

基于视听情感一致性的多模态人脸伪造检测

王利华^{1,2}, 严伟丹^{3,4}, 王新宇⁴, 吴忠骏⁴, 李松斌³, 张登银⁴

(1. 南京邮电大学 通信与信息工程学院, 江苏 南京 210003; 2. 联通物联网有限责任公司, 江苏 南京 210001; 3. 无锡学院 电子信息工程学院, 江苏 无锡 214105; 4. 南京邮电大学 物联网学院, 江苏 南京 210003)

摘要: 针对高质量人脸伪造中视听伪影弱化、情感不一致难以捕获的问题, 提出基于视听情感一致性的多模态检测方法 AVEmo, 从情感内容与动态变化两个层面挖掘跨模态伪造线索。该方法分别提取面部表情与语音情感特征, 并通过跨模态注意力建模情感内容一致性与时序同步关系; 进一步采用多任务学习联合优化音频模态、视频模态和视听融合分支的真伪分类, 增强模型对跨模态情感异常的判别能力。实验表明, AVEmo 在 HiFi-AVDF 所评估的五个生成子集上的平均 AP 和 AUC 分别达到 76.8% 和 78.9%, 其中平均 AUC 较 AVFF 和 HAVIC 分别提高 10.6% 和 2.1%, 表现出一定的跨数据集泛化能力。方法代码已公开: <https://anonymous.4open.science/r/AVEmo-00A2>。

关键词: 人脸伪造检测; 视听情感一致性; 多模态融合; 多任务学习

中图分类号: TP391

文献标志码: A

Multimodal Deepfake Detection Based on Audio-Visual Emotional Consistency

WANG Lihua^{1,2}, YAN Weidan^{3,4}, WANG Xinyu⁴, WU Zhongjun⁴, LI Songbin³, ZHANG Dengyin⁴

1. School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China

2. China Unicom Internet of Things Co., Ltd., Nanjing 210001, China

3. School of Electronic Information Engineering, Wuxi University, Wuxi 214105, China

4. School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210003, China

Abstract: To address the weakening audio-visual artifacts and the difficulty in capturing emotional inconsistencies in high-quality deepfake videos, this paper proposes AVEmo, a multimodal deepfake detection method based on audio-visual emotional consistency, which explores cross-modal forgery cues from both emotional content and dynamic variation. AVEmo extracts facial expression and speech emotion features, and models emotional content consistency and temporal synchronization through cross-modal attention. A multi-task learning strategy is further adopted to jointly optimize authenticity classification for the audio, visual, and audiovisual fusion branches, thereby enhancing the model's ability to identify cross-modal emotional anomalies. Cross-dataset evaluation on five selected generation subsets of HiFi-AVDF yields an average AP of 76.8% and an average AUC of 78.9%. The average AUC exceeds those of AVFF and HAVIC by 10.6% and 2.1%, respectively. These results suggest that AVEmo can generalize across datasets under the evaluated conditions. The code is available at: <https://anonymous.4open.science/r/AVEmo-00A2>.

Key words: Deepfake detection, Audio-visual emotional consistency, Multimodal fusion, Multi-task learning

收稿日期: XXXX-XX-XX; 修回日期: XXXX-XX-XX

通信作者: 张登银, zhangdy@njupt.edu.cn

基金项目: 国家自然科学基金资助项目 (No. 62471241); 江苏省研究生科研与实践创新计划 (No. SJCX250295, No. KYCX251222); 江苏省产学研合作项目 (No. BY20250356); 无锡学院高层次人才科研启动专项经费资助 (No. 2025r029)

Foundation Items: National Natural Science Foundation of China (No. 62471241), Postgraduate Research & Practice Innovation Program of Jiangsu Province (No. SJCX250295, No. KYCX251222), Jiangsu Industry-University-Research Cooperation Project (No. BY20250356), Wuxi University Research Start-up Fund for High-level Talents (No. 2025r029)

0 引言

随着人工智能生成内容 (Artificial Intelligence Generated Content, AIGC) 的快速发展, 人脸替换、表情编辑、语音合成和音视频驱动生成等深度伪造手段取得了显著进展^[1]。高质量伪造视频在身份保持、表情自然度和音画同步性等方面不断逼近真实视频, 给个人隐私保护、网络内容可信性和社会舆论安全带来潜在威胁。因此, 研究鲁棒、泛化的人脸伪造检测方法具有重要的现实意义^[2]。

现有检测方法按照输入类型可以分为: 图像级、视频级和多模态检测。图像级方法^[3-8]通常从单帧人脸中挖掘纹理异常、边界融合痕迹或在频率域辨别伪影; 视频级方法^[9-17]则进一步引入时序信息, 重点关注帧间运动、生物特征和时空不一致性。上述方法虽然在实验数据集上取得较好效果, 但多依赖视觉模态中的低层伪影或局部时序异常。当伪造模型质量提升, 或视频经过压缩、重采样等后处理操作后, 这类线索容易被削弱, 导致模型泛化能力不足。

为弥补上述单一模态方法的不足, 多模态人脸伪造检测方法^[18-31]逐渐受到关注。真实视频中的听觉与视觉信息通常具有较强的协同性, 如语音内容与唇部运动同步、说话人声音与面部身份匹配等。因此, 部分多模态方法围绕唇动与语音同步或身份一致性进行建模, 分别利用局部时间对齐关系和跨模态身份对应关系进行鉴伪。然而, 随着生成模型在音视频对齐和身份保持方面不断改进, 伪造视频中的口型偏差和身份差异逐渐减小, 使上述线索在高质量伪造场景下的区分能力可能受到限制。除上述线索外, 面部表情与语音韵律之间的情感关联也蕴含潜在的鉴伪线索, 但其判别价值仍有待进一步挖掘。在多数自然交流场景中, 面部表情与语音韵律通常共同反映人物的情感状态。例如, 当人物表现出愤怒、喜悦、惊讶或悲伤等情绪时, 视觉模态中的眉眼、嘴角和面部肌肉变化往往与音频模态中的音高、能量和节奏相互呼应。而在人脸替换、表情驱动、换声或音视频拼接等伪造场景中, 伪造算法即使能够较好地保持身份外观或唇形同步, 仍可能在视听情感内容及其动态变化之间产生不一致。围绕情感线索, Mittal 等人^[32]联合利用视听模态嵌入与情感嵌入的距离进行鉴伪; Zhang 等人^[33]结合情感感知与跨模态对比学习, 以文本引导视听特征

融合。上述研究主要通过嵌入匹配或文本引导融合建立视听联系。本文则侧重情感内容一致性与时序变化同步性的联合建模, 通过视频主导和音频主导的协同特征提取, 使面部动态与语音上下文相互引导, 并结合注意力瓶颈融合, 刻画二者的内容关联与动态变化关系。

于是, 本文考虑到将视听情感一致性作为区别于低层视觉伪影和唇动同步关系的高层语义判别线索, 能够为高质量人脸伪造检测提供补充的语义判别依据。进而提出一种基于视听情感一致性的多模态人脸伪造检测方法 AVemo (Audio-Visual Emotion Consistency based Deepfake Detection), 主要围绕“视听情感内容是否一致”和“情感时序变化是否同步”两个问题展开。具体而言, 所提方法首先对输入视频进行预处理, 提取人脸帧序列并将对应音频转换为梅尔频谱图, 分别获得视觉特征和音频特征。其次, 设计多模态情感特征提取器, 从视频主导和音频主导两个方向挖掘情感线索, 使面部动态与语音上下文在特征提取过程中相互引导。在视频主导分支中, 以音频特征为辅助, 结合局部时序建模与状态空间建模, 提取面部情感特征的局部变化及序列依赖; 在音频主导分支中, 以视频特征为辅助, 结合增强长短期记忆网络建模语音语境中的情感依赖关系, 从而刻画视听情感表达差异。在此基础上, 进一步提出基于注意力瓶颈的视听情感特征融合模块, 联合建模视听情感内容的一致性及其时序变化的同步关系。最后, 本文构建多任务学习策略, 联合优化视频模态、音频模态和视听融合分支的真伪分类任务, 使模型在学习伪造判别特征的同时保持对情感一致性线索的关注。

本文的主要贡献如下:

(1) 基于视听情感一致性的多模态人脸伪造检测方法 AVemo, 从面部表情与语音情感的内容一致性及其变化同步性角度挖掘伪造线索;

(2) 多模态情感特征提取器, 分别构建视频主导和音频主导的情感线索建模分支, 结合局部时序建模与状态空间建模提取面部动态情感特征, 并利用增强长短期记忆网络挖掘语音语境中的情感依赖关系;

(3) 基于注意力瓶颈的视听情感特征融合模块, 通过跨模态注意力实现视听情感特征对齐与交互, 增强模型对跨模态情感不一致和时序不同步现象的

感知能力;

(4) 多任务学习框架,联合视频模态、音频模态和视听融合分支的真伪分类任务,并通过域内、跨伪造方法和跨数据集实验评估所提方法的检测性能。

1 相关工作

1.1 单模态伪造检测

图像级人脸伪造检测方法作为特定领域的图像二分类问题,旨在从单帧人脸图像中提取判别特征。典型方法有 Xception^[3]和 EfficientNet^[4]等通用分类网络,其在人脸伪造检测中常作为基准模型。随后,研究者进一步从空间域、频率域及空频联合角度挖掘伪造痕迹。例如,SBIs^[5]通过自混合图像构造更具泛化性的伪造样本;F3-Net^[6]从频率感知特征出发捕捉隐蔽伪影;M2TR^[7]则利用多尺度 Transformer 融合空间与频率信息,以提升检测性能;Effort^[8]引入奇异值分解提升鉴伪泛化能力;李等人^[9]提出一种融合动态伪影追踪与空频交互分析的金字塔式交互双流网络;郭等人^[10]基于多特征协同置信感知,借助语义-取证双流框架实现深度伪造图像检测及边界精细定位。然而,随着生成模型质量提升,单帧视觉伪影逐渐弱化,仅依赖空间或频率特征的方法在跨数据集和高质量伪造场景下仍容易出现性能下降。于是为进一步利用视频中的时间信息,视频级检测方法重点引入了时序建模机制,旨在挖掘帧间运动、面部区域变化和时空不一致性。其中部分方法^[11-13]从身份特征、目光方向等生物线索出发,利用真实视频中较稳定的生理规律进行真伪判别,熊等人^[14]利用三维双流网络联合提取视频帧的高低频时空特征。此外,时序卷积^[15]、时序 Transformer^[16]或图神经网络^[17]也被用于建模连续帧之间的长程依赖关系。虽然视频级方法进一步弥补单帧检测对时序信息利用不足的问题,但其核心线索仍来自视觉模态,面对视觉质量较高且时序平滑的伪造视频时,检测能力仍存在一定局限。

1.2 多模态伪造检测

现有多模态人脸伪造检测方法主要包括跨模态一致性检测和多模态特征融合两类。跨模态一致性检测方法通常利用音频与视觉之间的对应关系发现异常。例如,MDS^[18]和 VFD^[19]分别从音视频片段

协调性和身份匹配性角度建模跨模态匹配关系,以识别伪造视频中音频与视觉信号的不一致;在模型学习方面,AVH-Align^[20]采用无监督方式,AVAD^[21]采用自监督学习方式在真实视频上学习音视频同步或模态内外时序一致性分布,并将偏离真实一致性模式的样本判定为潜在伪造;LipFD^[22]和 PIA^[23]都专注于挖掘细粒度唇形伪造线索;HAVIC^[24]与 ICSAV^[25]都考虑了模态内与跨模态一致性学习真实视频的整体音视频一致性先验,以提升方法泛化性。此类方法具有较强的物理可解释性,但当伪造视频经过高质量唇形驱动或身份保持处理后,单一的一致性线索可能不足以支撑稳定检测。多模态特征融合方法则通过联合建模音频特征与视觉特征,提高模型对复杂伪造样本的判别能力。例如,AV-DFD^[26]、AVFakeNet^[27]和 AVoiD-DF^[28]均提出设计网络对视听模态进行融合检测;为增强音视频表征的互补性,AVFF^[29]进一步引入自监督预训练、互补掩码,MRDF-CE^[30]则提出模态内外正则化增强方法;FoVB^[31]将音视频相关性建模为高斯潜变量,提出基于变分贝叶斯的伪造感知音视频自适应方法,以区分跨模态相关性异常。基于视听情感线索的鉴伪也已有研究。Mittal 等^[32]联合利用视听模态嵌入与情感嵌入的距离进行真伪判别;Zhang 等^[33]结合情感感知与跨模态对比学习,以文本引导视听特征融合。与主要依赖音画同步或身份一致性的方法相比,包括本文在内的情感鉴伪方法进一步从情感表达层面挖掘面部表情与语音韵律之间的跨模态不一致线索,为高质量人脸伪造视频检测提供补充判别依据。相较于上述以嵌入匹配或文本引导融合建立视听联系的方法,AVEmo 侧重情感内容一致性与变化同步性的联合建模:在特征提取阶段,通过视频主导和音频主导分支实现面部动态与语音上下文的协同建模;在融合阶段,通过注意力瓶颈联合刻画视听情感的内容关联与动态变化关系。

2 本文方法

本文提出一种基于视听情感一致性的多模态人脸伪造检测方法 AVEmo。该方法以真实视频中面部表情与语音情感在内容表达和时序变化上应保持一致为鉴伪基准,通过分别挖掘视觉模态中的面部情感线索和音频模态中的语境情感线索,进一步建

模二者之间的跨模态一致性差异,从而实现人脸伪造检测。

2.1 总体框架

如图1所示,所提方法主要包括数据预处理、多模态情感特征提取以及多任务伪造判别三个部分。对于给定的待检测视频片段 $x_v \in \mathbb{R}^{T_v \times H \times W \times 3}$,其中 T_v 表示视频帧数, H 和 W 分别表示单帧的高度和宽度。对于其对应的音频信号,本文首先采用分帧、加窗和短时傅里叶变换等预处理操作,将其转换为梅尔频谱图序列 $x_a \in \mathbb{R}^{T_a \times L}$,其中 T_a 表示音频帧数, L 表示梅尔频谱通道数。随后,分别采用视频编码器和音频编码器提取基础视听特征:

$$F_v = \mathbb{E}_v(x_v), F_a = \mathbb{E}_a(x_a) \quad (1)$$

其中, $\mathbb{E}_v(\cdot)$ 和 $\mathbb{E}_a(\cdot)$ 分别表示视频编码器和音频编码器, $F_v \in \mathbb{R}^{T \times d}$ 和 $F_a \in \mathbb{R}^{T \times d}$ 分别表示时间对齐后的视频帧级特征和音频片段级特征, T 为统一后的时间步数, d 为特征维度。

在此基础上,AVEmo通过多模态情感特征提取器 $\mathcal{M}(\cdot)$ 从视频主导和音频主导两个方向进行情感线索建模得到视频主导情感特征 E_v 和音频主导情感特征 E_a :

$$E_v, E_a = \mathcal{M}(F_v, F_a) \quad (2)$$

随后,视听情感特征融合模块 $\mathcal{F}(\cdot)$ 进一步实现跨模态情感特征的对齐与交互,生成最终的视听情感判别特征 Z_{a+v} :

$$Z_{a+v} = \mathcal{F}(E_v, E_a) \quad (3)$$

最后,模型通过多个分类器分别完成视频模态伪造预测、音频模态伪造预测和视听融合伪造判

别。本文所提AVEmo将视听情感内容一致性和情感变化同步性嵌入特征提取与融合过程,并通过多任务联合优化实现多模态人脸伪造检测任务。需要说明的是,视听情感一致性在本文中作为伪造检测的辅助线索,而非判断视频真假的充分条件或必要条件。模型的最终判别综合利用音频信息、视觉信息、跨模态交互信息及时序特征,并非仅依据单一的一致性得分。

2.2 多模态情感特征提取器

为充分挖掘视听模态中具有判别性的情感表征,所提方法不采用直接融合两种模态特征的方式,而是采用双分支结构实现双主导情感建模:视频主导分支在语音情感线索的引导下关注面部表情动态;音频主导分支则是在面部表情线索的引导下刻画语音情感语境。通过这种双向条件约束,模型能够在特征提取过程中捕捉视听情感表达之间的潜在一致性关系。在提取情感特征前,本方法首先将基础特征映射到统一维度,以实现跨模态适配:

$$E_v^0 = \phi_v(F_v), E_a^0 = \phi_a(F_a) \quad (4)$$

其中, $\phi_v(\cdot)$ 和 $\phi_a(\cdot)$ 分别表示视频特征映射层和音频特征映射层,用于将两种模态的基础特征投影到统一的特征空间。对于第 k 层多模态情感特征提取器,其输入为视频主导情感特征 E_{va}^k 和音频主导情感特征 E_{av}^k ,其中 E_{va}^1 和 E_{av}^1 初始化为 E_v^0 和 E_a^0 ,其输出则为更新后的 E_{va}^{k+1} 和 E_{av}^{k+1} ,用于第 $k+1$ 层提取器输入。

本节后续将在2.2.1节介绍用于挖掘视频情感特征的面部情感提取单元,在2.2.2节介绍用于挖掘音频情感特征的语境情感提取单元,并在2.2.3

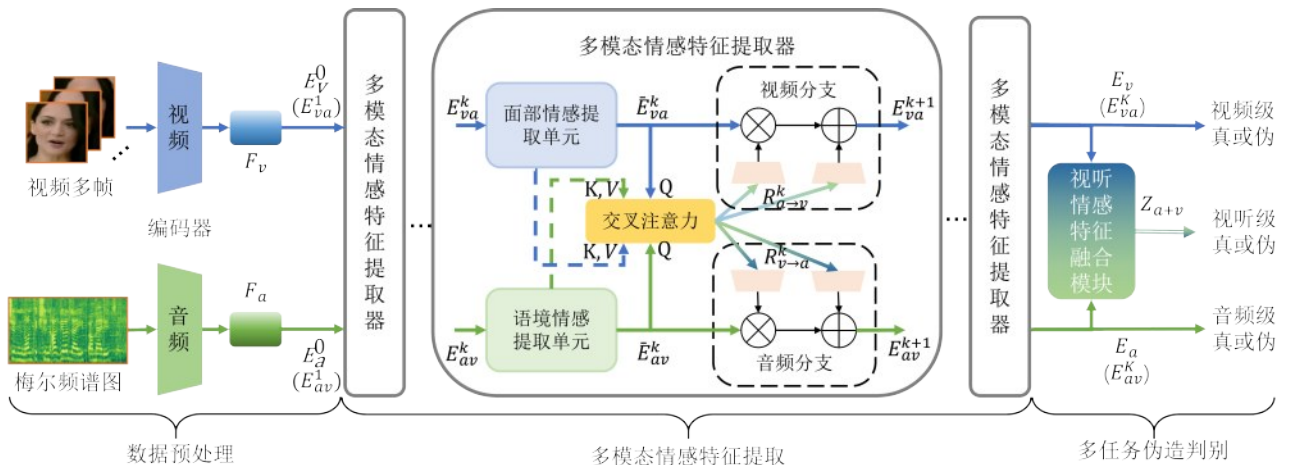


图1 AVEmo总体框架

节给出整体的双主导跨模态交互机制的细节。

2.2.1 面部情感提取单元

为提取面部情感特征及其动态变化, 本文设计面部情感提取单元, 将局部时序建模与状态空间建模[34]相结合。局部时序建模聚合邻近时间步的信息, 提取面部表情的局部变化; 状态空间建模进一步整合序列上下文, 建立不同时间步之间的依赖关系, 为后续跨模态情感交互提供视觉表征。

具体而言, 对于给定输入特征序列 $X = \{x^1, x^2, \dots, x^T\} \in \mathbb{R}^{T \times d}$, 首先通过局部特征变换操作得到相邻帧之间的面部表情变化特征 X_m :

$$X_m = \text{DWConv}(\text{LN}(X)) \# (5)$$

其中, $\text{LN}(\cdot)$ 表示层归一化, $\text{DWConv}(\cdot)$ 表示深度可分离卷积, 用于聚合局部时间邻域内的面部特征, 提取邻近时间步的表情变化信息。

随后, 利用状态空间模型 $\text{SSM}(\cdot)$ 对局部变化特征序列进行动态建模, 得到包含时序上下文的面部情感特征 X_s :

$$X_s = \text{SSM}(X_m) \# (6)$$

所采用状态空间模型中的第 t 个时间步, 其状态空间递推过程可表示为:

$$\begin{aligned} h_t &= Ah_{t-1} + Bx'_m \# (7) \\ x'_s &= Ch_t \end{aligned}$$

其中, x'_m 为 X_m 在第 t 个时间步的特征, h_t 为隐状态, x'_s 为状态空间模型的输出, A 、 B 、 C 分别表示状态转移、输入映射和输出映射矩阵。该递推结构能够在较低复杂度下捕捉长时间范围内的表情变化关系。

最后, 通过前馈变换 $\text{FFN}(\cdot)$ 得到面部情感提取单元 $\text{FE}(\cdot)$ 的输出:

$$\text{FE}(X) = X + \text{FFN}(\text{LN}(X_s)) \# (8)$$

该单元兼顾局部变化与序列上下文两个层面的信息。局部时序建模侧重表情变化的邻域特征, 状态空间建模则从更长的时间范围描述面部情感特征的演化, 使输出表征同时包含局部动态信息与时序上下文。这种设计为后续视听交互提供了具有时间关联的视觉情感信息, 也为分析两种模态在情感内容及动态变化方面的关系提供了基础。

2.2.2 语境情感提取单元

为了建模连续语音片段中的情感上下文, 挖掘语音信号中由语调、节奏、能量和韵律变化所体现

的情感线索。本单元考虑到听觉情感往往难以由单个音频帧独立判断, 而是体现在相邻语音片段之间的动态变化过程中。于是, 本节采用增强长短期记忆网络 xLSTM [35] 对音频特征序列进行时序建模, 以捕捉语音情感随上下文演化的动态特征。

具体而言, 对于给定输入特征序列 $X = \{x^1, x^2, \dots, x^T\} \in \mathbb{R}^{T \times d}$, 首先通过增强长短期记忆网络 $\text{xLSTM}(\cdot)$ 提取语音情感的上下文表示 X_c :

$$X_c = \text{xLSTM}(X) \# (9)$$

为了突出相邻时间片之间的情感变化关系, 本节在此基础上进一步引入记忆注意力机制, 根据相邻时间步的上下文特征计算注意力权重:

$$\begin{aligned} a_t &= \sigma(W_a[x_c^{t-1}; x_c^t]) \# (10) \\ \tilde{x}_c^t &= a_t \odot x_c^t \end{aligned}$$

其中, $[x_c^{t-1}; x_c^t]$ 表示相邻时间步特征拼接, a_t 表示记忆注意力权重, $\tilde{x}_c^t$ 表示增强后的当前语音情感特征, $\odot$ 表示逐元素乘法。

随后, 通过门控机制更新语境记忆:

$$\begin{aligned} \gamma_t &= \sigma(W_\gamma \tilde{x}_c^t) \\ \eta_t &= \sigma(W_\eta \tilde{x}_c^t) \\ \mu_t &= \tanh(W_\mu \tilde{x}_c^t) \# (11) \\ m_t &= \gamma_t \odot m_{t-1} + \eta_t \odot \mu_t \end{aligned}$$

其中, m_t 表示第 t 个时间步的语境记忆, γ_t 为保留

门, 用于控制历史情感信息的保留程度; η_t 为更新门, 用于控制当前情感信息的写入强度; μ_t 为候选记忆, 用于表示当前时间步待写入的情感内容。最终, 将所有时间步的语境记忆表示为 $X_g = \{m_1, m_2, \dots, m_T\}$, 并通过残差连接得到语境情感单元 $\text{AC}(\cdot)$ 的输出:

$$\text{AC}(X) = X + W_o X_g \# (12)$$

其中, W_o 表示输出映射矩阵。通过 xLSTM 、记忆注意力和门控更新的结合, 该单元能够建模语音情感在时间维度上的连续演化, 为后续跨模态情感一致性交互提供可靠听觉特征。

2.2.3 双主导跨模态交互机制

为了建立视听情感之间的关联与交互, 使每一模态的情感线索挖掘都非独立产生, 本节提出了双主导跨模态交互机制, 旨在以“一主一辅”的方式实现跨模态融合。

具体而言, 对于第 k 层输入的视频情感特征

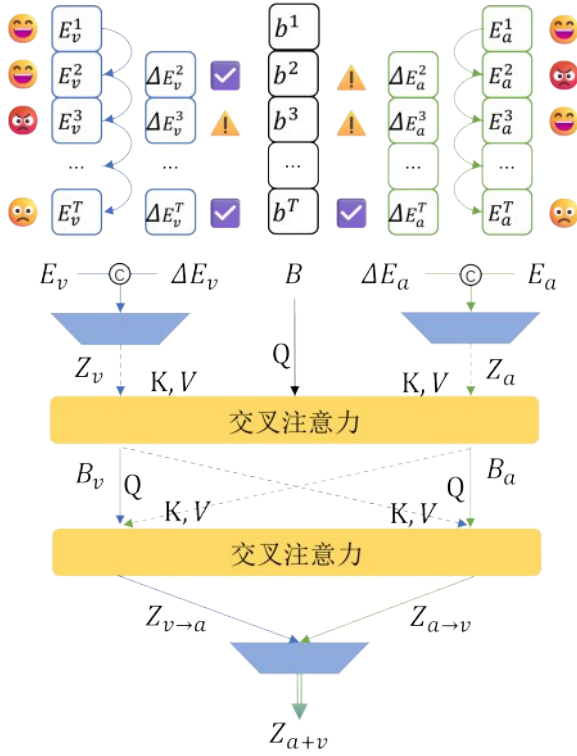


图2 视听情感特征融合模块架构图

E_{va}^k 和音频情感特征 E_{av}^k , 首先分别利用上两节所提情感提取单元获得单模态情感特征:

$$\bar{E}_{va}^k = FE(E_{va}^k), \bar{E}_{av}^k = AC(E_{av}^k) \quad (13)$$

其中, $\bar{E}_{va}^k$ 表示视频侧面动态情感特征, $\bar{E}_{av}^k$ 表示音频侧面情感动态特征。

随后, 本节采用交叉注意力 $CA(\cdot)$ 获得双向跨模态辅助线索, 交叉注意力为:

$$CA(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (14)$$

其中, Q 、 K 和 V 分别表示查询矩阵、键矩阵和值矩阵, d 表示特征维度。

视频主导分支以视频特征为查询, 从音频特征中选择与当前面部表情相关的语音情感线索。音频主导分支则以音频特征为查询, 从视频特征中选择与当前语音情感相关的面部表情线索:

$$\begin{aligned} R_{a \rightarrow v}^k &= CA(E_{va}^k, E_{av}^k, E_{av}^k) \\ R_{v \rightarrow a}^k &= CA(E_{av}^k, E_{va}^k, E_{va}^k) \end{aligned} \quad (15)$$

接着, 本节采用仿射变换将辅助线索有效纳入主导线索更新进程中:

$$\begin{aligned} [\alpha_a^k, \beta_a^k] &= \text{MLP}_a(R_{a \rightarrow v}^k) \\ [\alpha_v^k, \beta_v^k] &= \text{MLP}_v(R_{v \rightarrow a}^k) \\ E_{va}^{k+1} &= (1 + \alpha_a^k) \odot \bar{E}_{va}^k + \beta_a^k \\ E_{av}^{k+1} &= (1 + \alpha_v^k) \odot \bar{E}_{av}^k + \beta_v^k \end{aligned} \quad (16)$$

其中, α 与 β 分别表示尺度和偏移参数, $\text{MLP}(\cdot)$ 为多层感知器。

通过本节双向交互机制, 视频分支在语音情感线索的约束下建模面部表情变化, 音频分支在面部表情线索的约束下建模语音情感变化。当视听情感表达一致时, 跨模态辅助线索能够增强主导模态中的有效情感成分; 当伪造视频存在情感不匹配或响应异常时, 这种条件调制会产生更具判别性的差异表征。

经过上述 K 个多模态情感特征提取器后, 得到最终的视频主导情感特征 $E_v = E_{va}^K$ 和音频主导情感特征 $E_a = E_{av}^K$ 。

2.3 视听情感特征融合模块

如图2所示, 为进一步刻画面部表情变化与语音情感变化之间的时间同步关系, 本文提出基于注意力瓶颈的视听情感特征融合模块, 旨在以时间锚定的瓶颈特征作为跨模态交互中介, 对视频与音频中的情感状态及其变化趋势进行对齐交互, 从而突出视听情感动态之间的同步性判别信息。

具体而言, 本节首先初始化一组可学习的时间瓶颈特征作为时间锚点, 用于对不同时间步上的情感状态进行聚合与对齐:

$$B = \{b^1, b^2, \dots, b^T\} \in \mathbb{R}^{T \times d} \quad (17)$$

其中, b^t 表示第 t 个时间步的瓶颈特征。为了显式刻画情感变化同步性, 本节计算相邻时间步之间 ($t = 2, \dots, T$) 的情感变化量:

$$\Delta E_v^t = E_v^t - E_v^{t-1}, \Delta E_a^t = E_a^t - E_a^{t-1} \quad (18)$$

随后, 将情感状态与情感变化进行组合, 得到变化增强特征:

$$\begin{aligned} Z_v &= \text{MLP}([E_v; \Delta E_v]), Z_a = \\ &= \text{MLP}([E_a; \Delta E_a]) \end{aligned} \quad (19)$$

其中, $\Delta E_v = \{\Delta E_v^2, \Delta E_v^3, \dots, \Delta E_v^T\}$, $\Delta E_a = \{\Delta E_a^2, \Delta E_a^3, \dots, \Delta E_a^T\}$ 。

瓶颈特征分别从视频和音频中聚合情感状态与变化信息:

$$B_v = CA(B, Z_v, Z_v), B_a = CA(B, Z_a, Z_a) \quad (20)$$

在此基础上,生成视听情感判别特征:

$$Z_{v \rightarrow a} = CA(B_v, B_a, B_a), Z_{a \rightarrow v} = CA(B_a, B_v, B_v) \quad (21)$$

其中, $Z_{v \rightarrow a}$ 表示从视频情感变化到音频情感变化的对齐结果, $Z_{a \rightarrow v}$ 表示从音频情感变化到视频情感变化的对齐结果。

最后,将两个方向的交互结果进行融合,得到最终的视听情感判别特征 Z_{a+v} :

$$Z_{a+v} = \text{MLP}([Z_{v \rightarrow a}; Z_{a \rightarrow v}]) \quad (22)$$

通过所提双向瓶颈交互机制,模型能够对齐视听两种模态的情感变化过程,从而捕捉时间错位、响应滞后和变化强度不匹配等同步性异常。

2.4 多任务联合优化

为增强模型对情感线索的判别利用能力,本文所提方法采用多任务联合优化策略,同时约束单模态和多模态真伪分类。基于视频主导情感特征 E_v 和音频主导情感特征 E_a , 分别进行单模态真伪预测:

$$\begin{aligned} \hat{y}_v &= C_v(E_v) \\ \hat{y}_a &= C_a(E_a) \end{aligned} \quad (23)$$

其中, $C_v(\cdot)$ 和 $C_a(\cdot)$ 分别表示视频模态真伪分类器和音频模态真伪分类器。基于融合特征 Z_{a+v} , 进一步进行视听融合真伪预测:

$$\hat{y}_{a+v} = C_{a+v}(Z_{a+v}) \quad (24)$$

其中, $C_{a+v}(\cdot)$ 表示视听分类器。

采用交叉熵损失分别约束三个预测分支:

$$\begin{aligned} \mathcal{L}_v &= \text{CE}(y_v, \hat{y}_v) \\ \mathcal{L}_a &= \text{CE}(y_a, \hat{y}_a) \\ \mathcal{L}_{a+v} &= \text{CE}(y, \hat{y}_{a+v}) \end{aligned} \quad (25)$$

其中, y_v 、 y_a 和 y 分别表示视频模态真伪标签、音频模态真伪标签和视频整体真伪标签,取值均为0或1,其中0表示真实,1表示伪造。

于是,常规分类优化目标为:

$$\mathcal{L}_{cls} = \lambda_v \mathcal{L}_v + \lambda_a \mathcal{L}_a + \lambda_{a+v} \mathcal{L}_{a+v} \quad (26)$$

其中, λ_v 、 λ_a 和 λ_{a+v} 为不同分类任务的损失权重。模态级分类监督分别约束视频侧和音频侧的真伪判别,融合分类监督则约束视听样本的整体真伪判别,使模型联合利用不同模态的判别信息。

此外,为刻画视听表征在情感内容及动态变化方面的差异,本文定义情感内容差异分数 d_{con} 和变化同步差异分数 d_{sync} 。对于视频侧与音频侧情感特

征 E_v 和 E_a , 采用归一化一范数距离度量内容表征差异;对于时间瓶颈聚合得到的变化增强特征 B_v 和 B_a , 采用归一化二范数距离度量变化表征差异:

$$\begin{aligned} d_{con} &= \frac{\|E_v - E_a\|_1}{\|E_v\|_1 + \|E_a\|_1 + \varepsilon} \\ d_{sync} &= \frac{\|B_v - B_a\|_2}{\|B_v\|_2 + \|B_a\|_2 + \varepsilon} \end{aligned} \quad (27)$$

其中, ε 为避免分母为零的常数。 d_{con} 与 d_{sync} 分别刻画所学特征空间中的视听情感内容表征差异与变化增强表征差异,较小的分数表示相应视听表征更为接近。两个差异分数均位于[0,1]。

在此基础上,以视听样本整体真伪标签 y 对两个差异分数进行辅助监督,分别构建情感内容一致性损失和情感变化同步损失:

$$\mathcal{L}_{con} = (d_{con} - y)^2, \mathcal{L}_{sync} = (d_{sync} - y)^2 \quad (28)$$

其中, $y = 0$ 表示真实样本, $y = 1$ 表示伪造样本。最终,总体优化目标为:

$$L = \mathcal{L}_{cls} + \lambda_{con} \mathcal{L}_{con} + \lambda_{sync} \mathcal{L}_{sync} \quad (29)$$

其中, λ_{con} 和 λ_{sync} 分别表示内容一致性损失和变化同步损失的权重系数。两项损失与模态级分类目标和融合分类目标共同优化。对于RVRA样本,两项差异损失鼓励差异分数减小;对于RVFA、FVRA和FVFA样本,则鼓励差异分数增大,从而引导模型学习与鉴伪类别相关的跨模态差异表征。

3 实验

3.1 数据集与评价指标

区别于通过对抗样本干预防造过程的主动防御研究^[36],本文主要面向人脸伪造内容的被动检测任务。为此,首先采用FakeAVCeleb^[37]数据集进行训练与域内测试。FakeAVCeleb是公开的音视频多模态伪造检测数据集,包含真实音频-真实视频(RVRA)、伪造音频-真实视频(RVFA)、真实音频-伪造视频(FVRA)和伪造音频-伪造视频(FVFA)四种视听真伪组合。基于该数据集,分别开展域内实验和跨伪造方法实验,以评估模型在同分布场景下的检测性能及其对未知伪造类型的泛化能力。

鉴于该领域常用的经典KoDF^[38]数据集上的检测性能已逐渐趋于饱和,为进一步评估所提方法在

高质量伪造视频上的跨数据集泛化表现, 本文在 Peng 等人^[24]提出的 HiFi-AVDF 数据集中选取五个生成子集进行测试。所选子集涵盖不同生成来源的高保真伪造视频, 用于分析模型在不同生成来源下的跨数据集检测表现。

评价指标方面, 本文采用准确率 (Accuracy, ACC)、ROC 曲线下面积 (Area Under Curve, AUC) 和平均精度 (Average Precision, AP) 对模型性能进行综合评价。其中, ACC 用于衡量模型在固定判别阈值下的整体分类正确率; AUC 反映模型在不同阈值下区分真实样本与伪造样本的能力; AP 则基于精确率-召回率曲线衡量模型对伪造样本的检出性能, 在类别分布不均衡场景下具有更稳定的意义。

表 1 FakeAVCeleb 数据集上的域内评估结果 (ACC/%, AUC/%)

方法	模态	ACC	AUC
Xception ^[3]	V	67.9	70.5
LipForensics ^[13]	V	80.1	82.4
FTCN ^[16]	V	64.9	84
RealForensics ^[12]	V	89.9	94.6
AVFakeNet ^[27]	AV	78.4	83.4
VFD ^[19]	AV	81.5	86.1
MDS ^[18]	AV	82.8	86.5
AVoid-DF ^[28]	AV	83.7	89.2
MRDF-CE ^[30]	AV	94.1	92.4
AVFF ^[29]	AV	<u>98.6</u>	99.1
HAVIC ^[24]	AV	99.8	99.9
AVEmo	AV	98.1	<u>99.4</u>

3.2 实验设置

本章参照主流模型 AVFF^[29]和 HAVIC^[24]的实验设置, 首先将输入视频划分为 3.2 s 的视频片段, 并从每个片段中提取视觉帧及其对应的音频信号。其中, 视觉帧以 5 fps 进行采样, 随后采用 FaceX-Zoo 对人脸区域进行检测与裁剪, 并将裁剪后的人脸图像缩放至 224×224; 音频信号以 16 kHz 采样, 并转换为包含 128 个 Mel 频率通道和 1024 个时间帧的 Mel 频谱图。AVEmo 采用 MARLIN^[39]作为视频编码器, AudioMAE^[40]作为音频编码器, 以分别提取视觉模态和音频模态特征。训练过程中, 采用

AdamW 优化器训练 50 轮, 批次大小设置为 4, 初始学习率设置为 1e-4, 并使用余弦学习率调度策略进行优化。为缓解数据类别不均衡问题, 训练阶段采用加权采样策略。所有实验均以融合分类头的输出作为最终预测结果。

3.3 实验结果与分析

3.3.1 域内评估

为验证 AVEmo 在同分布场景下的检测性能, 本节首先在 FakeAVCeleb 数据集上进行域内评估实验。本文采用与 HAVIC^[24]一致的数据划分和评价协

议, 对比方法的实验结果也均参考 HAVIC。各方法在 FakeAVCeleb 上的训练与测试设置保持一致, 区别在于 AVFF 和 HAVIC 提前使用了真实视频进行了额外的视听表征预训练, 而 AVEmo 未设置该预训练阶段, 仅采用公开预训练权重初始化音频和视频骨干网络。表 1 给出了所提方法与现有方法在 ACC 和 AUC 指标上的比较结果。AVEmo 取得 98.1% 的 ACC 和 99.4% 的 AUC, 其中 AUC 较 AVFF 提高 0.3%, 仅低于 HAVIC; ACC 较 AVFF 低 0.5%。ACC 反映固定阈值下的分类结果, 而 AUC 衡量不同阈值下的整体排序能力, 因此两项指标之间的小幅差异主要与输出分数分布和分类阈值有关。说明与 AVFF 相比, AVEmo 虽然在固定阈值下的分类准确率略低, 但仍表现出略高的整体样本区分能力。

3.3.2 跨伪造方法评估

为评估模型对未知伪造类型的泛化能力, 本节进一步在 FakeAVCeleb 数据集上开展跨伪造方法实验。具体而言, 实验以真实视频-真实音频 (RVRA) 作为真实样本, 并在 RVFA、FVRA-WL、FVFA-FS、FVFA-GAN 和 FVFA-WL 等伪造子类中每次留出一种作为测试集, 其余伪造子类用于模型训练。实验结果如表 2 所示, 当测试样本来自训练阶段未见过的伪造方式时, 部分对比方法的性能出现明显下降, 说明依赖特定视觉伪影、音频伪影或固定音画同步关系的模型容易受到伪造类型变化的影响。相比之下, AVEmo 在四种 FV 伪造类型上的平均 AP 和 AUC 分别达到 99.8% 和 98.8%, 且整体保持较高的检测性能, 表明视听情感一致性能作为更具跨方法适应性的判别线索。

3.3.3 跨数据集评估

为评估模型的跨数据集泛化表现, 将基于

表2 FakeAVCeleb数据集上的跨伪造方法评估结果(AP/%,AUC/%)													
方法	模态	RVFA		FVRA-WL		FVFA-FS		FVFA-GAN		FVFA-WL		FV 平均值	
		AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC
Xception ^[3]	V	-	-	88.2	88.3	92.3	93.5	67.6	68.5	91	91	84.8	85.3
LipForensics ^[13]	V	-	-	97.8	<u>97.7</u>	99.9	99.9	61.5	68.1	98.6	98.7	89.5	91.1
FTCN ^[16]	V	-	-	96.2	97.4	100	100	77.4	78.3	95.6	96.5	92.3	93.1
RealForensics ^[12]	V	-	-	88.8	93	99.3	99.1	99.8	99.8	93.4	96.7	95.3	97.2
AV-DFD ^[26]	AV	74.9	73.3	97	97.4	99.6	99.7	58.4	55.4	100	100	88.8	88.1
AVAD (LRS2) ^[21]	AV	62.4	71.6	93.6	93.7	95.3	95.8	94.1	94.3	93.8	94.1	94.2	94.5
AVAD (LRS3) ^[21]	AV	70.7	80.5	91.1	93	91	92.3	91.6	92.7	91.4	93.1	91.3	92.8
AVFF ^[29]	AV	93.3	92.4	94.8	98.2	100	100	<u>99.9</u>	100	99.4	99.8	98.5	99.5
ICSAV ^[25]	AV	94	95.4	-	-	99.3	96.6	99.3	95.2	99	90.9	99.2	94.2
FoVB ^[31]	AV	99.8	<u>97.2</u>	99.7	92.6	99.2	99.6	99.4	99.5	100	99.8	<u>99.6</u>	97.9
AVEmo	AV	<u>99.1</u>	99.3	<u>99.1</u>	95.2	100	100	100	100	100	100	99.8	<u>98.8</u>

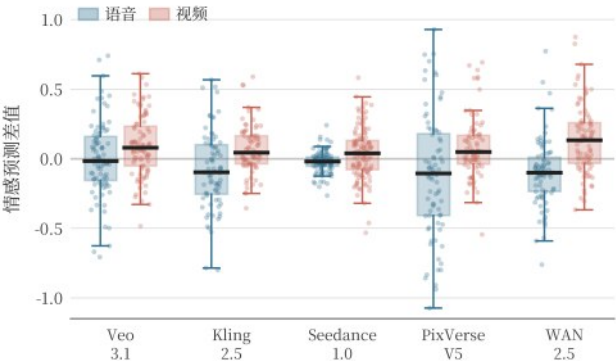


图3 HiFi-AVDF不同生成方法的单模态情感预测变化

FakeAVCeleb数据集训练得到的模型直接用于HiFi-AVDF中所选五个生成子集进行测试，实验结果如表3所示。AVEmo在五个生成子集上的平均

AUC达到78.9%，高于所比较的方法。其中，Seedance 1.0子集上的AUC达到98.5%；去除该子集后，AVEmo在其余四个子集上的平均AUC为74.0%，低于HAVIC的75.8%。上述结果表明，AVEmo在所评估场景下表现出一定的跨数据集泛化能力。其平均AUC优势主要受到Seedance 1.0子集结果的影响，模型在不同生成来源上的检测性能存在差异。

为描述不同生成来源下的单模态情感预测变化，本文采用参数固定的Wav2Vec2语音情感预测器^[41]和EmoNet视频情感预测器^[42]，分别计算生成片段相对于对应原始片段的语音与视频情感预测差值，结果如图3所示。Seedance 1.0的语音情感

预测差值主要集中在0附近，且分布范围明显

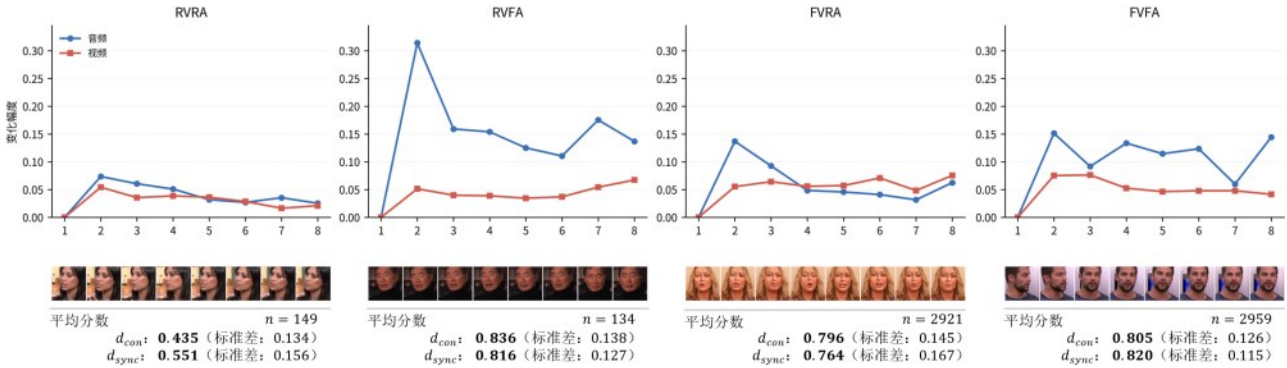


图4 不同视听真伪组合下的时序响应与差异分数统计

表3 HiFi-AVDF数据集上的跨数据集评估结果(AP/%,AUC/%)

方法	Veo 3.1		Kling 2.5		Seedance 1.0		PixVerse V5		WAN 2.5		平均值	
	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC
AVAD ^[21]	62.2	67.1	60.7	62.1	56.7	59.3	59.5	59.1	61.6	61.1	60.1	61.7
AVH-Align ^[20]	63.8	56.8	59.1	55.7	51.3	49.0	<u>74.9</u>	75.8	59.7	57.4	61.8	58.9
RealForensics ^[12]	61.3	58.2	61.5	62.9	64.4	65.7	57.4	60.5	61.7	64.9	61.3	62.4
LipFD ^[22]	56.4	59.5	62.5	68.5	61.4	67.6	57.2	60.2	54.3	61.9	58.4	63.5
AVFF ^[29]	63.9	60.6	<u>81.1</u>	78.3	70.4	74.4	59.2	53.9	67.1	74.4	68.3	68.3
Effort ^[8]	70.6	62.0	63.8	70.6	57.6	59.0	62.8	59.3	66.6	66.9	64.3	63.6
PIA ^[23]	58.9	65.1	66.4	73.1	64.3	65.0	53.7	57.2	59.2	65.2	60.5	65.1
HAVIC ^[24]	<u>68.7</u>	<u>65.9</u>	85.5	84.7	<u>76.8</u>	<u>81.1</u>	77.6	<u>74.4</u>	<u>76.5</u>	<u>78.0</u>	77.0	<u>76.8</u>
AVEmo	67.4	61.2	77.8	<u>83.9</u>	97.6	98.5	63.7	71.9	77.5	79.0	<u>76.8</u>	78.9

小于其他生成子集，这与其保留原始音频轨道、重新生成视觉内容的处理方式一致；其视频情感预测仍呈现一定变化。其他四种生成来源的语音与视频情感预测也出现了不同程度的变化。上述结果展示了不同生成来源下，音频与视频各自相对于原始片段的情感预测变化特征。此外，HiFi-AVDF还包含Sora 2子集，但该子集采用纯文本驱动的全生成方式，与本文关注的音视频篡改场景存在差异，且生成的音频与画面具有较强整体一致性，因此未纳入评估。

3.4 可视化分析

为分析不同视听真伪组合下的特征响应与差异分数，图4展示了RVRA、RVFA、FVRA和FVFA四类样本的个例时序响应，并补充各类测试视频的情感内容差异分数与变化同步差异分数统计。本节在FakeAVCeleb测试集上按式(27)计算两项分数。同一视频的多个片段先取平均，再统计各类有效视频的分数平均值与标准差。四类有效视频数量依次为149、134、2,921和2,959。图中曲线描述单个样本的模态内特征变化，下方统计描述对应类别的跨模态差异分数。

从情感内容差异来看，RVRA、RVFA、FVRA和FVFA的平均分数分别为0.435、0.836、0.796和0.805；从变化同步差异来看，四类平均分数分别为0.551、0.816、0.764和0.820。三类伪造样本的两项平均分数均高于RVRA，但不同真伪组合的表现有所差异：RVFA的内容差异平均分数最高，而FVRA的变化同步差异平均分数低于另外两类伪造样本，且标准差为0.167，高于RVFA的0.127和

FVFA的0.115，反映出其变化同步差异分数具有更大的离散程度。上述结果展示了整体真伪标签辅助监督下，模型能够学习与鉴伪任务相关的跨模态差异表征。图4从表征层面展示了该辅助监督下四类样本的分数特点，第3.6节则进一步通过损失消融考察其对检测性能的贡献。

3.5 多模态情感特征分析

为进一步验证AVEmo对情感相关信息的学习能力，参照文献[33]的实验设置，本文在CMU-MOSI^[43]情感数据集上开展迁移实验。CMU-MOSI将视频片段的情感倾向标注为-3至3，其中负值表示负向情感，正值表示正向情感，绝对值反映情感强度。实验加载完成鉴伪训练的AVEmo权重并冻结，仅在语音分支和视频分支后分别添加轻量级的分类头进行迁移训练。

为观察语音分支和视频分支对情感信息的响应，本文从测试集中随机选取样本，并绘制两路分支的情感预测结果，如图5所示。其中，红色实线和黄色虚线分别表示语音分支与视频分支的情感预测值，黑色虚线表示两路预测结果的标准差。可以看出，两路预测在多数样本上呈现相近的变化趋势，且标准差整体较小，说明两个分支能够对情感方向和强度的变化产生较为一致的响应。上述结果表明，AVEmo能够从语音和视频内容中学习情感相关特征，并对情感方向和强度的变化作出一致响应，从而证明模型具备一定的多模态情感信息提取能力。

3.6 消融实验

为验证多模态情感特征提取器中各组成部分的

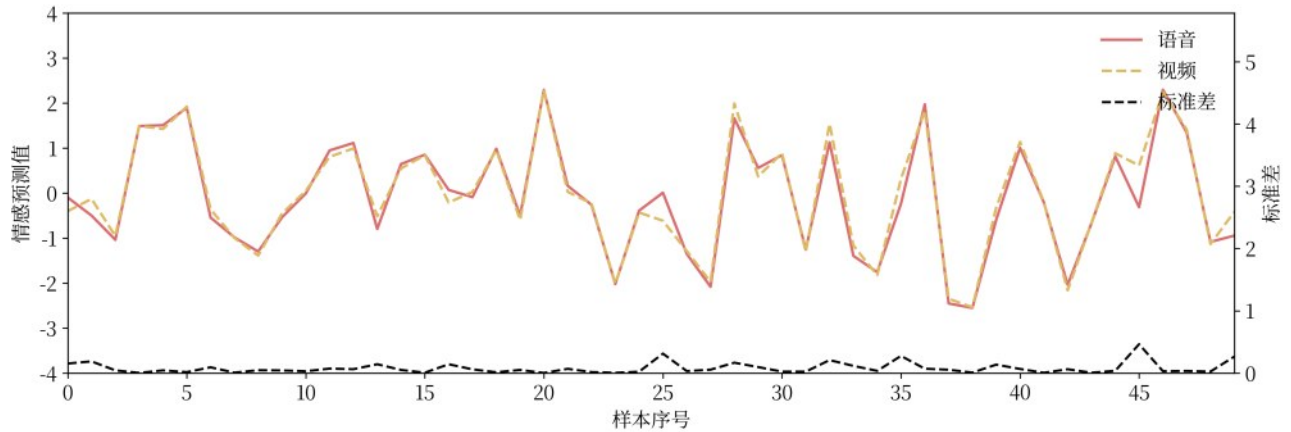


图5 CMU-MOSI测试集随机抽样的多模态情感预测结果

有效性, 本文分别对面部情感提取单元、语境情感提取单元以及双主导跨模态交互机制进行消融。其中, 移除面部情感提取单元或语境情感提取单元时, 直接将其输入特征作为输出; 关闭双主导跨模态交互时, 保留两路情感特征的提取过程, 但不进行跨模态辅助与仿射调制。其余网络结构和训练目标均保持不变, 实验结果如表4所示。与完整AVEmo相比, 移除面部情感提取单元和语境情感提取单元后, 模型在HiFi-AVDF上的AUC分别下降5.3%和3.8%, 说明两类情感提取单元均有助于检测性能的提升。关闭双主导跨模态交互后, AUC降至68.9%, 较完整模型下降10.0%, 表明在两路情感特征提取的基础上开展跨模态交互, 能够进一步提高模型的判别能力。总体来看, 三个组成部分均对模型性能具有积极作用, 其中关闭双主导跨模态交互带来的性能下降最为明显。

表4 各功能模块消融实验

方法	参数量/M	FakeAVCeleb AUC/%	HiFi-AVDF AUC/%
移除面部情感提取单元	258	95.5	73.6
移除语境情感提取单元	247	94.9	75.1
关闭双主导跨模态交互机制	263	92.1	68.9
AVEmo	285	99.4	78.9

进一步地, 为分析多任务联合优化中不同监督项的作用, 在保持完整网络结构不变的情况下, 对视频和音频模态辅助分类损失、情感内容一致性损失以及情感变化同步损失进行消融, 实验结果如表5所示。仅采用视听融合分类损失时, 模型在

FakeAVCeleb和HiFi-AVDF上的AUC分别为90.8%和65.3%。加入模态级辅助分类监督后, 分别提升至95.0%和69.1%。在此基础上, 加入情感内容一致性损失后, 两项AUC分别达到98.6%和76.6%; 加入情感变化同步损失后, 分别达到97.5%和75.1%。同时采用两项差异损失时, AUC进一步达到99.4%和78.9%。上述结果表明, 在表5的实验设置下, 以整体真伪标签监督的跨模态差异约束能够在模态级分类监督的基础上进一步提高检测性能, 为该辅助监督策略的任务有效性提供了支持。图4的补充分数分析则展示了训练后不同真伪组合下的表征特点。

表5 各训练目标的消融实验

$\mathcal{L}_{a+v}$	$\mathcal{L}_a + \mathcal{L}_v$	$\mathcal{L}_{con}$	$\mathcal{L}_{sync}$	FakeAVCeleb AUC/%	HiFi-AVDF AUC/%
√	×	×	×	90.8	65.3
√	√	×	×	95.0	69.1
√	√	√	×	98.6	76.6
√	√	×	√	97.5	75.1
√	√	√	√	99.4	78.9

表4和表5分别从功能模块和训练目标两个角度验证了模型各组成部分的有效性。为进一步比较不同视听建模机制, 本文在保持骨干网络、数据划分和训练配置一致条件下, 设置普通视听特征融合、普通视听同步、普通情感一致性和AVEmo基础模型四组对照实验, 结果如表6所示。其中, 普通视听特征融合直接聚合并拼接音频与视频特征; 普通视听同步逐段计算同一时刻音频特征与视频特

征的匹配程度；普通情感一致性仅比较音频与视频的全局情感表示；AVEmo基础模型则同时考虑视听同步和情感一致性，但不包含面部情感提取单元、语境情感提取单元及双主导跨模态交互机制。实验结果表明，引入情感一致性能够在普通特征融合和局部时序同步的基础上进一步提升检测性能，而完整AVEmo取得了最优结果，说明对视听情感内容及其动态变化关系进行联合建模能够更充分地挖掘伪造线索。

表6 不同机制的对比实验

方法	FakeAVCeleb AUC/%	HiFi-AVDF AUC/%
普通视听特征融合	80.4	57.3
普通视听同步	87.6	66.2
普通情感一致性	85.7	67.8
AVEmo基础模型	<u>92.5</u>	<u>71.4</u>
AVEmo	99.4	78.9

4 结束语

针对现有多模态方法对视听级语义关系利用不足的问题，提出了一种基于视听情感一致性的多模态人脸伪造检测方法AVEmo。该方法分别从面部表情和语音语境中挖掘情感线索，并通过双主导跨模态交互和注意力瓶颈进一步建模视听情感内容一致性及其变化同步关系，同时结合多任务联合优化增强模型的伪造判别能力。实验结果表明，AVEmo在FakeAVCeleb上的AUC达到99.4%，在跨伪造方法实验中取得99.8%的平均AP，并在HiFi-AVDF所评估的五个生成子集上取得78.9%的平均AUC。上述结果验证了所提方法的检测有效性，并体现出一定的跨数据集泛化能力。需要指出的是，视听情感一致性并非所有真实视频均需满足的严格条件。在情感抑制、刻意控制表情、表演或配音等场景中，面部表情与语音情感可能存在差异或时间偏移，从而带来误判风险。尽管AVEmo综合利用音频信息、视觉信息、跨模态交互信息及时序特征，上述风险仍不能完全消除。未来将进一步扩充包含弱情感、情感抑制和跨模态异步表达的真实样本，并探索视听情感一致性在全生成式视频及更加复杂跨域场景中的应用，以提升方法对新型高质量伪造内容的适应能力。

参考文献：

- [1] PEI G, ZHANG J, HU M, et al. Deepfake generation and detection: A benchmark and survey[J]. ACM Computing Surveys, 2026, 58(11): 1-41.
- [2] 王利华,王新宇,严伟丹,等.人脸深度伪造检测技术研究综述[J]. 计算机科学,2026,53(8):375-387.
WANG L H, WANG X Y, YAN W D, et al. Review of Research on Face Deepfake Detection Technology[J]. Computer Science, 2026, 53(8): 375-387.
- [3] CHOLLET F. Xception: Deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1251-1258.
- [4] TAN M, LE Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]//International conference on machine learning. PMLR, 2019: 6105-6114.
- [5] SHIOHARA K, YAMASAKI T. Detecting deepfakes with self-blended images[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 18720-18729.
- [6] QIAN Y, YIN G, SHENG L, et al. Thinking in frequency: Face forgery detection by mining frequency-aware clues[C]//European Conference on Computer Vision. Springer, Cham, 2020: 86-103.
- [7] WANG J, WU Z, OUYANG W, et al. M2TR: Multi-modal multi-scale transformers for deepfake detection[C]//Proceedings of the 2022 International Conference on Multimedia Retrieval. 2022: 615-623.
- [8] YAN Z, WANG J, JIN P, et al. Orthogonal Subspace Decomposition for Generalizable AI-Generated Image Detection[C]//Proceedings of the 42nd International Conference on Machine Learning. PMLR, 2025, 267: 70268-70288.
- [9] 李子龙,杨高明,韩栋宇,等.融合动态伪影追踪与空频交互分析的跨域深度伪造检测[J]. 电子与信息学报,2026,48(6):2743-2753.
LI Z L, YANG G M, HAN D Y, et al. Cross-domain Deepfake Detection with Dynamic Artifact Tracking and Spatio-frequency Interaction Analysis[J]. Journal of Electronics & Information Technology, 2026, 48(6): 2743-2753.
- [10] 郭祯,邱润尧,徐嘉,等.基于多特征协同置信感知的深度伪造图像检测与定位方法[J]. 通信学报,2026,47(8):156-175.
GUO Z, QIU R Y, XU J, et al. Deepfake Image Detection and Localization Method Based on Multi-feature Collaborative Confidence Awareness[J]. Journal on Communications, 2026, 47(8): 156-175.
- [11] PENG C, MIAO Z, LIU D, et al. Where deepfakes gaze at? Spatial-temporal gaze inconsistency analysis for video face forgery detection [J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 4507-4517.
- [12] HALIASSOS A, MIRA R, PETRIDIS S, et al. Leveraging real talking faces via self-supervision for robust forgery detection[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022: 14930-14942.
- [13] HALIASSOS A, VOUGIOUKAS K, PETRIDIS S, et al. Lips don't lie: A generalisable and robust approach to face forgery detection[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021: 5037-5047.
- [14] 熊礼治,曹梦琦,付章杰.基于三维双流网络的视频目标移除篡改取证[J]. 通信学报, 2021, 42(12): 202-211.

- XIONG L Z, CAO M Q, FU Z J. Forensic of video object removal tamper based on 3D dual-stream network[J]. Journal on Communications, 2021, 42(12): 202-211.
- [15] GU Z, CHEN Y, YAO T, et al. Delving into the local: Dynamic inconsistency learning for deepfake video detection[C]//Proceedings of the AAAI conference on artificial intelligence. 2022, 36(1): 744-752.
- [16] ZHENG Y, BAO J, CHEN D, et al. Exploring temporal coherence for more general video face forgery detection[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2021: 15024-15034.
- [17] YANG Z, LIANG J, XU Y, et al. Masked relation learning for deepfake detection[J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 1696-1708.
- [18] CHUGH K, GUPTA P, DHALL A, et al. Not made for each other: audio-visual dissonance-based deepfake detection and localization[C]//Proceedings of the 28th ACM international conference on multimedia. 2020: 439-447.
- [19] CHENG H, GUO Y, WANG T, et al. Voice-face homogeneity tells deepfake[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2023, 20(3): 1-22.
- [20] SMEU S, BOLDISOR D A, ONEATA D, et al. Circumventing shortcuts in audio-visual deepfake detection datasets with unsupervised learning[C]//2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2025: 18815-18825.
- [21] FENG C, CHEN Z, OWENS A. Self-supervised video forensics by audio-visual anomaly detection[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2023: 10491-10503.
- [22] LIU W, SHE T, LIU J, ET AL. Lips are lying: Spotting the temporal inconsistency between audio and visual in lip-syncing deepfakes[J]. Advances in Neural Information Processing Systems, 2024, 37: 91131-91155.
- [23] DATTA S K, RANGA T, SUN C, et al. PIA: Deepfake detection using phoneme-temporal and identity-dynamic analysis[C]//2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). IEEE, 2025: 1607-1617.
- [24] PENG J, WANG Y, LI Y, et al. Leave No Stone Unturned: Uncovering Holistic Audio-Visual Intrinsic Coherence for Deepfake Detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings. IEEE, 2026: 6655-6666.
- [25] ANSHUL A, GOPAL S, RAJAN D, et al. Intra-modal and cross-modal synchronization for audio-visual deepfake detection and temporal localization[C]//2025 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2025: 13826-13836.
- [26] ZHOU Y, LIM S N. Joint audio-visual deepfake detection[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2021: 14780-14789.
- [27] ILYAS H, JAVED A, MALIK K M. AVFakeNet: A unified end-to-end Dense Swin Transformer deep learning model for audio - visual deepfakes detection[J]. Applied Soft Computing, 2023, 136: 110124.
- [28] YANG W, ZHOU X, CHEN Z, et al. Avoid-df: Audio-visual joint learning for detecting deepfake[J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 2015-2029.
- [29] OORLOFF T, KOPPISETTI S, BONETTINI N, et al. Avff: Audio-visual feature fusion for video deepfake detection[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2024: 27092-27102.
- [30] ZOU H, SHEN M, HU Y, et al. Cross-modality and within-modality regularization for audio-visual deepfake detection[C]//ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024: 4900-4904.
- [31] NIE F, NI J, ZHANG J, et al. Towards Generalizable deepfake detection via forgery-aware audio-visual adaptation: A variational bayesian approach[J]. IEEE Transactions on Information Forensics and Security, 2026.
- [32] MITTAL T, BHATTACHARYA U, CHANDRA R, et al. Emotions Don't Lie: An Audio-Visual Deepfake Detection Method Using Affective Cues[C]//Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM, 2020: 2823-2832.
- [33] ZHANG T, LI G, XIAO Y, et al. Emotion-Aware Multimodal Deepfake Detection[J]. Neural Networks, 2026, 199: 108675.
- [34] GU A, DAO T. Mamba: Linear-time sequence modeling with selective state spaces[J]. arXiv preprint arXiv:2312.00752, 2023.
- [35] BECK M, PÖPPEL K, SPANRING M, et al. xlstm: Extended long short-term memory[J]. Advances in Neural Information Processing Systems, 2024, 37: 107547-107603.
- [36] 高帆, 严伟丹, 邵文泽, 等. 面向人脸伪造防御的属性感知对抗样本生成方法[J]. 电子与信息学报, 2026, 48(7): 2930-2942.
- GAO F, YAN W D, SHAO W Z, et al. Defending against Deepfakes by Attribute-Aware Attack[J]. Journal of Electronics & Information Technology, 2026, 48(7): 2930-2942.
- [37] KHALID H, TARIQ S, KIM M, et al. FakeAVCeleb: A novel audio-video multimodal deepfake dataset[J]. arXiv preprint arXiv: 2108.05080, 2021.
- [38] KWON P, YOU J, NAM G, et al. Kodf: A large-scale korean deepfake detection dataset[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2021: 10724-10733.
- [39] CAI Z, GHOSH S, STEFANOV K, et al. Marlin: Masked autoencoder for facial video representation learning[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2023: 1493-1504.
- [40] HUANG P Y, XU H, LI J, et al. Masked Autoencoders that Listen[C]//Advances in Neural Information Processing Systems. 2022, 35: 28708-28720.
- [41] WAGNER J, TRIANTAFYLLOPOULOS A, WIERSTORF H, et al. Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(9): 10745-10759.
- [42] TOISOUL A, KOSSAIFI J, BULAT A, et al. Estimation of Continuous Valence and Arousal Levels from Faces in Naturalistic Conditions[J]. Nature Machine Intelligence, 2021, 3(1): 42-50.
- [43] ZADEH A, ZELLERS R, PINCUS E, et al. Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages[J]. IEEE Intelligent Systems, 2016, 31(6): 82-88.



王利华（1975— ），男，南京邮电大学博士生，主要研究方向为人脸深度伪造检测。



严伟丹（1999— ），男，博士，无锡学院讲师，主要研究方向为图像复原、伪造检测与缺陷检测。