图片更多的细节，如前文所述，像图片一样，音频是由大趋势和微小细节构成的，因此使用 GAN 训练音频超分辨率网络也能增强网络还原音频细节的能力。生成器的结构和图 3 所示 SRCNN 一致，判别器的结构如图 4。

图4 SRGAN判别器网络结构图

由于判别器的拟合任务比生成器小很多，训练时判别器的 loss 会迅速降低，这种过于严格的判别器无法给生成器有效的指导。为了防止这种情况，当判别器 loss 低于生成器 loss 的一半时，程序会停止训练判别器。

生成器网络优化使用的 loss 由两部分组成：MSEloss 与判别器判定的 loss 的 0.5%。

```
TotalLoss = MSEloss + BCEloss( discriminator( G ),1 ) * 0.005
```

此处将 MSEloss 纳入的原因是标准 GAN 的输入值为随机噪声，无法用于回归任务，而此网络生成器输入的是低分辨率的音频信号，输出的是对应的高分辨率音频信号。因此需要使用普通损失函数来指导网络优化的大方向。训练开始后，

MSEloss 不断降低，生成器 loss 和判别器 loss 不断震荡，还原效果越来越好。判别器在生成器的优化中占的比重逐渐增大，代替 MSEloss 继续训练生成器还原音频细节的能力，最终实现还原高频细节的效果。

4.实验分析

4.1 实验数据集

本文的训练数据来自 1000 首各种风格的歌曲，格式为双声道、采样率 44100Hz、位宽 16bit 的无损 wav 文件。我们使用 C++编写了预处理程序，调用 lame 编码器将这些 wav 文件编码为 MP3 文件，再将这些 MP3 文件转换回 WAV 格式并去除 MP3 编码时在歌曲开头加入的空白，得到训练数据。预处理流程如图 5 所示。MP3 压缩前后的音频信号损失如图 2 所示。从图 2 中可以看出，MP3 压缩后的音频主要丢失了 15kHz 以上的高频信息。

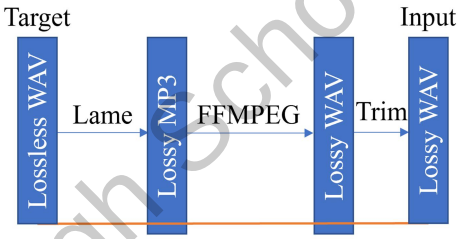


图5 数据预处理流程图

4.2 比较方法与消融方法

4.2.1 比较方法 1-循环神经网络(RNN)

RNN 通过记录中间层状态来结合时间信息，完成拟合目标的任务，适合用来处理音频等有时间连续性的数据。此网络由一个循环神经网络(RNN)和两个全连接层组成。网络的输入和输出均为 513 维，对应短时傅里叶变换(STFT)输出的 513 个频段。此网络使用 MSE loss 损失函数和 Adam 优化器进行训练。我们每次将隐藏层神经元数量增加 100，将隐藏层层数增加 1，然后对比 loss 下降到 0.0004 所需的时间。我们发现当 RNN 增加到 6 个隐藏层，每个隐藏层有 1500 个神经元时网络效率最高，收敛速度最快。

4.2.2 比较方法 2-频带预测卷积神经网络

卷积是从图片中提取特征的最好方法之一。由于频谱是一种音频的图形化形式，因而可以像处理图片一样使用卷积神经网络去处理频谱，从而获得较好的效果。我们的卷积编码器由卷积层、池化层、全连接层和各种激活函数构成。卷积层和池化层用于学习频谱的特征，全连接层用于应对 0 填充带来的伪影并提升拟合能力。由于音乐频谱是离散的，我们的卷积编码器采用带泄露线性整流函数(Leaky ReLU)作为卷积层的激活函数，能有效防止梯度消失并增强特征提取能力。BPCNN 的工作流程是先使用一个卷积编码器从音乐频谱中提取特征图，以寻找频谱中泛音的规律,然后用两个全连接层根据特征图推测下一个频段的振幅大小。其网络结构如图 6 所示：

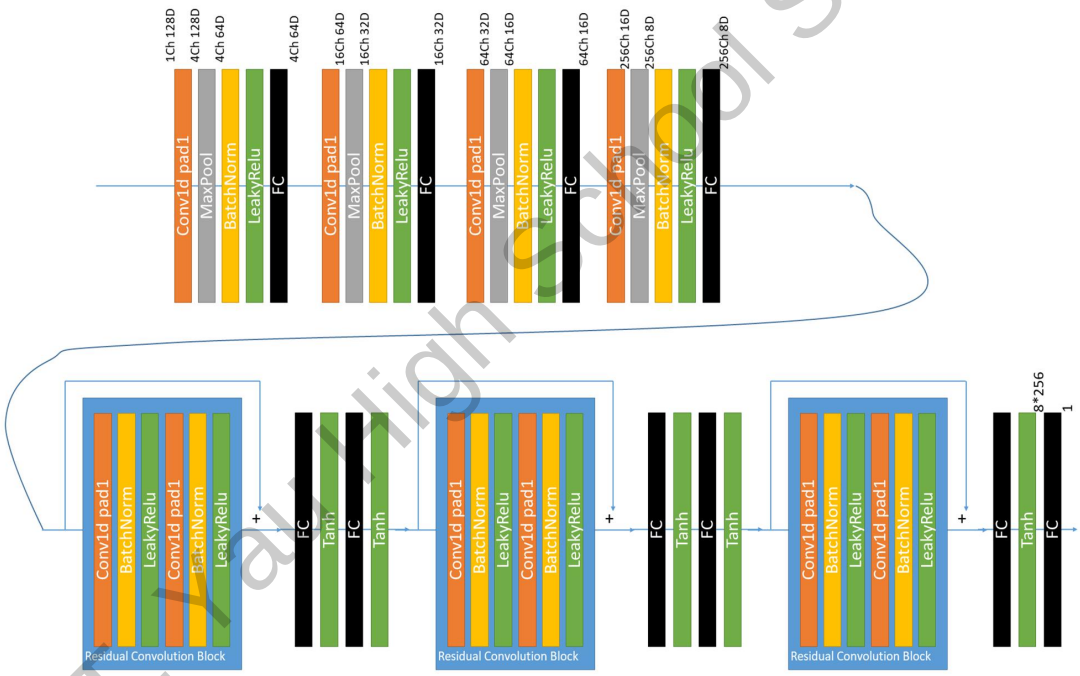


图6 BPCNN网络结构图

BPCNN 还原流程如图 7 所示。网络输入为 128 维，对应着 128 个连续无损频段。网络输出为 1 维，为其所输入的 128 个频段上方的第一个频段。每次将网络输入频段窗口向上移一个，直到还原全部缺失的频谱。

$$S[m+1] = \mathrm{BPCNN}(S[m-128:m])$$

S 是需要补全的振幅谱，S[m]代表第 m 个频段。此网络使用 MSEloss 损失函

数，使用 Adam 优化器进行优化，训练时使用的 STFT 频段数为 1025。该网络优点为对拟合能力要求较低，缺点为还原时将网络的输出重新作为输入会发生误差累积，经常出现高频爆音。经过 3000 epoch 训练后，训练集 loss 从 0.345 下降至 0.0000623，测试集 loss 从 0.8526 下降至 0.000691。在 2000 epoch 后，我们发现 loss 不断大幅度跳动，这是学习率过大的表现，于是将学习率由 0.0002 下调至 0.00001。

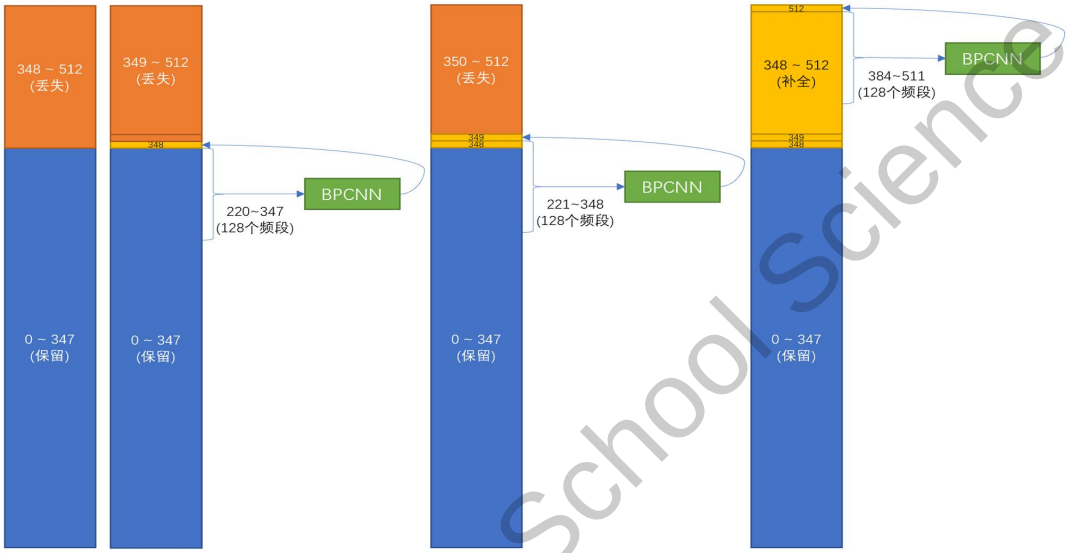


图7 BPCNN还原流程

4.2.3 比较方法 3-改进的 U-net

U-net 每层卷积都有残差结构[7]，这种结构可以帮助网络将大部分拟合能力用在还原丢失的高频上，而忽略差别较小的低频部分。由于音乐有时序性，而 RNN 能结合时序信息进行相应的处理，我们在 U-net 中间加入了一个 RNN 以提升其拟合能力，从而更精准地还原丢失的高频信息。改造后的网络结构如图 8 所示。

该网络的输入为 STFT 产生的频谱，频谱有 513 个频段，其中网络输入为后 512 个频段，输出为 513 个频段中最高的 165 个频段。在输入的数据中，最高的 165 个频段在 MP3 压缩过程中丢失，为了防止输入的高频部分全为 0 带来的梯度消失，采用随机噪声填充这 165 个频段，那么网络输出的 165 个频段与原来未缺失的 348 个频段进行拼接后即可得到完整的音乐频谱。

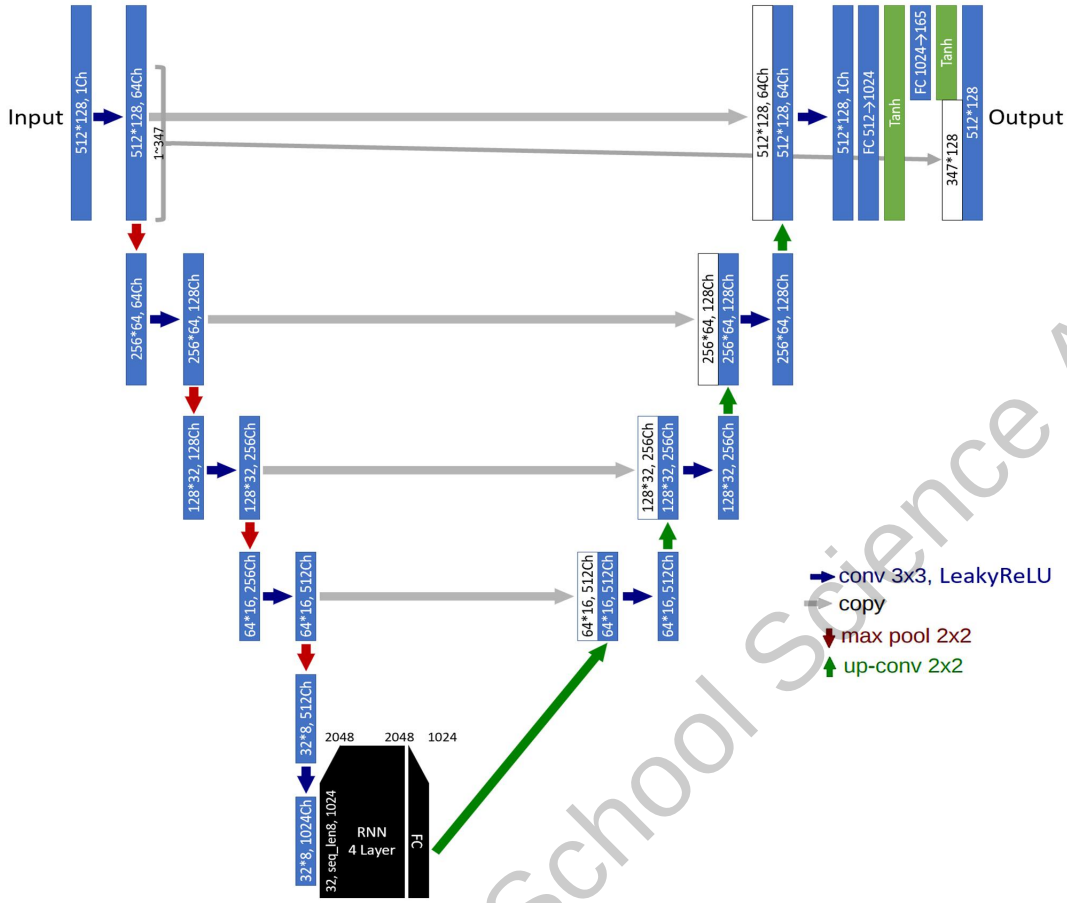


图8 调整后的U-net网络结构图

4.2.4 消融实验

为了验证我们超分卷积网络各结构的有效性，我们进行了两组消融实验，分别去除了网络中的全连接层结构和特征选择结构。网络其余部分、数据预处理方法、训练超参数等均保持不变。

4.2.5 人耳区分实验

为了验证我们的方法是否能有效地改善 MP3 音乐文件的听觉感受，我们使用问卷星平台进行了问卷调查和实验。被调查者需要从 3 个音频中分辨出哪个是无损、128k MP3 和 SRCNN 还原的音质。实验中人们无法看到音频的波形或频谱，仅能够通过声音进行判断。

4.3 实验结果

深度神经网络还原音频的质量是通过对比还原的频谱和原始无损频谱来评

估的。我们将音频信号标准化，然后以宽度为 1024 的 Hann 窗口和 512 的步长计算 STFT。本文各方法的还原效果如图 9 所示。

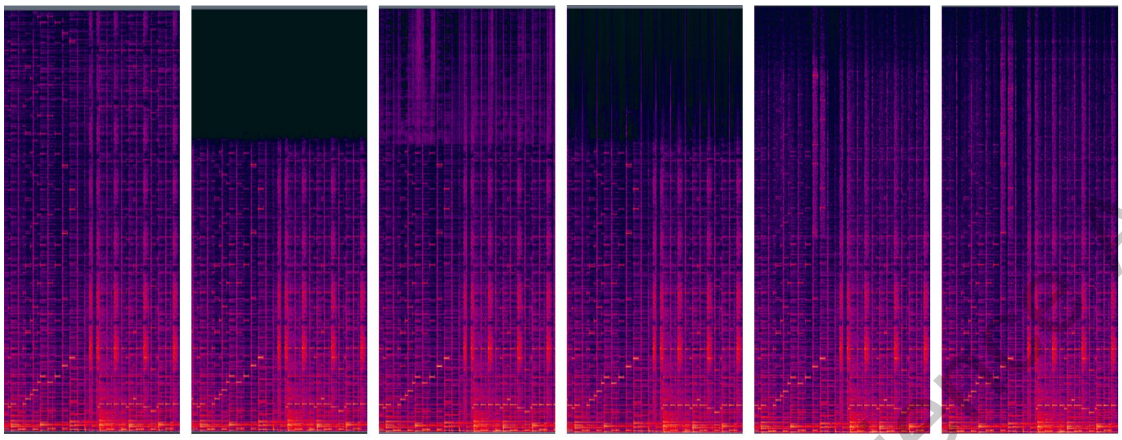


图9 不同网络还原效果对比

除了本文所设计的各种模型，我们还使用了其他研究人员的相关算法进行实现，包括 Marius 设计的 CNN 算法 [1]， PWCNN [2]和 Somesh 设计的 FWR [3]。分别在古典音乐和电子音乐数据集上进行测试。通过各种算法还原的频谱与无损频谱的 L1 loss 如表 1 所。

表 1 不同算法的还原效果（L1 loss）对比

音乐类型	MP3	SRCNN	SRGAN	SRGAN 无全连接层	SRGAN 无特征选择	RNN	BPCNN	改进的 U-net GAN	FWR	PWCNN	CNN Marius
电子音乐	4.990	1.0829	1.068	1.141	1.126	1.407	1.610	1.104	0.998	1.196	1.163
古典音乐	3.033	0.825	0.901	1.237	1.210	1.779	1.604	2.507	1.061	0.954	1.600

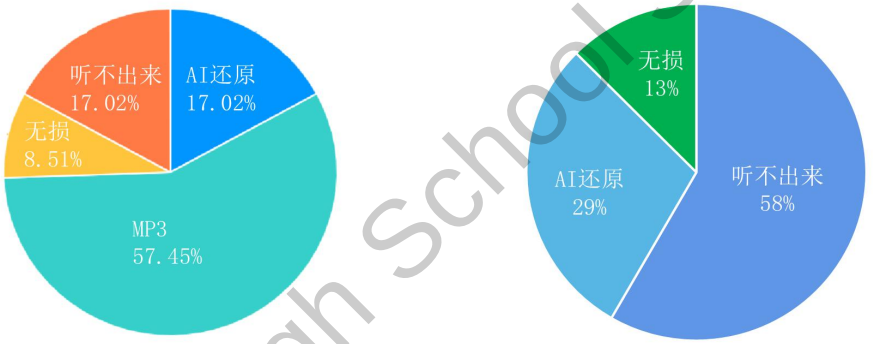
从表中可以看出，删除全连接层结构和删除特征选择结构均对卷积超分网络的效果有负面的影响。这说明我们添加的这些结构有着提升网络还原效果的作用。

本文提出的 SRCNN 和 SRGAN 增强古典音乐的效果在现有的方法中分别位于第一名和第二名。这说明我们的方法能够切实有效地补全缺失的高频信息。SRGAN 在判别器的帮助下解决了普通损失函数无法指导网络学习细节的问题，与

Marius 研究工作中的 CNN 相比能够还原更多的高频细节。

SRGAN 增强电子音乐的效果位于第二，仅次于 FWR 算法。这是因为电子合成器的声音频谱在低频部分和高频部分几乎一模一样，直接复制频谱即可获得很好的效果。但 FWR 对不同类型的音乐的复原结果差异较大。

人耳区分实验参与者共 48 位，实验结果如图 10 所示。问题 1 调查结果显示，57.45%的人正确分辨出了 MP3 有损音频，可以作为分析 AI 音频还原效果的有效数据。在这些人中，问题二的调查结果中能够区分无损音频和 AI 还原音频的人仅占 13%，判断错误的人占 29%，这部分人可以被归入无法区分 AI 还原音频和无损音频，再加上选择“听不出来”的 58%，共有 87%的人认为我们的音频增强算法是有效的，能够通过还原丢失的高频信息来切实增强听觉感受。



问题1：请问MP3（有损）音频是哪一个？　　问题2：请问无损音频是哪一个？

图10 人耳区分实验结果

5. 结论

本文提出了一种创新性的基于深度神经网络带宽扩展方法支持高质量的音乐复原方法，将有带宽限制的音频信号增强至 CD 质量（44.1kHz 采样率），以还原原始数据带来的听觉感受。通过分析音乐中人声和不同种类乐器在高频和低频的特征和相关性，经过长期探索，我们设计了一套基于 CNN 和 GAN 方法，从频域和时域进行带宽扩展。对于频域的带宽扩展，我们使用了类似图片补全的方法；对于时域的带宽扩展，我们使用了超分辨率的方法。我们将所提出的方法与 RNN、BPCNN 等方法进行比较。实验证明本文所提出的方法具有最低的频谱损失和最好的还原效果。人耳听力对比实验结果进一步证明了音频增强算法的有效性。

参考资料

[1] M. Miron, DAFX, Averio, Portugal, tech., 2018.

[2] B. Si, D. Luo, and J. Ju, Electronic Letters, tech., 2021.

[3] S. Ganesh, Georgia Tech Center for Music Technology, Atlanta, Georgia, rep., 2016.

[4] Brian McFee, "librosa/librosa: 0.8.1rc2". Zenodo, May 25, 2021. doi: 10.5281/zenodo.4792298.

[5] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," arXiv.org, 03-Dec-2019. [Online]. Available: https://arxiv.org/abs/1912.01703. [Accessed: 28-Aug-2021].

[6] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," arXiv.org, 25-May-2017. [Online]. Available: https://arxiv.org/abs//1609.04802. [Accessed: 28-Aug-2021].

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," arXiv.org, 18-May-2015. [Online]. Available: https://arxiv.org/abs/1505.04597v1. [Accessed: 28-Aug-2021].

[8] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv.org, 30-Jan-2017. [Online]. Available: https://arxiv.org/abs/1412.6980. [Accessed: 28-Aug-2021].

[9] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li, "High-Resolution image Inpainting USING Multi-scale Neural PATCH SYNTHESIS," arXiv.org, 13-Apr-2017. [Online]. Available: https://arxiv.org/abs/1611.09969. [Accessed: 28-Aug-2021].

[10] J. S. J. Ren, L. Xu, Q. Yan, and W. Sun, "Shepard Convolutional Neural Networks," Shepard convolutional neural networks, 2015. [Online].Available: https://academic.microsoft.com/paper/2184016288 [Accessed: 28-Aug-2021].

[11] S. Kim and V. Sathe, "Adversarial audio Super-resolution with unsupervised Feature Losses," OpenReview, 21-Dec-2018. [Online]. Available: https://openreview.net/forum?id=H1eH4n09KX. [Accessed: 28-Aug-2021].

[12] Perraudin, Nathanaël & Balazs, Peter & Søndergaard, Peter. (2013). A Fast Griffin–Lim Algorithm. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. 1-4. 10.1109/WASPAA.2013.6701851.

[13] W. Cai and M. Li, "A unified deep speaker embedding framework for mixed-bandwidth speech data," arXiv.org, 01-Dec-2020. [Online]. Available: https://arxiv.org/abs/2012.00486. [Accessed: 01-Sep-2021].

[14] Y. Li, M. Tagliasacchi, O. Rybakov, V. Ungureanu, and D. Roblek, "Real-time speech frequency bandwidth extension," arXiv.org, 09-Feb-2021. [Online]. Available: https://arxiv.org/abs/2010.10677. [Accessed: 01-Sep-2021].

[15] S. Kim and V. Sathe, "Bandwidth extension on raw audio via generative adversarial networks," arXiv.org, 21-Mar-2019. [Online]. Available: https://arxiv.org/abs/1903.09027. [Accessed: 01-Sep-2021].

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," arXiv.org, 10-Dec-2015. [Online]. Available: https://arxiv.org/abs/1512.03385. [Accessed: 31-Aug-2021].

[17] A. Castillo, M. Escobar, J. C. Pérez, A. Romero, R. Timofte, L. Van Gool, and P. Arbeláez, "Generalized real-world super-resolution through adversarial robustness," arXiv.org, 25-Aug-2021. [Online]. Available: https://arxiv.org/abs/2108.11505. [Accessed: 01-Sep-2021].

[18] Z. Liang, Y. Wang, L. Wang, J. Yang, and S. Zhou, "Light field image super-resolution with transformers," arXiv.org, 17-Aug-2021. [Online]. Available: https://arxiv.org/abs/2108.07597. [Accessed: 01-Sep-2021].

[19] M. Xu, Z. Wang, J. Zhu, X. Jia, and S. Jia, "Multi-attention generative adversarial network for remote sensing image super-resolution," arXiv.org, 14-Jul-2021. [Online]. Available: https://arxiv.org/abs/2107.06536. [Accessed: 01-Sep-2021].

[20] M. A. Farooq, A. A. Khan, A. Ahmad, and R. H. Raza, "Effectiveness of state-of-the-art super resolution algorithms in surveillance environment," arXiv.org, 08-Jul-2021. [Online]. Available: https://arxiv.org/abs/2107.04133. [Accessed: 01-Sep-2021].

致谢

本项目得到了北京大学董豪老师和北京市第一 0 一周宇辰老师的悉心指导，在此表示衷心的感谢！

1. 论文选题来源、研究背景；

论文的选题主要来源于本人作为音乐发烧友对音乐质量的深刻感受和对人工智能技术的浓厚兴趣。

目前很多音频都是以有损格式进行存储和传输的，造成了听觉感受的下降。因此，需要一种算法来实时将有损的音频进行带宽扩展，还原为无损的音频，以重建原曲的听觉感受。鉴于现有的音频带宽扩展方法无法较好地处理音乐，我们针对音乐信号设计了音乐超分辨率神经网络，实现了高质量实时还原有损音频。详细的研究背景见本文第一节。

本项目的研究工作到目前为止历时 15 个月，先后在大量音乐数据集上循序渐进地进行了 8 种模型的设计、验证与优化。早期 3 种基于 MLP 和不同音频特征组合实验的研究报告参加了丘成桐数学中心和腾讯联合组织的 2020 犀牛鸟中学科学培养计划，并获得全国三等奖（总排名第 4-7 名之间）。本项目研究报告是 2020 年 10 月之后对犀牛鸟阶段性研究成果的进一步延伸，将网络结构扩展到 CNN 和 GAN 等模型，并创新性地使用超分辨率进行频率扩展，得到了更好的实验效果。

2. 每一个队员在论文撰写中承担的工作以及贡献；

本项目为本人在导师指导下独立完成。

3. 指导老师与学生的关系，在论文写作过程中所起的作用，及指导是否有偿；

董豪老师是北京市第一 0 一中学北大前沿计算研究中心 AI 实验室校外指导老师。周宇辰老师是北京市第一 0 一中学北大前沿计算研究中心 AI 实验室负责人。本人为该实验室项目组成员。

4. 他人协助完成的研究成果

本文的工作均为本人在导师指导下独立完成。

团队成员介绍：

刘衍东，北京市第一 0 一中学高三学生。热爱科技创新，曾获得北京市中学生科技创新竞赛一等奖、丘成桐数学中心和腾讯联合举办的 2020 犀牛鸟中学科学人才培养计划全国三等奖(总排名第 4-7 名之间)、2019~2020 届 China Thinks Big 全球创新研究大挑战作为队长带领团队获得全球站团体二等奖。

指导教师介绍：

董豪，北京大学前沿计算研究中心助理教授，博士生导师。董博士毕业于帝国理工学院，研究领域为机器人与计算机视觉，当前方向为自监督/基于模型的机器人学习，以实现高效的机器人学习方法，降低学习的数据要求和系统的成本。他致力于推广人工智能技术，是鹏城实验室双聘人员，还是深度学习开源框架 TensorLayer 的创始人，并获得 ACM MM 2017 年度最佳开源软件奖。

周宇辰，博士，研究员，北京市第一 0 一中学英才学院人工智能导师。ACM 和 IEEE 高级会员，曾任中国计算机学会嵌入式系统专委会委员。IBM 二十年技术创新经验，曾任 IBM 中国研究院人工智能感知研究主管、IBM 科学院成员、IBM 发明大师等，3 次获杰出技术成就奖。学术专著 1 部，国际标准 2 项，国际专利近 50 项，学术论文 30 余篇。