软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2024,35(5):2165−2175 [doi: 10.13328/j.cnki.jos.007027]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

分层特征编解码驱动的视觉引导立体声生成方法*

王睿琦[1], 程皓楠[2], 叶 龙[2]

[1](媒介音视频教育部重点实验室 (中国传媒大学), 北京 100024)

[2](媒体融合与传播国家重点实验室 (中国传媒大学), 北京 100024)

通信作者: 叶龙, E-mail: yelong@cuc.edu.cn

摘 要: 视觉引导的立体声生成是多模态学习中具有广泛应用价值的重要任务之一, 其目标是在给定视觉模态信息及单声道音频模态信息的情况下, 生成符合视听一致性的立体声音频. 针对现有视觉引导的立体声生成方法因编码阶段视听信息利用率不足、解码阶段忽视浅层特征导致的立体声生成效果不理想的问题, 提出一种基于分层特征编解码的视觉引导的立体声生成方法, 有效提升立体声生成的质量. 其中, 为了有效地缩小阻碍视听觉模态数据间关联融合的异构鸿沟, 提出一种视听觉特征分层编码融合的编码器结构, 提高视听模态数据在编码阶段的综合利用效率; 为了实现解码过程中浅层结构特征信息的有效利用, 构建一种由深到浅不同深度特征层间跳跃连接的解码器结构, 实现了对视听觉模态信息的浅层细节特征与深度特征的充分利用. 得益于对视听觉信息的高效利用以及对深层浅层结构特征的分层结合, 所提方法可有效处理复杂视觉场景中的立体声合成, 相较于现有方法, 所提方法生成效果在真实感等方面性能提升超过 6%.

关键词: 立体声; 视觉引导的声音生成; 分层特征编解码; 多模态学习; 跳跃连接

中图法分类号: TP18

中文引用格式: 王睿琦, 程皓楠, 叶龙. 分层特征编解码驱动的视觉引导立体声生成方法. 软件学报, 2024, 35(5): 2165–2175. http://www.jos.org.cn/1000-9825/7027.htm

英文引用格式: Wang RQ, Cheng HN, Ye L. Visually Guided Binaural Audio Generation Method Based on Hierarchical Feature Encoding and Decoding. Ruan Jian Xue Bao/Journal of Software, 2024, 35(5): 2165–2175 (in Chinese). http://www.jos.org.cn/1000-9825/7027.htm

Visually Guided Binaural Audio Generation Method Based on Hierarchical Feature Encoding and Decoding

WANG Rui-Qi[1], CHENG Hao-Nan[2], YE Long[2]

[1](Key Laboratory of Media Audio & Video (Communication University of China), Ministry of Education, Beijing 100024, China)

[2](State Key Laboratory of Media Convergence and Communication (Communication University of China), Beijing 100024, China)

Abstract: Visually guided binaural audio generation is one of the important tasks with wide application value in multimodal learning. The goal of the task is to generate binaural audio that conforms to audiovisual consistency with the given visual modal information and mono audio modal information. The existing visually guided binaural audio generation methods have unsatisfactory binaural audio generation effects due to insufficient utilization of audiovisual information in the encoding stage and neglect of shallow features in the decoding stage. To solve the above problems, this study proposes a visually guided binaural audio generation method based on hierarchical feature encoding and decoding, which effectively improves the quality of binaural audio generation. In order to effectively narrow the heterogeneous gap that hinders the association and fusion of audiovisual modal data, an encoder structure based on hierarchical coding and fusion of audiovisual features is proposed, which improves the comprehensive utilization efficiency of audiovisual modal data in the encoding stage. In order to realize the effective use of shallow structural feature information in the decoding process, a decoder structure with a skip connection between different depth feature layers from deep to shallow is constructed, which realizes the full use of shallow

* 基金项目: 国家自然科学基金 (61971383, 62201524); 国家重点研发计划 (2021YFF0900504)
本文由“多模态协同感知与融合技术”专题特约编辑孙立峰教授、宋新航副研究员、蒋树强教授、王莉莉教授、申恒涛教授推荐.
收稿时间: 2023-04-10; 修改时间: 2023-06-08; 采用时间: 2023-08-23; jos 在线出版时间: 2023-09-11
CNKI 网络首发时间: 2023-12-15

detail features and depth features of audiovisual modal information. Benefiting from the efficient use of audiovisual information and the hierarchical combination of deep and shallow structural features, the proposed method can effectively deal with binaural audio generation in complex visual scenes. Compared with the existing methods, the generation performance of the proposed method is improved by over 6% in terms of realism.

Key words: binaural audio; visually guided audio generation; hierarchical feature encoding and decoding; multimodal learning; skip connection

视觉引导的立体声生成是指在给定视觉模态信息以及单声道音频模态信息的情况下, 生成符合视听空间感知一致性的立体声音频. 视觉引导的立体声生成作为多模态学习中一个关键的子任务, 主要针对难以获得声源精确位置和运动轨迹的立体声构建场景, 在历史影音资料修复、多媒体内容生成等领域具有广泛的应用[1−5].

现有视觉引导的立体声生成方法生成的立体声在包括内容、艺术氛围等方面已经可以在很大程度上满足观众思维认知一致性的要求, 但是在空间感知一致性上仍存在生成效果不理想的问题; 主要体现在面对复杂的真实场景时, 例如, 镜头快速运动、光影频繁变化的场景, 立体声生成效果会出现明显降低. 这种生成效果不佳的问题主要是由于两方面的原因造成的: (1) 视听信息利用率不足. 现有方法针对的场景相对简单, 真实场景中视听觉模态数据存在较多的变化, 使得模态间存在更为明显的异构鸿沟; 现有方法一般基于编解码器结构进行生成, 只将视觉模态信息嵌入到编码阶段的深层结构特征中, 这种只在深层引入的嵌入机制难以应对异构鸿沟更为明显的场景. (2) 忽视浅层结构特征的利用. 虽然现有的视觉引导的立体声生成研究还未涉足不同深度特征结构在生成中起到的作用, 但是在计算机视觉领域已有充分的相关研究证明, 浅层结构特征同样在生成任务中起到关键作用[6]. 现有方法聚焦深层结构特征, 忽略了浅层结构特征在立体声生成, 特别是生成音频质量提升中的潜在作用.

为了解决现有方法因编码阶段视听信息利用率不足、解码阶段忽视浅层特征导致的立体声生成效果不理想的问题, 本文提出了一种基于分层特征编解码的视觉引导的立体声生成方法, 有效提升了立体声生成的质量. 其中, 为了有效地缩小阻碍视听觉模态数据间关联融合的异构鸿沟, 提出了一种视听觉特征分层编码融合的编码器结构, 提高了视听模态数据在编码阶段的综合利用效率; 为了实现解码过程中浅层结构特征信息的有效利用, 构建了一种由深到浅不同深度特征层间跳跃连接的解码器结构, 实现了对视听觉模态信息的浅层细节特征与深度特征的充分利用. 得益于对视听觉信息的高效利用以及对深层浅层结构特征的分层结合, 本文方法可有效处理复杂视觉场景中的立体声合成, 在真实感生成效果方面, 本文方法相较于现有方法提升 6%. 同时, 实验证明了浅层结构特征在立体声生成上具有同样重要的作用, 多深度编解码结构在视觉引导的立体声生成领域具有良好的应用前景.

本文的贡献主要体现在以下 3 个方面.

(1) 提出了一种视听觉特征分层编码融合的编码器结构, 提高了视听模态数据在编码阶段的综合利用效率.

(2) 构建了一种由深到浅不同深度特征层间跳跃连接的解码器结构, 实现了对视听觉模态信息的浅层特征与深度特征的充分利用.

(3) 探索了不同深度结构对于视觉引导立体声生成效果的影响, 证明了浅层结构特征在立体声生成上具有和深层结构特征同样重要的作用.

本文第 1 节介绍了关于视觉引导的立体声生成的相关工作. 第 2 节叙述了本文提出的模型和方法. 第 3 节展示了本文方法与相关方法的对比实验结果和分析. 第 4 节提供了关于本文工作的总结和对未来工作的展望.

1 相关工作

本文聚焦的视觉引导的立体声生成作为一种跨模态学习研究, 不同于音频处理领域的立体声生成, 在输入信息中不包含声源运动信息, 只通过音频信息和对应的视频信息来构建双声道. 视觉引导的立体声生成是近年来刚刚兴起的课题, 具有较少受到音频类型影响, 适合处理互联网多媒体素材的特点, 在近些年的多模态学习研究中受到了广泛的关注. Gao 等人[7]为了解决立体声录音缺失的问题, 构造了双模态立体声数据集 FAIR-PLAY, 并设计了 Mono2Binaural 网络, 第 1 次实现了视觉信息引导下的基于单声道音频生成立体声. FAIR-PLAY 数据集也作为领域内最常用的数据集被广泛应用于之后的视觉立体声生成. 在此工作的基础上, Zhou 等人[8]提出了一种通过对视觉信息的整合来将声源分离引入, 实现了可以综合处理声音分离与生成的立体声生成模型. Li 等人[9]提出一种将立体声生成任务与翻转音频分类任务整合的多任务框架, 将双耳音频生成和翻转音频任务整合到一个统一的框架下, 生成较高质量的双耳音频. Parida 等人[10]研究了不同发声物体到麦克风的距离差异问题, 提出了一个基于层

级注意力机制的编解码器结构的视觉引导的立体声生成网络. Lu 等人[11]提出了一种自监督的音频空间化网络, 可以在给定相应的视频和单声道音频的情况下生成与 Gao 等人[7]相近的效果; Xu 等人[12]也提出了类似的方法. Lin 等人[13]基于前人的工作[14], 提出利用音频和视觉成分间的关系及左右声道的关联性, 使用自监督学习的方式对立体声进行视觉引导的生成, 以此减轻对真实数据标签的依赖. Garg 等人[15]通过搜索视频流中的几何特征来强化视觉引导的立体声生成效果. 上述方法在针对特定数据集的立体声生成上取得了一定的效果, 但是泛化性能相对有限, 难以应对光影变幻更为复杂的场景. 这是因为现有方法在面对具有异构鸿沟的视听觉模态数据时构建关联性的能力较差, 没有充分利用视听觉模态信息[16]. 现有方法只将视觉模态信息嵌入到编码阶段的深层特征中 (一般只嵌入到最后 1 层), 忽视了浅层特征在立体声生成中的作用. 这种视听觉模态信息关联性缺失、没有充分利用视听觉模态信息, 特别是浅层特征信息的问题, 限制了模型的生成性能.

2 本文方法

为了解决现有方法因视听信息利用率不足而导致的立体声生成效果不理想的问题, 本文提出了一种基于分层特征编解码的视觉引导的立体声生成方法, 有效提升了立体声生成的质量. 本文的方法框架如图 1 所示.

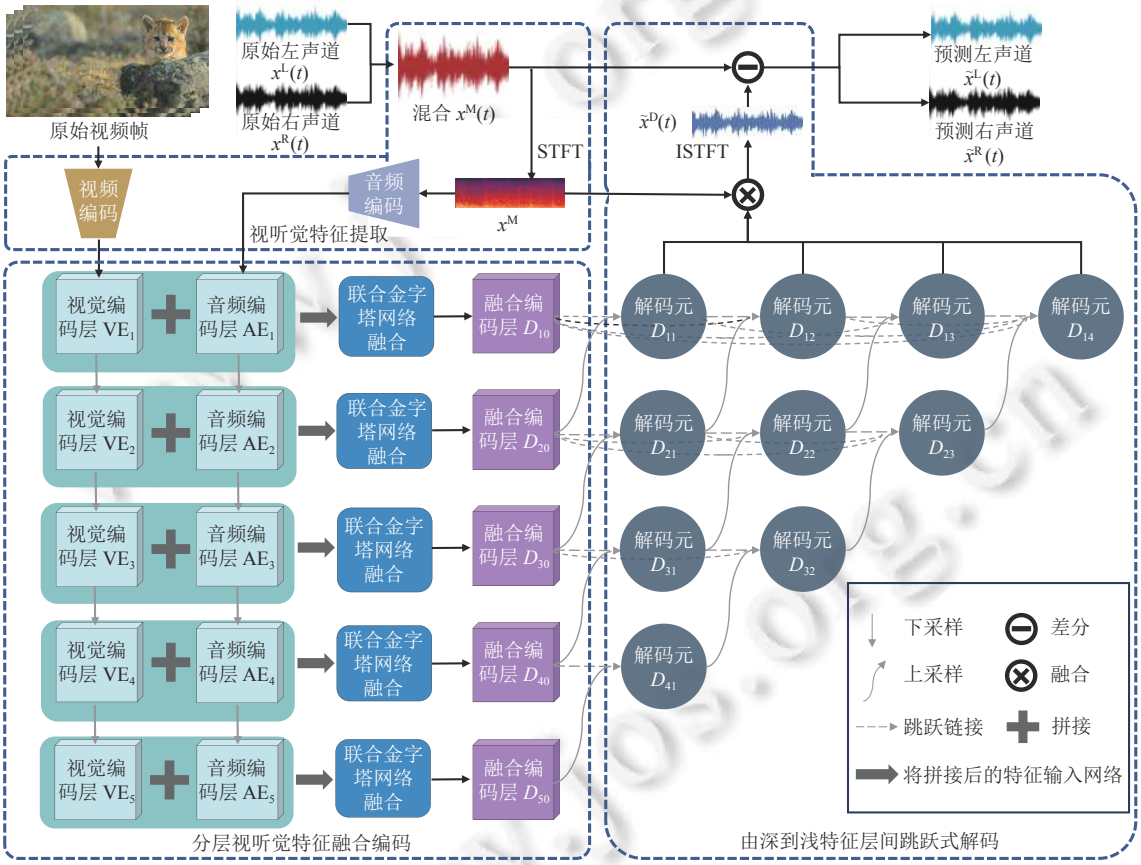

图 1 本文方法整体框架

图 1 中, 由输入的视频序列及单声道音频到最终结果输出的立体声音频, 共经历了包含视听觉特征提取、分层视听觉特征融合编码及由深到浅特征层间跳跃式解码在内的 3 个模块. 算法首先进行视听觉特征提取, 实现了对于输入的音视频序列的特征提取以及预处理. 接着, 在分层视听觉特征融合编码中, 对于提取到的视听觉模态特征进行了 5 层卷积操作, 并实现了对于这两种存在异构鸿沟的模态特征的多深度融合, 输出分层深度融合特征图. 最后, 在由深到浅特征层间跳跃式解码模块中, 通过一系列的反卷积及跳跃式融合操作, 实现了对于不同深度融合

模态特征信息的解码, 生成了最终的立体声音频.

2.1 视听觉特征提取

人耳对于立体声的感知主要基于双耳效应, 通过接收到的双声道音频的差异来定位空间中的声源. 因此, 视觉引导的立体声生成的关键是生成符合视听一致性的差异化的双声道音频. 已有的相关研究已经证明, 直接预测声源左右声道音频的效果较差; 因此本节采用了替代性方案, 通过左右耳声道共有部分来预测左右耳声道的差异部分, 进而求得左右声道信息. 同时, 由于时域音频信号信息与视觉信息难以对齐融合, 本节对于听觉模态信息进行了变换, 将其转换为频域图进行编码; 并对视觉模态信息的特征进行了初步的提取.

2.1.1 听觉特征提取

在经典的双工理论中[3], 人耳对于立体声的感知主要通过感知左右声道的时间差、相位差、声级差及音色差来为空间声源定位提供依据. 因此, 立体声生成的关键是如何基于单声道音频生成符合视听一致性的差异化的左右耳声道音频. 已有的研究表明, 直接合成双声道音频的效果一般不够理想[7–9]. 本文参考了文献 [17,18] 中的思路, 将声音到人耳的传播这一复杂的过程解构为两个独立阶段: (1) 声源发出单声道音频, 该声音经由介质传播到了听众双耳的位置. 在这一过程中, 相较于声源到听众的距离, 人耳距离是极为微小的, 因此此处的声音可以被视为混合的单声道音频. (2) 双声道音频被听众的双耳分别接收, 人脑根据双声道音频特性的细微差异进行编码. 建模这一双阶段的完整过程是极为复杂的, 因此本文采用简化的模型来实现这一过程. 具体过程如图 2 所示, 假设自然界中的声源的声音信号为 x_0, 人类左右耳听到的声音分别为 $\tilde{x}^{\mathrm{L}}$ 和 $\tilde{x}^{\mathrm{R}}$, 则 $\tilde{x}^{\mathrm{L}}$ 和 $\tilde{x}^{\mathrm{R}}$ 应该是由左右耳声道的共有部分 $\tilde{x}^{\mathrm{M}}$ 和其差异部分组成的; 其中, 左右耳声道的混合部分 $\tilde{x}^{\mathrm{M}}$ 也被称为黄金听感位置声音[17,18]. 左右耳声道的差异部分为本网络架构的预测目标.

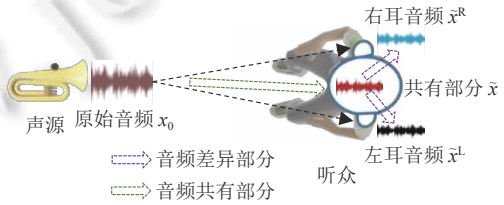

图 2 人耳对自然界音频的感知

本模块以样本中双耳音频的左右声道的混合部分作为网络输入, 并使端到端网络的优化目标为左右声道混合部分到左右声道差异部分的条件融合. 设输入的原始时域音频信号的左右声道的分别为 $x^{\mathrm{L}}(t)$ 与 $x^{\mathrm{R}}(t)$, 以左右声道的共同部分作为自监督任务的输入来求取预测的左右声道 $\tilde{x}^{\mathrm{L}}(t)$ 与 $\tilde{x}^{\mathrm{R}}(t)$, 即:

$$x^{\mathrm{M}}(t)=x^{\mathrm{L}}(t)\oplus x^{\mathrm{R}}(t) \tag{1}$$

其中, $\oplus$ 表示取相同部分. 考虑到直接使用 $x^{\mathrm{M}}(t)$ 预测左右声道依旧缺少必要的信息, 因此, 通过预测左右声道之间的差异 $\tilde{x}^{\mathrm{D}}(t)$ 来间接求取, $\tilde{x}^{\mathrm{D}}(t)$ 的计算方式为:

$$\tilde{x}^{\mathrm{D}}(t)=x^{\mathrm{L}}(t)\ominus x^{\mathrm{R}}(t) \tag{2}$$

其中, $\ominus$ 表示取差异部分.

为了更好地在同一层级将听觉模态特征与视觉模态特征对齐, 使用输入音频对应的复值时频谱作为音频编码网络的输入特征. 对于输入的音频 $x^{\mathrm{M}}(t)$, 首先通过短时傅里叶变换获得其时频谱 X^{M}. 为了更好地保留复值时频谱中的相位信息, 对复值时频谱的实部和虚部分别进行了处理. 接着, 为了满足卷积网络的维度要求, 在频谱图的特征矩阵最后一个维度上进行拼接操作, 并将该组特征作为输入传递给多深度视听觉特征融合编码模块, 用来求取左右声道差异 $\tilde{x}^{\mathrm{D}}(t)$ 的频谱图 $\tilde{X}^{\mathrm{D}}$. 最终, 求取左右声道的公式为:

$$\begin{cases}\tilde{x}^{\mathrm{L}}(t)=\dfrac{x^{\mathrm{M}}(t)\oplus\tilde{x}^{\mathrm{D}}(t)}{2}\\ \tilde{x}^{\mathrm{R}}(t)=\dfrac{x^{\mathrm{M}}(t)\ominus\tilde{x}^{\mathrm{D}}(t)}{2}\end{cases} \tag{3}$$

2.1.2 视觉特征提取

视觉特征提取的目的是提取视频流中的 RGB 色度空间信息, 用以指导声音空间化时的位置. 视觉特征提取如图 1 所示. 其中, 视觉编码通过在 ImageNet 数据集上训练的深度残差网络-18 (ResNet-18) 进行[19]. 通过检测输入网络的音频段的中心帧来定位需要提取的视觉特征的时间位置. ResNet-18 网络提取的特征的尺寸接着被调整到了与音频频谱特征相同的维度以方便后续的编码. 然后, 视觉特征被输入分层视听觉特征融合编码模块进行后续处理.

2.2 分层视听觉特征融合编码

视觉特征与听觉特征本身的差异导致其在构成上存在明显的异构鸿沟, 如何缩小视听觉模态间的异构差异, 促进不同模态间数据的融合是本节要解决的主要问题. 因此, 本节设计了分层视听觉特征融合编码模块, 通过将视听觉模态映射到不同维度, 在这些维度分别进行特征融合的方式来弥合异构鸿沟过大的问题.

2.2.1 特征编码网络

视觉引导的立体声生成任务可以被认为是一种特殊的声源分离任务[8]; 其本质是基于特定规则对音频信号频谱进行分解并进行重新组合. 虽然现有的视觉引导的立体声生成方法还未涉足从多重深度分层特征对音频信号进行分解重组的效果, 但是在对图像信息进行分解的图像分割领域已有充分的相关研究, 研究结果已经充分证明, 使用多重深度分层特征对信号进行分解重组的效果明显优于仅从单一的浅层特征或深层特征处理的效果[19]. 这是因为, 浅层结构可以提供频谱图的图像特征, 包括丰富细致的色彩、边界等; 而深层结构由于感受野的扩展, 可以通过更少的神经元来学习更宏观的组合特征; 在信号的分解与重组过程中, 浅层特征与深层特征拥有同样重要的地位, 综合使用才能达到更好的生成效果. 现有的视觉引导的立体声生成方法一般采用由 Ronneberger 等人[20]提出的 U-Net 网络的左侧收缩路径作为编码器结构, 这种结构只能提取特定深度听觉特征的感受野. 为了获取不同深度视听觉模态特征的感受野, 不同于 U-Net 网络, 本节设计了视觉流与听觉流并行处理的分层视听觉特征融合编码器模块. 在网络模型的编码过程中, 通过顺序进行多组堆叠卷积后接最大池化操作的方式, 不断增加输出的不同深度特征图数目, 增强特征层的感受野, 以此来记录并提取输入信息从浅层到深层的特征. 本节构建的多深度特征学习网络对于输入的视觉模态特征及听觉模态特征进行同样尺度的卷积操作, 以保证两种模态的特征被映射到了同样的维度. 整体来看, 编码器中的多重下采样可以保留多深度的特征信息, 实现对于浅层特征和深层特征的综合提取, 进一步整合视听觉的局部和整体特征信息, 更好地协助解码器生成立体声音频的细节信息, 保证生成音频具有更好的听觉表现.

本方案设计的编码器结构共包含 5 组以下采样编码为目的的视听觉数据收缩路径单元, 每组做包括步长为 2 的用于下采样的卷积层 (卷积核为 4×4)、批处理归一化层和 Sigmoid 激活函数. 采用 2×2 的最大池化层来去除信息冗余. 每组收缩路径单元对视听觉特征分别进行同样的处理.

2.2.2 特征融合

在完成不同深度视听觉特征层的提取之后, 接下来需要解决的是对这两种模态融合的问题. Sep-stereo[8]中的关联金字塔网络 (associative pyramid network) 的方法对于每一层的视听觉特征进行融合. 具体来说, 将视觉特征送入关联金字塔网络进行深度编码, 获得每个位置的视觉特征, 将其进行内核转换后, 与同层的听觉模态特征层进行拼接. 接着, 使用关联卷积对于拼接后的特征层进行处理, 并使用采用 2×2 的最大池化层来去除信息冗余, 得到最终的融合特征层, 将其输入多深度特征层间跳跃式解码.

2.3 由深到浅特征层间跳跃式解码

在获得了不同深度的视听觉模态融合信息之后, 如何更大限度地在立体声生成中综合利用这些信息是本节要解决的问题. 现有的视觉引导的立体声生成方法对于视觉信息的利用相对较少, 一般只将视觉信息使用在编解码器网络结构的部分解码层. 这种仅限于浅层或单一深层的信息映射, 导致这些方法在处理一些复杂多声源场景时的效果不够理想, 会出现视听觉信息的空间感不一致. 为此, 本节构建了一种基于不同深度特征层间跳跃连接的解码器结构, 实现了视听觉模态信息中不同分辨率特征图的细节特征与深度特征的融合.

UNet++结构右侧的展开路径 (expansive path)[21]在处理多层跳跃式解码时有比较好的表现. 但是, UNet++是处理医疗图像分割问题的方法, 结构不适用于本文的视觉引导的立体声生成任务. 不同于 UNet++, 本节设计了一个对拼接后的视听觉模态特征逐步扩张的解码器结构, 在第 1 列的解码元中引入了额外的降采样过程将视听觉融合特征压缩至与同层听觉特征相同的维度, 并在输出层进行了剪枝. 通过在上采样后引入解卷积操作来逐渐逼近最终生成的频谱图. 同时, 在同级别的特征图及下级特征图之间使用跳跃连接的方式来融合深层和浅层的特征信息, 使解码器网络可以更大化地利用编码层中的视听觉融合信息. 解码器的长链接式跳跃可以更好地保留在编码器卷积过程中损失的特征信息细节. 解码器的逐层上采样过程可以结合下采样过程所保存的各层信息, 并结合上采样及前序解码元的输入信息来还原细节信息, 逐步构建高精度频谱图. 解码器结构通过对深层信息和浅层信息的多重拼接, 充分利用视听觉模态的局部和整体特征信息, 保障生成声具有优异的空间表现.

具体来说, 本方案设计的解码器由 4 层共 10 个解码元组成. 第 i 列第 j 行的解码元 $D_{i,j}$ 都由其同级的解码元 $D_{i-1,j}$, 及更深层的解码元 $D_{i-1,j-1}$ 等通过残差链接得到. 具体来说, 可表现为:

$$D_{i,j}=\begin{cases}C(D_{i-1,j}), & j=0\\ C([[D_{i,k}]_{k=0}^{j-1},V(D_{i+1,j-1})), & j>0\end{cases} \tag{4}$$

其中, $C(\cdot)$ 表示由卷积层和线性修正单元激活函数组成的单元, $V(\cdot)$ 表示由接卷积层和最大池化层组成的单元, 以 $D_{1,2}$ 为例解释, $D_{1,2}$ 是由 $D_{1,0}$, $D_{1,1}$ 和解卷积后的 $D_{2,1}$ 沿第 2 个维度拼接之后, 再经过 1 次卷积和非线性修正单元得到. 每个解码元通过密集卷积块和密集跳跃连接构成, 采用 2×2 反卷积核实现上采样, 并卷积块内部包含 2 次 padding 为 1 的 4×4 卷积核和 1 个 ReLU 激活函数.

最终的解码器输出结果由最上层的 4 个解码元加权融合得到, 暨最终网络生成的频谱图为:

$$\bar{X}^{D}=\frac{\sum_{i=1}^{4}\mu_i\cdot D_{1,i}}{\sum_{i=1}^{4}\mu_i} \tag{5}$$

针对公式 (5) 中的解码元 $D_{1,1}$, $D_{2,1}$, $D_{3,1}$ 及 $D_{4,1}$ 本文选择将其加权后输出, 因为这样可以更好地保存在本文结构的频繁采样中损失的信息. 同时, 不同深度的特征层也会保留不同尺度的特征, 采用深浅层解码元结合的方式能够更好地保证对于这些不同特征的有效利用.

3 实验结果与分析

本文采用的实验平台为 Matlab 2019 和 PyTorch 1.13.1+cu117, 所有实验均在显卡型号为 NVIDIA Geforce RTX 3090, 24 GB 显存的服务器上进行. 为验证提出的跨模态环境声音合成算法的有效性, 本文在视觉引导的立体声生成方法最常用的 FAIR-PLAY 数据集[7]上进行训练与测试, 并与当前最新的视觉引导的立体声生成方法进行了定量与定性比较. 在本文立体声生成模型训练过程中, 采用了 Adam 优化算法作为优化策略, 初始学习率设为 5E–4. 在视觉编码网络中, 网络参数与在 ImageNet 上训练的原始 ResNet-18 模型参数设置相同.

3.1 训练细节及评价标准

3.1.1 数据集准备与训练参数

视觉引导的立体声生成方法最常用的数据集是开源数据集 FAIR-PLAY[7], 该数据集包含了约 100 GB 的音视频双模态数据, 其中的音频全部为立体声音频, 视频的帧率为 60 Hz, 分辨率为 1920×1080 . 所有数据被划分为了 1 871 条长度为 10 s 的视频片段及对应的立体声, 内容主要为室内的音乐场景. 考虑到 FAIR-PLAY 的音视频内容相对简单, 声源类型较为单一, 场景较为理想, 不适于训练出有更强鲁棒性、更好泛用性的模型, 因此, 本文从视频网站下载了可用于训练和测试的数据, 整合为声源类型及声源场景更全的音视频跨模态数据集. 下载数据包含比 FAIR-PLAY 更复杂的画面光影变幻以及镜头运动变化. 同时, 声音类型也更复杂. 所有下载的音视频数据都整合为了与 FAIR-PLAY 数据集相同的格式. 最终共有 2 200 条长度为 10 s 的视频片段及对应的立体声音频.

对于前述数据集, 本文随机选择 80% 的数据用于训练以及 20% 的数据用于测试和验证. 同时, 为了保证数据选择的随机性, 本文随机构建了 10 套数据集划分方法. 如非特别声明, 实验中所有结果为依次计算 10 套方法后的平均值. 接着, 在进行网络训练前, 对音视频数据样本进行了预处理. 对于视频数据, 对其按每秒 10 帧的帧率抽取图像, 并将其结合时间戳信息保存在对应的文件夹中. 对于音频数据, 考虑到目前模型无法处理长序列音频, 因此, 在网络训练加载数据时, 对其按照时间戳信息以 1.0 s 的时长进行切分, 按照时间顺序逐段处理. 同时, 选择该段音频中心时刻对应的视频图像帧组作为该段对应的视频数据. 另外, 为了保证音频信息的一致性, 对音频数据进行了归一化处理, 即对于当前处理的时长为 1 s 的立体声音频段 $x=[x^{\mathrm{L}},x^{\mathrm{R}}]$, 有:

$$\tilde{x}=\frac{k\cdot x}{\|x\|_2} \tag{6}$$

其中, k 为归一化后的期望均方值. $\tilde{x}$ 即为第 2.1.1 节的待进行短时傅里叶变换的音频时域信号. 另外, 在训练过程中, 为了避免浅层编码元出现过拟合, 公式 (5) 的 μ_1, μ_2, μ_3, μ_4 分别设定为: 0.2, 0.2, 0.3, 0.3.

3.1.2 评价标准

为了评价生成立体声的还原度, 本文依照 Mono2Binaural[7]、Ambisonics[22]等基于视觉的空间声生成规则里常用的做法测量了生成音频与原始音频的频谱距离 (STFT distance) 和包络线距离 (envelope distance). 其中, 频谱距离表示原始音频和生成音频的左右声道的复数频谱的欧氏距离, 计算方法为:

$$D_{\{\mathrm{STFT}\}}=\|X^{\mathrm{L}}-\tilde{X}^{\mathrm{L}}\|_2+\|X^{\mathrm{R}}-\tilde{X}^{\mathrm{R}}\|_2 \tag{7}$$

包络线距离表示原始音频和生成音频的左右声道的信号波形之间的欧氏距离, 计算方法为:

$$D_{\{\mathrm{ENV}\}}=\|E[x^{\mathrm{L}}(t)]-E[\tilde{x}^{\mathrm{L}}(t)]\|_2+\|E[x^{\mathrm{R}}(t)]-E[\tilde{x}^{\mathrm{R}}(t)]\|_2 \tag{8}$$

频谱距离和包络线距离可以分别从信号的频域和时域来衡量生成音频与原始音频相比的质量, 数值越大, 则相对于原始音频的质量越差.

另外, 为了评价立体声生成的声音质量, 本文选取了多种常用于音频质量评估的算法, 包括: 源失真比 (source-to-distortion ratio, SDR) 算法、源干扰比 (source-to-interference ratio, SIR) 算法、源伪影比 (source-to-artifact ratio, SAR) 算法、短时客观可懂度 (short-time objective intelligibility, STOI) 算法、扩展短时客观可懂度 (extended short-time objective intelligibility, ESTOI)[23]以及 MOSNet 方法[24]如前所述, 视觉引导的立体声生成可以被认为是一种特殊的声源分离任务, 因此, SIR、SDR、SAR 等常用于声源分离的评价方法也可以被应用到视觉引导的立体声生成算法评价中. MOSNet 方法[24]在音频生成领域使用较多, 该方法在验证音频质量上 (包括是否存在音频噪声、是否存在音频扭曲失真等) 具有与人类主观感受一致性较高的表现.

3.2 对比实验

为了验证本文模型生成的立体声音频的空间感, 本文与同领域内有代表性的 3 种最新方法进行了生成效果的对比, 包括: Mono2Binaural 方法[7]、Sep-stereo 方法[8]以及 BinauralAG 方法[9]. 所有方法均在本文数据集上进行了训练. 图 3 展示了本文生成的立体声与其他方法生成的立体声的对比, 同时, 为了更清晰对比, 原始音频的立体声也被展示在了最左侧. 这段音视频的主要内容为一位演员在画面右侧讲话, 其画面部分存在较大的光影变化, 同时, 拍摄的镜头并非固定镜头, 是一个持续进行位移的运动镜头. 可以看到, 这种声源的复杂运动及光影变幻会对现有方法的性能产生比较明显的影响; 而本文方法的鲁棒性更高, 保持了相对较好的生成效果. 表 1 展示了本文方法相较于其他 3 种方法在 $D_{\{\mathrm{STFT}\}}$ 以及 $D_{\{\mathrm{ENV}\}}$ 两项参数上的性能. 可以看到, 除 BinauralAG 方法外, 另外两种方法都与本文方法有较明显的差距. 本文方法相较于 BinauralAG 方法的差距较小, 但依旧有一定优势.

接着, 通过客观评价方法来进一步验证了本文方法生成音频的空间感. 在这个实验中, 原始音频被选为了参考, 通过各种不同的方法比较生成音频与参考音频间的差距. 如表 2 所示, 本文方法在大部分指标上相较于其他方法都有一定优势. 在 STOI 及 ESTOI 上, Sep-stereo 方法展现出了较好的性能, 这主要是源于 Sep-stereo 在整合了较为完整的声源分离模型, 在使用这些原本专门用来评价声源分离任务的指标时会在评分上有一定的优势. 整体来说, 本文方法相较于现有方法, 在客观指标对比上有大约 6% 的提升.

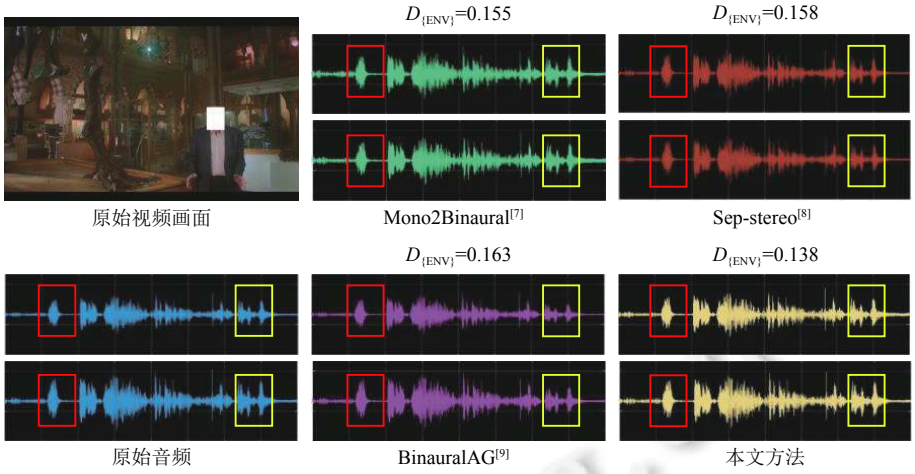

图 3 生成效果对比

表 1 基于频谱距离和包络线距离的音频生成评价实验

方法	频谱距离	包络线距离
Mono2Binaural[7]	0.959	0.141
Sep-stereo[8]	0.879	0.135
BinauralAG[9]	0.751	0.130
本文方法	**0.708**	**0.125**

表 2 基于客观评价方法的音频生成评价实验

方法	STOI	ESTOI	SAR	SIR	SDR
Mono2Binaural[7]	0.821	0.810	13.269	8.631	7.652
Sep-stereo[8]	**0.954**	**0.930**	16.215	10.543	9.619
BinauralAG[9]	0.935	0.924	**16.596**	**11.032**	**10.517**
本文方法	**0.952**	**0.947**	**18.491**	**12.871**	**11.849**

为了验证生成立体声音频的音频质量, 本文通过 MOSNet 这种评价方法对本文方法及相关方法进行了对比. 如图 4 所示, 各个方法在该项数值的差距相对较小. 这可以说明, 现有方法在生成不含有多余噪声、不存在明显信号失真等方面都有较好的表现. 在此基础上, 本文方法的效果相对较好.

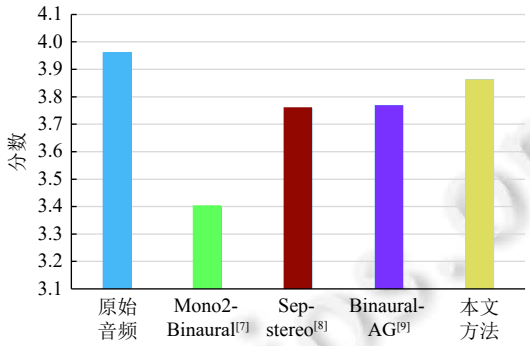

图 4 基于 MOSNet 评价方法的对比实验

接着, 通过用户学习的方式, 从主观评价的角度验证本文方法的有效性. 在具体实验中, 13 位听力正常的受试者被要求在完全安静的环境下佩戴耳机, 首先聆听原始的立体声音频, 接着聆听不同方法生成的立体声音频, 并被要求在聆听后以填写调查问卷的形式对于生成音频与画面的一致性, 以及与原始音频的相似性进行评价. 受试者在评价中可以选择认为效果较好的一种方法, 或者选择“中立”, 即难以分辨哪种方法更优秀. 综合之前的实验结果, 本节只选取了 3 种对比方法中性能更好的 BinauralAG 方法作为对比, 以保证受试者不会再聆听过多音频的过程中收到干扰. 统计结果如图 5 所示, 可以看到, 受试者整体对于立体声的评价有较为清晰的认知, 较少出现难以评价最好结果的情况. 综合来看, 本文方法在一致性与相似性上有较好的表现.

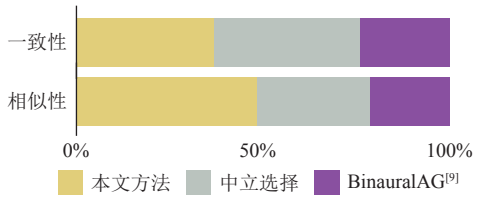

图 5 基于主观评价的音频生成效果对比实验

以上实验很好地说明了, 本文的分层编解码结构相较于现有方法的在各个方面都有优势, 也证明了浅层特征在立体声生成的优化中起到了与深层特征同样重要的作用, 综合利用浅层特征及深层特征对音频信号进行分解重组的效果明显优于仅从单一的深层特征处理的效果.

3.3 消融实验

为了验证本文方法各个环节的有效性, 本节进行了消融实验. 消融实验的验证包含了两个部分: 本文结构的必要性以及损失函数的有效性. 在验证本文结构的必要性方面, 主要研究的问题是现有结构的深度是否是当前任务的最优解. 对于本文的包含 5 层编码层和 10 个解码元的 4 层编码结构进行了扩展与删除, 构造了包含 3 层编码层和 3 个解码元的 2 层解码结构、包含 4 层编码层和 6 个解码元的 3 层解码结构以及包含 6 层编码层和 16 个解码元的 5 层解码结构, 如图 6 所示. 在验证损失函数的有效性方面, 主要研究的问题是现有损失结构是否是当前任务的最优解, 待对比的损失函数包括了 MSE 损失函数以及 L1 损失函数. 同时, 引入了在视觉引导的立体声生成中常用的遮掩结构[3], 来测试使用预测遮掩的方式能否获得生成效果的提升.

表 3 展示了本节消融实验的结果. 表 3 中的 MSE 和 L1 分别代表了损失函数的类型, U2 代表了 2 层解码结构. U4-M 代表了在 U4 结构的基础上在解码层加入遮掩. 可以看到, 随着网络深度的增加, 本文模型的立体声生成效果也在提升. 这种提升在超过 4 层解码结构时逐渐趋缓. 同时, 由于本文结构的特点, 5 层解码结构的计算复杂度相对过高, 对算力的要求过大, 这也导致了其对于生成音频效果的较小提升显得得不偿失. 引入遮掩并不能明显提升立体声生成质量, 这是因为, 本文方法全面地利用了深浅层结构特征, 更好地还原了双声道音频的细节特性, 这也使得加入遮掩的方式很难获得更明显的提升. 同时, 加入遮掩会明显提高网络计算量, 这也使得加入遮掩变得得不偿失. 另外, 可以看到, 在解码层数不变的前提下, MSE 损失函数的效果普遍优于 L1 损失函数的效果. 基于此实验, 本文的模型中选择了 4 层解码结构以及 MSE 损失函数.

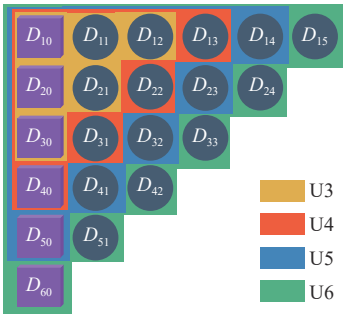

图 6 消融实验解码结构

表 3 针对不同结构及损失函数的消融实验

不同结构	频谱距离		包络线距离	
	L1	MSE	L1	MSE
U2	0.748	0.732	0.175	0.163
U3	0.695	0.717	**0.159**	0.156
U4-M	**0.705**	0.694	0.160	**0.150**
U5	0.711	**0.690**	0.153	0.152
U4	**0.704**	**0.692**	**0.158**	**0.151**

4 总 结

本文提出了一种基于分层特征编解码的视觉引导的立体声生成方法, 有效提升了立体声生成的质量. 其中, 为了有效地缩小阻碍视听觉模态数据间关联融合的异构鸿沟, 提出了一种视听觉特征分层编码融合的编码器结构, 提高了视听模态数据在编码阶段的综合利用效率; 为了实现解码过程中浅层结构特征信息的有效利用, 构建了一种由深

到浅不同深度特征层间跳跃连接的解码器结构, 实现了对视听觉模态信息的浅层细节特征与深度特征的充分利用. 得益于对视听觉信息的高效利用以及对深层浅层结构特征的分层结合, 本文方法可有效处理复杂视觉场景中的立体声合成, 相较于现有方法, 本文方法生成效果在真实感等方面有着超过 6% 的提升. 同时, 相关结论也证明了浅层结构特征在立体声生成上具有同样重要的作用, 多深度编解码结构在视觉引导的立体声生成问题上有较好的应用前景.

尽管本文方法可以生成质量较高的立体声音频, 但是仍存在一定的局限性. 首先, 本文方法计算复杂度相对较高, 对设备算力有着相对高的要求. 考虑到视觉引导的立体声生成在应用场景上一般不对实时性有要求, 本文方法通过可接受范围内的计算复杂度的提高, 获得了较为明显的生成质量提高. 另外, 本文方法在处理特定影视素材的立体声生成时还难以达成理想的效果, 例如在处理含旁白的立体声时, 本文方法无法提取并分离特定成分的音频. 这种对特定类型音频自适应的立体声化是本文接下来要解决的问题.

References:

[1] Yang Y, Zhan DC, Jiang Y, Xiong H. Reliable multi-modal learning: A survey. Ruan Jian Xue Bao/Journal of Software, 2021, 32(4): 1067–1081 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6167.htm [doi: 10.13328/j.cnki.jos.006167]

[2] Ge XL. Influence of audiovisual congruency on the auditory intensity change judgment [MS. Thesis]. Shanghai: Shanghai Jiao Tong University, 2011 (in Chinese with English abstract).

[3] Lv ZL. Study on generation of spatial audio using audio-visual cues [MS. Thesis]. Chongqing: Chongqing University of Posts and Telecommunications, 2021 (in Chinese with English abstract). [doi: 10.27675/d.cnki.gcydx.2021.000729]

[4] Cheng HN, Li SJ, Liu SG. Deep cross-modal synthesis of environmental sound. Journal of Computer-aided Design & Computer Graphics, 2019, 31(12): 2047–2055 (in Chinese with English abstract). [doi: 10.3724/SP.J.1089.2019.17906]

[5] Wang RQ, Cheng HN, Ye L, Qi QT. Reproduction transformation rule-based sound generation for film soundtrack. Journal of Computer-aided Design & Computer Graphics, 2022, 34(10): 1524–1532 (in Chinese with English abstract). [doi: 10.3724/SP.J.1089.2022.19727]

[6] Huang HM, Lin LF, Tong RF, Hu HJ, Zhang QW, Iwamoto Y, Han XH, Chen YW, Wu J. UNet 3+: A full-scale connected UNet for medical image segmentation. In: Proc. of the 2020 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. Barcelona: IEEE Computer Society Press, 2020. 1055–1059. [doi: 10.1109/ICASSP40776.2020.9053405]

[7] Gao RH, Grauman K. 2.5D visual sound. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE Computer Society Press, 2019. 324–333. [doi: 10.1109/CVPR.2019.00041]

[8] Zhou H, Xu XD, Lin DH, Wang XG, Liu ZW. Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer Press, 2020. 52–69. [doi: 10.1007/978-3-030-58610-2_4]

[9] Li SJ, Liu SG, Manocha D. Binaural audio generation via multi-task learning. ACM Trans. on Graphics, 2021, 40(6): 243. [doi: 10.1145/3478513.3480560]

[10] Parida KK, Srivastava S, Sharma G. Beyond mono to binaural: Generating binaural audio from mono audio with depth and cross modal attention. In: Proc. of the 2022 IEEE/CVF Winter Conf. on Applications of Computer Vision. Waikoloa: IEEE Computer Society Press, 2022. 2151–2160. [doi: 10.1109/WACV51458.2022.00221]

[11] Lu YD, Lee HY, Tseng HY, Yang MH. Self-supervised audio spatialization with correspondence classifier. In: Proc. of the 2019 IEEE Int'l Conf. on Image Processing (ICIP). Taipei: IEEE Computer Society Press, 2019. 3347–3351. [doi: 10.1109/ICIP.2019.8803494]

[12] Xu XD, Zhou H, Liu ZW, Dai B, Wang XG, Lin DH. Visually informed binaural audio generation without binaural audios. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE Computer Society Press, 2021. 15480–15489. [doi: 10.1109/CVPR46437.2021.01523]

[13] Lin YB, Wang YCF. Exploiting audio-visual consistency with partial supervision for spatial audio generation. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence, the 33rd Conf. on Innovative Applications of Artificial Intelligence, the 11th Symp. on Educational Advances in Artificial Intelligence. AAAI Press, 2021. 2056–2063. [doi: 10.1609/aaai.v35i3.16302]

[14] Rachavarapu KK, Aakanksha A, Sundaresha V, Rajagopalan AN. Localize to binauralize: Audio spatialization from visual sound source localization. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE Computer Society Press, 2021. 1910–1919. [doi: 10.1109/ICCV48922.2021.00194]

[15] Garg R, Gao RH, Grauman K. Geometry-aware multi-task learning for binaural audio generation from video. In: Proc. of the 32nd British Machine Vision Conf. BMVA Press, 2021. 1082–1092.

[16] Cao JJ, Nie ZB, Zheng QB, Lü GJ, Zeng ZX. State-of-the-art survey of cross-modal data entity resolution. Ruan Jian Xue Bao/Journal of

Software, 2023, 34(12): 5822–5847 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6764.htm [doi: 10.13328/j.cnki.jos.006764]

[17] Leng YC, Chen ZH, Guo JL, Liu HH, Chen JW, Tan X, Mandic DP, He L, Li XY, Qin T, Zhao S, Liu TY. BinauralGrad: A two-stage conditional diffusion probabilistic model for binaural audio synthesis. In: Proc. of the 36th Conf. on Neural Information Processing Systems. New Orleans: Curran Associates, 2022. 23689–23700.

[18] Wightman FL, Kistler DJ. The dominant role of low-frequency interaural time differences in sound localization. The Journal of the Acoustical Society of America, 1992, 91(3): 1648–1661. [doi: 10.1121/1.402445]

[19] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE Computer Society Press, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

[20] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Proc. of the 18th Int'l Conf. on Medical Image Computing and Computer-assisted Intervention. Munich: Springer Press, 2015. 234–241. [doi: 10.1007/978-3-319-24574-4_28]

[21] Zhou ZW, Siddiquee MMR, Tajbakhsh N, Liang JM. UNet++: A nested U-Net architecture for medical image segmentation. In: Proc. of the 4th Int'l Workshop and the 8th Int'l Workshop Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Granada: Springer Press, 2018. 3–11. [doi: 10.1007/978-3-030-00889-5_1]

[22] Morgado P, Vasconcelos N, Langlois T, Wang O. Self-supervised generation of spatial audio for 360° video. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montréal: Curran Associates, 2018. 360–370.

[23] Valin JM. A hybrid DSP/deep learning approach to real-time full-band speech enhancement. In: Proc. of the 20th IEEE Int'l Workshop on Multimedia Signal Processing. Vancouver: IEEE Computer Society Press, 2018. 1–5. [doi: 10.1109/MMSP.2018.8547084]

[24] Lo CC, Fu SW, Huang WC, Wang S, Yamagishi J, Tsao Y, Wang HM. MOSNet: Deep learning-based objective assessment for voice conversion. In: Proc. of the 20th Annual Conf. of the Int'l Speech Communication Association. Graz: Springer Press, 2019. 1541–1545. [doi: 10.21437/Interspeech.2019-2003]

附中文参考文献:

[1] 杨杨, 詹德川, 姜远, 熊辉. 可靠多模态学习综述. 软件学报, 2021, 32(4): 1067–1081. http://www.jos.org.cn/1000-9825/6167.htm [doi: 10.13328/j.cnki.jos.006167]

[2] 葛小立. 视听一致性对声音响度变化判断的影响 [硕士学位论文]. 上海: 上海交通大学, 2011.

[3] 吕柱良. 利用音视频信息的空间音频生成技术研究 [硕士学位论文]. 重庆: 重庆邮电大学, 2021. [doi: 10.27675/d.cnki.gcydx.2021.000729]

[4] 程皓楠, 李思佳, 刘世光. 深度跨模态环境声音合成. 计算机辅助设计与图形学学报, 2019, 31(12): 2047–2055. [doi: 10.3724/SP.J.1089.2019.17906]

[5] 王睿琦, 程皓楠, 叶龙, 齐秋棠. 基于还音转换规则的胶片音频生成方法. 计算机辅助设计与图形学学报, 2022, 34(10): 1524–1532. [doi: 10.3724/SP.J.1089.2022.19727]

[16] 曹建军, 聂子博, 郑奇斌, 吕国俊, 曾志贤. 跨模态数据实体分辨研究综述. 软件学报, 2023, 34(12): 5822–5847. http://www.jos.org.cn/1000-9825/6764.htm [doi: 10.13328/j.cnki.jos.006764]

王睿琦(1992－), 男, 博士生, 主要研究领域为电影胶片修复, 跨模态音频生成.

叶龙(1983－), 男, 博士, 教授, 博士生导师, CCF专业会员, 主要研究领域为智能音视频处理技术, 虚拟现实技术.

程皓楠(1994－), 女, 博士, 副研究员, 主要研究领域为声音建模, 跨模态声音合成.