Causal-ASD:复杂场景下基于因果推断的主动说话人检测方法

刘若玲[1],周斌[1*],胡波[2]

(1 中南民族大学 计算机学院(人工智能学院),湖北 武汉 430074;2 武汉东信同邦信息技术有限公司,湖北 武汉 430074)

摘 要 主动说话人检测(Active Speaker Detection)旨在从多说话人场景中精准定位当前发声者,然而现有方法在复杂现实环境中常因背景噪声干扰、视觉信息遮挡及跨模态数据偏差等问题面临性能瓶颈.其根本挑战源于现有模型倾向于拟合训练数据中的虚假相关性(如唇部运动与无关背景语音的偶然共现)及模态间隐式混杂因素,导致泛化能力显著受限.为突破此瓶颈,提出了一种基于因果推断的端到端框架Causal-ASD,该框架构建了音频-视觉特征与检测结果的因果图结构,引入干预操作(Intervention)与前门调整(Front-door Adjustment)策略,从特征生成层面阻断非因果捷径路径,仅保留直接因果效应相关的特征表示.进一步地,框架设计对抗判别器以动态消除残余混杂因子,并结合双向门控循环单元(Bi-GRU)与自注意力机制强化时空上下文建模能力.实验结果表明:Causal-ASD在AVA-ActiveSpeaker数据集和WASD数据集上分别取得0.8%和1.5%的绝对性能提升.该方法为多模态场景下的因果可解释性建模提供了新范式,尤其在低质量输入条件下展现出更强的环境适应性.

关键词 主动说话人检测;因果推断;跨场景泛化

中图分类号 TP391.4 **文献标志码** A **文章编号** 1672-4321(2026)05-0697-12

doi:10.20056/j.cnki.ZNMDZK.20260711

Causal-ASD:A causal inference approach for active speaker detection in complex scenarios

Liu Ruoling[1],Zhou Bin[1*],Hu Bo[2]

(1 College of Computer(College of Artificial Intelligence), South-Central Minzu University, Wuhan 430074, Hubei China; 2 Wuhan Dongxin Tongbang Information Technology Co., Ltd.,Wuhan 430074, Hubei China)

Abstract Active Speaker Detection (ASD) aims to accurately identify the current speaker in multi-speaker scenarios. However, existing methods often encounter performance bottlenecks in complex real-world environments due to background noise interference, visual occlusion, and cross-modal data biases. The fundamental challenge lies in the tendency of current models to fit spurious correlations in training data (e.g., the coincidental co-occurrence of lip movements and unrelated background speech) as well as implicit confounders across modalities, which significantly limits their generalization ability. To address this issue, an end-to-end framework based on causal inference is proposed, termed Causal-ASD. Specifically, Causal-ASD constructs a causal graph structure between audio-visual features and detection outcomes, and incorporates intervention and front-door adjustment strategies to block non-causal shortcut paths at the feature generation level, preserving only feature representations relevant to direct causal effects. Furthermore, an adversarial discriminator is designed to dynamically eliminate residual confounding factors, while a bidirectional gated recurrent unit (Bi-GRU) combined with the self-attention mechanism is employed to enhance spatio-temporal context modeling ability. Experimental results demonstrate that Causal-ASD achieves absolute performance gains of 0.8% and

收稿日期 2025-09-15

* **通信作者** 周斌(1971-),男,教授,博士,研究方向:大数据处理,E-mail: binzhou@mail.scuec.edu.cn

基金项目 中央高校基本科研业务费专项资金资助项目(CZY23006)

1.5% on the AVA-ActiveSpeaker and WASD datasets, respectively. This method introduces a new paradigm for causal interpretability modeling in multimodal scenarios and exhibits stronger environmental adaptability, particularly under low-quality input conditions.

Keywords active speaker detection; causal inference; cross-scene generalization

主动说话人检测(Active Speaker Detection, ASD)是计算机视觉与语音信号处理交叉领域的重要研究方向,其目标是在多说话人场景中准确识别并定位当前说话人.随着视频会议、在线教育、智能监控以及人机交互系统的快速发展,对高精度、实时化的说话人检测需求日益增长.与传统的语音识别或人脸检测任务不同,ASD不仅需要同时建模音频和视频两种模态的信息,还需要有效地捕捉两者之间的时序一致性.这种能力对于保证系统在实际场景中对说话行为的敏感度和响应准确性至关重要.

然而,现实环境往往复杂多变,ASD在应用中面临诸多挑战.首先,嘈杂的声学环境常混入大量背景噪声与混响,导致语音特征难以被准确提取,从而增加误检风险;其次,视觉模态同样具有不稳定性,面部遮挡、非语音相关的唇部动作都会干扰模型的判断,降低检测的鲁棒性.更为根本的是,现有方法主要依赖于基于相关性的学习机制,容易捕捉到音频与视频模态间的虚假相关性与隐式混杂因素,例如唇部运动与无关背景语音的偶然共现.这些非因果因素被模型误识为有效信号,不仅影响预测准确性,还显著限制了模型的泛化能力.

为解决上述问题,本研究提出了Causal-ASD,一种融合因果推断的多模态主动说话人检测框架.与现有的轻量级模型Light-ASD[1]相比,Causal-ASD的核心创新在于:

(1)首次在主动说话人检测中引入多模态因果图,显式刻画音频特征、视觉特征与说话状态之间的因果关系;

(2)通过干预机制、前门调整策略与对抗训练去除潜在混杂因素,从而获得具有真正因果关联的鲁棒表征;

(3)结合Bi-GRU[2]与自注意力机制实现跨模态时序建模,进一步提升检测精度.

通过这些设计,Causal-ASD在复杂多变的真实场景中展现更强的鲁棒性与跨场景泛化能力.

1 相关工作

1.1 主动说话人检测

主动说话人检测(ASD)的研究经历了从单模态方法到多模态融合,再到长时序建模的演进过程.

早期方法受限于小规模数据集和传统机器学习模型,ASD研究聚焦于单模态特征分析[3-4],如语音能量阈值、人脸运动等.Saenko等[3]首次提出精确定位唇部区域,通过视觉特征直接关联唇部运动与语音,为后续视觉模态的研究奠定了基础.Everingham等[4]虽然未直接聚焦唇部,但通过定位和裁剪面部区域进行特征计算,构建了早期面部特征提取的基本框架.然而,这类方法在复杂环境下的鲁棒性与泛化能力均较差.

随着Google发布大规模视听标注数据集AVA-ActiveSpeaker[5],以及基于深度学习的突破,ASD研究进入了多模态融合阶段,它们融合经典方法,比如3D卷积神经网络[6-8]、Transformer架构[9-11],这样可同时建模音频和视觉模态的时空信息,有效提高了检测精度.这一阶段的研究推动了时空联合建模的ASD框架[12-16]发展,使得模型在复杂场景下的检测精度显著提升.但是,它们普遍依赖大规模训练数据,计算开销较大,实际应用中受到一定限制.

近年来,ASD研究重心转向长时序特征建模,其核心思路是引入序列模型来捕捉长期时间上下文信息,从而更好地理解说话行为的动态特征[2,7,9,17-19].Tao等[9]利用交叉注意力和自注意力模块聚合音频特征和视觉特征,实现长期时间上下文建模,从而有效区分了说话人和非说话人,但对大规模时序数据依赖较重.Kopuklu等[7]则聚合短期特征,结合双向GRU实现长时序建模.Wang等[17]提出LoCoNet,该模型采用自注意力和交叉注意力机制,分别进行长期的单模态说话人内部建模(Long-term Intraspeaker Modeling, LIM)和短期跨说话人间建模(Short-term Inter-speaker Modeling, SIM).LoCoNet在ASD任务上表现优秀,已成为该领域的主流框架.这类时序建模方法虽然精度高,但过于依赖长时序序列,结构复杂且计算开销大,不适合实时或边缘设备.

为解决这些大规模模型计算开销大、难以满足实时性的问题，Liao 等提出了 Light-ASD，通过轻量化卷积网络提取视觉特征，并结合 Bi-GRU 进行时序建模，从而在保证精度的同时显著降低了参数量和计算复杂度．Light-ASD 在 AVA-ActiveSpeaker 等基准数据集上表现出较强的效率优势，特别适用于实时或资源受限的场景．即使如此，由于其核心机制仍基于相关性建模，Light-ASD 在嘈杂、多说话人、模态不同步等复杂环境下的鲁棒性有限，泛化能力也受到一定约束．这一不足说明，仅依赖轻量化和时序建模并不能充分解决 ASD 面临的混杂与虚假相关问题．

1.2　因果推断

因果推断是一种用于探究事件间因果关系的科学研究方法[20-21]．在日常生活中，两个事件可能频繁共现，却未必具有因果联系．如冰激凌销量与溺水事件同时增加，但这并非由于它们互为因果，而是受季节温度等混杂因素影响．因果推断通过实验设计、数据收集和模型构建，致力于区分真实因果与偶然关联．它在医学、经济学、社会科学等领域应用广泛，例如判断药物有效性或评估政策对社会现象的影响等．

从发展历程来看，因果推断起源于哲学思辨，亚里士多德与休谟等哲学家对因果本质的探讨为其奠定思想理论基础．至 20 世纪初，内曼的潜在结果模型为因果推断建立了数学表达框架，后经鲁宾等人进一步发展为鲁宾因果模型（RCM）[22]，逐步形成完整的理论体系．近年来，随着机器学习与深度学习的兴起，因果森林[23]、因果注意力[24]等融合因果推断思想的方法不断涌现，提升了因果效应估计的可靠性．

在人工智能领域，因果推断的应用日益广泛．Yang 等[24]提出因果注意力机制来应对视觉-语言模型中的混淆因素；Liu 等[25]利用因果图知识消除图像与文本的虚假关联；Qiu 等[26]提出的 CausalRec 模型利用干预和反事实推理降低了推荐任务中的视觉偏差；Kim 等[27]系统总结了因果推断在消除偏差、提升视觉语言任务性能方面的潜力．这些研究成果表明，因果推断能够有效抑制模态间的混淆因素、增强模型泛化能力．受此启发，本文将因果推断引入主动说话人检测（ASD）任务中，旨在通过构建因果机制抑制多模态学习中的虚假相关性，从而提升 ASD 模型在复杂环境下的准确性与鲁棒性．

2　方法

Causal-ASD 是基于 Light-ASD 提出的一个高效、轻量且适用于复杂场景的主动说话人检测框架．如图 1 所示，框架由编码器前端、因果推断模块和模态融合模块三部分组成．编码器前端包括视觉和音频特征编码器，分别获得视觉特征和音频特征．因果推断模块通过构建因果图，并结合干预操作和前门调整来消除音频和视觉模态间的混杂因素．模态融合模块采用门控融合和时序建模等策略融合两个模态的特征，并与因果特征进行深度整合，最终预测当前候选人是否为当前说话人．

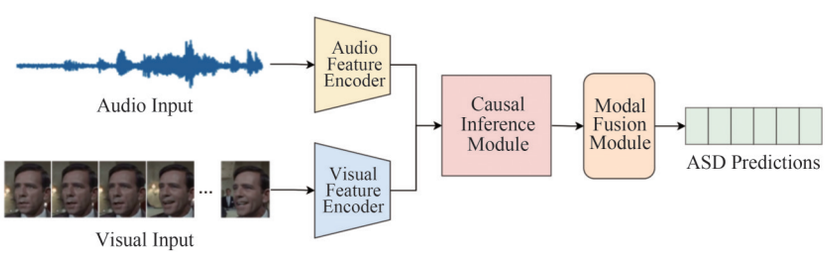


图 1　Causal-ASD 模型的总体架构图

Fig. 1　The overall architecture diagram of the Causal-ASD model

2.1　特征编码器

鉴于 Light-ASD 编码器部分的成熟度和有效性，本研究选择直接沿用它的音频编码器和视频编码器．

2.1.1　音频编码器

音频编码器的任务是从原始音频信号中提取音频特征，为模型提供辨别说话人的关键信息．本文的音频编码器将传统的 2D 卷积分解为两个 1D 卷积操作，分别用于建模音频信号在频率域与时间域的特征分布，如图 2（a）所示．具体来说，首先从原始音频信号中提取由 13 维梅尔频率倒谱系数（MFCCs）和时间信息组成的 2D 特征图，作为音频编码器的输入．随后，音频编码器通过两个 1D 卷积分别处理 MFCCs 维度和时间维度的信息，从而在不同域上捕

捉有效特征.

音频编码器包含3个音频块,每个音频块内设有两条并行路径,分别使用不同尺度的卷积核来提取不同感受野的特征,再利用1×1卷积对不同路径的特征进行融合.这样的设计能够有效捕捉音频信号中的时间动态变化和频率分布信息,同时保持模型的轻量化.此外,为保证跨模态特征对齐,音频编码器在前两层引入时间维度上的最大池化层操作,确保音频特征的时间维度与视觉特征相匹配.最后,通过全局平均池化操作在MFCCs维度上进行池化,得到候选说话人的音频特征F_a.这种音频编码器的设计在保证特征提取效果的同时,大幅减少了模型参数和计算量,提高了模型的运行效率.

2.1.2 视频编码器

视频编码器负责从候选说话人的面部序列中提取视觉特征,是主动说话人检测模型的关键部分之一.与传统的3D卷积神经网络不同,本文的视觉编码器采用将3D卷积分解为2D和1D卷积的方式,分别用于建模空间和时间信息.这种方式在显著降低模型参数数量和计算负担的同时,还能保持良好的性能.

视觉特征编码器如图2(b)所示,由3个视觉块组成,每个视觉块包含两条并行路径,利用不同尺度的卷积核提取多感受野的时空特征.在卷积操作设计上,步长大多为1,仅第一个视觉块中的2D卷积步长为2,这一设计有助于在特征提取过程中降低空间维度,从而减少后续产生的特征图的大小,进而降低内存占用,提高计算速度.最后,视觉编码器对特征进行全局最大池化操作,以获取候选说话人的视觉特征F_v.这种设计不仅显著降低了模型的复杂度,还提高了模型在不同场景下的适应性和鲁棒性,使其能够在资源受限的环境中高效运行,同时保持较高的检测精度.

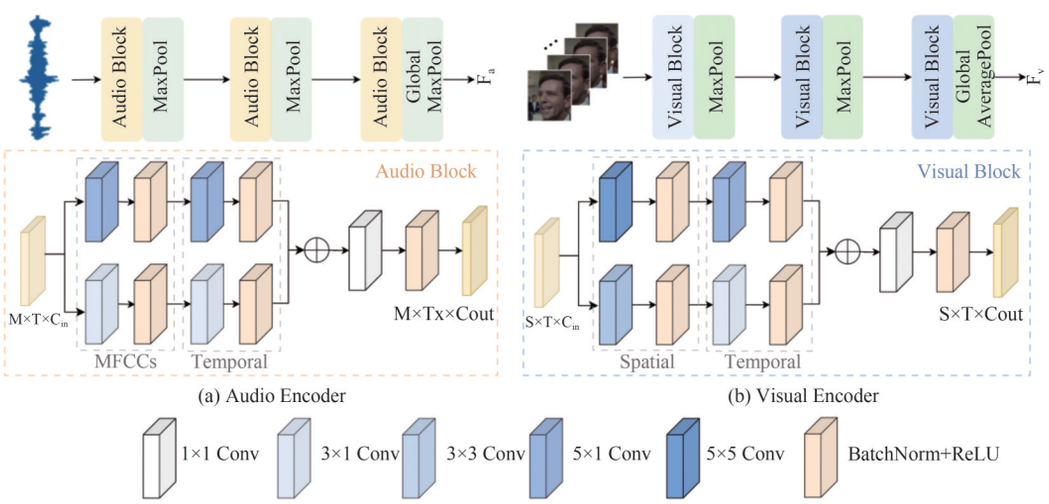


图2 编码器架构及内部模块

Fig. 2 The encoder architecture and internal modules

2.2 因果推断模块

2.2.1 因果图

在主动说话人检测这类复杂的多模态任务中,传统基于相关性的监督学习模型常常受到虚假关联(Spurious Correlations)的干扰.例如,当唇部运动与背景语音偶然同时出现时,模型可能错误地学习到两者的信息,从而在跨场景应用中表现出较差的泛化能力.这种现象在嘈杂环境、多说话人场景以及模态不同步情况下尤为突出,反映出单纯依赖相关性建模的ASD方法缺乏鲁棒性.

本文引入了因果推断的思想,从变量之间的因果机制出发,有效剥离由混杂因素带来的误导性关联.因果推断通常借助因果图(Causal Graph)来描述随机变量之间的因果关系.因果图是一种有向无环图(DAG, Directed Acyclic Graph),能够直观地表达因果结构与信息流向,从而为判别因果路径和非因果路径提供理论工具.通过分析因果图,可以明确哪些路径代表真正的因果影响,哪些路径则是由潜在混杂变量引入的伪相关.

因果图包含3种基础结构:链式结构(chain)、分叉结构(fork)和对撞结构(collider),图3为3个因果图示例,分别表示了这3种结构的因果关系.基于这些原理,本文针对ASD中的变量定义了特定的因果图,如图3(d),该因果图中,视觉特征F_v与音频特征F_a共同影响着结果Y.而隐藏混杂因素C会影响到视觉特征F_v和音频特征F_a,从而影响到结果Y.这种

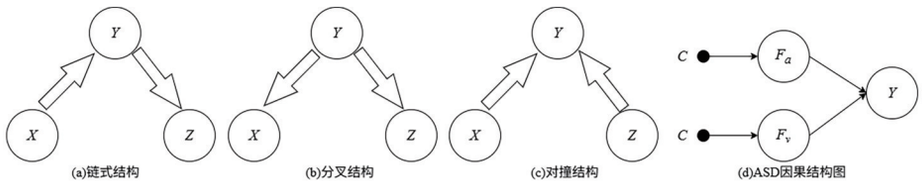


图3　3种因果结构图及ASD因果结构图

Fig. 3　Three causal structures diagrams and ASD causal structure diagram

后门路径可能会误导模型捕获非因果有效的关联，从而损害预测的稳健性和泛化能力.

2.2.2　干预

与传统的观察性推理不同，干预并非依赖观测数据中变量间的统计共现关系，而是通过主动“设定”某个变量的值，来评估其对结果变量的直接影响.这类干预操作通常用do算子(do-calculus)来表示.例如，$X = \mathrm{do}(X = x)$意味着主动将变量X固定为x，而非仅仅观测到它取值为x.结合图3(d)的因果结构，本文在模型中引入了干预算子$\mathrm{do}(F_a)$和$\mathrm{do}(F_v)$，分别对音频和视觉模态特征进行控制性干预.直观地说，这一操作相当于“人为设定”视觉或音频输入，从而断开其与其他潜在混杂源(如背景噪音、非语者唇动)之间的边.借助这一操作，模型能够更明确地识别出$F_v \rightarrow Y$和$F_a \rightarrow Y$的直接因果路径，而不受非因果信息的干扰.

基于以上思想，本研究设计了一个因果推断模块，如图4(a)所示.它用于在主动说话人检测模型中模拟“干预”，从原始联合特征中剥离出具有因果意义的纯净表征.具体而言，本文对输入模态特征进行结构化置零操作来实现“干预”：分别将视觉或音频通道置零，以模拟对另一模态的“外部控制”，并通过同一因果编码器生成单模态干预下的表征z_a和z_v.这两个表征由视觉特征F_v与音频特征F_a经因果干预得到的可学习中间特征，分别代表了在控制视觉或音频输入为零时的模型响应，即反映了各模态单独对预测结果的影响.为了进一步消除潜在混杂路径带来的伪相关，本研究在此基础上引入前门调整准则，其核心思想是从原始融合特征z中剔除由模态特有响应引入的非因果成分，得到因果净化后的特征z^*.

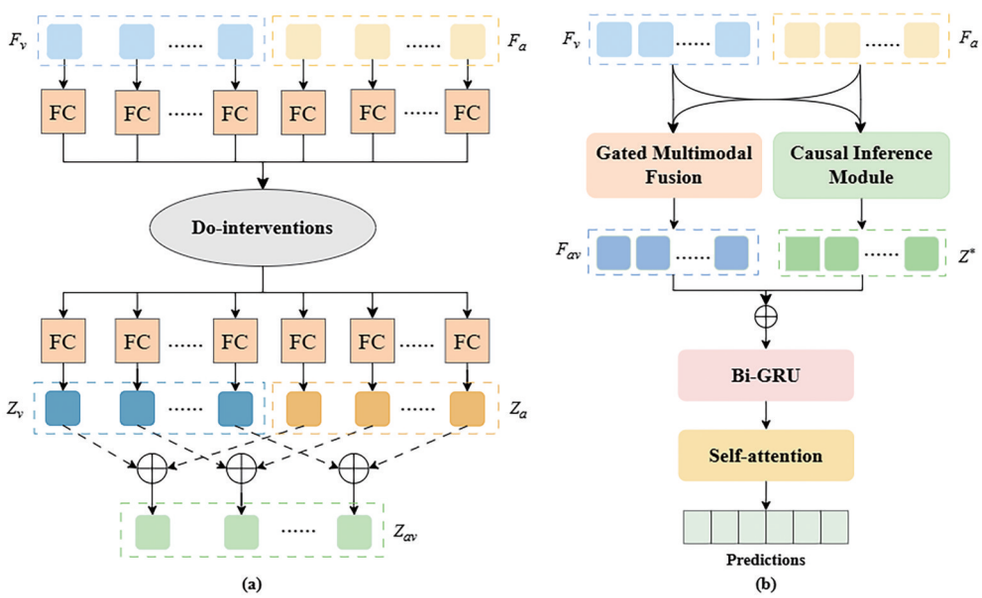


图4　因果推断模块和模态融合模块

Fig. 4　Causal inference module and modality fusion module

而本研究采用的前门调整遵循标准的可识别性条件如下：

(1) 所有从视觉特征F_v或音频特征F_a到Y的路径均通过z_v或z_a传递；

(2) 不存在直接从F_v或F_a到Y的未观测混杂；

(3) z_v或z_a阻断了所有后门路径.

在满足上述条件的情况下，干预分布$P(Y|\mathrm{do}(F_v))$或$P(Y|\mathrm{do}(F_a))$均可由观测分布识别，从而保证因果可识别性[20].基于这一理论假设，且因果路径在神经网络中无法显示计算，本文在网络结

构实现中采用近似形式的可微干预-前门调整操作.具体地,将模态融合表示z视为原始多模态联合特征,由视觉分支与音频分支分别生成中介响应z_v和z_a,前门调整公式近似为:

$$z^* = z - \alpha\cdot\left(z_v + z_a\right), \tag{1}$$

其中,z^*由视觉特征F_v和音频特征F_a融合而得,$\alpha \in [0,1]$为干预强度系数.该操作从原始特征中减去模态特有响应,去除由模态同步现象(如非语音的唇动、背景噪声)所引入的混杂信息,得到净化后的因果特征z^*.该因果特征更准确地保留了视觉路径$F_v \rightarrow Y$和音频路径$F_a \rightarrow Y$的直接因果效应,避免了模态间的伪相关性干扰.

因果推断模块的目标是从结构上消除源于数据生成过程中的系统性偏差,使模型学习到反映真实因果关系的特征,即侧重于结构性混杂.尽管经过因果建模后,表示空间中仍可能存在某些模态特有的伪相关模式,例如仅凭唇动或短暂的背景声音即可触发说话人判定.这类"捷径"特征属于局部残留的混杂信息,无法通过因果结构直接建模.

为此,引入了生成对抗网络(Generative Adversarial Networks,GAN)中的对抗判别器模块.判别器试图从因果模块输出的特征中重新预测标签,而主模型则通过对抗学习的方式抑制这些可被判别器识别的模态依赖信息.通过这种博弈机制,模型能够自适应地去除因果模块未能完全消除的残余混杂,从统计层面进一步净化表示.对抗判别器侧重于残余混杂,从特征统计分布上削弱模型对局部模态偏差的依赖.

2.3 模态融合模块

模态融合模块如图4(b)所示.该模块旨在将视觉特征F_v和音频特征F_a进行有效的动态融合,得到F_{av},从而综合利用两种模态的互补信息,增强模型对说话人行为的判断能力.本文引入门控融合机制,其核心思想是通过一个可学习的门控权重自适应地控制两种模态在不同场景下的贡献比例,具体计算公式如下:

$$g = \sigma\left(W_v F_v + W_a F_a + b\right), \tag{2}$$

$$F_{av} = g \odot F_v + \left(1 - g\right) \odot F_a, \tag{3}$$

其中,W_v和W_a为可学习的线性投影矩阵,b为偏置项,$\sigma(\cdot)$表示Sigmoid激活函数.Sigmoid激活函数将输入映射到$[0,1]$区间,使得g成为一个模态重要性权重,即表示当前时刻下视觉模态所占的相对比重.这种门控融合机制可自主学习到何时应强化视觉特征,何时应强化音频特征,在特征层面实现了模态间的自适应加权.在得到门控融合特征F_{av}后,模块将其和因果特征z^*进行拼接,并输入到配备自注意力的双向GRU中进行时序建模.该过程能够在时间维度上进一步挖掘模态之间的动态交互关系,使模型在面临复杂场景时,能够更准确地识别并定位当前说话人.

2.4 损失函数

本文的损失函数由3部分组成:主分类器损失、视觉辅助分类器损失和对抗判别器损失.主分类器损失通过交叉熵损失将预测的标签序列与真实标签序列进行比较,其函数式如下:

$$\mathrm{Loss} = -\frac{1}{T}\sum_{i=1}^{T}\left(y_i\cdot\log s_i + \left(1 - y_i\right)\cdot\log\left(1 - s_i\right)\right), \tag{4}$$

其中s_i和y_i是第i个视频帧的预测和真实ASD标签,$i \in [1,T]$,T指视频帧的数量.

视觉辅助分类器仅依赖于面部模态信息进行二分类判断,它同样采用交叉熵损失进行训练.对抗判别器损失函数如式(5)所示:

$$\mathrm{Loss}_{\mathrm{adv}} = -\frac{1}{T}\sum_{i=1}^{T}\left(y_i\cdot\log\widehat{y_i} + \left(1 - y_i\right)\cdot\log\left(1 - \widehat{y_i}\right)\right), \tag{5}$$

$\widehat{y_i} \in [0,1]$是判别器对第i帧的预测概率.完整的损失函数L为:

$$L = L_{\mathrm{av}} + \lambda_1 L_v + \lambda_2 L_{\mathrm{adv}}, \tag{6}$$

其中L_{av}和L_v表示主分类器和视觉辅助分类器的损失函数,权重系数λ_1和λ_2设定为0.5.

3 实验

3.1 实验设置

所有实验均在Intel(R) Xeon(R) Silver 4108 CPU、32GB内存以及NVIDIA RTX A6000 GPU的硬件环境下完成,保证了大规模数据处理和模型训练的高效性.为兼顾显存占用与收敛速度,实验采用帧级动态批采样:每次迭代加载约2000帧构成一个批次,并根据最长序列自动调整实际视频条数.优化器选用Adam,初始学习率设为0.001,并配合步长为1epoch的指数衰减,衰减系数0.95,使学习率在30个训练周期内平滑下降78%.

实验所用数据集为AVA-ActiveSpeaker和WASD,这两个数据集包含大量真实场景下的多模态视频片段,涵盖多种背景噪声、面部遮挡情况,能够全面评估模型在复杂环境下的鲁棒性与泛化能力.

在数据预处理方面,首先对视频帧中的人脸区

域进行裁剪，并统一调整为112×112像素，以确保输入数据的一致性．在数据增强方面，视觉模态采用随机缩放裁剪、水平翻转与旋转等操作，以提升模型对姿态变化和视角多样性的适应能力；在音频模态中，则通过从训练集中随机选取另一段音频并叠加为噪声的方式进行增强，从而模拟复杂的声学环境，增强模型在嘈杂背景下的鲁棒性．

3.1.1　数据集

AVA-ActiveSpeaker是ASD领域的标准基准测试数据集[5]，该数据集由262部好莱坞电影组成，包含约365万个标记帧，相当于约38.5 h的面部轨迹以及相应的音频数据．标注信息丰富多样，包括视频标识符、关键帧时间戳、人脸边界框坐标、标签序号和人物编号等．凭借其高质量的数据和详细的标注，AVA-ActiveSpeaker广泛用于多种场景，如发言者分类、会议视频重新定位、语音增强以及人机交互等．

WASD数据集[28]通过提升音频和面部信息的复杂性，专门针对ASD任务进行了优化，它包含164个视频，总计30 h的视频注释，并且分为训练集和验证集，划分比例与AVA-ActiveSpeaker数据集相似．WASD数据集分为5个类别，如图5所示，依次为最佳条件OC、语言障碍SI、面部遮挡FO、人声噪声HVN和监控设置SS，图像来源于WASD数据集官方网站．从最佳条件到监控设置，数据集逐步增加了对主动说话者检测的难度，同时还提供了人体姿势注释数据，以支持更全面的多模态分析．除了类别划分，还根据难易程度划分了简单和困难两个组，简单组包含OC和SI，这两个类别类似于AVA-ActiveSpeaker；而困难组包括FO、HVN和SS，这3个类别的检测难度比AVA-ActiveSpeaker高，混杂因素更多．

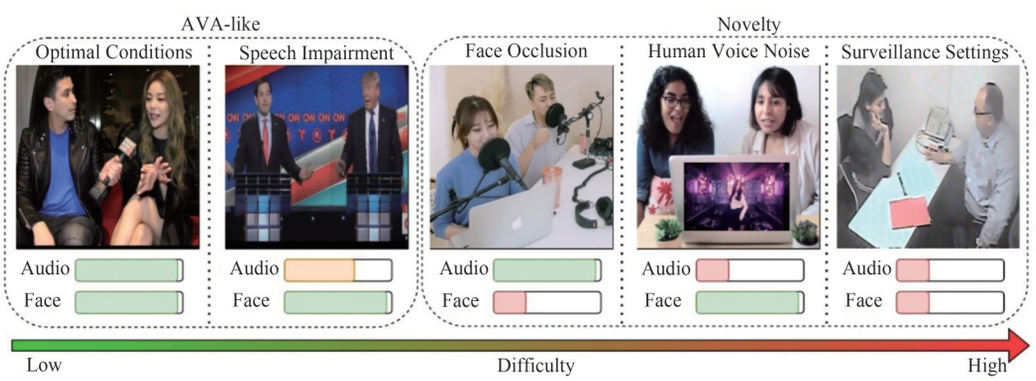


图5　WASD数据集的5个类别

Fig. 5　Five categories of the WASD dataset

与AVA-ActiveSpeaker数据集相比，WASD数据集更具挑战性，能够更好地模拟真实世界中的复杂场景，有助于推动ASD技术应用在更广泛的场景中．

3.1.2　评估指标

在评估指标方面，根据通用协议，AVA-ActiveSpeaker数据集和WASD数据集均采用平均值平均精度（mAP）作为核心评估指标，本文使用官方的ActivityNet评估工具进行计算．此外，为全面比较模型性能，还计算了模型参数量和浮点运算次数（FLOPs），以衡量不同方法在规模与复杂性上的差异．

3.2　性能对比

本文将所提出的Causal-ASD与其他端到端的主动说话人检测方法进行了对比．其中ASDNet[7]作为最早的端到端框架之一，为后续研究提供了基线参考．TalkNet[9]与TS-TalkNet[10]通过交叉与自注意力机制增强音视频交互建模，尤其在长时序依赖建模方面表现突出．EASEE-50[6]与Light-ASD[1]聚焦轻量化设计，在降低计算复杂度和适应资源受限环境方面具有优势．LoCoNet[17]则强调局部与全局时序建模的结合，而LR-ASD[29]在提升模型的鲁棒性方面提出了新的思路．通过涵盖从传统端到端、注意力机制到轻量化与鲁棒性优化等多种代表性方法，本研究能够在不同维度上充分验证所提方法的性能优势与适用性．

每一种方法的结果均是基于其开源代码复现而得，复现的数据与官方论文中的数据存在细微差异．其中，复现基线模型Light-ASD过程严格遵循开源代码，包括数据预处理和超参数设置等实验细节均与原代码一致，但其mAP比论文中公布的数据低．这种差异可能源于实验环境差异．

3.2.1 AVA-ActiveSpeaker数据集

从表1可以看出,本文提出的Causal-ASD在多个指标上均显著优于现有模型.与基线模型Light-ASD相比,本方法仅增加0.08M的参数量和0.3GFLOPs的计算量,就将mAP提升了0.82%,这一性能提升并非简单依赖更大规模的网络或预训练权重,而是源于模型结构中引入的因果推断机制.通过在特征层面施加干预操作并利用前门调整去除潜在混杂因素,模型能够更有效地捕捉视觉与音频模态的直接因果效应,避免了背景噪声、非语者唇动等伪相关信息的干扰,从而得到更纯净、更鲁棒的表征.相比之下,部分研究(如ASDNet、EASEE-50)依赖预训练模型和3D卷积结构来提升多候选输入下的表征能力,但这种方式对计算和存储资源要求较高.Causal-ASD则通过因果干预与模态融合策略,在单候选输入和端到端训练框架下,无需额外后处理即可实现高精度检测,兼顾了性能与效率.

表1 各种方法在AVA-ActiveSpeaker验证集上的性能比较

Tab. 1 Performance comparison for methods on the AVA-ActiveSpeaker validation set

Method	Single candidate	Pre-training	E2E	Params/M	FLOPs/G	mAP/%
ASDNet[7]	×	√	×	51.30	14.90	91.10
EASEE-50[6]	×	√	√	26.80	4.30	93.60
LoCoNet[17]	×	√	√	34.33	4.86	93.52
TalkNet[9]	√	×	√	15.70	1.50	91.80
TS-TalkNet[10]	√	×	√	36.80	2.30	92.70
Light-ASD[10]	√	×	√	1.02	0.60	93.41
LR-ASD[29]	√	×	√	0.84	0.51	93.74
Causal-ASD	√	×	√	1.10	0.90	94.23

3.2.2 WASD数据集

为了全面评估本文提出方法的泛化能力,本文在WASD数据集上进行了详细的性能分析,结果详见表2. WASD数据集的5个类别为:最佳条件(OC)、语音障碍(SI)、面部遮挡(FO)、人声噪声(HVN)、监控设置(SS),涵盖了从理想条件到极具挑战性的场景,可全面评估模型在不同环境下的鲁棒性.本文设计了两种模型训练设置:仅在AVA-ActiveSpeaker数据集上训练和仅在WASD数据集上训练,两种训练条件下,Causal-ASD的性能分别比基线模型Light-ASD高出1.46%和1.86%.这种提升表明本文提出的因果推断模块能够有效消除音视频模态间的虚假相关,使模型在面对语音缺陷、面部部分遮挡以及强背景噪声时,仍然能够捕捉到因果有效的信号路径.

表2 各种方法在WASD验证集上的性能比较

Tab. 2 Performance comparison for methods on the WASD validation set

Method	Train Set		Easy		Hard			mAP/%
	AVA	WASD	OC	SI	FO	HVN	SS	
ASDNet	√	×	91.12	90.45	78.20	74.91	48.12	79.25
TalkNet	√	×	91.60	92.98	86.41	77.26	64.66	85.03
TS-TalkNet	√	×	91.09	93.70	88.60	79.22	64.08	85.72
Light-ASD	√	×	92.72	93.53	88.56	81.16	65.87	86.24
Causal-ASD	√	×	93.94	94.54	89.72	83.42	67.45	87.70
Light-ASD	×	√	97.86	98.35	95.62	84.70	77.91	93.72
Causal-ASD	×	√	98.75	99.06	97.26	88.07	82.21	95.58

例如,在FO类别中,当面部受到遮挡时,因果干预帮助模型避免依赖唇动与语音的偶然相关关系,从而在该条件下依旧保持较高精度,两种训练条件下的mAP分别提升1.16%和1.64%;在HVN类别中,当声音质量较差时,模型通过剔除非因果噪声信息,更准确地识别出说话人与背景人声的差异,两种训练条件下的mAP改进分别达到2.26%和3.37%;而在SS类别中,在监控环境下,因果特征的抽取使模型能够抵抗监控场景下的低质量音视频输入,也取得了1.58%和4.3%的提升.综合来看,这些结果不仅验证了本文提出方法在理想条件下的有效性,更突出了其在复杂、跨场景环境下的鲁棒

性和泛化优势.

3.3　消融实验

本实验中，音频与视频编码器沿用了经充分验证的 Light-ASD 架构的编码器. 鉴于这些编码器模块在 Light-ASD 中已历经全面验证，且在本任务场景下未做任何修改或附加假设，故不对它们单独开展消融实验.

3.3.1　模块协同分析

通过分析不同模块组合对性能的影响，以明确各模块的有效性和协同作用，具体见表 3，可以观察到，完整的 Causal-ASD 模型在所有评估指标上均优于其去除子模块的变体. 移除因果推断模块会导致模型性能显著下降，这表明因果干预能够有效提升特征表示的可靠性. 此外，去除对抗判别器也会削弱模型的整体效果，进一步验证了该模块在去除潜在混杂因素以及增强模型泛化能力方面的有效性.

表 3　消融实验

Tab. 3　Ablation study

模块组合	Params/M	FLOPs/G	mAP/%
编码器+加权融合	1.02	0.60	93.08±0.03
编码器+因果推断模块+加权融合	1.02	0.60	93.41±0.10
编码器+贝叶斯推理+加权融合	1.02	0.60	92.15±0.28
编码器+因果推断模块+自注意力机制+加权融合	1.10	0.90	93.63±0.08
编码器+因果推断模块+交叉注意力机制+加权融合	1.10	0.90	93.62±0.05
编码器+因果推断模块+因果注意力机制+加权融合	1.07	0.90	93.31±0.06
编码器+因果推断模块+自注意力机制+门控融合	1.07	0.90	93.73±0.12
编码器+因果推断模块+自注意力机制+门控融合+对抗判别器	1.08	0.90	94.07±0.05
编码器+因果推断模块+自注意力机制+门控融合+对抗判别器+视频辅助分类器（Casual-ASD）	1.10	0.90	94.23±0.03

3.3.2　超参数分析

为评估模型对关键超参数的敏感性并确定最优设置，本研究在标准数据集上对主要超参数进行了系统性实验分析，包括学习率（Learning Rate）、输入序列长度（Input Length）. 在实验过程中，仅调整单个超参数，其余训练配置保持一致，并在 AVA-ActiveSpeaker 验证集上对模型性能进行评估.

学习率是影响模型收敛速度与稳定性的关键因素. 图 6 展示了不同学习率对模型性能的影响. 当学习率过小时，模型参数更新步幅受限，导致收敛缓慢且最终性能不佳；而当学习率过大时，梯度震荡明显，训练过程不稳定. 实验结果表明，当学习率设为 0.001 时，模型在收敛速度与稳定性之间取得了最佳平衡，并在验证集上获得最高的 mAP（94.23%）.

表 4 展示了不同输入序列长度对模型性能的影响. 输入序列长度决定了模型在时序建模中可利用的上下文信息范围. 当输入长度较短时，模型难以捕获语音与唇动之间的时序对齐关系，导致性能显著下降. 随着序列长度逐步增加至 1~4 s，模型的时序依赖建模能力显著增强，mAP 随之提升，并在 4 s 时达到峰值. 进一步增加长度虽提供更多帧信息，但会引入冗余上下文和静音片段，导致性能略有下降.

值得注意的是，大多数研究[6-10]采用可变长度输入，表 4 的实验结果验证了此方法：采用可变长度序列进行训练和测试，能够显著提升模型的时序建模能力和泛化性. 可变长度输入策略使模型在训练阶段接触到多样化的语音持续分布，从而更接近真实场景，减少了固定长度截取带来的分布偏移问题，并进一步提升对主动说话人检测的稳定性.

3.3.3　干预强度分析

在因果干预模块中，干预强度系数 α 控制了模型在特征干预与原始特征之间的平衡程度，其取值直接影响因果表示的纯净度与模型的泛化性能. 过小的 α 值会导致干预效果不足，模型仍受到混杂因素的影响；而过大的 α 值则可能破坏特征的语义一致性，导致表示空间偏离原始的模态分布. 因此，有必要对 α 的取值范围进行系统分析，以验证模型性能对干预强度的敏感性并确定最优设置.

本文在 AVA-ActiveSpeaker 验证集上进行实验，固定其余参数不变，仅调整干预强度 α 的取值，范围设定为 $[0.0, 1.0]$，步长为 0.1. 实验结果如图 7 所示. 整体趋势表明，随着 α 的增大，模型性能（以 mAP 为衡量指标）先上升后下降. 当 $\alpha < 0.3$ 时，干预效应较弱，模型仍部分依赖于混杂特征，导致性能提升有

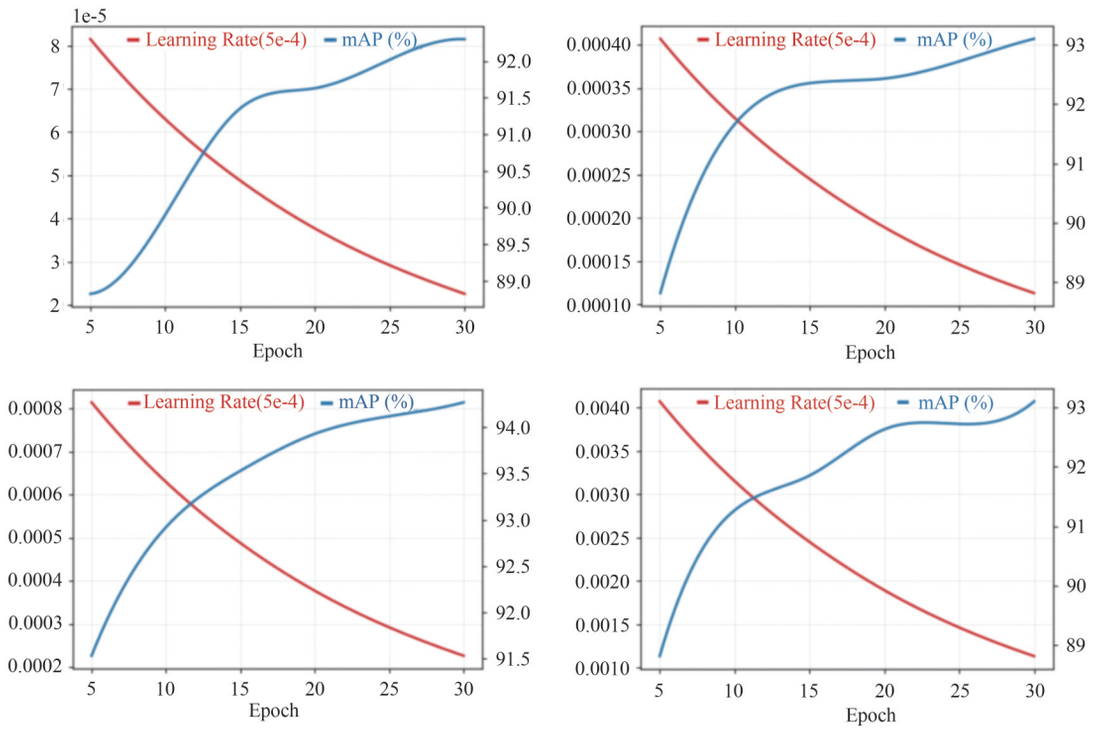


图6 不同学习率的曲线图

Fig. 6 Curves of different learning rates

表4 超参数实验

Tab. 4 Hyperparameter experiment

输入长度/s	mAP/%
0.2	85.20±0.30
0.4	87.65±0.68
1	90.92±0.24
2	92.23±0.28
4	93.26±0.08
6	92.57±0.16
8	92.32±0.06
变长(1-10)	94.23±0.05

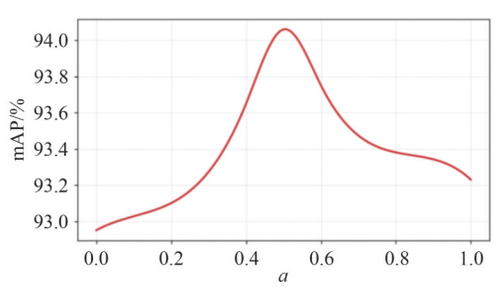


图7 干预强度与模型性能的关系曲线

Fig. 7 The curve of intervention strength versus model performance

限;当α增加至中等水平(约0.4~0.6)时,模型在去混杂与特征保持之间取得了良好平衡,mAP达到峰值;继续增大$\alpha > 0.6$后,过强的干预使原始特征分布发生偏移,削弱了多模态一致性,性能逐渐下降.

由此说明,适度的干预强度能够显著提升模型在因果特征空间中的判别能力,从而提高对主动说话人检测的总体性能.最优干预强度范围0.4~0.6对应的性能提升约为0.9%~1.2% mAP,验证了该超参数在模型中的关键作用.而极端值(过小或过大)均会造成性能退化.因此,在实验中,本研究将干预强度固定为$\alpha = 0.5$,作为经验上的最优选择.

3.3.4 对抗判别器分析

为证明对抗判别器与主模块在训练过程中能有效协同且保持平衡,本文引入了梯度范数比R作为训练平衡性指标,该指标量化了判别器损失对共享参数的梯度范数与主模型损失对共享参数的梯度范数之间的相对强度,其定义如下:

$$R = \frac{\| \nabla_\theta L_D \|}{\| \nabla_\theta L_M \|}, \tag{7}$$

其中,L_D代表判别器的损失函数,L_M代表主模型(或因果模块)的损失函数,而∇_θ表示对模型共享参数θ的梯度.理想情况下,一个稳定的对抗训练过程应使得R值在训练后期趋于稳定,表明判别器与主模型之间达到了动态均衡,避免了任何一方在训练中占据绝对主导地位.

图8展示了在整个训练过程中,R随训练迭代次数(Epoch)的变化曲线,可以看出,在训练初期(约前20个Epoch),R波动较大,这是对抗训练模型在探索最优策略时的常见现象;然而,随着训练的

进行，R值逐渐趋于稳定，并在大约30个Epoch之后保持在一个相对窄的区间内（如图中灰色区域所示，例如0.8到1.2之间），并围绕1.0上下波动.这清晰地表明判别器与主模型之间的训练达到了一个稳定的动态均衡.这一结果有力地证明了模型设计

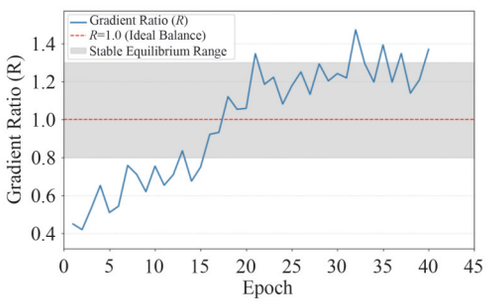


图8　训练平衡曲线

Fig. 8　Training balance curve

能够有效地平衡判别器与主模型的训练强度，使得两者能够各自发挥作用，共同实现去混杂的目标.

4　结语

本文提出了Causal-ASD框架，在主动说话人检测任务中有效融合了因果推断思想.具体而言，该框架构建了多模态因果图来建模音视频特征与结果之间的因果关系，并通过干预机制与对抗训练去除混杂因素，从而获得更具因果性的特征表示；同时，结合Bi-GRU与自注意力机制，实现对跨模态时序依赖的精准建模.在基准数据集AVA-ActiveSpeaker上的实验结果表明：Causal-ASD不仅在mAP上优于基线模型Light-ASD[1]，还在复杂场景下展现出良好的鲁棒性与跨场景泛化能力，验证了本文提出的因果推断建模、多模态干预与对抗去混杂、多模态时序建模三个方面创新设计的有效性.

参　考　文　献

[1] Liao J, Duan H, Feng K, et al. A light weight model for active speaker detection[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver: IEEE, 2023: 22932-22941.

[2] Chung J, Gulcehre C, Cho K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J].arXiv preprint arXiv:2014,1412.3555.

[3] Saenko K, Livescu K, Siracusa M, et al. Visual speech recognition with loosely synchronized feature streams[C]// Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1. Beijing: IEEE, 2005: 1424-1431.

[4] Everingham M, Sivic J, Zisserman A. Taking the bite out of automated Naming of characters in TV video[J]. Image and Vision Computing, 2009, 27(5): 545-559.

[5] Roth J, Chaudhuri S, Klejch O, et al. Ava active speaker: An audio-visual dataset for active speaker detection[C]//ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona: IEEE, 2020: 4492-4496.

[6] Alcázar J L, Cordes M, Zhao C, et al. End-to-end active speaker detection[C]//Computer Vision- ECCV 2022. Cham: Springer, 2022:1-15.

[7] Köpüklü O, Taseska M, Rigoll G. How to design a three-stage architecture for audio-visual active speaker detection in the wild[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal: IEEE, 2021: 1173-1183.

[8] Zhang Y H, Xiao J, Yang S, et al. Multi-task learning for audio-visual active speaker detection[C]//ActivityNet Large-Scale Activity Recognition Challenge Workshop at CVPR 2019.Long Beach: IEEE, 2019:1-8.

[9] Tao R, Pan Z, Das R K, et al. Is someone speaking?: Exploring long-term temporal features for audio-visual active speaker detection[C]//Proceedings of the 29th ACM International Conference on Multimedia. Virtual Event: ACM, 2021: 3927-3935.

[10] Jiang Y, Tao R, Pan Z, et al. Target active speaker detection with audio-visual cues[C]//The 24th INTERSPEECH Conference. Dublin: ISCA, 2023: 3152-3156.

[11] Datta G, Etchart T, Yadav V, et al. Asd-transformer: Efficient active speaker detection using self and multimodal transformers[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE, 2022: 4568-4572.

[12] Alcazar J L, Caba F, Mai L, et al. Active speakers in context[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020: 12462-12471.

[13] Alcázar J L, Heilbron F C, Thabet A K, et al. MAAS: Multi-modal assignation for active speaker detection[C]// 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal: IEEE, 2021: 265-274.

[14] Wuerkaixi A, Zhang Y, Duan Z, et al. Rethinking audio-visual synchronization for active speaker detection [C]//2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP). Xi'an: IEEE, 2022: 1-6.

[15] Xiong J, Zhou Y, Zhang P, et al. Look&listen: Multi-modal correlation learning for active speaker detection and speech enhancement [J]. IEEE Transactions on Multimedia, 2023, 25: 5800-5812.

[16] Zhang Y, Liang S, Yang S, et al. UniCon: Unified context network for robust active speaker detection [C]// Proceedings of the 29th ACM International Conference on Multimedia. Virtual Event: ACM, 2021: 3964-3972.

[17] Wang X, Cheng F, Bertasius G. LoCoNet: Long-short context network for active speaker detection[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2024: 18462-18472.

[18] Jung C, Lee S, Nam K, et al. TalkNCE: Improving active speaker detection with talk-aware contrastive learning[C]//ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul: IEEE, 2024: 8391-8395.

[19] Nguyen L T P, Yu Z, Lee Y J. LASER: Lip landmark assisted speaker detection for robustness[J]. arXiv preprint arXiv:2025, 2501.11899.

[20] Pearl J, Glymour M, Jewell N P. Causal inference in statistics[M].Chichester: John Wiley & Sons Ltd, 2016.

[21] Pearl J, Mackenzie D. The Book of Why[M]. New York: Basic Books, 2018.

[22] Rubin D B. Causal inference using potential outcomes: Design, modeling, decisions[J]. Journal of the American Statistical Association, 2005, 100(469): 322-331.

[23] Athey S, Wager S. Estimating treatment effects with causal forests: An application[J]. Observational Studies, 2019, 5(2): 37-51.

[24] Yang X, Zhang H, Qi G, et al. Causal attention for vision-language tasks[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA. IEEE, 2021: 9842-9852.

[25] Liu B, Wang D, Yang X, et al. Show, deconfound and tell: Image captioning with causal inference[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022: 18020-18029.

[26] Qiu R, Wang S, Chen Z, et al. CausalRec: Causal inference for visual debiasing in visually-aware recommendation[C]//Proceedings of the 29th ACM International Conference on Multimedia. Virtual Event: ACM, 2021: 3844-3852.

[27] Kim J, Kim J. A survey on causal inference in image captioning[C]//2023 International Conference on Electronics, Information, and Communication (ICEIC). Singapore: IEEE, 2023: 1-3.

[28] Roxo T, Costa J C, Inácio P R M, et al. WASD: A wilder active speaker detection dataset[J]. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2025, 7(1): 61-70.

[29] Liao J, Duan H, Feng K, et al. LR-ASD: Lightweight and robust network for active speaker detection[J]. International Journal of Computer Vision, 2025, 133(7): 4749-4769.

(责编 曹东,校对 雷建云)