基于时序对齐视觉特征映射的音效生成方法

摘　要　视觉引导的音效生成旨在提取无声视频的视觉特征，构建视觉特征引导模型，生成与视频内容一致、时序对齐的高质量声音效果。但是，目前方法生成的音效仍然存在保真度低、时序对齐效果差等问题。为此，本文提出了一种基于时序对齐视觉特征映射的音效生成方法，通过时序窗口聚合视觉特征，增强生成音效的对齐效果，并构建跨模态特征映射网络，实现视觉引导的高质量音效生成。首先，设计基于时序约束的特征聚合窗口，将视频序列滑动整合为视觉特征集合；然后，构建时空匹配的跨模态视音频特征映射网络，将视觉特征集合转换为多频段音频特征集合；最后，采用音频解码器将音频特征解码为梅尔频谱，再使用声码器将其转换为最终波形。实验结果表明，本文提出的新方法对比目前最先进的方法，在保真度和对齐效果方面，均有显著提升。

关键词　音效生成；跨模态；特征映射；自编码器；时序对齐

Sound Generation Method based on Timing-aligned Visual Feature Mapping

Abstract　Visually guided sound generation aims at generating high-quality sound matching to silent videos in content and timing alignment, through extracting the visual features and building a visual feature-guided model. However, existing methods can just generate sounds with low fidelity or poor timing alignment. In order to cope with these problems, in this paper, we propose a sound generation method based on timing-aligned visual feature mapping, which aggregates visual features through temporal windows to enhance the timing-alignment of generated sounds, and constructs a cross-modal feature mapping network to achieve high-quality sound generation. Firstly, we design a feature aggregation window based on timing constraints, which is used to integrate the video sequence into visual feature sets. Secondly, the visual feature sets are transformed into multi-frequency audio feature sets by a cross-modal feature mapping network for spatio-temporal matching. Finally, we use an audio decoder to obtain mel-spectrogram, and a vocoder to output the final waveform. The experimental results show that the proposed method in this paper has a significant improvement in fidelity and timing-alignment, compared with the state-of-the-art methods.

Key words　Sound Generation; Cross-modal; Feature Map; Autoencoder; Timing Alignment

1 引言

视觉引导的音效生成是指针对无声视频片段，构建视觉特征引导模型，生成时序匹配、内容一致的声音效果。例如，给定一段无声的烟花视频，通过视觉特征引导的音效生成技术，可以自动生成与其发声点匹配的爆炸音效。视觉引导的音效生成可广泛应用于影视制作、互动娱乐、幼儿教育等多种视音频内容创作场景，借助这种智能化处理方式能够进一步提升创作质量与效率。

2016 年，Owens 等[1]首次定义了视觉引导的音效自动生成任务，使用卷积神经网络（CNN）和长短期记忆网络（LSTM），提取视觉信息并传递为音频包络特征，结合基础样例自动生成匹配视频内容的鼓槌敲击、刮擦声音。自此，音效生成任务受到国内外研究者的广泛关注，目前的方法主要包括：基于跨模态编解码网络的方法和基于生成对抗网络的方法。

基于跨模态编解码网络的音效生成方法[2][3][4]通常构建视频编码模块，提取无声视频中的语义特征与时序特征，并传递至音频解码模块，结合基础样例或采用自回归模型将其解码为梅尔频谱或波形。这类方法可以推理生成连贯且保真度较高的音效，但大部分模型时序重建能力较差，生成音效与视频对齐效果不佳。基于生成对抗网络的音效生成方法[5][6][7][8]在视频编码模块提取视觉特征之后，构建生成对抗网络，预测音效的梅尔频谱。这类方法可以较为准确地传递视频的时序信息，但生成的音效噪声高、保真度较低。

为了实现时序匹配、高保真的音效生成，本文提出基于时序对齐视觉特征映射的音效生成方法。该方法首先利用 ResNet-18 模型[9]提取视频每帧对应的视觉特征；然后设计特征聚合窗口，按照固定长度和滑动步长，对视觉特征序列滑动分割并整合，从而获取每个窗口的视觉特征集合；接着，构建时空匹配的跨模态特征映射网络，将每个聚合窗口的视觉特征集合转换为多频段音频特征；最后，借助音频自编码器中的解码模块，将多频段音频特征集合解码为梅尔频谱，再采用预先训练的声码器转换为最终音频波形。本文在 VAS 数据集[6]上进行了大量实验，结果表明，新方法可以有效提升时序对齐效果及音效保真度。总之，本文的主要贡献如下：

（1）提出一种新的视觉引导音效自动生成框架，该框架由视觉特征聚合、视音频跨模态特征映射、音频解码三个模块构成，该框架可以有效提升生成音效的保真度及时序对齐效果；

（2）设计基于时序约束的特征聚合窗口，采用等长的滑动窗口对视觉特征序列进行分割与整合，建立视音频特征间的时序对齐约束，提升生成音效的对齐效果；

（3）构建时空匹配的视音频跨模态特征映射网络，将每个聚合窗口的视觉特征集合转换为多频段音频特征，并定义跨模态时空相关性损失函数，进一步增强网络的映射能力，进而提升生成音效的保真度。

2 相关工作

2.1 音效自动生成

现有基于深度学习的音效生成主要分为基于跨模态编解码网络和基于生成对抗网络两种。

2016 年，Owens 等[1]首次提出了一种音效自动生成方法，构建深度神经网络模型，生成与视频匹配的鼓槌敲击、刮擦声音。2018 年，Zhou 等[2]提出一种基于 SampleRNN[10]的方法直接从视频中提取内容信息生成音频波形。同年，Chen 等[11]提出基于感知优化分类的音频生成网络（POCAN），先识别声音的类别再生成对应的音效，并采用声音分类网络 SoundNet[12]计算感知损失，从而对齐语义信息。2020 年，Ghose 等[3]基于残差网络[9]（ResNet50）、全连接长短期记忆网络（FS-LSTM）提出两个不同的视频分类模型，在预测音效类别的同时计算各类别的基础样本，结合预测类别的基础样本和视觉信息生成梅尔频谱，再采用短时逆傅里叶变换（ISTFT）转换为音频波形。2021 年，Iashin 等[4]提出使用 GPT-2[13]完成视音频间的特征映射，引入音频自编码器将音频特征解码为梅尔频谱，采用声码器转换为最终波形。这些方法生成的音效保真度较高，但模型时序传递能力较弱，无法生成与视频对齐的音效。针对此问题，本文设计基于时序约束的特征聚合窗口，改善模型时序传递能力，从而提升生成音效的对齐效果。

2020 年，Ghose 等[5]在 AutoFoley 的基础上再次引入 BigGAN[14]提高了音效生成器的性能，提升了生成音效的质量与效率。同年，Chen 等[6]提出 Regnet 网络框架，使用时序分割网络（TSN）[15]从

视频帧中提取内容和运动信息，融合音频调节器，控制与视频内容不相关的声音分量，引入生成对抗网络（GAN）[16]生成与视频对齐的音效梅尔频谱，最后采用 WaveNet[17]将其转换为音频波形。2021 年，Liu 等[7]提出一个端到端的音效生成模型 V2RA，按视音频采样频率间的比例对视频特征进行抽取，再将其输入生成对抗网络（GAN）[16]中预测对应音效。同年，Ghose 等[8]提出基于分类条件生成对抗网络的音效生成框架 FoleyGAN，首先提取视频特征预测动作的类别及相似动作发生的概率，再将其传入 BigGAN[14]中预测生成音效的声谱图，最后采用短时逆傅里叶变换（ISTFT）转换为音频波形。这类方法可以较为准确地传递时序信息，但生成的音效保真度较低。本文构建时空匹配的跨模态特征映射网络，并融合音频自编码器，将提升生成结果的保真度。

2.2 音效生成数据集

2016 年，Owens 等[1]录制了 977 个视频，其中包含 46577 个鼓槌刮擦、敲击的动作，整理并公开为 The Greatest Hits Dataset 数据集。2018 年，Zhou 等[2]在 AudioSet[18]的基础上提出包含水流、烟花、直升机等 8 类音效的 VEGAS 数据集，共获得 28109 个视频片段，每个视频片段的时长约为 7 秒。2020 年，Ghose 等[3]提出了包含切菜、脚步、键盘等 12 类音效的 AFD 数据集，其中共包含 1000 个时长 5 秒的视频片段。Chen 等[6]在 AudioSet[18]和 VEGAS[2]的基础上提出包含烟花、枪声、喷嚏等 8 类音效的 VAS 视音频数据集，共 13008 个视频片段，每个视频片段的时长约为 7 秒。VAS 数据集是目前音效生成领域中动作音效类别最丰富的公开数据集，因此，本文将在此数据集上完成相关实验。

3 音频编解码器

本文音效生成方法需要预训练一个音频编解码器 VQGAN[19]，具体由编码器 E，解码器 G 和生成器 D 三个模块构成，网络结构如图 1 所示。

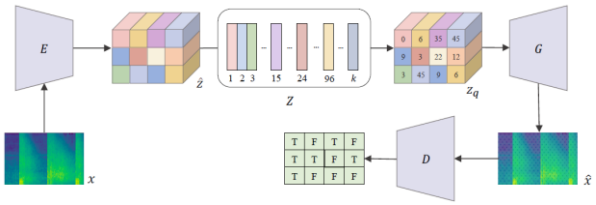

图 1 VQGAN 网络结构

首先，本文构建结构对称的编码器 E 和解码器 G，以及由 k 个音频特征矢量 z_q 组成的特征空间 Z。然后，使用编码器 E 将原始梅尔频谱 x 编码为特征向量集合 $\hat{z}$，通过量化计算 $q(\cdot)$，为每个特征向量 $\hat{z}$ 寻找其在特征空间 Z 中距离最近的特征矢量 z_q：

$$z_q = q(\hat{z}) := \left(\arg\min_{z_k \in Z} \left\| \hat{z}_{ij} - z_k \right\| \right) \in \mathbb{R}^{h \times w \times n} \tag{1}$$

其中，h 和 w 为特征向量集合 $\hat{z}$ 的高度和宽度；n 为特征向量的维度。最后，使用解码器将 z_q 解码为梅尔频谱 $\hat{x}$，利用局部鉴别器 D，对原始梅尔频谱 x 和生成梅尔频谱 $\hat{x}$ 分区域判别。

本文使用 VAS 数据集[6]中全部音频的梅尔频谱，对编码器 E、解码器 G、鉴别器 D 和特征空间 Z 进行联合训练，最终获得包含 128 个 256 维音频特征矢量 z_q 的特征空间 Z。根据该音频特征空间，可以将分辨率为 848×80 的梅尔频谱编码为 53×5 个有序特征矢量 z_q。本文将利用预先训练的编码器 E，对原始音频的梅尔频谱 x 进行编码，获得跨模态特征映射网络的音频特征真实值，并采用解码器 G，将跨模态特征映射网络的预测结果，解码为梅尔频谱 $\hat{x}$。

4 视觉引导的音效生成

4.1 整体网络框架

基于时序对齐视觉特征映射的音效生成方法主要由视觉特征聚合、时序对齐特征映射及音频解码三个模块组成，整体网络框架如图 2 所示。

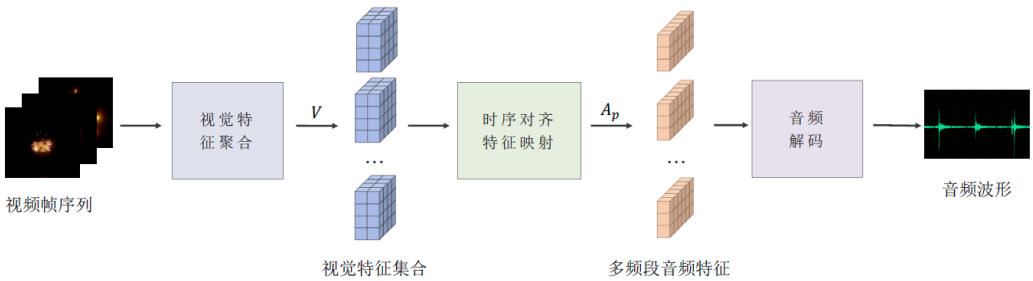

图 2 基于时序对齐视觉特征映射的音效生成方法整体框架

首先，使用 ResNet-18 预训练模型[9]提取视频每一帧的视觉特征，设计时序等长的特征聚合窗口，按照固定的步长对视觉特征序列滑动分割，并将窗口中的相邻视频帧特征整合为一组视觉特征集合V 。

接着，将视觉特征集合V 输入时空匹配的跨模态视音频特征映射网络。先利用映射网络的空域特征整合模块，获取视觉特征集合V 中目标对象的运动信息；再采用映射网络的频域特征转换模块，将物体的运动信息按照频率高低分别转换为多个频段的音频特征，整合并输出为多频段音频特征 A_p 。

最后，将音频特征集合 A_p 输入音频解码模块，从音频特征矢量空间中，查找出与多频段音频特征最邻近的音频矢量；并采用预先训练的 VQGAN 解码器，将完整视频对应的音频矢量集合解码为梅尔频谱；利用预先训练的声码器，将梅尔频谱解码为最终音频波形。

4.2 基于时序约束的视觉特征聚合

由于视频和音频的采样率不同，过去的工作通常复制视频帧，实现视音频间的时序一致。这种方法增加了模型的计算量，因此，本文根据视音频持续时间相同的特点，设计基于时序约束的特征聚合窗口，设置时序等长的滑动窗口对视觉特征序列进行分割与整合，建立视音频特征间的时序对齐约束。

首先，使用在 ImageNet 数据集[20]上预先训练的 ResNet-18[9]，提取视频每帧对应的视觉特征，保存其中第四层维度为 512×7×7 的特征信息，在提取视频语义信息的同时保留了帧内的空间信息。考虑到音效特征与局域视觉信息相关性较强，因此，设计长度为d ，步长为l的特征聚合滑动窗口，对视觉特征序列 F 进行分割与整合，获得 n 组视觉特征 V_t：

$$V_t = F_{It:(It+d)}, t \in [0, n] \tag{2}$$

再将每一组特征集合中的帧特征按照帧序由上至下依次拼接，获得 $512\times7\,d\times7$ 特征集合，视觉特征聚合方法如图 3 所示。

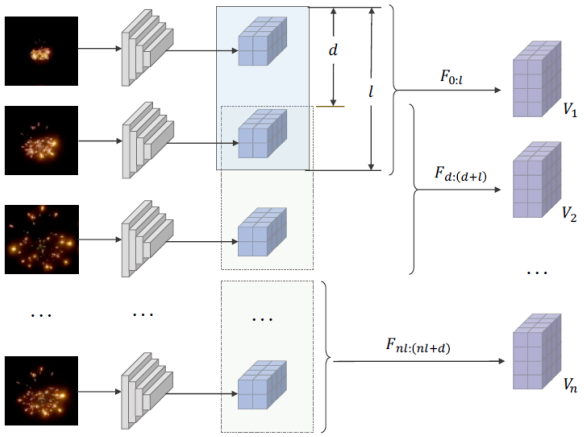

图 3 基于时序约束的视频特征聚合

在本文中，设置每个特征聚合窗口长度为 0.188 秒，窗口重叠长度为 0.94 秒。由于输入视频的帧率为 21.5 帧/秒，设置视觉特征聚合滑动窗口 d 长度为 8，步长 l 为 4；对视觉特征序列分割聚合后，可以获得 53 组时序约束下的视觉特征集合。

4.3 时空匹配的视音频跨模态特征映射

梅尔频谱的横纵坐标分别代表着音频的时间信息与频率特征，因此特征聚合时序窗口内的音频特征也应对应着着由低至高的多频段特征。频率的高低可对应运动的快慢，也对应着视觉特征时序聚合后的高低频特征。

因此，本文构建了一个时空匹配的视音频跨模态特征映射网络，结构如图 4 所示。首先使用空域特征整合模块，提取视觉特征集合V_t 中物体的运动信息；再利用频域特征转换模块，构建不同的频率分支，从视觉特征中抽取不同频率的运动信息并映

射到对应频段音频特征，最后整合为多频段音频特征集合 A_{pt} 。

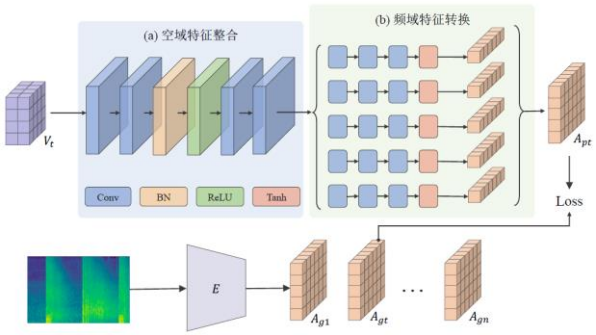

图 4 时空匹配的视音频跨模态特征映射

特征映射网络由空域特征整合和频域特征转换两个模块组成。如图 4（a）所示，空域特征整合模块由 4 个 2 维卷积层组成，其中，第二个卷积层后带有 1 个批归一化（BN）层和 1 个线性整流激活函数（ReLU）。该模块快速整合视觉特征集合内相邻帧特征间的差异，获取视频中物体的运动信息。如图 4（b）所示，频域特征转换模块的 5 个分支均由 3 个 2 维卷积层和一个双曲正切激活函数（Tanh）组成。该模块将物体的运动信息，按照频率的高低转换为多个频段的音频特征，并整合输出为多频段音频特征 A_{pt} 。

利用预先训练的 VQGAN 编码器，获取原始音频特征 A_{gt} ，并以此作为真实值，完成对整个特征映射网络的训练。

为较好地实现视音频模态深度特征间的映射，本文提出了跨模态时空相关性损失：

$$\mathcal{L}_{APL} = \alpha \mathcal{L}_S + \beta \mathcal{L}_D \tag{3}$$

其中， $\mathcal{L}_S$ 为重建损失：

$$\mathcal{L}_S = \begin{cases} 0.5 * \mathcal{L}_L, |\mathcal{L}_L| < 1 \\ |\mathcal{L}_L| - 0.5, \ \mathcal{L}_L < -1 \text{ or } \mathcal{L}_L > 1 \end{cases}, \tag{4}$$

其中， $\mathcal{L}_L$ 为特征距离损失：

$$\mathcal{L}_L = \frac{1}{nm} \sum_{i=1}^{m} \sum_{j=1}^{n} \left| A_{pt}^{ij} - A_{gt}^{ij} \right|, \tag{5}$$

A_{pt} 为预测音频矢量， A_{gt} 为真实音频矢量， m 为音频矢量维度， n 为每次预测音频矢量的数量。 $\mathcal{L}_D$ 为感知损失：

$$\mathcal{L}_D = \frac{1}{nm} \sum_{i=1}^{m} \sum_{j=1}^{n} \left(A_{pt}^{ij} - A_{gt}^{ij} \right)^2. \tag{6}$$

使用跨模态时空相关性损失评估预测音频矢量与真值间的听觉感知差异，进一步提升映射网络的回归能力。经过优化与调整，设置 m 为 256， n 为 5， α 为 0.1、 β 为 0.9，使得该网络可以较好地完成视音频模态间的深度特征映射，有效传递音效生成所需要的视觉信息。

4.4 音频解码

本文采用预先训练的 VQGAN 解码器完成音频解码的任务。音频解码模块的流程及方法如图 5 所示，查找音频特征 A_p 在音频特征空间 Z 中的最邻近矢量 A_p' ，再利用解码器 G 将音频矢量 A_p' 解码为梅尔频谱 $\hat{x}$ 。最后，使用 MelGAN 声码器 V_o [21]，将梅尔频谱 $\hat{x}$ 转换为最终的音频波形。

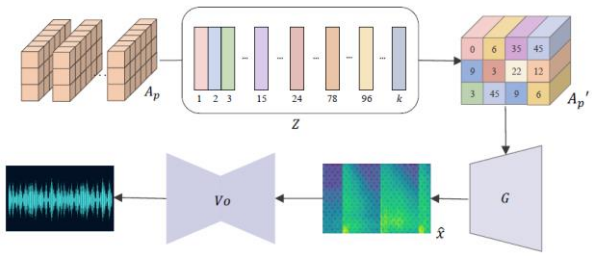

图 5 音频解码模块流程图

5 实验

为验证新方法的有效性，本文进行大量的实验，采用多种评价方式，详细分析和对比实验结果，进一步验证本文方法的有效性。

5.1 数据预处理

本文使用 VAS 数据集[6]完成相关实验，VAS 数据集包含 13008 个视频，每个视频平均时长为 6.73 秒，含有烟花、狗叫、鼓声、婴儿哭声、喷嚏、枪声、咳嗽、锤子敲击等 8 个类别，每个类别的视频数量如表 1 所示，其中，前四类中保存 128 个视频作为测试集，后四类中保存 64 个视频作为测试集。

表 1 VAS 数据集样本统计

数据类别	样本数量
烟花	3114
狗叫	2784
鼓声	2605
婴儿哭声	2060
枪声	865
锤子敲击	382
咳嗽	378
喷嚏	345

在实验开始前，对数据集中的每个视频及其音频进行相应的预处理。将每个视频裁剪为 10s，分辨率调整至 224×224，帧率调整至 21.5fps，为满足滑窗计算，重复最后一帧，共获得 216 帧视频图像。提取每个视频中的音频文件，将其采样率调整至 22.5kHz，采用短时傅里叶变换（STFT）将音频信号变换为声谱图，设置移动窗口大小为 1024，步长为 256；再使用 80 个梅尔窗口将 125～7600Hz 的原始频率映射至梅尔标度，获得分辨率为 860×80 的梅尔频谱，使用 CenterCrop 图片填充算法将其大小裁剪至 848×80。

5.2 评估方法

本文从音频质量、视觉对齐、整体效果三个方面对生成的音效进行量化评估及定性评估。

音频质量：音频质量感知评估（PESQ）参数评价生成音效的质量，使用原始音效作为参考样本，计算生成音效与其时频域或变换域特征参数的差异，再将特征参数差异送入神经网络模型中获得客观的音质分值。

时序对齐：对原始音效与生成音效的发声点进行人工标注，计算二者间的平均偏移量 Δt：

$$\Delta t = \frac{1}{m}\sum_{i=1}^{m}\frac{n_i}{(n_i - b_i)^2}\sum_{j=1}^{n_i - b_i}\left|t_{pj} - t_{gj}\right| \tag{7}$$

其中，t_p 为生成音效的发声时间，t_g 为原始音效的发声时间，w 为生成音效中缺失与多余发声点之和，m 为测试视频数量，n 为每个测试视频原始音效中出现的发声点数量。

人工评估：将生成音效与无声视频剪辑合成，邀请 20 位听众对音效整体效果评分（0～5 分），计算测试集内音效评分的均值，评估生成音效的整体效果。

5.3 实验结果

本文实验采用了 Intel Xeon E5-260 CPU，内存为 64GB，NVIDIA TITAN XP GPU，利用 Adam 优化器在进行了 800 个周期的训练，初始学习率为 0.0001，衰减度为 0.99，衰减步长为 5。

在现有方法中，Chen 等[6]提出的 Regnet 方法和 Iashin 等[4]提出的 SpecVQGAN 方法，在音效保真度和时序对齐方面效果最优。因此，本文与 Regnet 和 SpecVQGAN 进行对比实验，验证新方法的有效性。

实验结果如图 6 所示，图中展示了烟花、鼓声、喷嚏、枪声四组生成结果。每一组中，从上到下分别为通过（a）Regnet、（b）SpecVQGAN 和（c）本文方法生成的音效梅尔频谱，以及（d）原始音频的真实梅尔频谱。

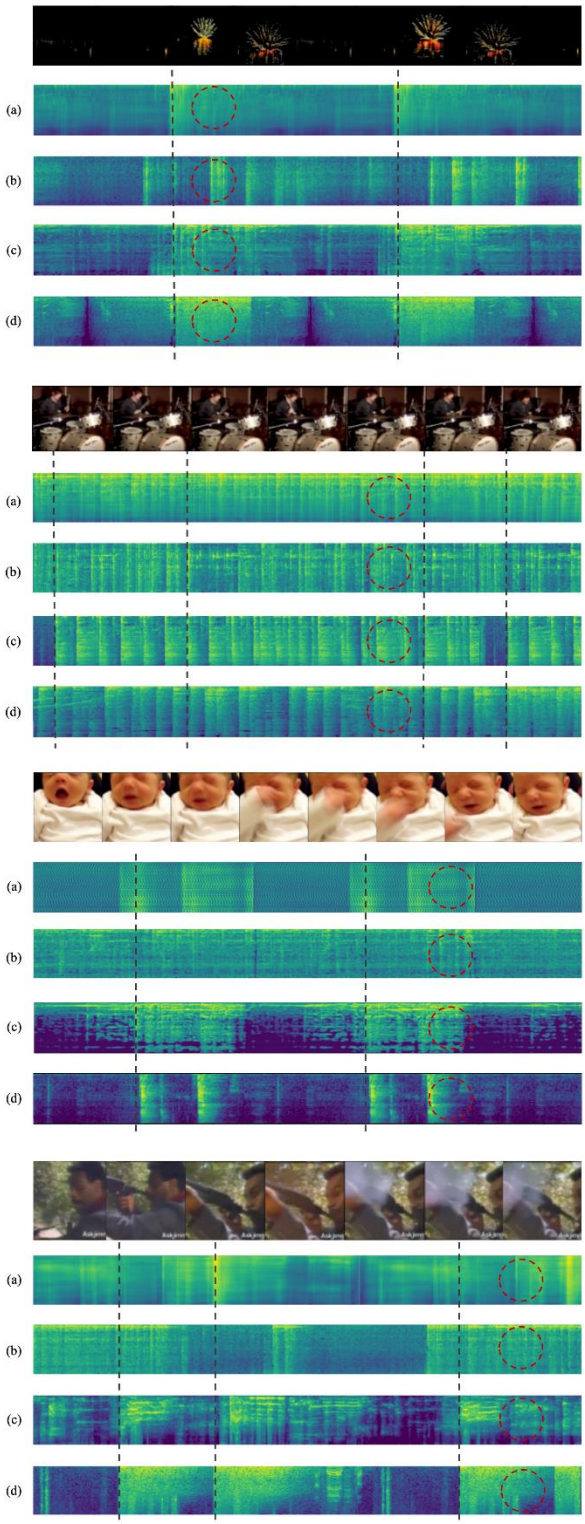

图 6 本文方法与 Regnet, SpecVQGAN 结果对比

对于不同场景的音效生成任务，新方法在时序对齐与音效质量上都展现出了显著的优势。图 6 中采用黑色虚线标记了时序对齐的发声点，结果显示：Regnet 方法生成的音效在发声点处存在时序偏差；SpecVQGAN 方法在时序重建方面表现较差，生成音效与真实值发声点无法对应；采用本文方法，可以生成与原始视频精确对齐的音效。从图 6 中红色虚线框可以观察到，Regnet 方法生成的梅尔频谱较为平滑，音频特征模糊，解码获得的音效波形保真度较低；而本文方法与 SpecVQGAN 方法生成的梅尔频谱较为清晰，解码后的音效波形具有较高的保真度。

在烟花示例中，相比于 Regnet 和 SpecVQGAN，新方法生成的结果音量较大且持续时间较长，整体效果与真实值最为接近。SpecVQGAN 对于整体视觉信息的依赖较强，生成的音效较为离散，而 Regnet 的生成结果缺失细粒度的音频特征。

在鼓声示例中，采用本文方法，可较好地生成时序一致且节奏饱满的鼓声，而 Regnet 和 SpecVQGAN 难以捕捉快速连续有节奏的视觉特征，生成的音效杂乱、缺乏时序特征。

在喷嚏示例中，动作在画面中占比较小，产生的视觉特征变化较少，但本方法仍可根据视觉特征，准确预测发声点，生成与真实值一致的音效结果。而 Regnet 和 SpecVQGAN 的生成结果在时序对齐和保真度两个方面均表现较差。

在枪声示例中，需要同时捕捉人物主体的动作与目标对象发生的变化。本方法可以成功预测出开枪的时间点，生成与真实值接近的音效，效果优于 Regnet 和 SpecVQGAN。

表 2 不同方法的 PESQ 结果对比

数据类别	Regnet[6]	SpecVQGAN[4]	本文
烟花	1.15	1.35	**1.41**
狗叫	1.05	1.26	**1.34**
鼓声	0.91	1.36	**1.87**
婴儿哭声	1.16	1.17	**1.31**
枪声	1.09	1.27	**1.45**
锤子敲击	0.97	**1.25**	1.16
咳嗽	1.03	1.14	**1.31**
喷嚏	1.06	1.21	**1.35**

注：粗体表示最优结果。

本文利用音频质量感知评估（PESQ）参数，评价生成音效的保真度。如表 2 所示，Regnet 的生成结果音质较差，SpecVQGAN 生成的音效质量较高，但其时序传递能力较差，缺乏对于发声点的重建能力。而采用本文方法生成的音效，具有较高的保真度，在 7 个类别的质量评估中获得最高分数。

通过计算发声点平均偏移量 Δt，评估时序对齐效果。实验结果如表 3 所示，SpecVQGAN 的时序重建能力较差，生成音效与原始音效的发声点无法匹配，时间偏移约为 1 秒。Regnet 和本文方法生成的音效发声点平均偏移量较低，时序对齐效果较好。与 Regnet 和 SpecVQGAN 相比，本文方法在所有类别上，均获得最小平均偏移量。实验结果展示了本方法在时序重建方面的显著优势。

表 3 不同方法的平均偏移量 Δt 对比

数据类别	Regnet[6]	SpecVQGAN[4]	本文
烟花	0.17	1.26	**0.15**
狗叫	0.16	1.15	**0.13**
鼓声	0.28	0.97	**0.24**
婴儿哭声	0.19	1.04	**0.18**
枪声	0.27	1.21	**0.23**
锤子敲击	0.33	0.94	**0.32**
咳嗽	0.25	1.17	**0.22**
喷嚏	0.23	1.02	**0.17**

注：粗体表示最优结果。

为了进一步验证本文方法的有效性，采用人工评估方法，获取听众对音效整体效果的主观评价。如表 4 所示，本文方法生成的音效，在 7 个类别上整体效果最佳，在锤子敲击类别上效果稍差于 SpecVQGAN。

表 4 不同方法的人工评估结果对比

数据类别	Regnet[6]	SpecVQGAN[4]	本文
烟花	1.60	1.35	**2.25**
狗叫	1.65	1.25	**2.40**
鼓声	2.00	1.70	**3.05**
婴儿哭声	2.15	1.40	**2.95**
枪声	2.05	1.55	**3.00**
锤子敲击	1.10	**1.85**	1.80
咳嗽	2.05	1.45	**2.55**
喷嚏	1.75	1.65	**2.30**

注：粗体表示最优结果。

5.4 超参数与消融实验

本文采用特征聚合窗口，对视音频特征进行时

序等长的分割与整合，提升了生成音效的时序对齐效果。通过超参数对比实验，选取最合适的特征窗口长度。设置时间窗口长度为 0.94、1.41、1.88、2.35 秒，即视觉特征窗口长度为 4、6、8、10 帧，滑动步长保持不变，实验结果如表 5 所示。

表 5 不同窗口长度时间偏移量Δt 对比

数据类别	4	6	8	10
烟花	0.32	0.27	**0.15**	0.24
狗叫	0.25	0.18	**0.13**	0.21
鼓声	0.32	0.30	**0.24**	0.33
婴儿哭声	0.26	0.23	**0.18**	0.25
枪声	0.34	0.28	**0.23**	0.30
锤子敲击	0.48	0.40	**0.32**	0.37
咳嗽	0.36	0.29	**0.22**	0.31
喷嚏	0.31	0.25	**0.17**	0.27

注：粗体表示最优结果。

实验结果表明，采用长度为 8 的视觉特征聚合窗口，对视音频特征进行时序约束，发声点平均偏移量最低。

表 6 消融实验 PESQ 参数结果对比

数据类别	本文（无频域特征转换模块）	本文
烟花	1.29	**1.41**
狗叫	1.25	**1.34**
鼓声	1.65	**1.87**
婴儿哭声	1.21	**1.31**
枪声	1.32	**1.45**
锤子敲击	1.05	**1.16**
咳嗽	1.20	**1.31**
喷嚏	1.26	**1.35**

注：粗体表示最优结果。

表 7 消融实验平均偏移量 Δ*t* 对比

数据类别	本文（无频域特征转换模块）	本文
烟花	0.18	**0.15**
狗叫	0.15	**0.13**
鼓声	0.26	**0.24**
婴儿哭声	0.21	**0.18**
枪声	0.27	**0.23**
锤子敲击	0.33	**0.32**
咳嗽	0.26	**0.22**
喷嚏	0.22	**0.17**

注：粗体表示最优结果。

为了验证频域特征转换模块的有效性，对比了直接利用空域特征整合模块的预测结果，并利用音频质量感知评估（PESQ）参数及发声点平均偏移量 Δ*t* 进行评价。如表 6 和表 7 所示，实验结果表明，利用频域特征转换模块对不同频段特征进行转换，提升了生成模型的性能，改善了生成音效的保真度。

6 结论

本文提出了一种基于时序视觉特征映射的音效生成方法，实现了视觉引导下的高质量音效生成。首先利用基于时序约束的特征聚合模块对视觉特征序列进行滑动整合，随后构建时空匹配的跨模态特征映射网络将其转换为多频段音频特征，最后采用预先训练的音频解码器及声码器解码为最终波形。实验结果表明，本文提出的方法可以生成保真度较高、时序对齐效果较好的音效。

同时，该方法存在一定的局限性，需要根据音效的类别对特征映射网络进行训练。为了进一步提升模型的泛化性与适应性，未来的研究工作将对构建多类别特征映射网络进行探索。此外，目前公开的音效生成数据集包含的音效种类较少，且存在声画不同步、噪声大等问题，增加了音效生成的难度。因此，扩充高质量音效生成数据集也将成为未来音效生成研究工作的重点。

参 考 文 献

[1] Owens A, Isola P, McDermott J, et al. Visually indicated sounds[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2405-2413.

[2] Zhou Y, Wang Z, 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. 2018: 3550-3558.

[3] Ghose S, Prevost J J. Autofoley: Artificial synthesis of synchronized sound tracks for silent videos with deep learning[J]. IEEE Transactions on Multimedia, 2020, 23: 1895-1907.

[4] Iashin V, Rahtu E. Taming Visually Guided Sound Generation[J]. arXiv preprint arXiv:2110.08791, 2021.

[5] Ghose S, Prevost J J. Enabling an IoT system of systems through auto sound synthesis in silent video with DNN[C]//2020 IEEE 15th International Conference of System of Systems Engineering (SoSE). IEEE, 2020: 563-568.

[6] Chen P, Zhang Y, Tan M, et al. Generating visually aligned sound from

videos[J]. IEEE Transactions on Image Processing, 2020, 29: 8292-8302.

[7] Liu S, Li S, Cheng H. Towards an End-to-End Visual-to-Raw-Audio Generation with GAN[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021.

[8] Ghose S, Prevost J J. FoleyGAN: Visually Guided Generative Adversarial Network-Based Synchronous Sound Generation in Silent Videos[J]. arXiv preprint arXiv:2107.09262, 2021.

[9] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[10] Mehri S, Kumar K, Gulrajani I, et al. SampleRNN: An unconditional end-to-end neural audio generation model[J]. arXiv preprint arXiv:1612.07837, 2016.

[11] Chen K, Zhang C, Fang C, et al. Visually indicated sound generation by perceptually optimized classification[C]//Proceedings of the European Conference on Computer Vision (ECCV) Workshops. 2018: 0-0.

[12] Aytar Y, Vondrick C, Torralba A. Soundnet: Learning sound representations from unlabeled video[J]. Advances in neural information processing systems, 2016, 29.

[13] Lagler K, Schindelegger M, Böhm J, et al. GPT2: Empirical slant delay model for radio space geodetic techniques[J]. Geophysical research letters, 2013, 40(6): 1069-1073.

[14] Brock A, Donahue J, Simonyan K. Large scale GAN training for high fidelity natural image synthesis[J]. arXiv preprint arXiv:1809.11096, 2018.

[15] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: Towards good practices for deep action recognition[C]//European conference on computer vision. Springer, Cham, 2016: 20-36.

[16] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.

[17] Oord A, Dieleman S, Zen H, et al. Wavenet: A generative model for raw audio[J]. arXiv preprint arXiv:1609.03499, 2016.

[18] Gemmeke J F, Ellis D P W, Freedman D, et al. Audio set: An ontology and human-labeled dataset for audio events[C]//2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2017: 776-780.

[19] Esser P, Rombach R, Ommer B. Taming transformers for high-resolution image synthesis[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 12873-12883.

[20] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[21] Kumar K, Kumar R, de Boissiere T, et al. Melgan: Generative adversarial networks for conditional waveform synthesis[J]. Advances in neural information processing systems, 2019, 3