

# 基于时序对齐视觉特征 映射的音效生成方法



2022 GUIYANG  
CHINA M

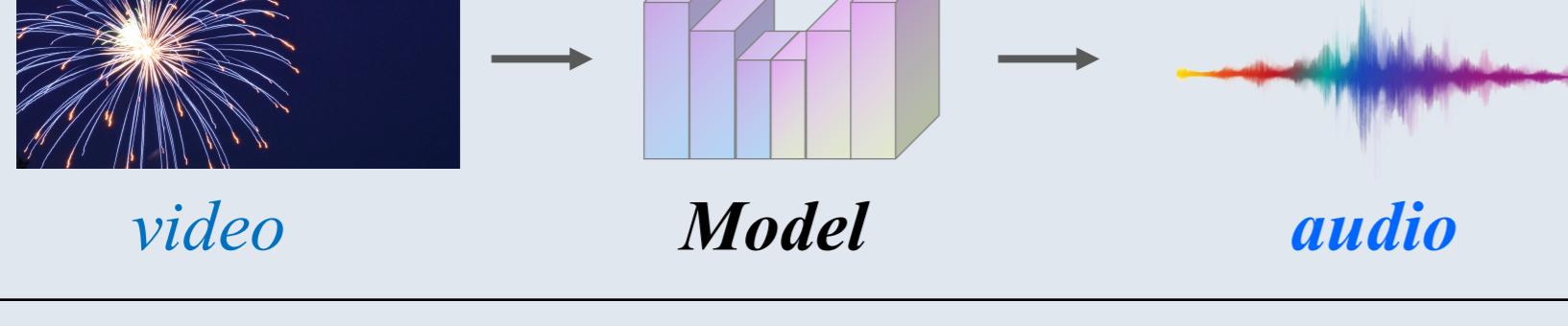
谢志峰<sup>1,2</sup> 孙络袆<sup>1</sup> 孙郁洲<sup>1</sup> 余椿鹏<sup>1</sup> 马利庄<sup>2,3</sup>  
<sup>1</sup>上海大学影视工程系 <sup>2</sup>上海电影特效工程技术研究中心  
<sup>3</sup>上海交通大学计算机科学与工程系

## Poster Report

### 任务描述

- 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。

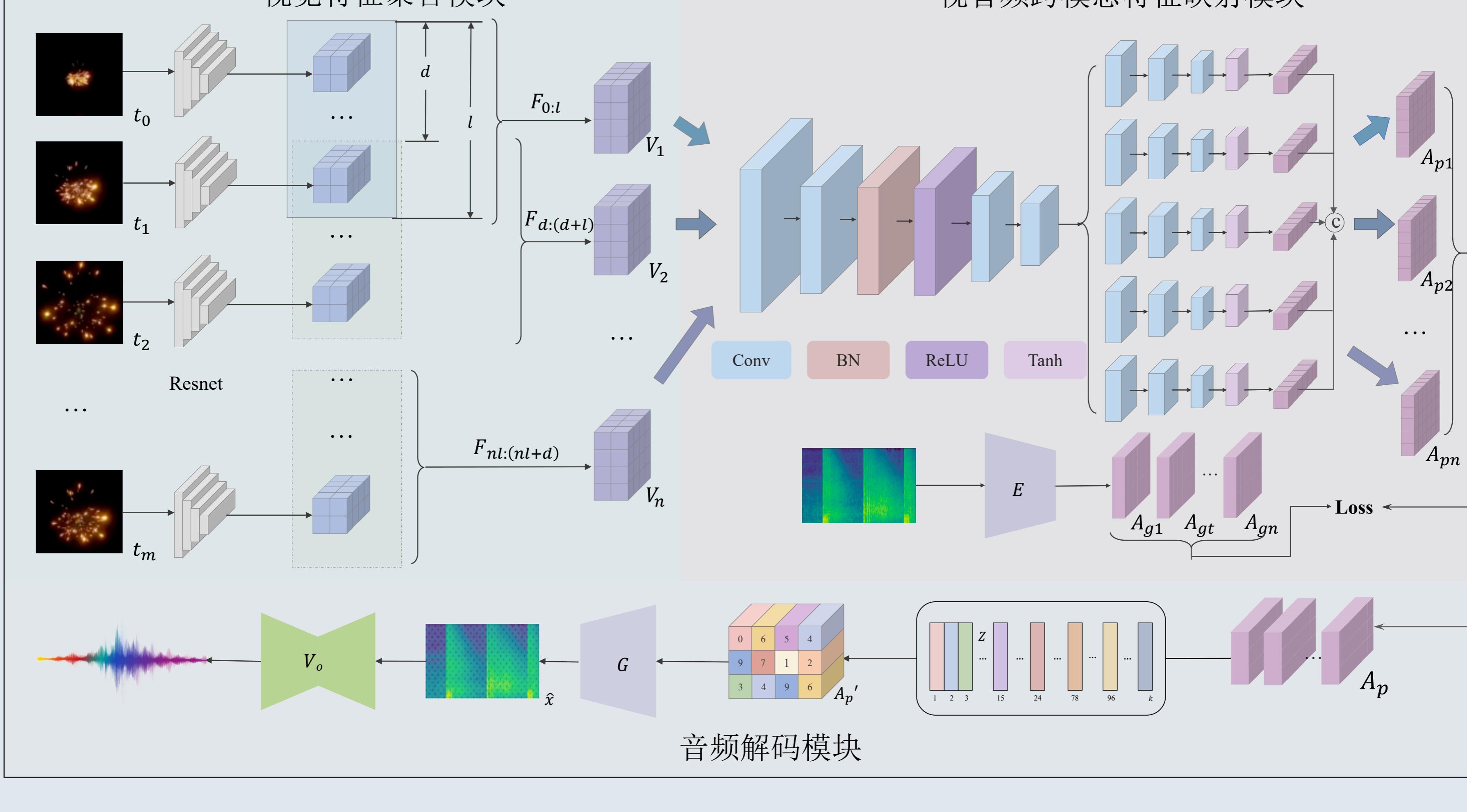
$$\text{Audio} \Leftarrow \mathcal{F}_t(\text{video})$$



### 研究意义

- 推动电影智能剪辑技术发展，为音效艺术家提供初始参考，提升配音效率；
- 提升互动娱乐、幼儿教育等多种视音频内容质量与创作效率，降低创作成本；
- 提升生成音效的多样性与创意性，避免作品间的音效重复，提升自动配音体验。

### 视觉引导的音效生成架构



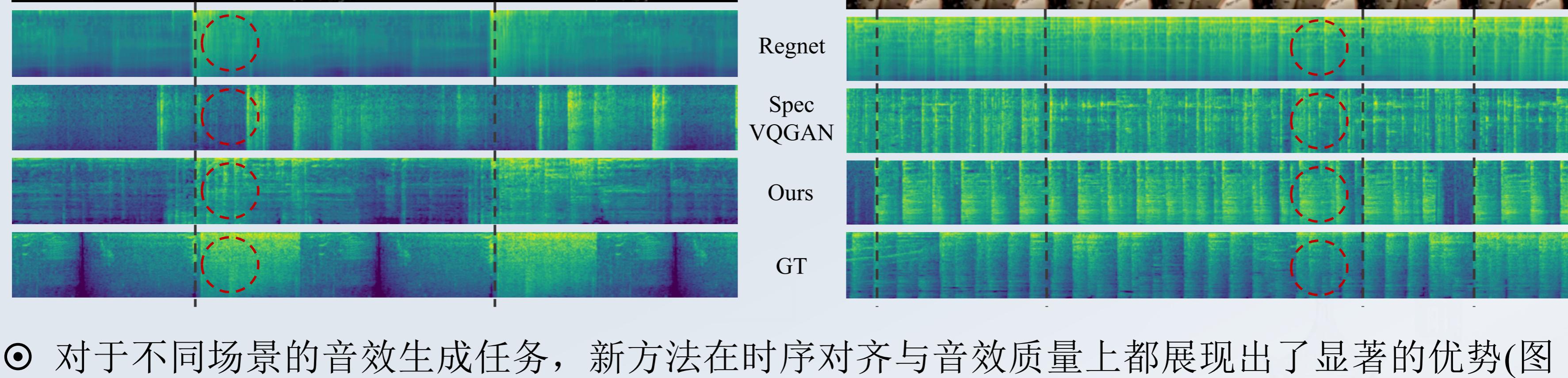
### 创新点与主要贡献

- 提出基于视觉特征聚合、跨模态特征映射与音频解码的视觉引导音效生成框架；
- 设计基于时序约束的特征聚合窗口，分割整合视觉特征，提升音效的对齐效果；
- 构建视音频跨模态映射网络，通过时空特征聚合与映射提升生成音效的保真度。

### 评价指标

- 音频质量感知评估(PESQ)参数评估音质
- 发声点平均偏移量( $\Delta t$ )评估对齐效果
$$\Delta t = \frac{1}{m} \sum_{i=1}^m \frac{n_i}{(n_i - b_i)^2} \sum_{j=1}^{n_i - b_i} |t_{pj} - t_{gj}|$$
- 人工评估：将音效与原视频剪辑合成，邀请20位听众评估其整体效果(0~5分)

### 实验结果与对比评估



- 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(图中采用黑色虚线标记时序对齐的发声点，红色虚线框标记梅尔频谱的精细程度)；
- 使用音频质量感知评估(PESQ)评估音频质量，建立发声点平均偏移量( $\Delta t$ )函数评估音效时序对齐效果，设计人工评估实验评价音效整体效果；在上述评估方法中均获得最优结果。



上海大学  
SHANGHAI UNIVERSITY



上海电影学院  
SHANGHAI FILM ACADEMY



上海交通大学  
SHANGHAI JIAO TONG UNIVERSITY