arXiv:2303.12745v2 [cs.CV] 4 Aug 2023

视听欺骗检测：DOLOS 数据集与参数高效跨模态学习

郭小宝[12*] Nithish Muthuchamy Selvaraj[3*] 于子童[3†] AdamsWai-Kin Kong[2]沈炳权[4] AlexKot[31]
南洋理工大学 Rapid-Rich Object Search (ROSE) 实验室、跨学科研究生项目[2]南洋理工大学计算机科学与工程学院 [3]南洋理工大学电气与电子工程学院[4]新加坡 DSO 国家实验室{xiaobao001, ms

.nithish, zitong.yu, adamskong, eackot}@南洋理工大学.edu.sg sbingqua@dso.org.s g

摘要

对话中的欺骗检测是一项具有挑战性但十分重要的任务，在商业可信度评估、多媒体反欺诈和海关安全等诸多领域具有关键应用。尽管如此，欺骗检测研究仍受到高质量欺骗数据集匮乏，以及难以有效学习多模态特征等问题的制约。为解决这一问题，我们推出了 *DOLOS*[1]，这一规模最大的游戏节目欺骗检测数据集，包含丰富的欺骗性对话。*DOLOS* 包含 *1,675* 个视频片段，涵盖 *213* 名受试者，并附有视听特征标注。我们提供训练-测试、时长和性别方案，以探究不同因素的影响。我们在此前提出的欺骗检测方法上对该数据集进行基准测试。为进一步通过微调更少参数来提升性能，我们提出参数高效跨模态学习（*PECL*），其中，均匀时序适配器（*UT-Adapter*）探索基于 *Transformer* 的架构中的时序注意力，而即插即用音视频融合（*PAVF*）跨模态融合模块则结合音视频特征中的跨模态信息。基于 *DOLOS* 丰富的细粒度视听标注，我们还采用多任务学习，通过同时预测欺骗和音视频特征来提升性能。实验结果证明了 *DOLOS* 数据集的理想质量以及 *PECL* 的有效性。*DOLOS* 数据集和源代码可在此处获取。

*共同贡献
†通讯作者[1]“DOLOS”这个名称源自希腊神话。

1. 引言

欺骗是一种普遍且复杂的现象，存在于生活的各个领域。理解其本质和影响对于防止负面后果至关重要。在边境安全、反欺诈、商务谈判等领域，有效的欺骗检测具有重要意义。[14, 12, 1, 13, 9, 27, 44]深度学习算法在许多复杂任务中的表现已达到甚至超过人类水平 [22, 41, 40, 7, 38]。人们或许会期待人工智能模型也能在欺骗检测方面带来重大突破。尽管计算机视觉 [48, 10, 31, 46] 和音频表征学习 [5, 18, 28], 已取得丰硕进展，但如何高效地探索人工智能处理多模态信息的能力，以感知和预测人类的欺骗行为，仍是一项重大挑战。

人工智能模型在欺骗检测中的表现高度依赖于能否获得真实世界中真实且有效的欺骗样本。欺骗者必须表现得自然且具有动机 [8, 49] ，这样某些行为线索（例如音高和抬下巴）才会更加明显。来自法庭审判 [36], 、电视游戏节目 [42], 和实验室场景 [15], 的公开基准数据集，推动了人们对欺骗检测研究的兴趣，并促进了相关研究的进展。在多种现实场景中，欺骗会受到诸多因素的影响。因此，有必要针对不同场景创建欺骗数据集。然而，由于数量和质量方面的局限，现有数据集仍不足以推动进一步发展并激发新思路。这些局限包括：（1）欺骗样本和受试者数量较少；（2）缺乏丰富的带标注的视觉和语音属性；（3）实验方案多样。因此，亟需构建一个规模更大、内容更丰富的欺骗检测数据集。尤其需要更多欺骗样本和受试者，为面部动作、手势和音频属性提供更完善的标注，

并增加更多类型的实验方案，以研究影响欺骗检测的因素。

除了高质量的数据集，有效的方法对于欺骗检测同样重要。许多研究致力于利用视频中的视觉和声学信息进行欺骗检测 [42, 15, 36]。现有方法可分为两类：单模态学习和多模态融合。然而，这两种方法都存在局限，因为它们未能充分利用单模态特征，也未能整合多个模态的互补信息。为了更好地利用现有信息，对大型预训练模型进行微调已展现出良好的效果 [39, 26, 20, 47]。然而，对预训练模型（如 W2V2 [5]和 ViT [10]，）进行全量微调可能效率低下并导致过拟合，尤其是在下游数据集规模有限时。因此，在微调预训练模型时，考虑参数效率至关重要。适配器 [17, 21] 为模型微调提供了一种高效的方法。适配器最初是为语言 [17]或视觉 [21] 任务提出的，但尚未应用于多个模态的时序特征提取。

贡献。 本文建立了一个新的欺骗检测数据集，并提出了用于视听欺骗检测的参数高效跨模态学习方法。

作为首项贡献，我们推出了 DOLOS，这是一个用于视听欺骗检测的新型游戏节目数据集。与当前公开数据集相比，DOLOS 具有多项优势。首先，游戏节目是收集欺骗检测数据的可靠来源，因为所有参与者都有作弊动机，而且真实标签是已知的。其次，我们提出的数据集采用对话场景，其中的欺骗行为呈现得更加自然。第三，按受试者数量计算，我们的数据集规模最大；按视频片段数量计算，它也是规模最大的不基于实验室的数据集。该数据集经过标注，包含细粒度的视听标注。我们还使用涉及单模态和多模态特征的既有欺骗检测方法，对该数据集进行了基准测试。此外，我们提供了三种不同的实验方案，即训练-测试、时长和性别，以探究可能影响欺骗检测的各种因素。

我们的第二项贡献是参数高效跨模态学习（ PECL），这是一种欺骗检测方法，仅需微调少量额外的可学习权重即可实现高性能。具体而言，我们提出了一种均匀时序适配器（UT-Adapter），无需进行复杂的修改，即可探索视觉和音频模态的输入嵌入之间的时间注意力机制。对于多模态融合，我们提出了即插即用音视频融合（PAVF），利用视觉和音频特征之间的互补信息来提升整体性能。PECL 旨在实现参数高效，仅将 UT-Adapter、PAVF 模块和分类器设为可训练参数。此外，我们还探索了多任务学习的优势；事实证明，这种方法能够提升音频和视觉任务的表现 [54, 45]。通过同时预测欺骗、面部动作和语音特征，我们提出的方法能够进一步改进。

为了比较并展示我们数据集和所提方法的优势，我们进行了大量实验。在与当前游戏节目基准数据集的交叉测试中， [42], DOLOS 在不同场景的若干测试集上表现更佳。DOLOS 上的实验结果也表明，P ECL 在不同实验方案下表现出色。通过这些实验，我们揭示了多模态欺骗检测方面的宝贵见解。我们相信，DOLOS、所提方法以及基准测试实验将为其他研究人员推进该领域的研究提供宝贵资源。

2. 相关工作

2.1. 欺骗检测数据集

欺骗检测研究的早期工作探讨了欺骗的心理学基础， [8, 16, 50, 30, 49]并识别出潜在线索，例如生理、视觉、声音、语言和行为线索。多模态欺骗检测的开创性工作随着真实生活试验（RLT）数据集的引入而出现， [36] 该数据集采集自法庭审判。法庭环境提供了一个有价值且真实的说谎场景，但 RLT 规模较小，其性能瓶颈很快便显现出来。P´erez-Rosas 等人 [37] 和 Kamboj 等人 [23] 后来提出了另外两个新数据集，分别采集自街头采访和政治演讲。然而，很难核实这些数据集中受试者回答的真实性。为克服这些不足，研究者提出了若干基于实验室的多模态欺骗检测数据集。这些数据集通常设置在允许说话者撒谎的情境中，例如描述某件物品、 [15], 某种性格， [32], 以及与受试者进行面对面采访 [43]。然而，欺骗动机可能较弱，因此捕捉到的欺骗线索较少。

以真实或谎言游戏为主题的真人秀节目，为收集对话欺骗的实例提供了途径。它们也恰好兼顾了激励、真实性和样本量方面的要求。参与者常常为了赢得比赛而相互说谎，并在回合结束时揭晓真相。网上有大量游戏节目录像。首个基于游戏节目的欺骗检测数据集是在 [42]中提出的，且主要关注两种模态：文本和人工标注的视觉特征。然而，直接使用该数据集，借助音频和视觉模态开展多模态欺骗检测会面临若干挑战，我们将在第3节进行说明。

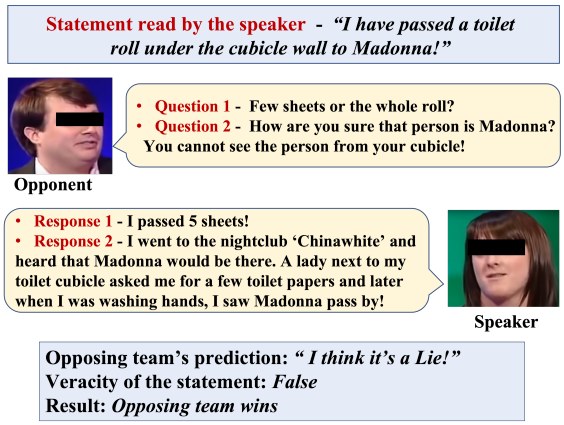


图 1. 游戏节目形式。欺骗回合示例。

2.2. 多模态欺骗检测

以往的多模态欺骗检测研究包括使用单模态和多模态特征，并探索有效的融合机制 [33, 24, 25, 4, 9, 51, 13]。Gogate等人 [13] 提出了一种深度模型，将音频线索与视觉和文本模态相结合，以提高欺骗预测的准确率。Karimi等人 [24] 以端到端的方式，从 RGB 图像和原始音频中探索欺骗线索。Wu等人 [51] 使用了多种特征，包括 RGB 图像中的微表情和改进密集轨迹（IDT）特征、音频中的梅尔频率倒谱系数（MFCC）特征以及转录文本。Avola等人 [4] 直接使用视频帧中的面部动作单元（AUs），并通过支持向量机（SVM）对其进行分类，以解决欺骗检测问题。Ding等人 [9] 尝试结合面部表情和身体动作，以提取更多视觉欺骗线索。他们还尝试通过加入音频特征和转录文本来提升性能。Mathur 和 Matarić [33] 聚焦于用于欺骗检测的面部表情，并研究了视觉、声音和语言模态中的可解释特征。Karnati等人 [25] 提出了能够学习用于欺骗检测的音频、视觉和脑电图表征的深度网络。

大多数以往研究试图通过引入更多模态、更多特征以及采用更好的特征提取方法来提升性能。然而，这些研究过度依赖空间特征，而非模态内部及模态之间的时间信息，因此可能会丢失潜在的重要线索，导致结果不够理想。

在融合方法方面，最常见的方法是特征级融合：将多模态嵌入拼接起来，并通过线性层提取跨模态信息 [24, 4, 9, 51,13]。其他研究则采用决策级融合，例如分数级融合 [25, 13]。在所有情况下，可靠且高效的融合方法都能提升性能。

3. DOLOS 数据集

3.1. 游戏节目形式

我们从 YouTube 上播出的英国真人喜剧游戏节目中收集数据 [2]，节目中六位参与者分成两支队伍进行比赛。每轮由一名队员从卡片上读出一则关于自己生活的陈述。这则陈述可能是真，也可能是假，只有发言者知道真相。随后，另一支队伍向发言者提出一系列问题，发言者必须作答，为自己的陈述辩护。发言者的目标是说服对方队伍相信自己的陈述属实，而对方队伍则试图仅根据发言者的回答，判断陈述究竟是真是假。每轮结束时，陈述的真实性会揭晓；如果对方队伍正确猜出陈述是真是假，就赢得该轮，否则发言者所在队伍获胜。游戏会进行数轮，直到所有参与者都轮流上场。值得注意的是，如果陈述属实，那么发言者的回答就会被视为真实视频片段，反之亦然。游戏节目的形式见图1。

3.2. 数据收集与标注流程

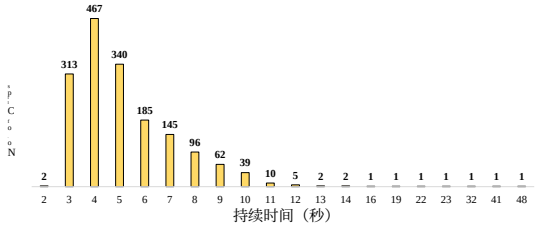
从游戏节目视频的每一集中，提取符合以下要求的视频片段：

- 参与者只清晰地说出相关内容（即，说真话或谎话），且没有明显的背景噪声。
- 参与者的面部清晰可见，未被遮挡。

我们从总计84集中提取了1675个视频片段，涉及213名参与者（141名男性和72名女性）。视频片段的时长为2至19秒。我们使用广为采用的MUMIN编码方案，对提取的视频片段中的非语言欺骗线索进行人工标注[2], ，重点关注视觉（25种面部）和语音（5种语音）特征。MUMIN编码方案中的非语言特征以及视频片段的分布如图2所示。

为消除人工标注员之间的偏差，六名人工标注员都对数据集中的同一子集进行了标注，以此完成校准，并使用科恩卡帕系数评估其表现 [6]。如果发现标注员之间的视觉标注和音频标注存在显著差异，他们会通过讨论协调分歧，然后重新标注该子集。重复这一迭代过程，直至达到标注者间一致性。校准前，科恩卡帕系数的平均值为 0.5（音频标注为 0.53，视觉标注为 0.47）。校准后

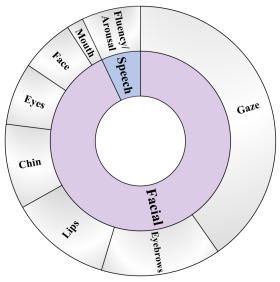
[2]视频依据合理使用政策从YouTube收集



(a) 片段分布（按时长）

MUMIN编码方案			
Lips 嘴角上扬 嘴角下垂 突出 后缩	**Gaze** 朝向交谈对象 Up Down 向侧面	**Chin** 向上移动 向下移动 **眉毛** 皱眉 抬起	**Face** 微笑 大笑 怒容
Eyes 频繁眨眼 夸张张开 夸张闭合	嘴巴 夸张张开 闭合（尴尬的停顿）	流畅度 非"啊、嗯"类言语障碍 词语和短语重复 静默停顿	唤醒度 响度 紧张度

(b) 标注 - 面部和言语特征



(c) 特征分布

	欺骗	真实	总计
Male	746	578	1324
女性	153	198	351
总计	899	776	1675

(d) 片段分布（按真实性和性别划分）

图2. 用于多模态欺骗检测的 DOLOS 数据集。放大查看效果更佳。

数据集	人工标注特征	受试者数量	样本 总计	欺骗性	真实	欺骗/真实比率	场景
真实生活试验[36]	✓	56	121	61	60	1.02	法庭
现实生活中的欺骗[37]		-	56	31	25	1.24	街头采访
政治欺骗[23]		88	180	-	-	-	政治演讲
谎言之袋[15]		35	325	162	163	0.99	Lab
MU3D[32]		80	320	160	160	1	Lab
欺骗检测和远程光电容积描记[43]		70	1680	630	1050	0.6	Lab
谎言盒子[42]	✓	26	1049	862	187	4.61	游戏节目
DOLOS（我们的方法）	✓	213	1675	899	776	1.16	游戏节目

表格 1. 多模态欺骗检测数据集的比较

集

平均分数提高至 0.65（音频标注为 0.67，视觉标注为 0.63）。

根据图2中的统计数据，我们将 DOLOS 数据集划分为三种不同的实验方案。首先，我们提供了一种包含 3 折划分的训练-测试协议，用于评估欺骗检测器在 DOLOS 数据集上的性能。其次，为反映真实场景中说话时长的差异，我们根据数据集中的统计数据，提供了包含短片段（2-4 秒）和长片段（5-10 秒）的时长协议。此外，我们还提供了性别（男性和女性）实验方案，以研究欺骗检测任务中的不同因素。

3.3. 与谎言盒子及其他数据集的比较

我们将 DOLOS 与谎言盒子（BOL）进行比较， [42]，后者同样采集自游戏节目视频。从数量上看，BOL 的视频片段和受试者都较少。BOL 仅包含 26 名受试者，占 DOLOS 的 17.9%。从质量上看，BOL 的 1,049 个视频片段中，有 537 个缺少说话者面部。在其他视频片段中，部分片段里的说话者与音频不匹配。BOL 中，游戏节目的主持人对应的视频片段比其他受试者更多，因此造成了较大的偏差。相比之下，第3.2 节中提到的数据采集要求从根本上避免了这些问题。因此，DOLOS 更适用于视听欺骗检测

研究（见第5.3节）比 BOL 更为丰富。

我们还将 DOLOS 与其他公开可用数据集进行比较。表1总结了比较结果。简而言之，DOLOS 具有以下优势：

- DOLOS 的受试者数量最多。在非实验室场景中，其视频片段数量也最多。
- 与谎言盒子游戏节目数据集相比，DOLOS 包含更多视频片段，且欺骗样本与真实样本的比例更加均衡。
- DOLOS 为所有视频片段提供了人工标注的 MUMIN 特征。这些标注可直接用于欺骗检测任务，也可应用于其他相关任务（例如，面部表情预测）或多任务学习。

4. 方法论

为高效利用来自大规模通用音频和视觉任务的先验知识，我们提出了用于音视频欺骗检测的参数高效跨模态学习方法（PECL）。PECL 的整体框架如图3 所示，其中包括主网络架构（图3 (a)）以及若干具有详细结构的组件（图3 (b)-(e)）。

具体而言，在图 3(a) 中，视觉输入由二维卷积模块进行标记化，音频输入则由

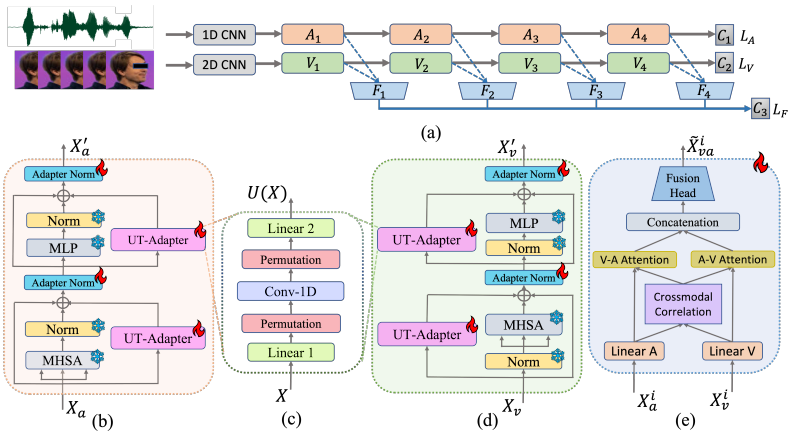


图3. 参数高效跨模态学习（PECL）框架。(a) 为主体架构。(b) 为音频特征的编码器层 A_i。(c) 为音频和视觉特征共享结构的 UT-Adapter。(d) 为视觉特征的编码器层 V_i。(e) 为 PAVF 模块 F_i。可训练模块和层以“火焰”标记，冻结层以“雪花”标记。

一维卷积模块（见第 5.1 节）进行标记化，二者的维度相同，分别记为 $\boldsymbol{X_v} \in \mathbb{R}^{L\times D}$ 和 $\boldsymbol{X_a} \in \mathbb{R}^{L\times D}$，其中 L 是序列长度，D 是每个序列嵌入的维度。V_i 和 A_i 分别表示基于 Transformer 的视觉模态和音频模态编码器层；我们采用 W2V2 [5] 作为音频模态的骨干网络，并采用 ViT [10] 作为视觉模态的骨干网络。F_i 是每对单模态编码器层之间的融合模块，且 $i = 1$ = 2、3、4。C_1、C_2 和 C_3 分别是音频、视觉和融合网络的分类器。L_A、L_V 和 L_F 分别是音频、视觉和融合模态的交叉熵损失。

4.1. 均匀时序适配器（UT-Adapter）

为提高参数效率并缓解过拟合问题，我们在预训练视觉和音频模型中插入可学习适配器层 [17, 21]。训练期间，除适配器层和分类层外，所有权重均被冻结。不同于用于语言任务的 NLP-Adapter [17] 和用于视觉任务的 Conv-Adapter [21]，我们将适配器层命名为均匀时序适配器（UT-Adapter），旨在探索视觉和音频模态的时间注意力机制。

令 F_V 和 G_A 分别表示图3中的视觉和音频编码器层。它们的输出表示为

$$\boldsymbol{X_v^{'}} = F_V(\boldsymbol{X_v})\ and\ \boldsymbol{X_a^{'}} = G_A(\boldsymbol{X_a}), \tag{1}$$

其中 $\boldsymbol{X_v^{'}} \in \mathbb{R}^{L\times D}$ 和 $\boldsymbol{X_a^{'}} \in \mathbb{R}^{L\times D}$。$F_V$ 和 G_A 由多头自注意力（MHSA）模块、UT-Adapter、多层感知器（MLP）模块和层归一化（LN）构成。为简化记号，令 $\boldsymbol{X}$为一个输入特征，即即，$\boldsymbol{X_v}$ 和 $\boldsymbol{X_a}$（图3（b）和（d））。$\boldsymbol{X}$ 被映射为查询 $\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W_Q}$、键 $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{W_K}$和值 $\boldsymbol{V} = \boldsymbol{X}\boldsymbol{W_V}$，其中 $\boldsymbol{W_Q}$，$\boldsymbol{W_K}$和 $\boldsymbol{W_V} \in \mathbb{R}^{D\times D}$ 是预训练的投影权重。对于

对于单个注意力头 H_j，$j \in \{1, 2, \cdots, N\}$，自注意力运算可表示为

$$H_j(\boldsymbol{X}) = Softmax\left(\frac{\boldsymbol{Q_j}\boldsymbol{K_j}^{\intercal}}{\sqrt{D}}\right)\boldsymbol{V_j}, \tag{2}$$

随后，将所有 H_j、$j \in \{1, 2, \cdots, N\}$ 的输出拼接起来，并投影为单个输出 $H(\boldsymbol{X})$。

UT-Adapter 与多头自注意力和多层感知机模块并行放置。具体而言，UT-Adapter 由线性层和一维卷积层堆叠而成。如图3 (c) 所示，UT-Adapter 的输出按如下方式运算

$$U(\boldsymbol{X}) = L_2(P(C(P(L_1(\boldsymbol{X}; \boldsymbol{W_1})); \boldsymbol{W_C})); \boldsymbol{W_2}), \tag{3}$$

其中，L_1 和 L_2 分别是图 3 (c) 中的线性层1和线性层2，P和 C分别是置换和一维卷积。$\boldsymbol{W_1} \in \mathbb{R}^{D\times 128}$ 和 $\boldsymbol{W_2} \in \mathbb{R}^{128\times D}$ 分别是L_1 和 L_2的可训练权重。$\boldsymbol{W_C}$ 是一维卷积权重，其卷积核大小为 3。L_1 将输入 $\boldsymbol{X} \in \mathbb{R}^{L\times D}$ 投影至 $\boldsymbol{X} \in \mathbb{R}^{L\times 128}$，置换层 P 将 $\boldsymbol{X} \in \mathbb{R}^{L\times 128}$ 移至 $\boldsymbol{X} \in \mathbb{R}^{128\times L}$，而 C 沿时间维度应用，以捕捉时间动态。P层和 L_2 随后将$\boldsymbol{X} \in \mathbb{R}^{128\times L}$ 投影至 $\boldsymbol{X} \in \mathbb{R}^{L\times D}$。

具体而言，对于 F_V，多头自注意力和 UT- 适配器的输出可表示为：

$$\boldsymbol{X_v^{''}} = AN(\boldsymbol{X_v} + H(LN(\boldsymbol{X_v})) + U(\boldsymbol{X_v})), \tag{4}$$

和 $\boldsymbol{X_v^{'}}$ 在公式 1 中由以下方式得到：

$$\boldsymbol{X_v^{'}} = AN(\boldsymbol{X_v^{''}} + MLP(LN(\boldsymbol{X_v^{''}})) + U(\boldsymbol{X_v^{''}})), \tag{5}$$

其中，AN 是 UT- 适配器之后引入的归一化层，该层的参数可训练。

G_A 与 F_V具有相似的结构。输出 $\boldsymbol{X_a}$的计算方式为

$$\begin{aligned}\boldsymbol{X_a^{''}} &= AN(\boldsymbol{X_a} + LN(H(\boldsymbol{X_a})) + U(\boldsymbol{X_a})),\\ \boldsymbol{X_a^{'}} &= AN(\boldsymbol{X_a^{''}} + LN(MLP(\boldsymbol{X_a^{''}})) + U(\boldsymbol{X_a^{''}})).\end{aligned} \tag{6}$$

输入 $\boldsymbol{X_v}$ 和 $\boldsymbol{X_a}$ 是来自视觉和音频的序列嵌入。Transformer 编码器中的 MHSA 模块能够实现全局注意力，使模型有效地从输入中学习空间和时间注意力。然而，这种方式可能并不利于学习局部时间注意力和局部空间信息。因此，我们提出 UT-Adapter，与 MHSA 和 MLP 模块并行捕获局部时间注意力。二维卷积神经网络捕获局部空间信息，一维卷积神经网络捕获时间信息。实验表明，这种适配器-*Transformer* 架构能够以参数高效的方式进行微调，并取得更好的性能。

4.2. 即插即用音视频融合（PAVF）

为探索音视交互，我们提出了一个即插即用音视频融合（PAVF）模块，用于学习融合所需的跨模态注意力机制。

我们将来自 V_i A_i $\boldsymbol{X_v^i} \in$ and $\mathbb{R}^{L\times D}$ 的输出特征记为 和 $\boldsymbol{X_a^i} \in \mathbb{R}^{L\times D}$ and，并提出在低维嵌入空间中学习跨模态注意力，这样效率更高且计算成本更低。$\boldsymbol{X_v^i}$ and $\boldsymbol{X_a^i}$ 首先通过线性投影层投影到 $\boldsymbol{X_v^i} \in \mathbb{R}^{L\times D^{'}}$ and$\boldsymbol{X_a^i} \in \mathbb{R}^{L\times D^{'}}$, $D^{'} < D$。学习跨模态相关性 $\boldsymbol{P^i} \in \mathbb{R}^{L\times L}$，以表示视觉序列与音频序列之间的重要性。在一个视频片段中，某个视觉序列可能与特定音频序列具有更高的相关性，这为欺骗检测任务提供了重要线索。为计算$\boldsymbol{P^i}$，我们引入一个可训练的跨模态相关性权重矩阵 $\boldsymbol{W_P^i} \in \mathbb{R}^{\boldsymbol{D^{'}}\times \boldsymbol{D^{'}}}$，$\boldsymbol{P^i}$的计算方式为

$$\boldsymbol{P^i} = \boldsymbol{X_a^i} \boldsymbol{W_P^i} \boldsymbol{X_v^{i\intercal}}. \tag{7}$$

我们进一步利用学习到的跨模态相关性 $\boldsymbol{P^i}$，执行 $\boldsymbol{X_v^i} \rightarrow \boldsymbol{X_a^i}$ 和 $\boldsymbol{X_a^i} \rightarrow \boldsymbol{X_v^i}$ 注意力机制。经过注意力机制处理的特征为

$$\begin{aligned} \tilde{\boldsymbol{X}}_{\boldsymbol{v}}^{\boldsymbol{i}} &= Softmax(\boldsymbol{P^i}) \boldsymbol{X_v^i} + \boldsymbol{X_v^i}, \\ \tilde{\boldsymbol{X}}_{\boldsymbol{a}}^{\boldsymbol{i}} &= Softmax(\boldsymbol{P^{i\intercal}}) \boldsymbol{X_a^i} + \boldsymbol{X_a^i}, \end{aligned} \tag{8}$$

其中，$\boldsymbol{P^{i\intercal}}$是 $\boldsymbol{P^i}$、$\tilde{\boldsymbol{X}}_{\boldsymbol{v}}^{\boldsymbol{i}} \in \mathbb{R}^{L\times D^{'}}$和$\tilde{\boldsymbol{X}}_{\boldsymbol{a}}^{\boldsymbol{i}} \in \mathbb{R}^{L\times D^{'}}$的转置。Softmax 按列计算。将经过注意力机制处理的特征拼接后输入融合头。融合头由线性投影L_p、归一化层 LN和非线性激活函数 $ReLU$组成，其中线性投影将嵌入维度进一步缩减至 D''。它可以表示为

$$\tilde{\boldsymbol{X}}_{\boldsymbol{va}}^{\boldsymbol{i}} = ReLU(LN(L_p(\tilde{\boldsymbol{X}}_{\boldsymbol{v}}^{\boldsymbol{i}} \oplus \tilde{\boldsymbol{X}}_{\boldsymbol{a}}^{\boldsymbol{i}}))), \tag{9}$$

其中，$\tilde{\boldsymbol{X}}_{\boldsymbol{va}}^{\boldsymbol{i}} \in \mathbb{R}^{L\times D^{''}}$ 和 $\oplus$ 表示拼接操作。

PAVF 模块被插入到每对 V_i 和 A_i中以进行融合，从而能够学习到更丰富的融合信息。随后，将每个 PAVF 模块的输出 $\tilde{\boldsymbol{X}}_{\boldsymbol{va}}^{\boldsymbol{i}}$ 拼接起来用于预测。

4.3. 多任务学习

多任务学习已在各种视听任务中展现出良好的性能 [19, 45], ，通过相互学习有效提升多个任务的表现。借助 DOLOS 的 MUMIN特征（图 2 (b)），我们通过多任务学习提升性能，同时使用融合的多模态特征执行两项预测任务。具体而言，对于 K 个 MUMIN特征标签和一个欺骗-真实标签，我们预测 $K+1$ 个标签。请注意，所有标签均为二元标签。我们将˜融合音视嵌入记作 $\boldsymbol{X_{va}}$。分类˜层以 $\boldsymbol{X_{va}}$ 为输入，生成预测分数 $S \in \mathbb{R}^{1\times(K+1)}$。给定真实标签 $Y \in \mathbb{R}^{1\times(K+1)}$，我们使用交叉熵损失进行多任务学习。其形式化表示如下：

$$\mathcal{L}_M = -\sum_{k=1}^{K+1} \left(Y_k \log(S_k) + (1-Y_k)\log(1-S_k)\right), \tag{10}$$

其中，$\mathcal{L}_M$ 是 S 和 Y中各元素交叉熵损失的总和。

5. 实验结果

5.1. 实现细节

数据预处理。对于每个视频片段，我们均匀采样了 $L = 64$ 张图像，并使用 MTCNN 人脸检测器裁剪出面部区域 [53]。这些图像经过归一化处理，并调整至 160×160 像素。对原始语音音频进行重采样，使 W2V2 特征提取器输出 $L = 64$ 个词元。两种模态均未进行数据增强。

模型细节。我们使用在 ImageNet 上预训练的 ViT 作为视觉模态的骨干网络。我们使用二维卷积模块对面部图像进行词元化，得到维度为 64×256的特征。对于音频模态，我们采用了预训练的 W2V2模型。原始音频由一维卷积模块进行词元化，每个音频样本的特征大小为 64×512 。通过线性投影层，将视觉和音频词元投影到 64×768维空间。我们根据经验使用 ViT 和 W2V2模型的前 4 个 Transformer编码器层提取特征，因为欺骗检测可以从低级特征中获益 [44, 29, 13]。UT-Adapter 和 PAVF 模块被插入这四个编码器层中（见图 3(a)）。视觉和音频编码器中 UT-Adapter 的一维卷积层（见图3 (c)）的卷积核大小为 3，步幅为 1。

训练与评估。模型采用交叉熵损失进行训练。我们使用Adam优化器，以3e-4的学习率和16的批次大小训练模型20轮。我们使用准确率（ACC）、F1分数（F1）和曲线下面积（AUC）评估了所提出的方法。

模态	三折平均			时长协议			性别协议		
	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
视觉	61.44	69.42	58.89	61.03	**72.01**	56.51	59.37	64.19	54.94
音频	59.19	**73.46**	52.54	58.24	71.83	52.38	52.62	63.22	51.08
拼接	61.62	70.2	60.5	60.8	69.63	57.79	58.0	66.22	53.75
融合（PAVF）	64.75	71.2	62.71	62.43	70.04	59.92	58.28	65.41	53.31
PAVF +多任务	**66.84**	73.35	**64.58**	**64.48**	71.09	**62.44**	**59.04**	**66.84**	**55.1**

Tab表 2. DOLOS 数据集上的多模态欺骗检测。指标为准确率（%）、F1分数（%）和曲线下面积（% ）.

	方法	训练数据 - DOLOS		训练数据 - 谎言盒子	
		BgOL 测试	RLT 测试	BgOL 测试	RLT 测试
V	RN18+长短期记忆网络	55.08 (+4.62)	56.62 (+3.85)	50.46	52.77
	三维残差网络	55.69 (+3.38)	**57.87** (+**5.53**)	52.31	52.34
	我们的方法（视觉）	**56.92** (+4.30)	55.32 (+2.55)	52.62	52.77
A	RN18 †	**55.84** (+**4.91**)	53.53 (+1.64)	50.93	51.89
	多层感知机（梅尔频率倒谱系数）	52.31 (+2.46)	52.34 (+0.42)	49.85	51.92
	我们的方法（音频） ‡	55.39 (+4.00)	**59.15** (+**4.68**)	51.39	54.47
V+A	三维残差网络+RN18	50.77 (+1.23)	54.47 (+1.70)	49.54	52.77
	我们的方法 (PAVF)	**57.54** (+**4.31**)	**56.17** (+**2.98**)	53.23	53.19

表格 3. 谎言之袋 (BgOL) 和真实生活试验 (RLT) 数据集上的交叉测试。度量指标为准确率 (%)。RN 表示 ResNet。所有视觉特征均来自面部帧。 † 使用梅尔频谱图进行训练， ‡ 使用原始音频进行训练。

	特征	方法	准确率（%）
V	Open Face [33, 29]	LSTM	56.81
	面部动作单元 [33, 4]	LSTM	57.47
	面部情感 [33, 24]	LSTM	57.67
	MUMIN特征 [36, 37]	MLP	58.84
	面部 [25, 9, 51, 13, 29, 24]	RN18+长短期记忆网络	59.55
	Face	我们的方法（视觉）	**61.44**
A	梅尔频率倒谱系数 [51]	MLP	56.86
	Open SMILE [33, 13, 52]	MLP	57.49
	W2V2 音频 [25, 9, 29, 24]	我们的方法（音频）	**59.19**
V+A	Open Face +Open SMILE [33]	分数融合	59.75
	面部 +Open SMILE [13, 29, 52]	分数融合	60.00
	面部 +W2V2 音频 [25, 24, 9]	我们的方法（拼接）	**61.62**
	面部 +W2V2 音频	我们的方法（PAVI	**64.75**

表格 4. 在视觉和音频特征上对 DOLOS 数据集进行基准测试。分数融合采用加权平均。RN18 代表 ResNet-18。

5.2. DOLOS 中的视听欺骗检测

PECL模型在DOLOS上的性能如表2所示。平均结果根据训练-测试协议中定义的3折计算得出。时长和性别协议的结果分别是训练/测试时长组合为（长/短 + 短/长）和训练/测试性别组合为（男性/女性 +女性/男性）所得结果的交叉测试平均值。对于单模态结果，在所有实验方案中，视觉模态的表现都略优于音频模态。为了研究简单特征级融合的效果，我们将视觉编码器和音频编码器的最后一层输出嵌入拼接起来进行预测。拼接通常只能达到单模态预测器的平均水平。为解决这一问题，所提出的PAVF模块从不同编码器层中学习视觉和音频特征之间的相关性，从而提升了性能。多任务学习进一步提高了所有实验方案下的准确率。然而，通过比较整体性能，我们发现时长和性别因素会对性能产生不利影响。这些结果凸显了时长和性别因素在欺骗检测中的重要性。

5.3. 与谎言盒子的比较

为了将 DOLOS 与现有游戏节目数据集 BOL 进行比较，我们进行了交叉测试：使用 DOLOS 和 BOL 进行训练，并使用其他欺骗检测数据集，即谎言之袋和真实生活试验，进行测试。这项任务颇具挑战，因为这些数据集采集自不同的欺骗场景。如表3所示，我们比较了若干音频、视觉和融合方法，发现使用 DOLOS 训练的模型总体上准确率更高。使用 DOLOS 训练的 PECL模型的测试准确率也高于使用 BOL 训练的模型（例如，融合（ PAVF）在 BgOL 和 RLT 上分别提升了 4.31% 和 2.98%）。这些结果表明，与使用 BOL 训练的模型相比，使用 DOLOS 训练的模型能够更有效地识别其他场景中的欺骗性内容。

5.4. DOLOS 基准测试

我们按照欺骗检测领域的研究惯例，在表格 4中对 DOLOS 进行基准测试。值得注意的是，现有欺骗检测研究在实验方案、数据处理和特征提取方法方面尚未标准化。为涵盖并突出欺骗检测领域最具代表性的研究，我们使用此前用于欺骗检测的常见视觉和音频特征，对 DOLOS 进行了基准测试。OpenFace 特征（二维面部标志点、眼动方向和头部姿态）以及动作单元（AU）特征均使用 OpenFace 工具包提取 [3]。面部情感（情绪）特征则使用 AffectNet [34] 模型提取。对于音频，梅尔倒谱系数（MFCC）和 openSMILE 特征由 openSMILE 工具包提取 [11]。

性能。对于视觉模态，RGB人脸图像的表现最佳，因为其中包含更多欺骗线索。其次是人工标注的MUMIN特征、情感特征和AU特征，它们都体现了高层次的人类表情。低层次的面部标志点特征表现不佳。对于音频模态，包含丰富欺骗线索的原始语音音频表现最佳。openSMILE特征的表现优于梅尔频率倒谱系数，因为它提取的音频特征数量几乎是梅尔频率倒谱系数的三倍

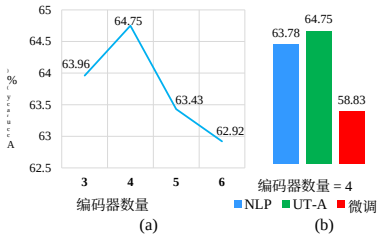


图4. 编码器深度和适配器类型的消融实验。

方法	可训练参数 适配器	PAVF	总计	总计	可训练参数/总参数
微调	–	1.982	70.180	70.180	100%
自然语言处理 [17]	5.345	1.982	7.327	73.347	9.99%
UT-A	3.077	1.982	**5.059**	71.079	**7.12%**

表5. 可训练参数数量与总参数数量的比较（百万）。

位置	∥ 多头自注意力 ∥前馈网络	∥ 多头自注意力	∥ 前馈网络	多头自注意力 △前馈网络
准确率 (%)	**64.75**	63.88	63.37	64.39

表6. UT-Adapter 位置的消融研究。 ∥ 表示“并行”， △ 表示“位于.....之间”。

特征。对两种模态中表现最佳的特征进行了分数级融合。所提出的基于Transformer的检测器在单模态和多模态设置下均取得了最佳表现。

5.5. 消融实验

基于训练-测试协议，我们针对 PECL模型的不同组件开展了消融研究。

编码器深度和适配器类型的影响。图4 (a) 表明，所提出的 PECL 在使用四层时表现最佳，增加层数后可能会过拟合。图4 (b) 显示，UT-Adapter 的表现优于 NLP 适配器 [17] 和全量微调，这可能是因为 UT-Adapter 中的一维卷积层能够实现更好的时序建模。

参数量对比。表格5 表明，与 NLP 适配器和全量微调方法相比，我们提出的方法以最少的可训练参数取得了最佳表现。请注意，UT-Adapter 在维度为 128 时表现最佳，而 NLP 适配器在维度为 256 时表现最佳。

适配器位置的影响。表格6 中的结果 [21], 表明，当 UT-Adapter 与 MHSA 层和 FFN 层并行时，效果最佳，因为这两者都能捕捉时间注意力。仅与其中任一层并行时，性能会下降。将 UT-Adapter 放置在 MHSA 层与 FFN 层之间时，性能也较为接近。

PAVF模块的影响。图中的 PAVF 模块数量 5 对应于与这些模块配合运行的音视频编码器的最后一个编号。早期阶段融合的优势 [35] 显而易见：使用四个 PAVF 模块时，性能大幅提升，此时

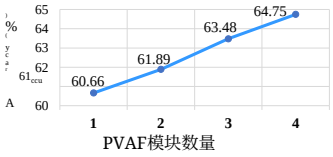


图 5. PAVF 模块数量的消融实验。

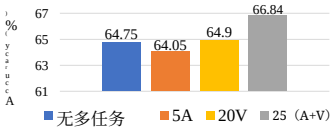


图 6. 使用不同特征进行多任务学习的结果。

捕捉到了低层次的跨模态交互。结果表明，四个 PAVF 模块尤其有助于捕捉音视频欺骗线索。

多任务学习的影响。在图6中，我们展示了第4.3节所述多任务学习的消融研究，涵盖不同的音视频特征（参见图2(b)）。结果表明，使用 20 个视觉特征（20V）比使用 5 个音频特征（5A）效果更好，而使用 25 个音视频特征则进一步提升了准确率。这些结果证明，DO-LOS 中人工标注的特征有助于提升欺骗检测性能。

6. 结论

本文介绍了 DOLOS，这是规模最大的游戏节目欺骗数据集，包含细粒度的音视频特征标注。我们还提出了参数高效跨模态学习，其中 UT-Adapter 学习视觉和音频模态的时序注意力机制，PAVF模块则捕获音频与视觉模态之间的相关性信息。我们使用现有的欺骗检测方法对数据集进行了基准测试。大量实验表明，DOLOS 的质量优于 Box of Lies，且 PECL模型在 DOLOS 上取得了更好的结果，也可应用于其他欺骗性内容。

伦理考量。欺骗检测研究对于可信度评估、反欺诈、边境安全等至关重要。然而，滥用欺骗检测可能造成隐私侵犯、歧视和／或有害监控等潜在负面影响。本研究在获得伦理审查批准（IRB-2022-901）后开展。数据依据合理使用政策从 YouTube 收集，受试者均为公众人物。

致谢。 本研究在南洋理工大学电子与电气工程学院的快速丰富物体搜索（ROSE）实验室完成。本研究部分由南洋理工大学—北京大学联合研究院（由南洋理工大学与北京大学合作成立，并获黄廷方慈善基金捐赠资助）以及新加坡 DSO 国家实验室通过项目协议 No. DSOCL21238 资助。

参考文献

[1] Mohamed Abouelenien、Ver´onica P´erez-Rosas、Rada Mihalcea 和 Mihai Burzo。通过整合多种模态的判别特征检测欺骗行为。《IEEE 信息取证与安全汇刊》，12(5):1042–1055，2016。1

[2] Jens Allwood、Loredana Cerrato、Laila Dybkjaer、Kristiina Jokinen、Costanza Navarretta 和 Patrizia Paggio。mumin 多模态编码方案。NorFA 年鉴，2005:129–157，2005。3

[3] Brandon Amos、Bartosz Ludwiczuk、Mahadev Satyanarayanan 等。OpenFace：一款通用人脸识别库，支持移动应用。卡内基梅隆大学计算机科学学院，6(2):20，2016。7

[4] Danilo Avola、Luigi Cinque、Gian Luca Foresti 和 Daniele Pannone。使用面部动作单元自动检测 RGB 视频中的欺骗行为。载于《第13届国际分布式智能摄像机会议论文集》，第1–6页，2019。3, 7

[5] Alexei Baevski、Yuhao Zhou、Abdelrahman Mohamed 和 Michael Auli。wav2vec 2.0：语音表征自监督学习框架。《神经信息处理系统进展》，33:12449–12460，2020。1, 2, 5

[6] Mousumi Banerjee、Michelle Capozzoli、Laura McSweeney 和 Debajyoti Sinha。超越卡帕系数：评估者间一致性度量综述。《加拿大统计学杂志》，27(1):3–23，1999。3

[7] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。语言模型是少样本学习者。《神经信息处理系统进展》，33:1877–1901，2020。1

[8] Bella M DePaulo、James J Lindsay、Brian E Malone、Laura Muhlenbruck、Kelly Charlton 和 Harris Cooper。欺骗线索。《心理学公报》，129(1):74，2003。1, 2

[9] Mingyu Ding、An Zhao、Zhiwu Lu、Tao Xiang 和 Ji-Rong Wen。用于视频欺骗检测的面部聚焦跨流网络。载于《IEEE/CVF 计算机视觉与模式识别会议论文集》，第7802–7811页，2019。1, 3, 7

[10] Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly、Jakob Uszkoreit 和 Neil Houlsby。一张图像胜过 16×16 个词：用于大规模图像识别的 Transformer。国际学习表征会议，第0–7页，2021。1, 2, 5

[11] Florian Eyben、Martin W¨ollmer 和 Bj¨orn Schuller。Opensmile：慕尼黑通用快速开源音频特征提取器。载于《第18届 ACM 多媒体国际会议论文集》，第1459–1462页，2010。7

[12] Tommaso Fornaciari 和 Massimo Poesio。意大利法院案件中的自动欺骗检测。《人工智能与法律》，21(3):303–340，2013。1

[13] Mandar Gogate、Ahsan Adeel 和 Amir Hussain。由深度学习驱动的多模态融合，用于自动欺骗检测。载于《2017 IEEE计算智能研讨会系列（SSCI）》，第1–6页。IEEE，2017。1, 3, 6, 7

[14] Martin Graciarena、Elizabeth Shriberg、Andreas Stolcke、Frank Enos、Julia Hirschberg 和 Sachin Kajarekar。结合韵律、词汇和倒谱系统进行欺骗性语音检测。载于《2006 IEEE国际声学、语音与信号处理会议论文集》，第1卷，第0–7页。IEEE，2006。1

[15] Viresh Gupta、Mohit Agarwal、Manik Arora、Tanmoy Chakraborty、Richa Singh 和 Mayank Vatsa。谎言之袋：用于欺骗检测的多模态数据集。载于《IEEE/CVF计算机视觉与模式识别会议研讨会论文集》，第0–7页，2019。1, 2, 4

[16] Julia Bell Hirschberg、Stefan Benus、Jason M Brenier、Frank Enos、Sarah Friedman、Sarah Gilman、Cynthia Girand、Martin Graciarena、Andreas Kathol、Laura Michaelis 等。区分欺骗性与非欺骗性语音。2005。2

[17] Neil Houlsby、Andrei Giurgiu、Stanislaw Jastrzebski、Bruna Morrone、Quentin De Laroussilhe、Andrea Gesmundo、Mona Attariyan 和 Sylvain Gelly。自然语言处理的参数高效迁移学习。载于《国际机器学习会议》，第2790–2799页。PMLR，2019。2, 5, 8

[18] Wei-Ning Hsu、Benjamin Bolte、Yao-Hung Hubert Tsai、Kushal Lakhotia、Ruslan Salakhutdinov 和 Abdelrahman Mohamed。HuBERT：通过隐藏单元掩码预测进行自监督语音表征学习。IEEE/ACM《音频、语音与语言处理汇刊》，29:3451–3460，2021。1

[19] Ronghang Hu 和 Amanpreet Singh。UniT：基于统一型Transformer的多模态多任务学习。载于《IEEE/CVF国际计算机视觉会议论文集》，第1439–1449页，2021。6

[20] Chao Jia、Yinfei Yang、Ye Xia、Yi-Ting Chen、Zarana Parekh、Hieu Pham、Quoc Le、Yun-Hsuan Sung、Zhen Li 和 Tom Duerig。利用噪声文本监督扩展视觉及视觉语言表征学习。载于《国际机器学习会议》，第4904–4916页。PMLR，2021。2

[21] Shibo Jie 和 Zhi-Hong Deng。卷积旁路是更好的视觉Transformer适配器。arXiv预印本 arXiv:2207.07039，2022。2, 5, 8

[22] John Jumper、Richard Evans、Alexander Pritzel、Tim Green、Michael Figurnov、Olaf Ronneberger、Kathryn Tunyasuvunakool、Russ Bates、Augustin ˇZ´ıdek、Anna Potapenko 等。利用AlphaFold进行高精度蛋白质结构预测。《自然》，596(7873):583–589，2021。1

[23] Manvi Kamboj、Christian Hessler、Priyanka Asnani、Kais Ri-ani 和 Mohamed Abouelenien。多模态政治欺骗检测。IEEE MultiMedia，28(1):94–102，2020。2, 4

[24] Hamid Karimi、Jiliang Tang 和 Yanen Li。迈向视频中的端到端欺骗检测。载于《2018 IEEE国际大数据会议（Big Data）》，第1278–1283页。IEEE，2018。3, 7

[25] Mohan Karnati、Ayan Seal、Anis Yazidi 和 Ondrej Krejcar。LieNet：用于检测欺骗的深度卷积神经网络框架。IEEE认知与发展系统汇刊，2021。3, 7

[26] Alexander Kolesnikov、Lucas Beyer、Xiaohua Zhai、Joan Puigcerver、Jessica Yung、Sylvain Gelly 和 Neil Houlsby。大规模迁移（BiT）：通用视觉表征学习。载于《计算机视觉—ECCV 2020：第16届欧洲计算机视觉会议，英国格拉斯哥，2020年8月23—28日，论文集，第V部分16》，第491–507页。施普林格，2020。2[27] Chenqi Kong、Kexin Zheng、Shiqi Wang、Anderson Rocha 和 Haoliang Li。超越像素世界：一种新颖的面向智能手机的基于声学的人脸防伪系统。IEEE 信息取证与安全汇刊，17:3238–3253，2022。1[28] Qiuqiang Kong、Yin Cao、Turab Iqbal、Yuxuan Wang、Wenwu Wang 和 Mark D Plumbley。PANNs：用于音频模式识别的大规模预训练音频神经网络。IEEE/ACM《音频、语音与语言处理汇刊》，28:2880–2894，2020。1[29] Gangeshwar Krishnamurthy、Navonil Majumder、Soujanya Poria 和 Erik Cambria。一种多模态欺骗检测的深度学习方法。arXiv预印本 arXiv:1803.00344，2018。6, 7[30] Timothy R Levine 和 Steven A McCornack。欺骗理论化。《语言与社会心理学杂志》，33(4):431–440，2014。2[31] Ze Liu、Yutong Lin、Yue Cao、Han Hu、Yixuan Wei、Zheng Zhang、Stephen Lin 和 Baining Guo。Swin Transformer：采用移位窗口的分层视觉变换器。载于《IEEE/CVF国际计算机视觉会议论文集》，第10012–10022页，2021。1[32] E Paige Lloyd、Jason C Deska、Kurt Hugenberg、Allen R Mc-Connell、Brandon T Humphrey 和 Jonathan W Kunstman。迈阿密大学欺骗检测数据库。《行为研究方法》，51(1):429–439，2019。2, 4[33] Leena Mathur 和 Maja J Matari´c。将面部情感表征引入自动化多模态欺骗检测。载于《2020年国际多模态交互会议论文集》，第305–314页，2020。3, 7[34] Ali Mollahosseini、Behzad Hasani 和 Mohammad H Ma-hoor。AffectNet：用于自然场景下面部表情、效价和唤醒度计算的数据库。IEEE《情感计算汇刊》，10(1):18–31，2017。7[35] Arsha Nagrani、Shan Yang、Anurag Arnab、Aren Jansen、Cordelia Schmid 和 Chen Sun。用于多模态融合的注意力瓶颈。《神经信息处理系统进展》，34:14200–14213，2021。8[36] Ver´onica P´erez-Rosas、Mohamed Abouelenien、Rada Mihal-cea 和 Mihai Burzo。利用真实案件审判数据进行欺骗检测。载于《2015年ACM国际多模态交互会议论文集》，第59–66页，2015。1, 2, 4, 7[37] Ver´onica P´erez-Rosas、Mohamed Abouelenien、Rada Mihal-cea、Yao Xiao、CJ Linton 和 Mihai Burzo。真实场景欺骗检测中的言语与非言语线索。载于《2015年自然语言处理经验方法会议论文集》，第2336–2346页，2015。2, 4, 7[38]Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。学习——从自然语言监督中学习可迁移的视觉模型。在国际机器学习会议上发表，页码 8748–8763。PMLR，2021。1[39]Alec Radford、Karthik Narasimhan、Tim Salimans、Ilya Sutskever 等。通过生成式预训练提升语言理解能力。https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf，2018。2[40] David Silver、Aja Huang、Chris J Maddison、Arthur Guez、Laurent Sifre、George Van Den Driessche、Julian Schrittwieser、Ioannis Antonoglou、Veda Panneershelvam、Marc Lanctot 等。利用深度神经网络和树搜索掌握围棋。《自然》，529(7587):484–489，2016。1[41] David Silver、Thomas Hubert、Julian Schrittwieser、Ioannis Antonoglou、Matthew Lai、Arthur Guez、Marc Lanctot、Laurent Sifre、Dharshan Kumaran、Thore Graepel 等。通过自我对弈掌握国际象棋、将棋和围棋的通用强化学习算法。《科学》，362(6419):1140–1144，2018。1[42] Felix Soldner、Ver´onica P´erez-Rosas 和 Rada Mihalcea。谎言盒子：对话中的多模态欺骗检测。载于《2019年计算语言学协会北美分会会议论文集：人类语言技术》，第1卷（长篇与短篇论文），页码 1768–1777，2019。1, 2, 4[43] Jeremy Speth、Nathan Vance、Adam Czajka、Kevin W Bowyer、Diane Wright 和 Patrick Flynn。欺骗检测与远程生理监测：数据集及基线实验结果。载于2021年IEEE国际联合生物识别会议（IJCB），页码 1–8。IEEE，2021。2, 4[44] Anastasis Stathopoulos、Ligong Han、Norah Dunbar、Judee K Burgoon 和 Dimitris Metaxas。利用鲁棒面部特征进行视频欺骗检测。载于《未来技术会议论文集》，页码 668–682。施普林格，2020。1, 6[45] Fei Tao 和 Carlos Busso。基于多任务学习的端到端视听语音识别系统。《IEEE多媒体汇刊》，23:1–11，2020。2, 6[46] Ilya O Tolstikhin、Neil Houlsby、Alexander Kolesnikov、Lucas Beyer、Xiaohua Zhai、Thomas Unterthiner、Jessica Yung、Andreas Steiner、Daniel Keysers、Jakob Uszkoreit 等。MLP-Mixer：用于视觉的全多层感知机架构。《神经信息处理系统进展》，34 24272，2021 24261 1 : – 。[47] Nik Vaessen 和 David A Van Leeuwen。微调 wav2vec2 以进行说话人识别。载于ICASSP 2022-2022 IEEE国际声学、语音与信号处理会议（ICASSP），页码 7967–7971。IEEE，2022。2[48] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。注意力机制即一切所需。《神经信息处理系统进展》，30，2017。1

[49] Aldert Vrij 和 P¨ar Anders Granhag。引出欺骗与真实的线索：关键在于所提问题。《记忆与认知应用研究杂志》，1(2):110–117，2012。1，2[50] Gemma Warren、Elizabeth Schertler 和 Peter Bull。从情绪线索和非情绪线索中检测欺骗。《非语言行为杂志》，33(1):59–69，2009。2[51] Zhe Wu、Bharat Singh、Larry Davis 和 V Subrahmanian。视频中的欺骗检测。载于*AAAI*人工智能会议论文集，第32卷，2018。3，7[52] Jun-Teng Yang、Guei-Ming Liu 和 Scott C-H Huang。通过分析情绪状态特征进行视频中的多模态欺骗检测。*arXiv*预印本 *arXiv:2104 08373*，20217。。[53] Kaipeng Zhang、Zhanpeng Zhang、Zhifeng Li 和 Yu Qiao。使用多任务级联卷积网络进行联合人脸检测与对齐。*IEEE*信号处理快报，23(10):1499–1503，2016。6[54] Zixing Zhang、Bingwen Wu 和 Bj¨orn Schuller。通过注意力增强的端到端多任务学习，从语音中预测情绪。载于*ICASSP 2019-2019 IEEE*国际声学、语音与信号处理会议（*ICASSP*），页码 6705–6709。IEEE，2019。2