计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(3):133−142 [doi: 10.15888/j.cnki.csa.009785] [CSTR: 32024.14.csa.009785]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

面向视音频事件定位的跨模态时间对齐网络①

王志豪, 訾玲玲

(重庆师范大学 计算机与信息科学学院, 重庆 401331)
通信作者: 王志豪, E-mail: 2022210516095@stu.cqnu.edu.cn

摘　要: 视音频事件定位 (audio-visual event localization, AVEL) 任务通过观察音频信息和相对应的视觉信息来定位视频中的事件. 本文针对 AVEL 任务设计了一种跨模态时间对齐网络 CMTAN, 该网络包含预处理、跨模态交互、时间对齐和特征融合这 4 个部分. 具体而言, 在预处理部分, 通过一种新的跨模态音频指导模块和一种噪音弱化模块的处理, 模态信息中的背景和噪音被减少. 随后, 在跨模态交互部分, 使用基于多头注意力机制的信息强化和信息补充模块进行跨模态交互, 单模态信息得到全局信息优化. 在时间对齐部分, 本文设计了一种聚焦于跨模态交互前后单模态全局信息的时间对齐模块, 用于执行模态信息的特征对齐. 最后, 在特征融合过程中, 通过一种多阶段融合模块, 两种模态信息被从浅入深地融合, 且融合后的模态信息最终将被用于事件定位. 大量实验表明 CMTAN 在弱监督和全监督 AVEL 任务中都具有优秀的性能.

关键词: 跨模态; 视音频事件定位; 弱监督和全监督; 特征对齐

引用格式: 王志豪,訾玲玲.面向视音频事件定位的跨模态时间对齐网络.计算机系统应用,2025,34(3):133–142. http://www.c-s-a.org.cn/1003-3254/9785.html

Cross-modal Time Alignment Network for Audio-visual Event Localization

WANG Zhi-Hao, ZI Ling-Ling

(College of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China)

Abstract: The audio-visual event localization (AVEL) task locates events in a video by observing audio information and corresponding visual information. In this paper, a cross-modal time alignment network named CMTAN is designed for the AVEL task. The network consists of four parts: preprocessing, cross-modal interaction, time alignment, and feature fusion. Specifically, in the preprocessing part, the background and noise in the modal information are reduced by the processing of a new cross-modal audio guidance module and a noise reduction module. Then, in the cross-modal interaction part, the information reinforcement and information complementation modules based on the multi-head attention mechanism are used for cross-modal interaction, and the unimodal information is optimized with global information. In the time alignment part, a time alignment module focusing on the unimodal global information before and after cross-modal interaction is designed to perform feature alignment of modal information. Finally, in the feature fusion process, two kinds of modal information are fused from shallow to deep by a multi-stage fusion module. The fused modal information is ultimately used for event localization. Extensive experiments demonstrate that CMTAN has excellent performance in both weakly and fully supervised AVEL tasks.

Key words: cross-modal; audio-visual event localization (AVEL); weakly and fully supervised; feature alignment

① 基金项目: 重庆市教育科学规划重点课题 (K22YE205098); 重庆师范大学博士启动基金 (21XLB030, 21XLB029)
收稿时间: 2024-08-24; 修改时间: 2024-09-19; 采用时间: 2024-09-24; csa 在线出版时间: 2025-01-16
CNKI 网络首发时间: 2025-01-17

视音频事件定位 (audio-visual event localization, AVEL) 任务使用音频信息和相对应的视觉信息定位每个视频片段中的视听事件[1]. 当且仅当视频中同一片段的音频部分和视觉部分都出现了能代表同一事件的信息时, 该视频片段才能被认为具有此事件. 由此可见 AVEL 任务需要同时考虑两种模态信息, 因此该任务面临着以下 3 个挑战: (1) 音频和视觉信息中的噪音和背景会影响事件的分类; (2) 不同步的音频和视频信息会影响事件边界的预测; (3) 两种模态信息需要被融合进行事件定位, 但简单的融合会丢失模态的重要信息.

早期工作聚焦于第 1 个和第 3 个挑战, 通常独立处理两种模态信息, 再直接进行简单的特征融合[1,2]或基于注意力机制的特征融合[3]. 后续工作将不同步音频和视觉信息造成的事件边界预测问题纳入考虑, 使用基于注意力机制[4–7]和相似度机制[8]的跨模态信息交互方法去优化事件边界的预测. 然而, 这些方法忽略了跨模态交互后单模态全局信息的重要性, 这将降低事件边界预测的精确性. 与以往工作不同, 本文针对 AVLE 任务设计了一种跨模态时间对齐网络 (cross-modal time alignment network, CMTAN), 该网络利用跨模态交互前和跨模态交互后的特征生成注意力图, 并用该图使模态特征在时间维度上进行了对齐, 使事件边界预测的性能得到提升. 此外, 为缓解噪音和背景对事件分类的影响, 本文设计了一种基于注意力机制的跨模态音频指导模块和基于相似度机制的噪音弱化模块. 跨模态音频指导模块用于突出视觉信息中与音频信息具有紧密联系的部分视觉信息, 噪音弱化模块用于降低不匹配音频和视觉信息对音频信息指导的影响.

本文的主要贡献如下.

(1) 为优化特征对齐的效果, 本文设计了一种利用单模态全局信息的时间对齐模块进行特征对齐, 以此提高网络预测事件边界的能力.

(2) 本文设计了一种跨模态音频指导模块和一种噪音弱化模块. 这些模块在强化了模态内重要信息的同时, 降低了噪音和背景对事件分类的影响.

(3) 基于上述模块, 本文提出一种 CMTAN 网络, 该网络在弱监督和全监督 AVEL 任务中都取得优秀效果.

1 相关工作

1.1 视音频学习

近年来, 视音频学习获得了越来越多的关注, 该方向工作能为 AVEL 任务、视音频视频解析和动作识别等视频相关任务提供良好的特征学习方法. 一些视音频学习的工作聚焦于视音频表征融合[9–14]. 早期工作通常分别从音频和视觉模态中学习特征, 然后通过拼接、加法等方法进行融合[9]. 但简单融合可能会丢失模态内信息, 因此文献[10]使用长短期记忆网络、注意力机制和概率方法分别进行特征融合, 以此比较不同融合方法在模态间交互过程中的效果.

其他工作聚焦于视听表征的学习, 他们使用注意力机制、聚类和对比等方法去学习模态表征[15–21]. 例如文献[16]提出一种共同注意力方法, 其在模态融合前进行了模态间交互. 文献[21]设计了一种基于双编码器表示学习的多模态情感分析模型, 其通过双编码器结构学习模态信息的不变特征和模态私有的模态特定特征.

1.2 视音频事件定位

AVEL 任务的描述首次由文献[1]于 2018 年提出. 图 1 展示了 AVEL 任务的一个示例, 该例子具有 5 个包含音频和视觉信息的视频片段. 在第 3 个和第 4 个片段中, 音频事件和视觉事件都是“飞行器”, 因此这两个片段被定位为“飞行器”. 其他片段中, 音频事件和对应的视觉事件不同, 因此这些片段最终被定位为“背景”.

针对 AVEL 任务, 早期方法通常基于 LSTM, 倾向于独立处理两种模态信息, 如多模态残差结构 DMRN (multimodal residual network)[1]和视音频序列到序列双重网络 AVSDN (audio-visual sequence-to-sequence dual network)[2]. 然而, 这些方法仅在片段级上处理特征, 特征对齐容易受到不同步视音频信息的影响, 因此后续工作将全局信息纳入考虑, 例如双层注意力匹配 DAM (dual attention matching)[3]模型, 该模型根据全局信息使用一种全局交叉检查机制去查询单模态的局部信息. 针对模态信息中噪音和背景对 AVEL 任务的干扰, 一些方法使用音频信息去指导视觉信息, 并取得了优秀的效果, 如跨模态关系感知网络 CMRAN (cross-modal relation-aware network)[5]、跨模态背景抑制网络 CMBS (cross-modal background suppression)[7]以及文献[22]设计的视听融合方法. 与以上方法不同, 文献[23]提出了一种联合标签降噪 JoMoLD (joint modal label denoising) 方法, 该方法降低了噪音标签的干扰, 对模型的事件定位作用效果显著.

近年来, 大量的工作聚焦于跨模态交互, 如跨模态注意网络 CMAN (cross-modal attention network)[4]、AVT (audio-visual Transformer)[24]、多模态平行网络

MPN (multimodal parallel network)[25]以及掩码共同注意力 MCA (masked co-attention)[26]. 这些方法通过桥接音频和视觉信息中的语义鸿沟来提高事件边界的预测能力, 其中文献[8]提出的一种正样本传播 PSP (positive sample propagation) 方法取得了优秀的效果, 该方法通过建立两种模态不同片段之间的相似度图, 再根据指定的阈值保留了具有相似语义的片段并舍弃了具有不相似语义的片段. 此外, 一些方法聚焦于模态特征的融合, 如双向特征融合网络 BMFN (bi-directional modality fusion network)[6]、多模态金字塔注意力网络 MM-Pyramid (multimodal pyramid attentional network)[27]、视音频交互网络 AVIN (audio-visual interacting network)[28]和文献[29]提出的 AVRB 方法. 这些方法通过充分利用融合特征进行模态特征优化.

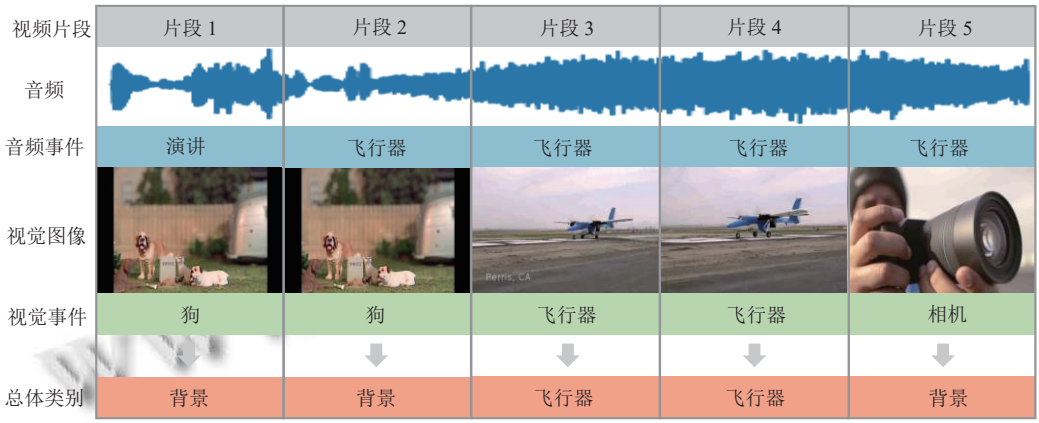


图 1 AVEL 任务的示例

根据以上相关工作, 本文将针对 AVEL 任务的主要方法划分为 3 个主要部分: 预处理、跨模态交互以及特征融合. 预处理部分包括音频和视觉特征的提取, 以及模态特征的优化等步骤, 该部分主要针对视频片段中事件的分类; 跨模态交互主要进行模态信息间的特征交互, 针对视频中事件边界的预测; 特征融合部分的结果将用于最终的事件定位, 因此该部分在进行特征融合的过程中应尽量保留两种模态的重要信息. 为进一步提高事件边界的预测准确率, 本文在提出的 CMTAN 的跨模态交互部分和特征融合部分之间加入了一种时间对齐模块, 其充分利用单模态全局信息, 使模态特征得到了优化. 实验结果证明, 模态内和模态间的特征得到了有效对齐, 并且特征融合的效果也得到了优化.

2 CMTAN 网络

本文提出的 CMTAN 包含预处理、跨模态交互、时间对齐和特征融合等 4 个部分, 其具体结构如图 2 所示.

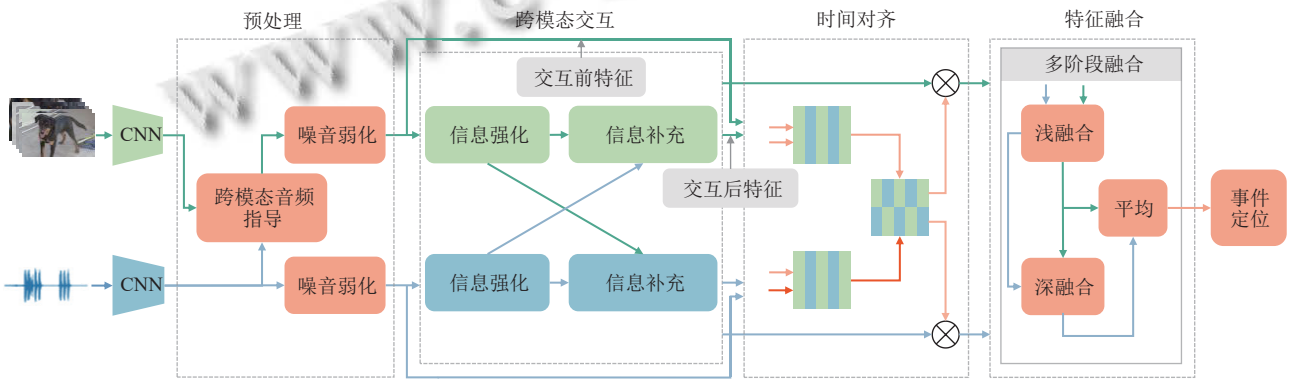


图 2 CMTAN 的总体结构图

在预处理部分中, 被卷积神经网络提取的音频特征和视觉特征作为输入, 在经过跨模态音频指导模块和噪音弱化模块的处理后, 两种模态特征中的噪音和背景得到一定程度的削弱. 然后, 跨模态交互部分的信

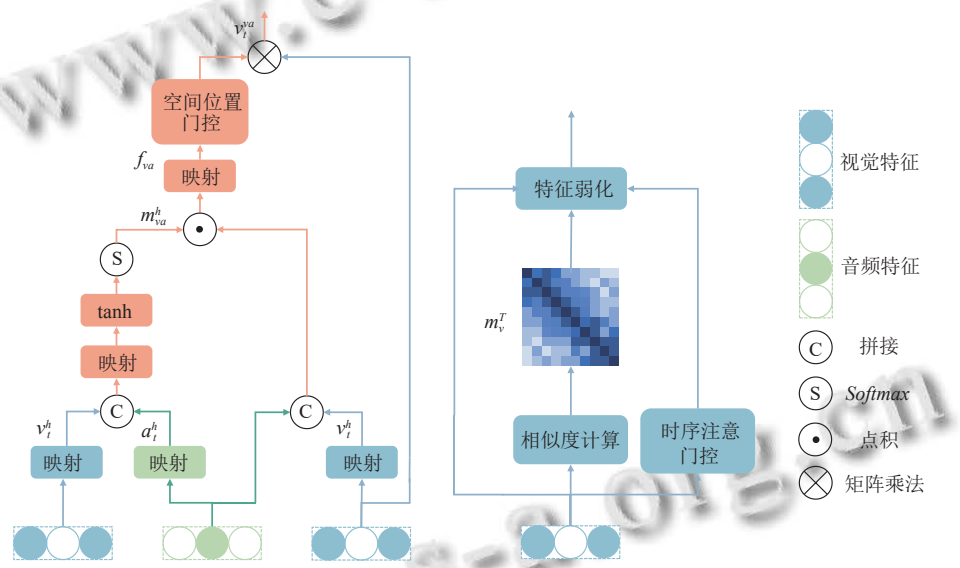
息强化和信息补充模块进行音频和视觉特征间的交互, 将两种模态的全局信息纳入单个模态信息中. 随后, 时间对齐部分的时间对齐模块通过交互前特征和交互后特征进一步调整两种模态特征, 提升网络的事件边界预测精准度. 最后, 通过含有浅融合和深融合过程的多阶段融合模块, 两种模态特征得到融合并用于最终的事件定位. 值得注意的是, 在预处理和特征融合部分中, CMTAN 将根据真实标签自动生成标签 *flag*, 网络会根据 *flag* 值侧重于不同模块.

2.1 问题描述

AVEL 任务通常被定义为: 给定一个具有 T 个不重复片段的视频$S=(v_t,a_t)_{t=1}^{T}$, 其中$v_t\in\mathbb{R}^{H\times W\times d_v}$和$a_t\in\mathbb{R}^{d_a}$分别代表从第 t 个视频片段中视觉和音频信息的模态特征, H 和 W 分别代表视觉特征映射图的高和宽. d_v和d_a分别代表视觉和音频特征的特征维度. 经过网络处理, 得到每个视频片段的事件概率$p_t=\{p_t^n|p_t^n\in(0,1),n=1,2,\cdots,n,\sum_n^C p_t^n=1\}$, 其中 C 代表事件类别的数量. p_t^n代表在第 t 个视频片段中第 n 个事件类别的概率, 且所有类别的概率和$\sum_n^C p_t^n=1$. 弱监督 AVEL 任务中, 训练阶段的真实标签仅包含视频的整体事件类别, 而在全监督 AVEL 任务中, 真实标签包含了视频中所有片段的事件类别.

2.2 预处理部分

在预处理部分中, 本文提出了一种新的跨模态音频指导模块和噪音弱化模块并用于优化模态特征, 图 3 展示了两种模块的主要结构.



(a) 跨模态音频指导模块的全局空间指导部分 (b) 噪音弱化模块

图 3 预处理部分的主要结构

以往的工作简单使用音频的空间信息和通道信息指导视觉特征, 但忽略了指导过程中两种模态之间空间信息的联系[5,6]. 这就会导致视觉信息中的背景部分被音频信息中的噪音强化, 进而降低了视频中事件分类的准确性. 受非局部块[30]和 CMRAN[5]的启发, 本文提出了一种新的跨模态音频指导模块. 通过结合全局空间指导和音频信息指导, 模块将全局的空间信息纳入考虑, 视觉信息中与真实标签相对应的部分得到强化, 事件分类更加准确. 图 3(a) 展示了跨模态音频指导模块的全局空间指导部分. 具体而言, 音频特征a_t和视觉特征v_t首先被映射到潜在空间中获得a_t^h和v_t^h. 随后, 通过拼接操作和映射函数结合a_t^h和v_t^h, 从而获得特征映射图m_{va}^h.

$$m_{va}^h=Softmax\left(\tanh\left(W_h cat\left(v_t^h,a_t^h\right)\right)\right) \tag{1}$$

其中, W_h是可训练的参数. 此外, 为减少映射操作中音频信息的损失, 本文将输入的音频特征a_t^h和经过另一个特征映射的视觉特征拼接, 其拼接结果将进一步和m_{va}^h进行点积, 以获得全局视觉特征$f_{va}\in\mathbb{R}^{(HW\times 2)\times d_v}$:

$$f_{va}=W_v\left(cat\left(v_t^h,a_t\right)\odot m_{va}^h\right) \tag{2}$$

其中, W_v是可训练的参数. 随后, f_{va}经过空间位置门控的处理并与v_t结合, 获得经过全局空间信息指导的视觉

特征v_t^{va}. 本文将v_t^{va}与简单基于音频空间信息和通道信息指导的视觉特征v_t^{cs}结合, 获得最终视觉特征$V \in \mathbb{R}^{T\times d}$.

图 3(b) 展示了噪音弱化模态的具体结构, 该模块是一个轻量级模块, 其被用于降低跨模态音频信息指导过程中不同步音频和视觉信息对指导过程带来的不利影响. 在不匹配的音频和视觉信息中, 视频片段的所有视觉特征都被认作是背景, 与真实标签的事件无关. 因此, 音频信息将不能指导出任何有效的视觉特征, 这些无效特征会影响事件分类. 受相似度机制的启发, 本文设计了一种基于相似度机制的噪音弱化模块, 该模块通过弱化 T–1 个视觉片段的特征来降低噪音和背景带来的不利影响. 具体而言, 以视觉特征为例, 该模块通过余弦相似度计算并获得了同一视频中不同片段间视觉特征的相似图$m_v^T \in \mathbb{R}^{T\times T}$. 同时, 通过时序注意门控预测每个视频片段中事件存在的概率, 并根据概率值得到最可能存在事件的片段, 假设为片段 k. 随后, 该模块基于片段 k 选取m_v^T中第 k 个片段中相似度最大值的行向量 (通常为第 k 行), 并将该向量的值与该视频中的视觉特征进行对应片段的乘法操作. 最终, 除了第 k 个片段, 其余 T–1 个片段的视觉特征都被整体削弱, 特征中的噪音也随之被弱化. 针对音频模态的噪音弱化操作同样如此. 需注意, 噪音弱化模块依赖时序注意门控的预测结果. 因此, 当训练阶段使用包含所有片段事件的真实标签时, CMTAN 更倾向使用噪音弱化模块.

2.3 跨模态交互部分

跨模态交互旨在完成模态间的信息补充, 以强化两种模态的信息表达能力. 受 Transformer[31]的启发, 本文使用基于多头注意力机制的信息强化模块和信息补充模块进行模态间的特征交互.

具体而言, 经过预处理的视觉特征 V 和音频特征 A 被送入各自的信息强化模块中, 获得自我关注后的强化特征V^{sr}和A^{sr}. 为了使每个模态都能从另一模态中获得补充信息, 本文将 V 和A^{sr}送入信息补充模块中, 获得视觉模态交互特征V^{cr}, 并根据相同方法获得音频模态交互特征A^{cr}, 其公式如下:

$$A^{cr} = Decoder1\left(A, Encoder2\left(V\right)\right) \tag{3}$$

$$V^{cr} = Decoder2\left(V, Encoder1\left(A\right)\right) \tag{4}$$

其中, *Encoder*1 和 *Encoder*2 是信息强化模块, *Decoder*1 和 *Decoder*2 是信息补充模块. 信息强化模块和信息补充模块基于多头注意力机制, 因此能使两种模态信息在时序上进行交互. 因此, 由于不同步音频和视觉信息带来的事件边界定位问题得到部分解决.

2.4 时间对齐

针对特征对齐, 当前的工作聚焦于两种模态信息的交互[5]. 然而, 经过跨模态交互处理后, 单个模态信息与另一个模态信息的联系非常紧密, 单模态内部不同片段之间特征的联系变弱, 使得单模态信息可能会在跨模态交互后丢失自身原有的全局信息, 这不利用后续的特征融合. 因此, 受解纠缠多模态非局部块[32]的启发, 本文在时间对齐部分内设计了一种基于注意力机制的时间对齐模块. 该模块计算了跨模态交互前和跨模态交互后单模态片段间的注意力图, 以及跨模态交互后特征的自注意力图, 并根据两张图对跨模态交互后的特征进行模态内特征对齐. 最终得到的模态特征在保留了部分互补模态全局信息的同时, 也因为适度降低了模态间的联系而保留了自身一定的全局信息. 时间对齐模块的结构如图 4 所示.

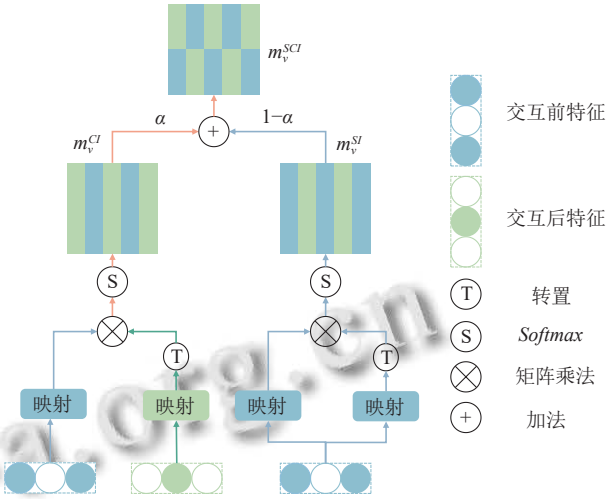


图 4 时间对齐模块的主要结构

以视觉模态特征为例. 首先, 交互前特征 V (预处理部分的输出) 和交互后特征V^{cr} (跨模态交互部分的输出) 被映射到同一潜在空间. 随后, 通过注意力机制计算出V^{cr}的自注意力图$m_v^{SI} \in \mathbb{R}^{T\times T}$, m_v^{SI}中的元素$m_{i,j\in T}^{SI}$代表第 i 个片段和第 j 个片段之间的关注程度. 同理, 该模块计算出V^{cr}和 V 的注意力图$m_v^{CI} \in \mathbb{R}^{T\times T}$, 并将$m_v^{CI}$和$m_v^{SI}$合并获得$m_v^{SCI} \in \mathbb{R}^{T\times T}$. 最终, 通过结合$m_v^{SCI}$和$V^{cr}$获得时间对齐特征$V^{SCI} \in \mathbb{R}^{T\times d_h}$. 通过相似的操作能获得音频的时间对齐特征$A^{SCI} \in \mathbb{R}^{T\times d_h}$. 具体过程表达如下:

$$m_v^{SI} = Softmax\left(W_V^{SI1}V^{cr} \otimes \left(W_V^{SI2}V^{cr}\right)^{\top}\right) \tag{5}$$

$$m_v^{CI} = Softmax\left(W_V^{CI1}V^{cr} \otimes \left(W_V^{CI2}V\right)^{\top}\right) \tag{6}$$

$$V^{SCI} = \left((1-\alpha)m_v^{SI} + \alpha m_v^{CI}\right) \otimes V^{cr} \tag{7}$$

其中, W_V^{SI1}、W_V^{SI2}、W_V^{CI1} 和 W_V^{CI2} 是可训练的参数, α 是控制两个注意力图结合比例的超参数, $(\cdot)^{\top}$ 代表转置操作. 此外, 本文将两种模态的时间对齐模块进行了参数共享, 使两种模态特征在经过时间对齐模块处理后能保持模态间特征的对齐, 并减少了训练参数.

2.5 特征融合

本文在特征融合部分设计了一种多阶段融合模块, 该模块用于有效融合两种模态特征, 其通过结合浅融合和深融合结果, 充分利用了两种模态的全局信息. 具体而言, 两种模态特征 V^{SCI} 和 A^{SCI} 首先通过浅融合获取浅融合特征, 其包含了两种模态的全局特征. 随后, 浅融合特征和音频特征进行通过相似的操作进行深融合, 从而获得融合特征, 深融合过程中充分利用了音频信息去指导浅融合特征, 进一步发挥了音频信息的指导作用. 最终, 该模块将浅融合特征和深融合特征进行加法和平均操作, 降低音频在指导浅融合特征过程中可能带来的指导错误等负面影响. 最终融合特征 $F_{sd} \in \mathbb{R}^{T\times d_h}$ 的生成过程如公式 (8) 所示.

$$\begin{aligned} F_{sd} = &\frac{1}{2}\times fuse\left(A^{SCI}, fuse\left(V^{SCI}, A^{SCI}\right)\right) \\ &+ \frac{1}{2}\times fuse\left(V^{SCI}, A^{SCI}\right) \end{aligned} \tag{8}$$

其中, $fuse(\cdot)$ 代表单个融合模块. 请注意, 多阶段融合模块会基于标签 $flag$ 的值自动选择合适的融合方法, 该模块使用了以下两种融合方法:

$$fuse(M,N) = \begin{cases} AVIM(M,N), & flag = 0 \\ SigmoidMD(M,N), & flag = 1 \end{cases} \tag{9}$$

其中 M 和 N 代表输入到融合方法中的两种模态特征. 当 $flag$ 为 0 时, 模块使用文献[5]设计的 $AVIM$ 模块进行特征融合, $AVIM$ 模块基于多头注意力机制, 因此具有良好的融合效果. 当 $flag$ 为 1 时, 本文设计了一种基于 $Sigmoid$ 的融合方法, 其具体过程如下:

$$SigmoidMD(M,N) = M + Dropout(Sigmoid(N)\odot M) \tag{10}$$

2.6 事件定位

与 CMRAN[5]相似, 本文使用事件存在分数 $S^j \in \mathbb{R}^{T\times 1}$ 和事件类别分数 $S^e \in \mathbb{R}^{T\times C}$ 定位每个视频片段的事件.

在弱监督训练中, 本文通过点乘获得 S^j 和 S^e 的联合分数并进行最大池化, 最终获得视频级分数 $S^E \in \mathbb{R}^C$, 并将其纳入损失函数的计算中. 总体损失是 S^E 和视频级标签 Y^E 之间的多标签交叉熵损失. 在全监督训练中, 本文将噪音弱化模块中的时序门控分数 (m_v^w, m_a^w) 和时间对齐模块中的注意力图 m_v^{SCI} 纳入损失函数的计算中. 总体损失函数如下:

$$L_f = L^S + L^J + L_t^N + L_t^{SCI} \tag{11}$$

其中, L^S 代表片段级真实标签 Y^S 和 S^e 之间关于事件类别的交叉熵损失. L^J 代表 Y^S 和 S^j 之间的二进制交叉熵损失. L_t^N 代表 Y^S 和 m_{va}^w 之间的二进制交叉熵损失, 其中 m_{va}^w 是 m_v^w 和 m_a^w 的乘积. L_t^{SCI} 代表 Y^S 和 m_v^{SCI} 之间事件相关度的二进制交叉熵损失.

3 实验分析

3.1 数据集和评价指标

与大部分聚焦 AVEL 任务的工作相同[1−8], 本文使用文献[1]提出的 AVE 数据集作为实验数据集. AVE 数据集是 AudioSet 数据集的子集, 其包含 4 143 个视频且每个视频长度都为 10 s. 除背景事件外, 每个视频中只有 1 类事件且在视频中连续出现. 所有视频一共涵盖与人类活动、动物活动和载具等相关的 28 个事件类别. 与基准模型 CMRAN[5]相同, 本文将 AVE 数据集划分成训练、验证和测试 3 个部分, 其分别包含 3339、402 和 402 个视频数据, 并采用总体准确率 (overall accuracy) 作为弱监督和全监督 AVEL 任务的评价指标.

3.2 实验设置

本文实验基于 PyTorch 实现, 硬件配置环境为 NVIDIA GeForce RTX 3060、内存 24 GB 和处理器 i5. 本文使用在 ImageNet 数据集上预训练的 VGG-19 和 ResNet-151 网络分别提取尺寸为 7×7×512 和 7×7×2048 的视觉特征, 并采用在 AudioSet 数据集上预训练的 VGG-like 网络提取尺寸为 128 的音频特征. 训练参数中, batch size 为 64 并且采用 Adam 作为优化器, 迭代次数为 200. 学习率为 0.000 6 并且分别在第 10 轮、20 轮和 30 轮训练时衰减为原本的一半. 超参数 α 值为 0.5.

3.3 对比实验

本节中, 本文将 CMTAN 与基准方法 CMRAN[5]以及 PSP[8]和 BMFN[6]等优秀方法在弱监督和全监督 AVLE 任务中进行了相同实验. 其中, CMRAN 通过注

意力机制建立音频和视觉模态信息之间的联系, PSP 通过构建两种模态间的相似度图处理正样本对和负样本对, 两种方法都具有视音频特征对齐的作用且视音频事件定位效果优秀, 具有较高的对比价值. 为保证对比公平, 所有对比方法都分别使用 VGG-19 和 VGG-like 提取视频的视觉特征和音频特征, 且都采用总体准确率作为实验指标.

图 5 展示了弱监督 AVEL 任务中 CMTAN 和其他方法的对比结果. CMTAN 的总体准确率达到了 74.9%, 相较于基准方法 CMRAN[5]提升了 2%. 相较于其他优秀的方法 PSP[8]和 BMFN[6], CMTAN 的总体准确率分别提高了 1.4% 和 0.9%. 实验对比结果可以看出, 本文提出的 CMTAN 在弱监督 AVEL 任务中的效果优秀.

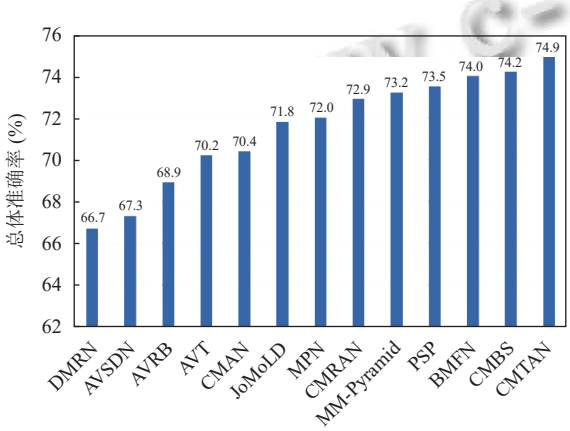


图 5 弱监督 AVEL 任务中的实验结果

图 6 展示了全监督 AVEL 任务中 CMTAN 和其他方法的对比结果. CMTAN 的总体准确率达到了 79.2%, 相较于基准方法 CMRAN[5], 其实验结果提升了 1.8%. 相较于其他优秀方法 PSP[8]和 BMFN[6], CMTAN 的总体准确率分别提高了 1.4% 和 0.5%.

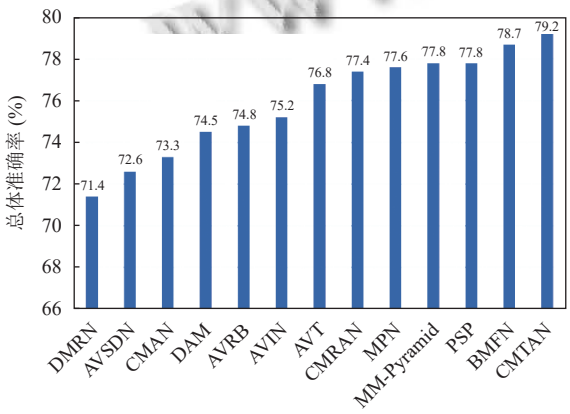


图 6 全监督 AVEL 任务中的实验结果

为研究 CMTAN 的通用性, 本文分别使用 RestNet-151 和 VGG-like 提取了视频的视觉特征和音频特征, 并和其他方法进行了相同的实验, 实验结果如表 1 所示. CMTAN 在弱监督和全监督 AVEL 任务中的实验结果并没有因为采用了其他特征提取器提取的特征而出现剧烈的波动, 实验结果略高于基准方法. 实验表明, CMTAN 对源于不同特征提取器提取的特征能保持稳定的事件定位效果.

表 1 弱监督和全监督 AVEL 任务中的实验结果 (%)

方法	弱监督	全监督
Visual	63.4	65.0
AVEL	71.6	74.0
AVSDN	74.2	75.4
CMRAN	75.3	78.3
CMTAN	75.4	78.5

3.4 消融实验

为验证 CMTAN 中每个部分的必要性和有效性, 本文通过移除或替换相关模块进行消融实验.

w/o 预处理: 使用一个能进行视觉信息的空间信息自我关注模块替换该部分的所有模块. 在验证预处理部分效果的同时, 使视觉模态的特征维度与后续部分特征对齐.

w/o 跨模态交互: 去除特征交互模块, 仅保留了特征强化, 以研究特征交互的作用.

w/o 时间对齐: 直接删除该部分的时间对齐模块.

w/o 特征融合: 使用平均方法替换特征融合部分的多阶段融合模块.

表 2 展示了消融实验的实验结果. 从表中可知, 通过移除或替换不同部分的模块, CMTAN 的总体准确率都出了不同程度的下降. 其中, 预处理部分对 CMTAN 的总体准确率影响最大, 去除时间对齐部分也对网络的总体准确率造成了比较大的影响.

表 2 消融实验结果 (%)

消融部分	弱监督	全监督
w/o 预处理	72.3	76.5
w/o 跨模态交互	71.8	77.0
w/o 时间对齐	73.1	77.8
w/o 特征融合	73.0	77.0
完整网络	74.9	79.2

此外, 为研究预处理部分中噪音弱化模块的实验效果, 本文在表 3 中展示了针对该模块的消融实验. 从表 3 可知, 在 VGG-19 和 ResNet-151 提取的视觉特征

上进行的实验中, 噪音弱化模块使 CMTAN 的总体准确率分别提升了 0.4% 和 0.3%. 由于该模块增加的训练参数非常少且计算成本低, 因此其带来的性能提升是可接受的.

表 3 噪音弱化模块的消融实验

模块	特征提取器	总体准确率 (%)
w/o 噪音弱化	VGG-like, VGG-19	78.8
完整网络	VGG-like, VGG-19	79.2
w/o 噪音弱化	VGG-like, ResNet-151	78.2
完整网络	VGG-like, ResNet-151	78.5

3.5 超参数实验

为探索时间对齐模块中超参数α对网络性能的影响, 本文通过调整α值进行了超参数实验. α的值越大, 时间对齐模块中交互前和交互后特征的注意力图在合成的注意力图中占比越大, 模块对交互前单模态全局信息越重视. 图 7 展示了超参数α的实验结果. 结果显示, 当$\alpha<0.5$时, 随着α的上升, CMTAN 的总体准确率呈上升的趋势. 当$\alpha=0.5$时, CMTAN 的性能达到最佳. 当$\alpha>0.5$时, CMTAN 的性能出现下降. 由此可见, 均衡地合成两张注意力图能显著提高网络的定位效果, 这代表过多或过少考虑模态交互前的模态全局信息都会影响模态间的信息交互.

3.6 质量分析

图 8 展示了 CMTAN 对两个视频的事件定位效果. 图 8(a) 中所有视频片段的视觉信息中都出现了钟, 但仅有最后 5 个视频片段的音频信息中出现了钟声. 因此, CMTAN 很容易联合两种模态信息定位到最后 5 个视频片段中的“教堂钟”事件. 图 8(b) 的例子更加复杂, 第 6 个和第 7 个视频片段的视觉信息中狗的图像几乎没有改变, 但仅有第 6 个片段的音频信息中出现了狗叫声, 因此 CMTAN 通过音频信息成功定位到具有“狗吠”事件的视频片段.

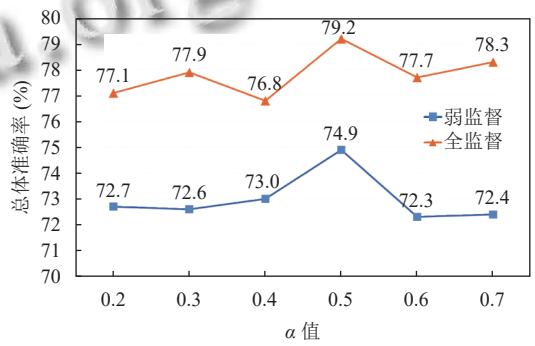


图 7 超参数α的实验结果

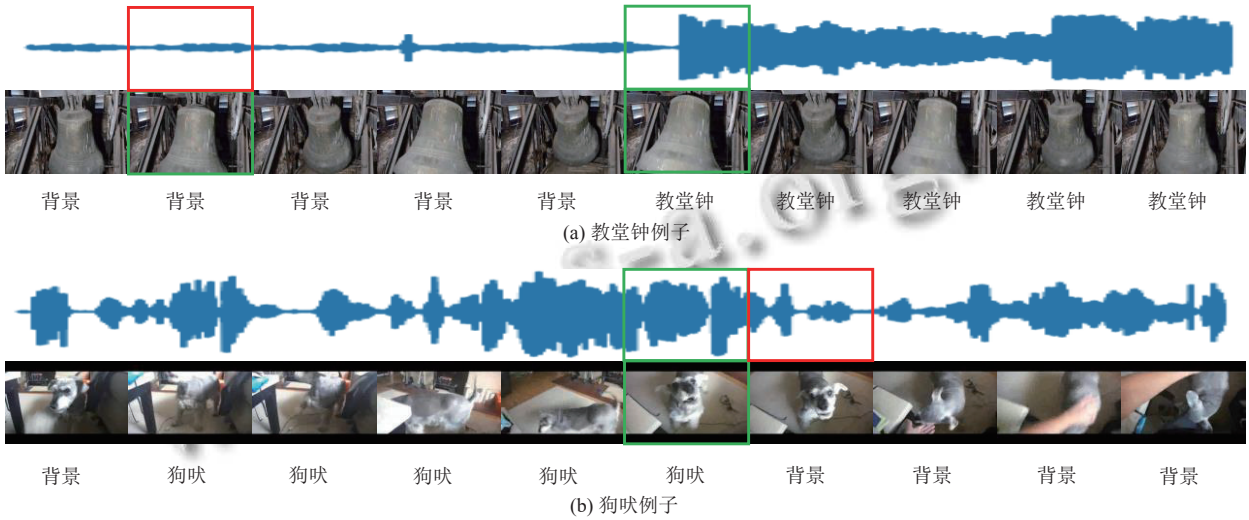


图 8 CMTAN 的事件定位效果

值得注意的是, 图 8(b) 中第 3 个视频片段的视觉信息出现了狗, 但是音频信息的狗叫声比较微弱. 如果第 3 个视频片段因为仅有微弱的狗叫声被定位为“背景”, 由于 AVE 数据集中事件仅连续出现一次, 那么第 2 个视频片段就很可能会被网络定位为“背景”, 产生误判, 这对 AVEL 任务的方法而言是一个很大的挑战. 然而, 本文设计的 CMTAN 模型基于时间对齐模块, 通过对模态全局信息的分析, 准确预测了事件边界.

此外, 本文根据文献[6,8]的代码对 PSP 和 BMFN 进行复现, 并对图 8(b) 中例子进行事件定位, 其定位效果如图 9 所示.

图 9 中, PSP 和 BMFN 能够定位到该复杂例子中

事件的大概位置, 对少数音频和视觉信息匹配不明显的视频片段的定位效果欠佳, 但本文提出的 CMTAN 通过对齐两种模态信息精准定位到了所有视频片段的事件.

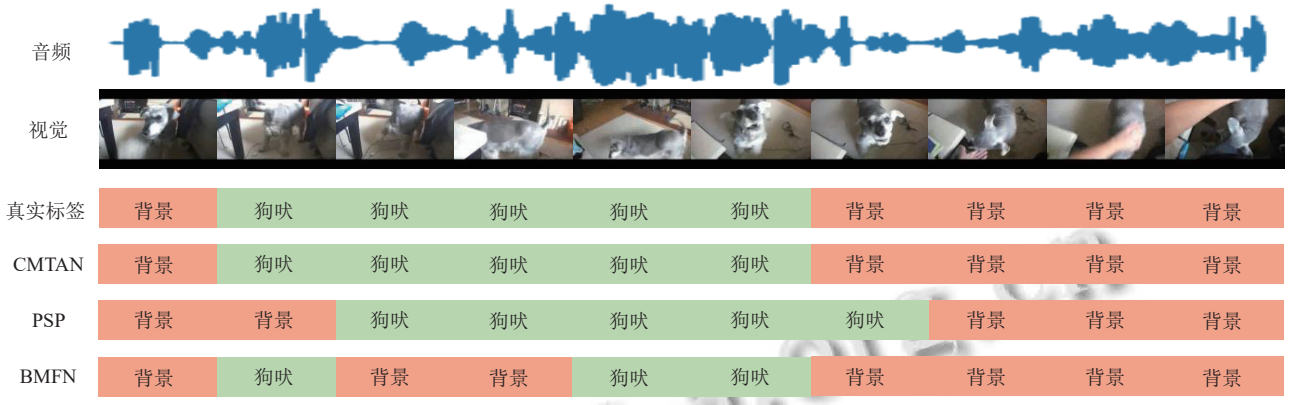


图 9 不同方法对相同视频中事件定位的效果比较

4 结论与展望

本文针对 AVEL 任务设计了一种跨模态时间对齐网络. 该网络利用注意力机制聚焦于不同模块输出的全局特征进行了特征对齐, 这能帮助网络对事件边界的预测. 此外, 基于注意力机制和相似度机制, 本文设计了一种新的跨模态音频指导模块和一种噪音弱化模块, 这些模块帮助网络精确分辨事件的类别. 实验证明, 本文提出的网络在弱监督和全监督 AVEL 任务中都具有优秀的效果. 未来工作将聚焦于跨模态交互, 探寻缓解 AVEL 任务过拟合问题的方法.

参考文献

1 Tian YP, Shi J, Li BC, *et al*. Audio-visual event localization in unconstrained videos. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 252–268.

2 Lin YB, Li YJ, Wang YCF. Dual-modality Seq2Seq network for audio-visual event localization. Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton: IEEE, 2019. 2002–2006.

3 Wu Y, Zhu LC, Yan Y, *et al*. Dual attention matching for audio-visual event localization. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 6291–6299.

4 Xuan HY, Zhang ZY, Chen S, *et al*. Cross-modal attention network for temporal inconsistent audio-visual event localization. Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York: AAAI, 2020. 279–286.

5 Xu HM, Zeng RH, Wu QY, *et al*. Cross-modal relation-aware networks for audio-visual event localization. Proceedings of the 28th ACM International Conference on Multimedia. Seattle: ACM, 2020. 3893–3901.

6 Liu S, Quan WZ, Liu Y, *et al*. Bi-directional modality fusion network for audio-visual event localization. Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore: IEEE, 2022. 4868–4872.

7 Xia Y, Zhao Z. Cross-modal background suppression for audio-visual event localization. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 19957–19966.

8 Zhou JX, Zheng L, Zhong YR, *et al*. Positive sample propagation along the audio-visual event line. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 8432–8440.

9 Arandjelovic R, Zisserman A. Look, listen and learn. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 609–617.

10 Long X, Gan C, de Melo G, *et al*. Multimodal keyless attention fusion for video classification. Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans: AAAI, 2018. 7202–7209.

11 Hori C, Hori T, Wichern G, *et al*. Multimodal attention for fusion of audio and spatiotemporal features for video description. Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City: IEEE, 2018. 2528–2531.

12 Nawaz S, Janjua MK, Gallo I, *et al*. Deep latent space learning for cross-modal mapping of audio and visual signals. Proceedings of the 2019 Digital Image Computing:

Techniques and Applications. Perth: IEEE, 2019. 1–7.
13 Lee J T, Jain M, Park H, *et al*. Cross-attentional audio-visual fusion for weakly-supervised action localization. Proceedings of the 9th International Conference on Learning Representations. Vienna: OpenReview.net, 2021. 1–17.
14 Li XY, Liu J, Xie YR, *et al*. MAGDRA: A multi-modal attention graph network with dynamic routing-by-agreement for multi-label emotion recognition. Knowledge-based Systems, 2024, 283: 111126. [doi: 10.1016/j.knosys.2023.111126]
15 Gao RH, Oh TH, Grauman K, *et al*. Listen to look: Action recognition by previewing audio. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 10454–10464.
16 Cheng Y, Wang RZ, Pan ZH, *et al*. Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning. Proceedings of the 28th ACM International Conference on Multimedia. Seattle: ACM, 2020. 3884–3892.
17 Hu D, Nie FP, Li XL. Deep multimodal clustering for unsupervised audiovisual learning. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2018. 9240–9249.
18 Alwassel H, Mahajan D, Korbar B, *et al*. Self-supervised learning by cross-modal audio-video clustering. Proceedings of the 34th Conference on Neural Information Processing Systems. Vancouver, 2020. 9758–9770.
19 Zhang JR, Xu X, Shen FM, *et al*. Enhancing audio-visual association with self-supervised curriculum learning. Proceedings of the 35th AAAI Conference on Artificial Intelligence. AAAI, 2021. 3351–3359.
20 Morgado P, Vasconcelos N, Misra I. Audio-visual instance discrimination with cross-modal agreement. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 12470–12481.
21 冼广铭, 阳先平, 招志锋. 基于双编码器表示学习的多模态情感分析. 计算机系统应用, 2024, 33(4): 13–25. [doi: 10.15888/j.cnki.csa.009461]
22 齐泽华. 基于 Transformer 的音视频事件定位的研究 [硕士学位论文]. 成都: 电子科技大学, 2023.
23 Cheng HY, Liu ZY, Zhou H, *et al*. Joint-modal label denoising for weakly-supervised audio-visual video parsing. Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022. 431–448.
24 Lin YB, Wang YCF. Audiovisual Transformer with instance attention for audio-visual event localization. Proceedings of the 15th Asian Conference on Computer Vision. Kyoto: Springer, 2021. 274–290.
25 Yu JS, Cheng Y, Feng R. MPN: Multimodal parallel network for audio-visual event localization. Proceedings of the 2021 IEEE International Conference on Multimedia and Expo. Shenzhen: IEEE, 2021. 1–6.
26 Liu HW, Gu XD. Masked co-attention model for audio-visual event localization. Applied Intelligence, 2024, 54(2): 1691–1705. [doi: 10.1007/s10489-023-05191-2]
27 Yu JS, Cheng Y, Zhao RW, *at al*. MM-pyramid: Multimodal pyramid attentional network for audio-visual event localization and video parsing. Proceedings of the 30th ACM International Conference on Multimedia. Lisboa: ACM, 2022. 6241–6249.
28 Ramaswamy J. What makes the sound? A dual-modality interacting network for audio-visual event localization. Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona: IEEE, 2020. 4372–4376.
29 Ramaswamy J, Das S. See the sound, hear the pixels. Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision. Snowmass: IEEE, 2020. 2959–2968.
30 Wang XL, Girshick R, Gupta A, *et al*. Non-local neural networks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7794–7803.
31 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.
32 Lee S, Kim D, Han B. CoSMo: Content-style modulation for image retrieval with text feedback. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 802–812.

(校对责编: 张重毅)