

## 深度跨模态环境声音合成

程皓楠, 李思佳, 刘世光\*

(天津大学智能与计算学部 天津 300350)  
(lsg@tju.edu.cn)

**摘要:** 随着计算机图形学技术的不断发展, 用户对视频及动画的声音质量提出了更高的要求. 针对现有方法中存在的算法复杂度高, 可扩展性不强等问题, 提出一种基于 CGAN 和 SampleRNN 的深度学习的环境声音合成算法, 采用 VGG 网络模型提取视频深度特征. 并将视频深度特征通过一个时序同步网络模型, 实现具有更高同步性的视频到音频的跨模态特征转换; 通过音色增强网络模型对合成声音的音色进行增强, 以提高网络结构的可扩展性, 并最终与视频同步的、真实感较强的环境声. 通过对音视频跨模态数据集中 12 类不同类别视频进行训练与测试, 结果的主观与客观评价表明, 文中算法所生成的结果真实感强, 提高了现有算法的可扩展性.

**关键词:** 环境声音合成; 深度学习; 跨模态; 时序同步网络模型; 音色增强网络模型  
**中图分类号:** TP391.41 **DOI:** 10.3724/SP.J.1089.2019.17906

## Deep Cross-Modal Synthesis of Environmental Sound

Cheng Haonan, Li Sijia, and Liu Shiguang\*

(Division of Intelligence and Computing, Tianjin University, Tianjin 300350)

**Abstract:** With the continuous development of computer graphics technology, users put forward higher requirements for accompanied sound of video and animation. Aiming at the problem that current methods usually are high complexity and poor scalability, this paper proposed a novel deep environment sound synthesis algorithm which is based on generative adversarial network and sample recurrent neural network. First, the deep features of the video are extracted based on the visual geometry group network model. Then, a novel synchronous sequential network model is proposed to realize the cross-modal feature transformation with higher synchronization rate from visual to audio. Finally, the generated sound is enhanced through the timbre enhancement network model for scalability improvement. Through training and testing 12 different types of video in the audio-video cross-modal data set, the subjective and objective evaluation of the results shows that the generated results are realistic and the proposed method is scalable.

**Key words:** environmental sound synthesis; deep learning; cross-modal; synchronous sequential network model; timbre enhancement network model

随着计算机图形学的不断发展, 用户对于动画、游戏的真实感不再仅仅局限于视觉效果, 对

于听觉的真实感也提出了更高的要求. 因此, 越来越多的研究者开始关注与视觉动画匹配的环境

收稿日期: 2019-06-17; 修回日期: 2019-07-22. 基金项目: 国家自然科学基金(61672375, 61170118). 程皓楠(1994—), 女, 博士研究生, 主要研究方向为计算机图形学、虚拟现实; 李思佳(1997—), 女, 在校学生, 主要研究方向为计算机图形学、深度学习; 刘世光(1980—), 男, 博士, 教授, 博士生导师, CCF 高级会员, 论文通讯作者, 主要研究方向为计算机图形学、图像/视频处理、可视化、虚拟现实.

声音合成算法,力求生成真实感较强的声音效果。

根据视觉内容生成对应的声音是一个跨模态声音合成问题。自动的跨模态环境声音合成方法可以分为基于物理的方法和基于学习的方法。近年来,研究者针对不同场景提出了一些基于物理的跨模态声音合成方法,如利用谐波气泡合成水声<sup>[1-4]</sup>,基于火焰模型的燃烧声<sup>[5-7]</sup>、褶皱声、滚动声和滑动声<sup>[8-9]</sup>,这些方法取得了令人满意的合成结果。然而,它们存在 2 个主要局限性:(1) 间接性,即这些方法为了建立视觉和听觉之间的联系,首先需要构建动画模型,然后从动画模型中提取合适的参数构造声学模型来得到最终同步的视频和音频结果。(2) 扩展性差,是基于物理方法的另一个局限性,即这些方法只能处理特定的场景,合成单一种类的声音。例如,燃烧声音的合成算法不能用于液体声音合成。为解决上述问题,本文提出一种基于深度学习的跨模态环境声合成算法,来探索神经网络能否更加直接地学习视觉和声音之间的关系。

近几年,卷积神经网络(convolutional neural network, CNN)、循环神经网络(recurrent neural network, RNN)以及生成对抗网络(generative adversarial network, GAN)等的出现,为基于学习的跨模态转换提供了可行方案。文献[10-11]将音频转换成不同的特征谱图作为输入,利用特征匹配在数据库中查找相应的声音来合成最终结果。这些方法的一个主要问题是它们不是以端到端的方式建立视频和音频之间的连接,而是变为由图像到谱图的转换。由于不同类别的环境声所需要的特征谱图不同,因此基于特征谱图的方法同样存在扩展性差的问题,如文献[10]方法仅适用于敲击声。为了提高其扩展性,文献[12]首次设计了一种端到端的基于 RNN 的跨模态环境声合成方法,虽然它可以利用同一网络框架合成 10 种不同类别的声音,但其无法有效地捕捉视频内容,导致视频与音频的同步性较低。因此,虽然基于深度学习的跨模态合成方法解决了基于物理的方法存在的间接性问题,可以直接根据视频画面生成声音,但是扩展性差仍然是一个亟待解决的问题。同时,同步性也成为目前基于学习的方法发展的一个主要瓶颈。针对上述问题,本文设计了一种端到端的跨模态环境

声合成算法,提高了现有基于深度学习方法的扩展性和同步性。

然而,端到端的跨模态环境声合成更具挑战性,这是因为图像和音频在采样率和特征结构等方面都存在较大的差异。视频的采样率通常是 30 帧/s,而声音的采样率则是 44 100 Hz,因此视频与音频对应的特征向量长度往往相差较大。此外,图像特征通常为矩形特征图,而音频特征则是条形信号。这些差异使得原有的一些网络结构无法使用,并且对于网络自身的学习能力也提出了较高的要求。为了解决上述问题,本文提出了一种利用视频编码技术来弥补时空尺度差距的端到端的网络框架。该框架包含一个视频编码器,一个保证时序同步性的网络模型和一个用于音色增强的网络模型。为了训练和测试本文提出的网络模型,搜集并搭建了一个包含数百个不同自然场景视频的数据集,称为音视频跨模态数据集。实验表明,本文所提出的网络架构和训练结果能够生成具有更高同步性的高质量、可扩展的跨模态环境声。

## 1 相关工作

### 1.1 基于物理的跨模态环境声合成算法

传统的计算机动画和电影中的音效制作往往需要专业的配音工程师,为了实现声音的自动生成,研究人员对动画物理建模与声音建模之间的关联性进行了探索。目前基于物理的声音合成技术在不同场景类别中都有着广泛的应用,包括水声<sup>[1-4]</sup>、风声<sup>[13]</sup>、火焰燃烧声音<sup>[5-7]</sup>、刚体运动以及形变产生的声音<sup>[8-9]</sup>等。由于这些方法大多是基于声音样本或声学公式,因此合成结果的质量通常可以达到令人满意的水平。然而,基于物理的合成方法存在 2 个主要问题:即不可扩展性和间接性。相比之下,基于学习的声音合成方法是一种更直接的方法,这类方法试图探索计算模型是否可以像人类感知一样直接学习视觉和声音之间的联系。因此本文采用基于学习的方式来训练跨模态环境声合成模型,可以根据输入的无声视频自动地生成与之同步的环境声。

### 1.2 基于深度学习的跨模态环境声合成算法

深度学习的发展为一些跨模态转换任务提供

环境声通常指非语音和音乐的其他类型声音,典型的环境声有雨声、海浪声、敲击声等。

了新思路. 视频到音频的跨模态任务的难点在于视频与音频分属 2 种不同的信号, 它们无论是从表现形式还是特征分析的角度都存在较大的差异, 人为寻找二者之间的关联难度较大. 而基于深度学习的方法可以自动地学习输入与输出之间的转换关系, 使得此任务具备了可行性. Owens 等<sup>[10]</sup>提出了一种基于 CNN 和长短时记忆网络的网络框架, 实现了自动的敲击声合成. Chen 等<sup>[11]</sup>基于 GAN 设计了一种适用于乐器声合成的网络结构. 上述 2 种声音合成算法分别基于耳蜗图和梅尔频谱, 因此不能直接生成原始音频信号. 本质上, 这些方法仍将一幅图像(视频帧)转换为另一幅图像(声音特征谱图).

由于音频和视频具有不同的时空尺度以及不匹配的特征结构, 构建端到端的跨模态环境声音合成系统虽然更直接, 但也更具挑战性. Zhou 等<sup>[12]</sup>基于 SampleRNN 模型, 通过对户外采集的视频进行训练, 合成了真实感更强的环境声音. 然而, Owens 等<sup>[10]</sup>的实验表明, RNN 对视觉内容的学习存在局限性, 因此该方法不能实现理想的同步效果. 为此, 本文提出了一种直接合成音频的解决方案, 不仅具有多类别的可扩展性, 同时可以实现音频和视频动作的更好的同步.

近年来, GAN 在不同应用中都展示了良好的性能<sup>[14-15]</sup>. 一些工作已经证明了利用 GAN 合成声音的可行性. 此外, Donahue 等<sup>[14]</sup>证明 GAN 能够在无监督的条件下合成音频. 然而, 本文所研究的问题既需要视频帧, 也需要音频作为输入进行训练, 因此传统的 GAN 并不适用于本文的任务. 而在过去的几年里, 条件生成对抗网络(conditional generative adversarial nets, CGAN)<sup>[16]</sup>引起了越来越多关注, 例如, 文本转换<sup>[16-18]</sup>、图像合成与检索<sup>[19-20]</sup>等均表现出较强的学习能力. 因此, 本文设计了一种基于 CGAN 的时序同步网络模型来提高算法的同步性. 但是, 跨模态环境声合成任务与上述工作有很大不同, 因此, 需要调整编码器并设计适当的滤波器和损失函数.

音频具有很强的时间依赖性, 考虑这种固有的时序关系, 本文利用 RNN 对合成的音频做进一步处理, 以增强其音色质量. 由于考虑了时间因素, RNN 非常适合于处理序列建模问题, 在许多问题上也得到了广泛的应用<sup>[21-22]</sup>. 但是, 音频信号的维度很高, 直接利用 RNN 建模比较困难, 因此很多方法都是首先对其降维, 再利用 RNN 处理. 而本

文期望直接对音频样本进行操作, 且更好地捕获音频的音色. 因此, 本文基于 SampleRNN<sup>[21]</sup>设计音色增强网络的模型. 其层次结构可以使它对长序列建模并学习其不同尺度的特征; 但基本的 SampleRNN 是无条件自回归生成模型. 而本文试图利用已有的低质量声音合成相应的高质量声音, 因此还需要对其增加条件控制信息.

## 2 本文算法具体实现

针对跨模态环境声音合成所面临的巨大挑战, 即扩展性与同步性问题, 本文提出了一种基于 CGAN 和 SampleRNN 的跨模态环境声音合成方法, 其框架如图 1 所示. 由视频到最终结果的输出经历视觉特征编码器、时序同步网络模型和音色增强网络模型 3 个模块. 算法开始于视觉特征编码器, 通过输入视频帧得到一个 44 100 维的向量作为输出. 然后, 此向量作为输入进入到时序同步网络模型的生成器中进行一系列的卷积和反卷积操作, 最终生成一个新的 44 100 维的向量作为生成器的输出. 随后, 此向量作为输入进入到音色增强网络模型中并得到最终的生成结果.

### 2.1 视觉特征编码器

视频和音频本身采样率和维度上的差异使得其无法通过网络直接训练, 而相同维度音频向量与视觉特征向量更便于网络学习二者之间的转换关系. 因此本文设计了一个视觉特征编码器来生成与音频序列长度相等的视频特征序列. 由此产生一个 44 100 维的向量与声音的维度相匹配. 首先, 本文将视频到音频的转换任务定义为

$$G(y_1, \dots, y_m) \rightarrow x_1, \dots, x_n, x \in \{\text{音频}\}, y \in \{\text{视频}\}.$$

其中,  $y_1, \dots, y_m$  代表输入视频帧的颜色通道信息, 每一个通道都是由 0~255 的数组成的矩阵;  $G(y_1, \dots, y_m)$  表示基于视频帧生成的声音信号的值, 其取值范围为 -1~1;  $x_1, \dots, x_n$  表示视频对应的声音信号的值, 其变化范围为 -1~1. 对于不同输入的视频, 为了减少计算量以及统一管理, 本文将输入图像缩小成大小为  $256 \times 256 \times 3$  的图像. 对于任意视频帧  $y_i$ , 提取其在 VGG19<sup>[23]</sup>网络下的特征向量  $v_i$ , 其维度为  $1 \times 44\ 100 \times 1$ . 本文定义  $S_{\text{video}}$  和  $S_{\text{audio}}$  分别为视频和音频的采样率(本文中分别为 30 和 44 100). 对于第  $t$  秒的视频, 对应的视觉特征向量为

$$V_t = v_{t,q} \oplus v_{t,q*2} \oplus \dots \oplus v_{t,q*p}.$$

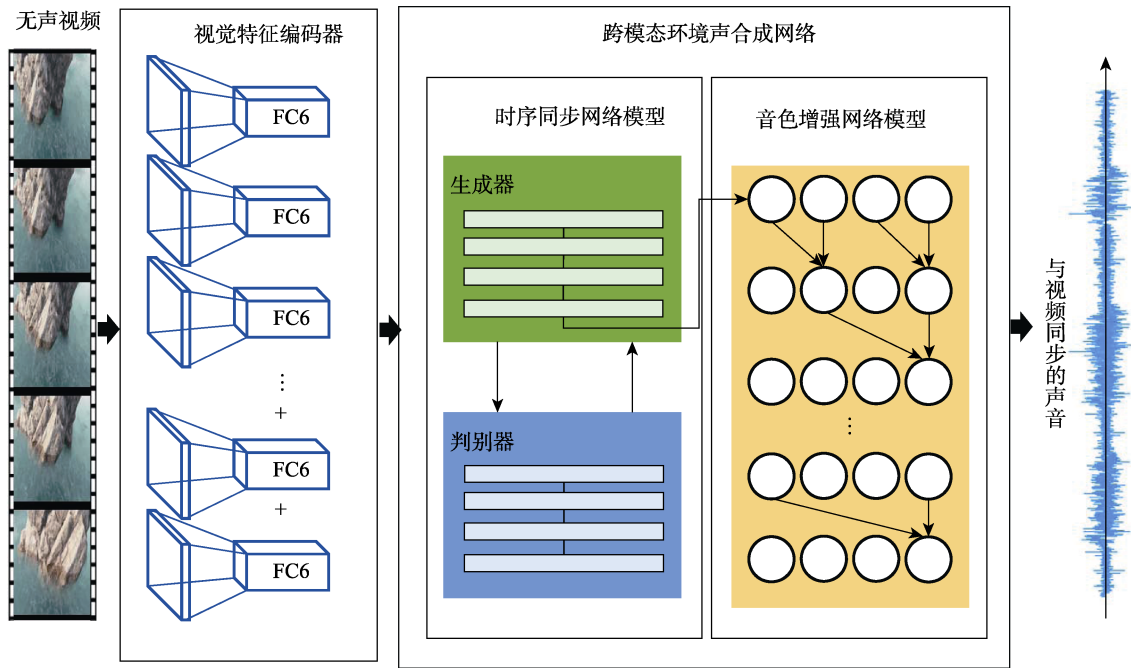


图 1 本文算法框架

其中,  $p = \lfloor S_{\text{audio}}/4096 \rfloor$ ,  $q = \lfloor S_{\text{video}}/p \rfloor$ ,  $v_{t,q}$  为第  $t$  秒第  $q$  帧视频的特征向量;  $\oplus$  为串联操作. 这样, 本文可以通过视觉特征编码器, 获得与音频向量  $X_t$  维度相同的视频向量  $V_t$ .

### 2.2 时序同步网络模型

为了建立视频序列和音频序列之间的同步映射, 本文基于 CGAN 网络模型, 设计了一种适用于音视频跨模态的时序同步网络模型. CGAN 网络由生成器和判别器组成, 其中生成器用来学习音视频特征之间的映射关系并生成具有时序同步性的声音向量, 而判别器则用来判断生成器的结果真假以提高生成器的性能.

与图像不同, 音频序列往往序列更长. 因此, 传统的 CGAN 中  $3 \times 3$  的感受野将不再适用于声音生成器. 因此, 本文通过更大的感受野来完成生成器和判别器的卷积操作, 并使用更适于音频向量处理的一维滤波器进行卷积. 为了避免在卷积和反卷积过程中出现维度近似的情况产生, 本文设计了一种变化的卷积核变化机制, 以确保网络的稳定性. 受文献[14]启发, 本文在生成器网络最后添加了一层滤波层来优化声音的合成. 具体卷积核变化以及网络参数配置如图 2 所示; 其中卷积核大小所对应的 3 个参数分别为感受野大小、该层输入通道数和输出通道数. 每层所对应的 3 个参数分

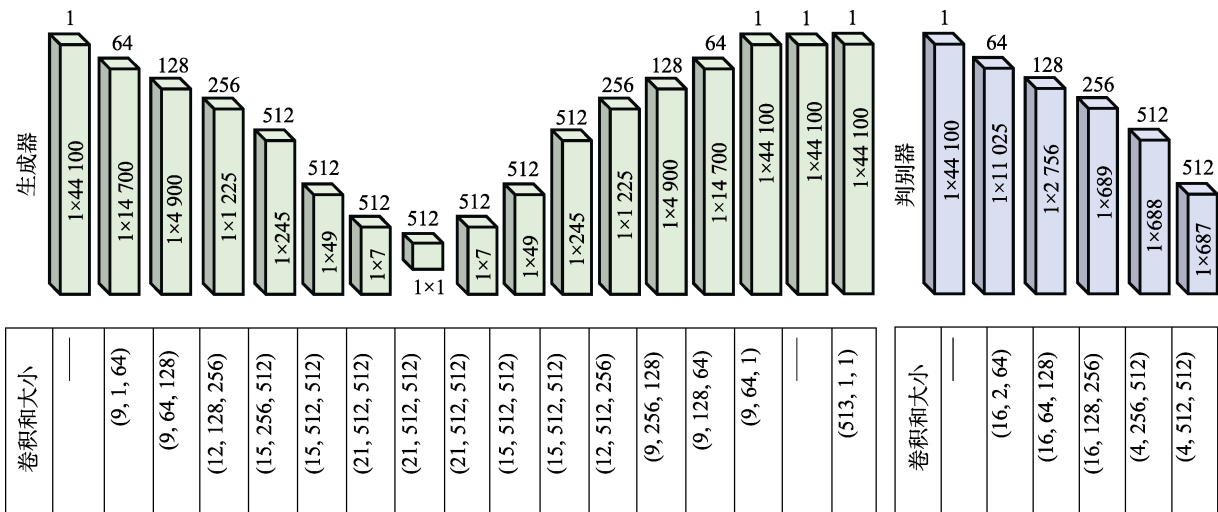


图 2 时序同步网络模型参数配置

别为该层 Batch 大小、输入向量维度和通道数.

传统的 CGAN 使用的损失函数往往是交叉熵损失函数, 其具体形式为

$$\min_G \max_D L(D, G) = E_{X, V \sim p_{\text{data}}(X, V)} [\log(D(X, V))] + E_{z \sim p_z(z), V \sim p_{\text{data}}(V)} [\log(1 - D(G(z, V), V))].$$

其中,  $D(\cdot)$  为判别器的输出值;  $X$  为音频向量;  $V$  为通过视频特征编码器的视频向量;  $z$  为随机噪声向量. 为了避免在使用交叉熵损失函数过程中的梯度消失现象, 本文对原有损失函数进行了调整, 采用最小二乘损失函数替代了交叉熵损失函数. 此外, 由于  $L_1$  损失函数可以捕获到低频信息, 因此本文在最小二乘损失函数的基础上添加了  $L_1$  损失函数. 最终, 本文所使用的损失函数为

$$\begin{cases} \min_D L(D) = \frac{1}{2} E_{X, V \sim p_{\text{data}}(X, V)} [(D(X, V) - 1)^2] + \frac{1}{2} E_{V \sim p_{\text{data}}(V)} [(D(G(V), V))^2] \\ \min_G L(G) = \frac{1}{2} E_{V \sim p_{\text{data}}(V)} [(D(G(V), V) - 1)^2] + \lambda E_{X, V \sim p_{\text{data}}(X, V)} [\|X - G(V)\|_1] \end{cases}$$

其中,  $G(\cdot)$  为生成器的输出值;  $\lambda$  为  $L_1$  损失函数的权重控制参数, 实验中  $\lambda = 100$ .

### 2.3 音色增强网络模型

音色增强网络保证了本文算法具有较高的可扩展性. 通过时序同步网络合成的声音有时还会存在一些噪声, 音色也会略显沉闷; 因此需要对此声音做进一步处理, 以降低噪声并增强音色. 传统的降噪算法在许多情境下有着较好的效果, 但是对于一些环境声音, 如雨声、海浪声等, 直接应用则可能会破坏声音原本的结构, 而且对音色也并没有提升作用. 因此, 本节采取可以在更多种情境中应用的神经网络方法, 实现对声音的降噪与音色的增强, 使本文算法具有更强的扩展性. 受文献[12]启发, 本文采用近年来提出的 SampleRNN 模型<sup>[21]</sup>作为音色增强网络, 实现了对波形样本的直接建模, 避免了复杂的音频信号处理过程, 而且可以学习到长序列中不同尺度的依赖信息.

然而, SampleRNN 是一个无条件的自回归生成模型, 多应用于直接生成任意长度的波形样本, 如音乐生成任务. 而本文需要解决的问题是通过给定的低质量的样本生成干净的音色较好的样本, 因此需要为原模型增加额外的条件信息, 指导其生成符合要求的样本. 近年来也有许多工作在利用 SampleRNN 进行合成的同时为其添加了条件控

制信息<sup>[12,24-26]</sup>. 文献[24]为原始模型添加了语言特征约束, 将其作为 SampleRNN 每一层的额外输入信息, 在保留了源说话者的内容信息的同时, 也学习到了目标说话者的音色, 实现了由源说话者到目标说话者声音的转换, 同时也表明了 SampleRNN 可以较好地学习到声音的音色. 文献[25-26]则是输入额外的声学特征, 实现了利用声学特征控制语音的合成. 文献[12]为模型添加了视觉信息, 通过 2 种方法利用视觉信息指导声音的合成: 第 1 种方法是将视觉特征与最顶层的结点连接起来, 本节主要受此启发; 第 2 种方法则是使用序列到序列模型, 最终通过视觉信息的输入, 合成了与画面对应的音色较好的音频. 因此, 本文通过 SampleRNN 学习真实音频的音色结构, 并将时序同步网络生成的样本直接作为额外的条件输入, 以获得音色增强后的样本.

用于音色增强的 SampleRNN 结构如图 3 所示. 其最底层是多层感知器, 其余层均为 RNN. 它的每一层都由不同的工作时钟所驱动, 最底层工作在单个样本之上, 其上各层每个时间步处理的样本数逐层增加, 每层的输入除了音频样本之外, 还包含上一层的输出. 这种层次结构使它可以学习到不同尺度的依赖信息; 限于篇幅, 更多具体的技术细节可以参考文献[21].

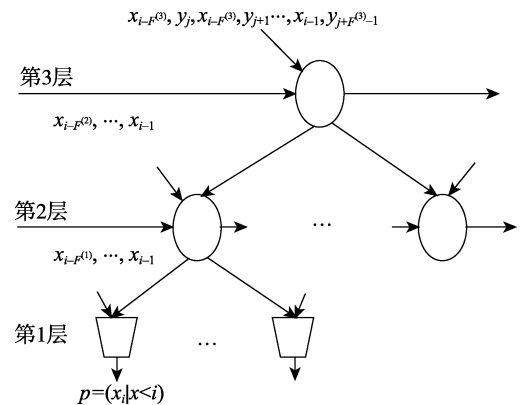


图 3 音色增强网络结构示意图

在基本 SampleRNN 的模型基础上, 本文加入了额外的输入条件. 由于最顶层控制着其下各层, 因此本文只需将条件信息输入至最顶层即可. 本文用  $A = \{a_1, \dots, a_T\}$  表示时序同步网络生成的音频样本, 用  $B = \{b_1, \dots, b_T\}$  表示真实的音频样本. 在模型的训练阶段, 原始的 SampleRNN 各层的输入本应为样本  $B$ , 经改动后本文将样本  $A$  与样本  $B$  连接起来作为最顶层的输入, 其余各层输入的样本

仍然为  $B$ . 现假设 SampleRNN 共有  $K$  层, 其中第  $k$  层每次向前回看  $F^{(k)}$  个样本, 本文将这些样本记作帧  $f^{(k)}$ . 若在某时刻最顶层应输入的  $B$  样本帧为  $f_b^{(K)} = [b_{i-F^{(K)}}, \dots, b_{i-1}]$ ; 则与其预测的下一帧相对应的为  $t$  时刻的样本帧  $A$ . 其中  $t = \lfloor i/F^{(K)} \rfloor$ , 该帧展开为  $f_a^{(K)} = [a_j, \dots, a_{j+F^{(K)}-1}]$ ; 其中,  $j = t \times F^{(K)}$ . 接下来本文将 2 帧连接起来, 即得到总的输入帧

$$f^{(K)} = \begin{bmatrix} [b_{i-F^{(K)}}, a_j], \\ \vdots \\ [b_{i-1}, a_{j+F^{(K)}-1}] \end{bmatrix}.$$

之后, 本文将其展开成一个向量作为最顶层的输入帧, 即  $f^{(K)} = [b_{i-F^{(K)}}, a_j, b_{i-F^{(K)}+1}, a_{j+1}, \dots, b_{i-1}, a_{j+F^{(K)}-1}]$ .

由于第  $K$  层是最顶层, 因此其输入只有音频样本帧, 即  $i^{(K)} = f^{(K)}$ . 除顶层外, 其余各层的输入不仅包含音频样本  $B$ , 还包括上一层网络的输出, 它以条件向量的形式输入至本层. 因此, 总输入可以表示为  $i^{(k)} = W_x^{(k)} f^{(k)} + c^{(k+1)}$ . 其中,  $W_x^{(k)}$  为用于线性组合的系数矩阵;  $f^{(k)}$  为只包含样本  $B$  的大小为  $F^{(k)}$  的样本;  $c^{(k+1)}$  为上一层输出的条件向量. 在生成阶段, 本文只需将时序同步网络生成的结果  $A$  以同样的方式输入至训练好的网络中, 最终模型将通过自回归的方式生成音频, 得到音色增强后的样本.

这种修改后的方法在利用 SampleRNN 的层次结构学习音色的同时, 又为其增加了额外的条件信息, 建立了低质声音与高质声音之间的联系, 最终使低质声音的音色得到增强.

### 3 实验结果与分析











本文采用的实验平台为 Matlab2015 和 PyTorch 0.2.0.post3, 所有实验均在显卡型号为 NVIDIA Quadro 5000, 8GB 显存以及 64GB 内存的服务器上进行. 对实验所用数据集中的每一类视频, 其训练时长约 20h. 为验证提出的跨模态环境声音合成算法的有效性, 本文在不同数据集上进行训练与测试, 并与当前最新基于物理和基于学习的方法进行了定量与定性比较.

### 3.1 训练细节及评价标准

#### 3.1.1 数据集准备与训练参数

现有的视频数据库 AudioSet<sup>[27]</sup>并不适用于视频到声音的跨模态转换任务. 这是由于这类视频数据库中的视频混入了太多的诸如人讲话声音等类似的噪声, 在使用此类视频进行训练时, 往往会由于背景噪声的存在而无法取得理想的训练结果. 因此本文对现有视频集<sup>[10,12]</sup>进行筛选, 并从视频网站下载可用于训练的数据, 整合为包含分类更全的音视频跨模态数据集. 本数据集共包含 12 类环境声音, 共有 1697 段视频, 总时长为 81462 s. 表 1 展示了本文所使用的训练数据库的详细信息, 其中前 7 类具有明显的同步性(例如在狗张嘴时才会产生叫声). 对于每一类数据, 本文均随机选择 75% 的数据用于训练以及 25% 的数据用于测试. 在时序同步网络模型训练过程中, 优化策略采用 Adam 优化算法, 初始学习率设为 0.001. 在音色增强网络中, 网络参数与原始 SampleRNN 模型参数设置相同, 共包含 3 层 RNN 网络, 其中每层网络采用单层包含 1024 个隐藏单元的 GRU 网络, 最底层为具有 3 层网络的多层感知器.

表 1 音视频跨模态数据集统计信息

种类	视频样例	视频时长 t/s
烟花		10 128
狗吠		8 400
鸟鸣		3 792
布料		6 816
木头		5 184
碎石		5 616
塑料		7 584
海浪		7 200
雨声		5 721
飞机		7 007
火车		6 374
电锯		7 640

<http://www.videezy.com>, <https://www.youtube.com/>

### 3.1.2 评价标准

为评价合成音频的声音质量, 本文采用常用于 Internet 协议的音频质量评估算法: 音频质量感知评估(perceptual evaluation of audio quality, PEAQ)算法. PEAQ 算法利用人耳主观感知特性计算出声音信号的失真阈值, 然后采用人工神经网络融合出客观差异等级(objective difference grade, ODG)作为音频质量度量参数. ODG 具体计算方式为

$$ODG = h_{\min} + (h_{\max} - h_{\min})\text{sig}(D_I).$$

其中,  $h_{\min} = -3.98$ ;  $h_{\max} = 0.22$ ;  $\text{sig}(\cdot)$  为 Sigmoid 函数;  $D_I$  为失真指数, 由文献[28]中网络模型计算. ODG 值越大, 表示人耳对于其可接受程度越高, 即声音的音色质量越高.

为了评估声音结果的同步性, 本文通过用户调查的方式对同步性进行评价. 用户根据听觉感受对每一段视频及其对应音频进行打分, 分数为

1~10. 参与用户调查的总人数为 30 人, 每位参与者均具有正常听力.

### 3.2 实验对比

#### 3.2.1 基于物理的声音合成方法对比

基于物理的声音合成算法可以合成真实感较强的声音结果. 本文选择了具有代表性的 2 类声音: 具有明显同步性的撞击声<sup>[9]</sup>和不具有明显声音结构的雨声<sup>[4]</sup>进行比较. 图 4 展示了每组结果的视频帧序列和声音波形图. 从图 4a 可以看出, 通过本文算法合成的声音与通过基于物理方法合成的声音均可以与视频内容同步; 而文献[9]方法需要通过动画建模得到匹配的声音结果, 本文算法仅需视频内容即可生成相似同步效果的声音. 此外, 文献[4,9]均为单一声音合成方法, 而本文算法可以生成多类声音, 在很大程度上提高了可扩展性.

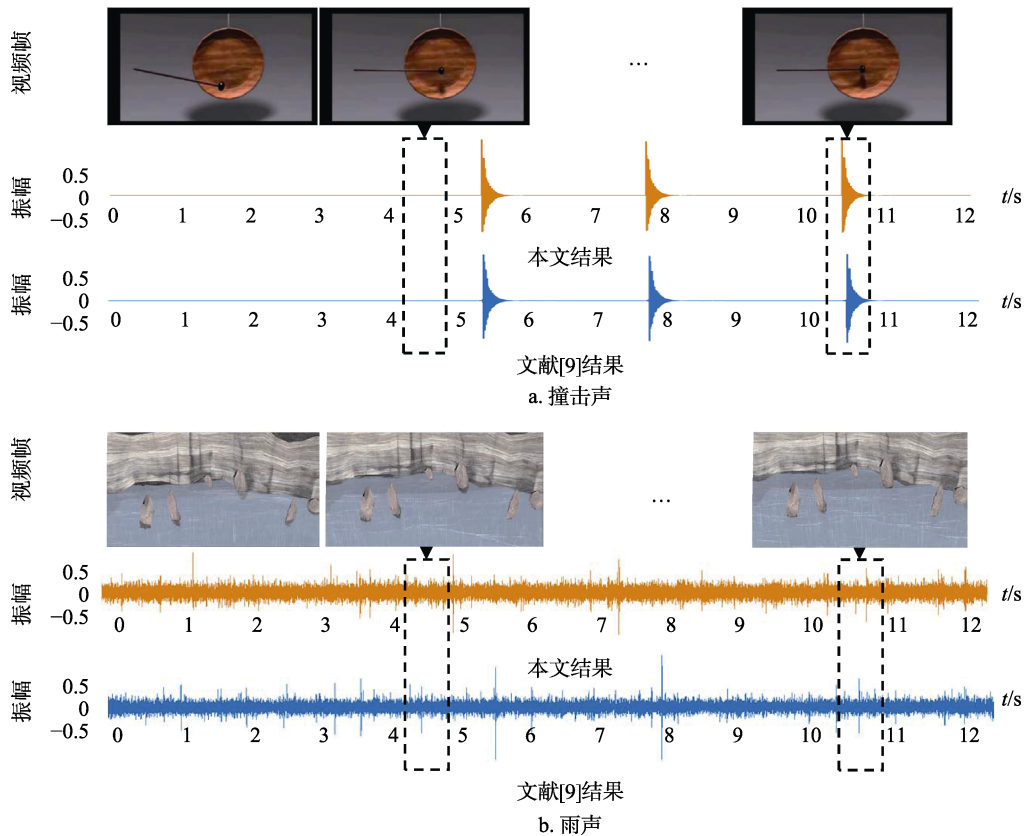


图 4 本文算法与不同基于物理的声音合成方法结果对比

#### 3.2.2 基于学习的声音合成方法对比

为了验证本文算法的同步性, 本文与 2 类最新的基于学习的声音合成算法<sup>[10,12]</sup>进行对比. 文献[10]中的声音合成算法无法直接合成声音, 只能通过网络合成声音特征谱图; 因此其扩展性较差, 仅适用于敲打声合成. 文献[12]实现了端到端的跨模

态声音合成, 然而其在同步性方面存在明显的局限性. 与上述 2 类方法的比较, 分别基于数据集 Greatest Hits<sup>[10]</sup>和 VEGAS<sup>[12]</sup>. 图 5 展示了每组结果的声音波形图. 如图 5a 所示, 本文算法在 Greatest Hits 数据集上训练可以获得与文献[10]相似的结果. 图 5b 展示了本文算法与文献[12]的对比结果, 可

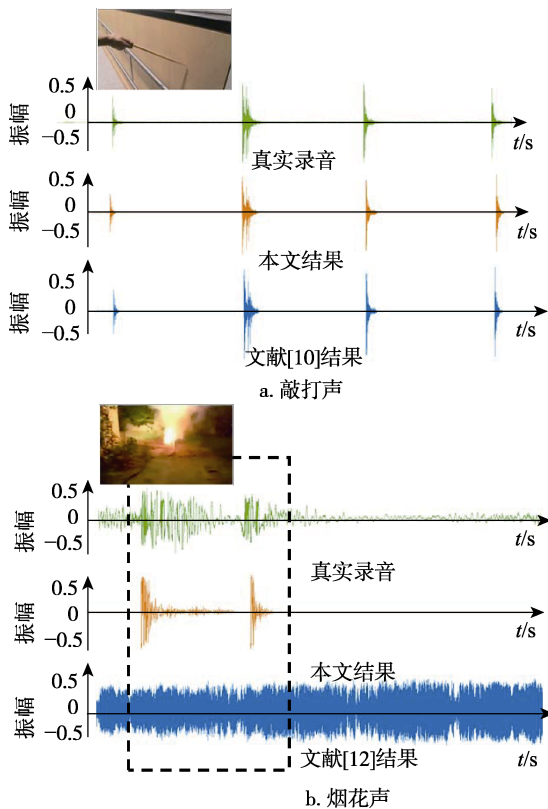


图 5 本文算法与真实声音和不同基于学习的声音合成方法结果对比

以看到本文算法合成的声音结果与真实声音的吻合度更高, 而文献[12]产生的声音结果则存在明显的异步性(如虚线框中所示).

表 2 展示了本文算法与上述 4 类方法在扩展性、同步性以及声音质量 3 个方面的对比结果, 可以观察到本文算法不仅可以合成多种类别声音, 同时在声音质量和同步性方面, 都可以与单一声音合成方法取得相似的性能表现.

表 2 不同方法扩展性与同步性对比

方法	合成声音类别	平均 ODG 值	平均主观同步得分
文献[9]	1(撞击声)	-0.1914	9.21
文献[4]	1(雨声)	-0.1917	8.97
文献[10]	1(敲打声)	-0.2057	8.36
文献[12]	10	-0.2210	6.89
本文	12	-0.2124	8.42

### 3.2.3 实验结果的局限性

本文算法存在一定的局限性, 即只能合成单一类别声音. 本文算法无法自动判别输入视频内容所属的类别, 因此需要对每类视频单独进行训练, 也只能合成与该类视频相符的单一类别声音. 当视频中同时存在多个发声主体时, 本文算法只能对其中一种声音进行合成, 无法对此种情况进

行有效处理. 本文的一个失败案例如图 6 所示, 当视频中同时存在海浪声和鸟叫声时, 难以对其声音进行有效的合成. 其中, 图 6b 所示为本文算法合成的声音, 只合成了海浪声; 图 6c 为真实声音, 可以看出除了海浪声外, 同时包含本文算法未能合成的鸟叫声. 因此, 本文算法只能合成某一特定种类声音, 无法有效地处理视频中存在多个发声主体的情况.

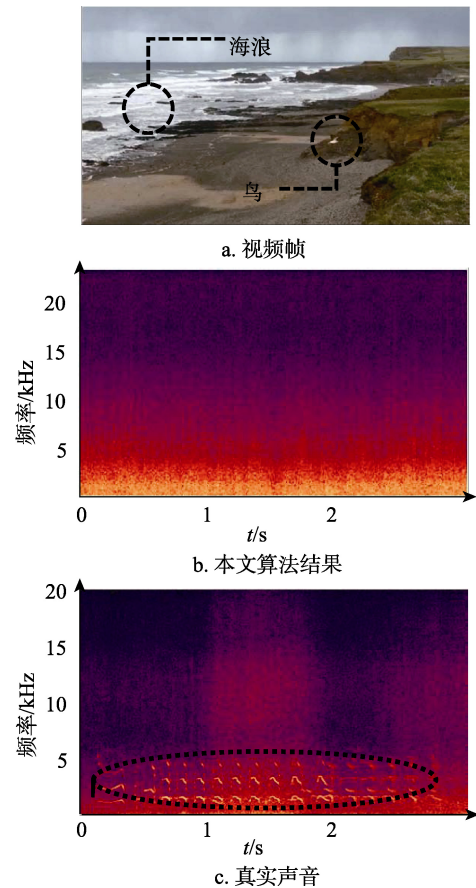


图 6 本文算法的一个失败案例

## 4 结 语

本文提出一种基于深度学习的环境声音合成算法, 实现了从视觉到听觉的跨模态环境声合成. 本文提出的框架由视觉特征编码器, 时序同步网络和音色增强网络共同组成, 可以端到端地合成与视频内容同步的音频. 实验结果表明, 通过本文算法合成的声音有较强的真实感. 与现有方法相比, 其具有更高的可扩展性和同步性.

尽管本文算法可以合成具有较强真实感的音频结果, 但是仍具有一些局限性. 首先, 本文的网络框架不具有类别识别功能, 因此需要对每一类



数据分别训练, 即本文算法无法自动理解视频内容所属类别. 因此, 如何将分类器与本文算法进行融合是未来的一个研究方向. 此外, 本文算法无法有效地处理多声源视频, 即视频中同时存在 2 个或以上的发声物体; 并且, 对于复杂声学效果(如混响、回声等)仍无法有效仿真. 因此, 如何基于深度学习技术合成复杂声场下的音频结果是接下来的研究方向.

## 参考文献(References):

- [1] Zheng C X, James D L. Harmonic fluids[J]. *ACM Transactions on Graphics*, 2009, 28(3): Article No.37
- [2] Moss W, Yeh H, Hong J M, *et al.* Sounding liquids: automatic sound synthesis from fluid simulation[J]. *ACM Transactions on Graphics*, 2010, 29(3): Article No.21
- [3] Langlois T R, Zheng C X, James D L. Toward animating water with complex acoustic bubbles[J]. *ACM Transactions on Graphics*, 2016, 35(4): Article No.95
- [4] Liu S G, Cheng H N, Tong Y Y. Physically-based statistical simulation of rain sound[J]. *ACM Transactions on Graphics*, 2019, 38(4): Article No.123
- [5] Dobashi Y, Yamamoto T, Nishita T. Synthesizing sound from turbulent field using sound textures for interactive fluid simulation[J]. *Computer Graphics Forum*, 2004, 23(3): 539-545
- [6] Chadwick J N, James D L. Animating fire with sound[J]. *ACM Transactions on Graphics*, 2011, 30(4): Article No.84
- [7] Yin Q, Liu S G. Sounding solid combustibles: non-premixed flame sound synthesis for different solid combustibles[J]. *IEEE Transactions on Visualization and Computer Graphics*, 2018, 24(2): 1179-1189
- [8] van den Doel K, Kry P G, Pai D K. Foleyautomatic: physicallybased sound effects for interactive simulation and animation[C]//*Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*. New York: ACM Press, 2001: 537-544
- [9] Ren Z M, Yeh H, Lin M C. Example-guided physically based modal sound synthesis[J]. *ACM Transactions on Graphics*, 2013, 32(1): Article No.1
- [10] Owens A, Isola P, McDermott J, *et al.* Visually indicated sounds[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2016: 2405-2413
- [11] Chen L L, Srivastava S, Duan Z Y, *et al.* Deep cross-modal audio visual generation[C]//*Proceedings of the Thematic Workshops of ACM Multimedia*. New York: ACM Press, 2017: 349-357
- [12] Zhou Y P, Wang Z W, Fang C, *et al.* Visual to sound: generating natural sound for videos in the wild[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2018: 3550-3558
- [13] Dobashi Y, Yamamoto T, Nishita T. Real-time rendering of aerodynamic sound using sound textures based on computational fluid dynamics[J]. *ACM Transactions on Graphics*, 2003, 22(3): 732-740
- [14] Donahue C, McAuley J, Puckette M. Synthesizing audio with generative adversarial networks[OL]. [2019-06-17]. <https://arxiv.org/abs/1802.04208v1>
- [15] Pascual S, Bonafonte A, Serra J. SEGAN: speech enhancement generative adversarial network[OL]. [2019-06-17]. <https://arxiv.org/abs/1703.09452>
- [16] Mirza M, Osindero S. Conditional generative adversarial nets[OL]. [2019-06-17]. <https://arxiv.org/abs/1411.1784>
- [17] Denton E L, Chintala S, Szlam A, *et al.* Deep generative image models using a Laplacian pyramid of adversarial networks[C]//*Proceedings of NIPS*. Cambridge: MIT Press, 2015: 1486-1494
- [18] Reed S, Akata Z, Yan X C, *et al.* Generative adversarial text to image synthesis[C]//*Proceedings of the 33rd International Conference on Machine Learning*. New York: ACM Press, 2016: 1060-1069
- [19] Isola P, Zhu J Y, Zhou T H, *et al.* Image-to-image translation with conditional adversarial networks[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2017: 1125-1134
- [20] Liu Yujie, Dou Changhong, Zhao Qilu, *et al.* Sketch based image retrieval with conditional generative adversarial network[J]. *Journal of Computer-Aided Design & Computer Graphics*, 2017, 29(12): 2336-2342(in Chinese)  
(刘玉杰, 窦长红, 赵其鲁, 等. 基于条件生成对抗网络的手绘图像检索[J]. *计算机辅助设计与图形学学报*, 2017, 29(12): 2336-2342)
- [21] Mehri S, Kumar K, Gulrajani I, *et al.* SampleRNN: an unconditional end-to-end neural audio generation model[OL]. [2019-06-17]. <https://arxiv.org/abs/1612.07837>
- [22] Li Yang, Liang Wei, Zhang Yinlong, *et al.* Automatic lumbar vertebrae recognition in intraoperative X-ray images based on hierarchical recurrent neural network[J]. *Journal of Computer-Aided Design & Computer Graphics*, 2019, 31(1): 132-140(in Chinese)  
(李杨, 梁伟, 张吟龙, 等. 基于层级循环神经网络的术中 X 线图像腰椎自动识别[J]. *计算机辅助设计与图形学学报*, 2019, 31(1): 132-140)
- [23] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[OL]. [2019-06-17]. <https://arxiv.org/abs/1409.1556>
- [24] Zhou C, Horgan M, Kumar V, *et al.* Voice conversion with conditional SampleRNN[OL]. [2019-06-17]. <https://arxiv.org/abs/1808.08311v1>
- [25] Ai Y, Wu H C, Ling Z H. SampleRNN-based neural vocoder for statistical parametric speech synthesis[C]//*Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. Los Alamitos: IEEE Computer Society Press, 2018: 5659-5663
- [26] Sotelo J, Mehri S, Kumar K, *et al.* Char2Wav: end-to-end speech synthesis[OL]. [2019-06-17]. <https://openreview.net/forum?id=B1VWyySKx>
- [27] Gemmeke J F, Ellis D P W, Freedman D, *et al.* AudioSet: an ontology and human-labeled dataset for audio events[C]//*Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. Los Alamitos: IEEE Computer Society Press, 2017: 776-780
- [28] de Lima A A, Freeland F P, de Jesus R A, *et al.* On the quality assessment of sound signals[C]//*Proceedings of the IEEE International Symposium on Circuits and Systems*. Los Alamitos: IEEE Computer Society Press, 2008: 416-419