HierCon：用于音频深度伪造检测的层次对比注意力

Zhili Nicholas Liang
nnliang@student.unimelb.edu.au 墨尔本大学 墨尔本，维多利亚州 澳大利亚

Soyeon Caren Han
care
n.han@unimelb.edu.au 墨尔本大学 墨尔本，维多利亚州 澳大利亚

Qizhou Wang
mi
ke.wang@unimelb.edu.au 墨尔本大学 墨尔本，维多利亚州 澳大利亚

Christopher Leckie
caleckie
@unimelb.edu.au 墨尔本大学 墨尔本，维多利亚州 澳大利亚

arXiv:2602.01032v1 [cs.SD] 1 Feb 2026

摘要

现代文本转语音和语音转换系统生成的音频深度伪造越来越难以与真实语音区分，对安全和在线信任构成了严重风险。尽管最先进的自监督模型能够提供丰富的多层表征，现有检测器却将各层独立处理，忽略了识别合成伪影所必需的时间和层次依赖关系。我们提出 HierCon，这是一种层次化层注意力框架，结合基于间隔的对比学习，对时间帧、相邻层和层组之间的依赖关系进行建模，同时促进域不变嵌入。在 ASVspoof 2021 DF 和 In-the-Wild 数据集上的评估结果表明，我们的方法达到了最先进的性能（等错误率分别为 1.93% 和 6.87%），相比独立层加权，性能分别提升了 36.6% 和 22.5%。实验结果和注意力可视化证实，层次建模增强了模型对跨域生成技术和录音条件的泛化能力。源代码见 https://github.com/adlnlp/HierCon。

CCS 概念

• **计算方法** → **语音识别**。

关键词

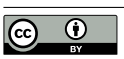
音频深度伪造检测、反欺骗、自监督学习

ACM 参考格式：
Zhili Nicholas Liang、Soyeon Caren Han、Qizhou Wang 和 Christopher Leckie。2026。HierCon：用于音频深度伪造检测的层次对比注意力。收录于 *ACM Web Conference 2026* 论文集（*WWW '26*），*2026*年*4*月*13*日至*17*日，迪拜，阿拉伯联合酋长国。ACM，纽约，美国纽约州，5 页。
https://doi.org/10.1145/3774904.3792873

1 引言

文本转语音（TTS）和语音转换（VC）系统的最新进展使合成语音愈发逼真，让攻击者能够生成与真实人类语音极为相似的语音 [7]。这些进步给语音认证、数字取证和在线通信带来了重大的安全与信任问题 [4]。早期的深度伪造检测依赖手工设计的频谱或倒谱特征 [11], ，但该领域迅速转向深度学习架构 [4, 8] 和自监督学习（SSL）模型 [1, 3], ，这些方法能够提供更丰富的多层语音表征。在这些方法中，基于 XLS-R 的方法取得了出色的效果；例如，敏感层选择（SLS）通过学习 Transformer 层之间的标量权重，在 ASVspoof 2021 DF 上实现了 2.09% 的等错误率 [13]。然而，SLS 等方法将每个 Transformer 层独立处理，忽略了两个重要的判别信息来源。第一，它们忽略了层内的时间动态，使每一帧都获得相同权重，尽管只有某些片段包含合成伪迹。第二，它们没有对层间依赖关系建模——浅层编码声学线索，而深层捕捉语义或韵律信号；两者的交互往往是检测现代深度伪造的关键。忽略这些依赖关系可能导致特征同质化，并降低对跨域生成流程的泛化能力 [12]。

为解决这些局限，我们提出一种结合基于间隔的对比学习的分层层注意力框架。我们的方法显式建模三个层级的依赖关系：每一层内的时间注意力、相邻层之间的组内注意力，以及较大层簇之间的组间注意力。我们引入对比正则化，以促使嵌入几何具有域不变性，从而更好地泛化到多种深度伪造生成方法。在 ASVspoof 2021 DF 和 In-the-Wild 基准测试上的评估结果表明，我们的方法达到了最先进的性能，将等错误率分别降至 1.93% 和 6.87%，相较于独立层加权，分别实现了 36.6% 和 22.5% 的相对提升。我们的贡献总结如下：**1）分层层注意力框架**：我们提出了一种三阶段注意力架构，用于建模时间、层内和层间依赖关系，使模型能够利用整个表征层级中的互补线索，而非将各层独立处理。**2）用于域不变学习的对比正则化**：我们在微调过程中引入基于间隔的对比学习，对层表征施加明确的几何约束。不同于以往基于自监督学习的检测器 [12, 13], ，我们的方法联合优化分类和对比目标，从而提升跨域鲁棒性，并避免依赖数据集特定伪迹。**3）可解释性与稳健泛化**：我们的层次注意力结构能够自然地揭示哪些时间区域和层组对预测起主导作用。在 ASVspoof 2021 DF 和 In-the-Wild 上的实验表明，该方法具有很强的泛化能力，并优于以往的层融合方法。



本作品采用知识共享署名 4.0 国际许可协议进行许可。*WWW '26,* 迪拜, 阿拉伯联合酋长国© 2026 版权归所有者/作者所有。ACM ISBN 979-8-4007-2307-0/2026/04 https://doi.org/10.1145/3774904.3792873

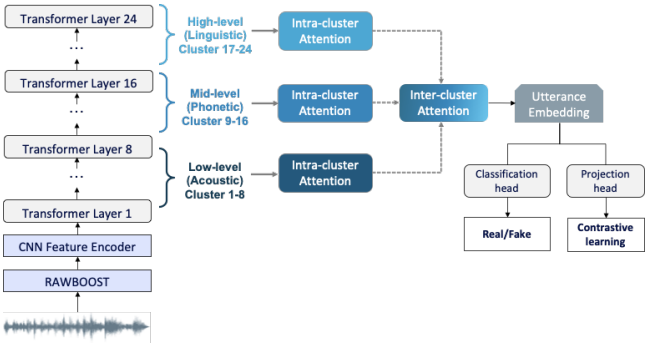


图 1：所提出的 HierCon 整体架构——一种用于可靠音频深度伪造检测的层次化对比注意力机制。

2 提出的方法

2.1 预备知识

基于自监督学习的音频深度伪造检测。一类广受欢迎的检测方法利用预训练的自监督学习音频模型，并展现出强劲的性能。其中，XLS-R 是最常见的选择。它是一种基于 Wav2Vec 2.0 构建的大规模跨语言表征模型，在 128 种语言的数据上训练，以捕捉通用的声学和语言特征。为编码原始波形，它首先按预定义间隔提取帧级表征，再由一个 L 层 Transformer 网络处理，以获得上下文化的隐藏状态 $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots\ldots, \mathbf{h}_L]$。

利用多层表征。现有基于自监督学习的检测器通常会利用所有隐藏状态，以捕获跨多个层级、更全面的音频特征，包括浅层声学特征、中层韵律特征和深层语义特征。然而，这些方法的一个关键局限在于，它们假设所有隐藏状态携带同等重要的信息，且可通过统一加权进行汇总。这可能会削弱判别性信号，而忽略层间依赖关系则会限制模型捕获各层级内部及层级之间互补关系的能力。因此，我们提出 HierCon，以动态建模不同粒度间的这些依赖关系。

2.2 HierCon

为充分捕捉 XLS-R 特征中的判别性信息，我们提出层次对比注意力（HierCon）。其核心是一个层次注意力模块，将 XLS-R 的 24 层划分为 8 组，每组包含连续的 3 层。此分组方式基于这样一个观察：相邻的 Transformer 层捕捉相似的抽象层级 [12]，从而使组内注意力能够建模局部依赖关系，并使组间注意力能够整合层次结构中的互补证据（早期声学、中层韵律、高层语义）。结合基于间隔的对比学习，这种设计能够学习具有泛化性的欺骗模式，而非数据集特定伪迹。

1）层次注意力。HierCon 通过三个阶段进行层次化建模，以捕捉依赖关系：在每一层内应用时间注意力，在相邻层之间应用组内注意力，并在不同层组之间应用组间注意力。

在阶段 *1*：时间注意力中，为了突出能够捕捉判别性伪迹的信息帧，同时降低无关内容的权重，每一层 l 都会对 T 时间帧应用可学习的注意力。给定帧级隐藏状态 $\mathbf{h}_t \in \mathbb{R}^{1024}$（对应 $t = 1 \ldots T$ 的„），我们通过一个具有可学习参数 $\mathbf{W}_1 \in \mathbb{R}^{128\times1024}$、$\mathbf{b}_1 \in \mathbb{R}^{128}$ 和 $\mathbf{w}_2 \in \mathbb{R}^{128}$ 的两层多层感知机计算注意力权重：

$$\mathbf{e}_t = \tanh(\mathbf{W}_1\mathbf{h}_t + \mathbf{b}_1), \quad \alpha_t = \mathrm{softmax}(\mathbf{w}_2^\top \mathbf{e}_t), \quad \mathbf{z}_l = \sum_{t=1}^{T} \alpha_t \mathbf{h}_t \quad (1)$$

得到层令牌 $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_{24}] \in \mathbb{R}^{24\times1024}$。在阶段 *2*：组内注意力中，为利用相似抽象层级内的互补信息，我们将 24 层划分为八组，每组三层（$g = 3$）：$\mathbf{z}'_k = \mathrm{AttnPool}(\mathbf{G}_k) + \mathrm{MLP}(\mathrm{AttnPool}(\mathbf{G}_k))$，其中 $\mathbf{G}_k \in \mathbb{R}^{3\times1024}$ 包含组 $k \in \{1, \ldots, 8\}$ 中的相邻层令牌。较早的组侧重于声学伪迹，而较晚的组则捕捉语义线索。

在阶段 *3*：组间注意力中，不同的生成技术会在不同的抽象层级呈现伪迹。聚合精炼后的组向量 $\{\mathbf{z}'_k\}_{k=1}^{8}$，得到话语嵌入：

$$\mathbf{u} = \mathrm{AttnPool}(\{\mathbf{z}'_k\}) + \mathrm{MLP}(\mathrm{AttnPool}(\{\mathbf{z}'_k\})) \quad (2)$$

自适应权重可识别出哪些抽象层级为每个样本提供判别性证据。

2）分类器与投影头 分类器是一个经过正则化的多层感知机（随机失活 + 层归一化），将 $\mathbf{u} \in \mathbb{R}^{1024}$ 映射为逻辑值 $\mathbf{y} \in \mathbb{R}^2$。与此同时，投影头将 $\mathbf{u}$ 映射到一个 256 维空间，以进行对比学习，并生成嵌入 $\mathbf{f}_i \in \mathbb{R}^{256}$。低维投影构成了一个几何瓶颈，有助于形成紧凑且具有判别性的表示，同时避免分类目标与对比目标之间相互干扰。推理时会移除投影头。

3）损失函数我们联合使用二元交叉熵（BCE）损失和基于间隔的对比损失，以稳定训练并塑造嵌入几何。BCE 项推动模型准确地区分真实与伪造样本，而对比项则缓解基于熵的训练导致的特征坍塌，并减少对数据集特定伪迹的过拟合。给定一个包含 N 个样本及其话语嵌入 $\{\mathbf{f}_i\}_{i=1}^{N}$ 的批次，我们首先将其投影到一个 256 维空间，并计算余弦相似度。对于每个锚点 i，我们将同一类别（真实或伪造）的样本定义为正样本集 $\mathbf{P}_i$，将相反类别的样本定义为负样本集 $\mathbf{N}_i$。然后计算平均相似度

$$\bar{s}_i^{(+)} = \frac{1}{|\mathcal{P}_i|} \sum_{j\in\mathcal{P}_i} \cos(\mathbf{f}_i, \mathbf{f}_j), \quad \bar{s}_i^{(-)} = \frac{1}{|\mathcal{N}_i|} \sum_{j\in\mathcal{N}_i} \cos(\mathbf{f}_i, \mathbf{f}_j), \quad (3)$$

并强制在正负相似度之间施加间隔 m 度：

$$\mathcal{L}_{\mathrm{con}} = \frac{1}{N} \sum_{i=1}^{N} \max\big(0,\, m + \bar{s}_i^{(-)} - \bar{s}_i^{(+)}\big). \quad (4)$$

整体训练目标为

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{con}}\mathcal{L}_{\mathrm{con}}, \quad (5)$$

其中，L_{CE} 表示 BCE 损失，λ_{con} 控制对比正则化的强度。我们根据验证实验设定 $\lambda_{\mathrm{con}} = 0.1$：较大的取值（0.5–1.0）会过度强调对比项，降低分类准确率；较小的取值（<

0.05）则只能带来微小的改进。我们保持该权重不变，而不采用调度策略，以便从训练开始就由这两个目标共同塑造表征空间，避免收敛到次优几何结构。

4）可解释性 层次注意力机制通过揭示哪些时间帧、层和组对检测贡献最大，提供可解释性。注意力权重（α_t、β_k、γ）可视化为热力图，以识别判别性时间片段和层级结构。

3 实验

我们从三个方面评估 HierCon：（i）与最先进基线相比的整体检测性能；（ii）消融分析，用于单独评估层次注意力和对比学习的贡献；（iii）通过多阶段注意力可视化进行可解释性分析。

实验设置 所有模型均使用 ASVspoof 2019 LA 子集进行训练，并在三个跨域难度各异的基准测试上进行评估：ASVspoof 2021 LA（148,176 条语音）、ASVspoof 2021 DF（533,928 条语音，涵盖 100 多种深度伪造生成流程）以及 In-the-Wild（ITW）（31,779 个真实场景样本）。检测性能采用音频反欺骗研究中的标准指标等错误率（EER%）报告。我们采用 XLS-R 300M 作为特征骨干网络，从全部 24 个 Transformer 层提取隐藏状态，并将音频切分为 4 秒窗口。HierCon 将层划分为每组三层的簇（共 8 组），并应用时间注意力、组内注意力和组间注意力；注意力和前馈组件的维度分别为 128 和 512。训练采用 Adam 优化器，学习率为 1×10^{-6}，批次大小为 16，并在 50 个训练轮次内进行早停。我们采用 RawBoost 数据增强 [10]，并对三个随机种子的所有报告结果取平均值。对比学习采用批次级采样，其中正样本对具有相同的欺骗类别（真实 $\longleftrightarrow$ 真实或伪造 $\longleftrightarrow$ 伪造），负样本对则类别不同。平均相似度分数构成对比间隔目标函数中的批次级正负约束。

3.1 总体性能

表 1 展示了三个数据集上的汇总等错误率，并与近期的检测器系列进行了比较，包括基于 Wav2Vec/WavLM 的系统、单类检测框架以及自监督 XLS-R SLS [13] 基线。HierCon 在跨域评估下展现出新的最先进性能趋势：

21 DF（深度伪造）：HierCon 达到 1.93% 的等错误率，优于强大的 XLS-R + SLS 基线（2.09%），并超过 OCKD（2.27%）和基于 WavLM 的系统。鉴于深度伪造生成技术的多样性，这一改进尤为显著。

21 LA：HierCon 的等错误率为 2.46%，缩小了与高度专业化的伪迹感知检测器（如 AASIST）之间的差距，同时较 XLS-R 基线（3.88%）实现了大幅提升。

In-the-Wild：HierCon 将等错误率降至 6.87%，相较于 XLS-R + SLS（8.87%）实现了 22.5% 的相对提升。

表 1：ASVspoof 2021 基准测试上的合并等错误率（%）。结果为 3 次运行的平均值。粗体表示最佳性能。† 表示结果引自原始论文。

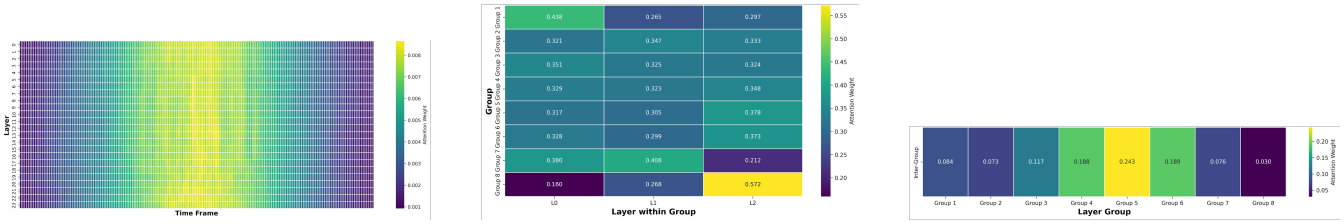
模型	21 逻辑访问	21 DF	ITW
Wav2Vec+逻辑回归 [2]†	-	-	7.20
WavLM+ASP [3]†	3.31	4.47	-
Wav2Vec+线性 [2]†	3.63	3.65	16.17
WavLM+注意力机制 [3]†	3.50	3.19	-
Wav2Vec+AASIST [4]†	0.82	2.85	-
FTDKD [9]†	2.96	2.82	-
Wav2Vec+AASIST2 [4]†	1.61	2.77	-
WavLM+MFA [3]†	5.08	2.56	-
OCKD [6]†	0.90	2.27	7.86
OC+ACS [5]†	1.30	2.19	-
XLS-R + SLS [13]	3.88	2.09	8.87
XLS-R + HierCon（我们的模型）	**2.46**	**1.93**	**6.87**

表2：用于分别评估层次注意力和对比学习贡献的消融实验。粗体表示最佳结果。

模型	21 逻辑访问	21 DF	ITW
XLS-R +SLS [13](基线)	3.88	2.09	8.87
+层次注意力	2.97	2.13	8.81
+层次注意力 +对比学习	**2.46**	**1.93**	**6.87**

3.2 消融实验

为量化各组件的贡献，我们从 XLS-R + SLS 基线开始，逐步启用 HierCon 模块。结果汇总于表2。基线模型在域内表现尚可，但在分布偏移下，尤其是在 ITW 上，表现不佳，说明仅依靠固定的层级权重，难以捕捉稳健的生成伪影。在不使用对比学习的情况下加入层次注意力，可显著提升跨域数据集上的表现：LA 上相对提升 23.5%（2.97% 对比 3.88%），In-the-Wild 上则小幅提升 0.7%（8.81% 对比 8.87%）。然而，DF 上的表现略有下降（2.13% 对比 2.09%），这表明层次建模能够有效捕捉层间依赖关系以促进泛化，但若缺乏正则化，也可能过拟合数据集特定的统计线索。这一现象说明，结构建模虽能增强特征抽象能力，但还需要额外的稳定机制，才能在未见过的攻击范式下保持稳健。将层次注意力与基于间隔的对比学习相结合后，模型在所有数据集上均取得最佳表现，并在最具挑战性的泛化基准上实现显著提升：在 In-the-Wild 上较 SLS 相对提升 27.7%（6.87% 对比 8.87%）。值得注意的是，对比学习也缓解了 DF 上的性能下降（1.93% 对比 2.13%），表明它有助于形成更具可迁移性的表示结构。这些提升意味着，对比正则化不仅增强了类别分离能力，还能防止层次机制依赖于在不同欺骗算法或录制环境下不一致的浅层伪影。这些发现验证了双目标设计的有效性：层次注意力能够捕捉分布于各 Transformer 层的互补线索，而对比学习则促使模型学习更具判别性且更少依赖



（a）时间注意力（阶段 1） （b）组内注意力（阶段 2） （c）组间注意力（阶段 3）

图 2：对 200 个 DF样本的 HierCon 注意力取平均值：（a）时间注意力聚焦于中间帧（40–70%）；（b）组内注意力从浅层（L0：0.438）转移至深层（L2：0.572）；（c）组间注意力在第 5 组达到峰值（L12–14：0.243）。

数据集依赖型表征。表现持续提升表明，架构设计与训练组件共同增强了模型的泛化能力，从而实现可靠的深度伪造检测。

3.3 注意力分析与可解释性

为评估模型所学习的行为是否符合预期设计，我们分析了三个 HierCon 阶段的注意力分布。如图 2 所示，时间注意力始终优先关注信号的中间区域，同时降低对边界区域的权重；这一模式与已有的取证观察结果一致，即合成伪迹更可能出现在持续语音过程中，而不是出现在起始端或尾部边缘。不同运行结果中该行为的一致性表明，模型学习到的是稳定的时间线索，而不是依赖优化噪声或特定样本的对齐偏差。在组内阶段，随着处理过程跨越不同组，我们观察到注意力逐渐从强调浅层的声学驱动层转向更深层的语义导向层。这一变化反映了现代生成系统如何在多个抽象层级引入伪迹：浅层往往暴露出细微的频谱失真或噪声整形异常，而深层则捕捉由合成流程引入的不自然韵律、节奏或语言不稳定性。与此同时，组间注意力最关注中层 Transformer 层。这表明，最具判别性的证据产生于纯声学表征与纯语义表征之间的交互区域，与多层自监督特征融合研究中的既有发现相一致。除了结构层面的检视之外，我们还观察到，在不同的欺骗机制下，包括基于声码器、基于扩散模型以及语音转换模型的机制，注意力模式都保持高度一致。这种趋同表明，模型并非只是记忆合成特定特征，而是在识别尽管生成流程存在差异却始终存在的重复性伪迹结构。这些模式在不同领域中的稳定性支持了这样一个假设：HierCon 学习到的是根植于合成语音基本属性的可迁移线索。综合来看，可解释性分析表明，HierCon 并非通过不透明的捷径特征提升性能，而是通过结构化且连贯地利用与已知欺骗信号特征相一致的多尺度表征来实现性能提升。观察到的行为表明，模型能够在不同抽象层级之间进行推理，而不是完全依赖局部伪迹，从而增强了我们对其稳健性以及超越受控基准数据集的泛化能力的信心。

4 结论

我们提出了 HierCon，这是一种结合对比学习的层次注意力框架，用于音频深度伪造检测。该方法通过建模层间依赖关系并强制构建领域不变的嵌入结构，解决了以往基于自监督学习的方法的局限性。在各项基准测试中，HierCon 均取得了业界领先的性能，包括在 DF 上达到 1.93% 的等错误率，以及相较于 XLS-R+SLS [13] 提升 22.5%。消融实验确认，层次注意力与对比学习带来了互补收益，同时提升了泛化能力并实现了透明的决策行为。未来的工作将研究适用于实时取证应用的跨模态变体和高效部署变体。

5 致谢

本研究获得韩国工业技术规划与评估院（KEIT）在韩国产业通商资源部资助下的支持（编号：RS-2025-25458052，制造业基础模型核心技术开发），并获得信息通信技术规划与评估院（IITP）在韩国政府（MSIT）资助下的拨款支持（编号：RS-2025-02217259，基于国际多学科治理的自进化人工智能偏见检测、纠正与解释平台开发）。

Qizhou Wang 和 Christopher Leckie 的部分研究工作得到了澳大利亚互联网观测站（AIO）的支持。该观测站是一个国家级研究基础设施，支持数字平台和智能数据研究。AIO 获得了澳大利亚研究数据共享平台（ARDC）通过国家协作研究基础设施战略（NCRIS）提供的投资。

参考文献

[1] A. Babu、C. Wang、A. Tjandra、K. Lakhotia、Q. Xu、N. Goyal、K. Singh、P. von Platen、Y. Saraf、J. Pino、A. Baevski、A. Conneau 和 M. Auli。2022。XLS-R：大规模自监督跨语言语音表征学习。载于*Interspeech*。[2] A. Baevski、Y. Zhou、A. Mohamed 和 M. Auli。2020。wav2vec 2.0：用于语音表征自监督学习的框架。载于*NeurIPS*。12449–12460。[3] S. Chen、C. Wang、Z. Chen、Y. Wu、S. Liu、Z. Chen、J. Li、N. Kanda、T. Yoshioka、X. Xiao、J. Wu、L. Zhou、S. Ren、Y. Qian、Y. Qian、J. Wu、M. Zeng、X. Yu 和 F. Wei。2022。WavLM：用于全栈语音处理的大规模自监督预训练。*IEEE*信号处理精选主题期刊 16，6（2022）。[4] J. Jung、H.-S. Heo、H. Tak、H. Shim、J. S. Chung、B.-J. Lee、H.-J. Yu 和 N. Evans。2022。AASIST：利用集成时频图注意力网络进行音频反欺骗。载于*ICASSP*。6367–6371。[5] G. Lavrentyeva、S. Novoselov、A. Malykh、A. Kozlov、O. Kudashev 和 V. Shchemelinin。2019。面向ASVspoof2019挑战赛的STC反欺骗系统。载于*Interspeech*。1033–1037。

[6] J. Lu、Y. Zhang、W. Wang、Z. Shang 和 P. Zhang。2024。用于欺骗语音检测的单类知识蒸馏。载于*ICASSP*。11251–11255。[7] N. M. Müller、P. Czempin、F. Dieckmann、A. Froghyar 和 K. Böttinger。2022。音频深度伪造检测能否泛化？*arXiv*预印本 *arXiv:2203.16263*（2022）。[8] M. Ravanelli 和 Y. Bengio。2018。基于原始波形和SincNet的说话人识别。载于*SLT*。1021–1028。[9] M. Sahidullah、T. Kinnunen 和 C. Hanilçi。2015。合成语音检测特征比较。载于*Interspeech*。2087–2091。

[10] H. Tak、J. Patino、M. Todisco、A. Nautsch、N. Evans 和 A. Larcher。2022。使用 RawNet2 进行端到端反欺骗。在 *ICASSP*。6369–6373。[11] M. Todisco、H. Delgado、K. A. Lee、M. Sahidullah、N. Evans、T. Kinnunen 和 J. Yamagishi。2018。集成呈现攻击检测与自动说话人验证：共同特征与高斯后端融合。在 *Interspeech*。[12] Y. Xiao 和 N. T. Vu。2025。使用 XLS-R 进行伪造音频检测的逐层决策融合。在 *Interspeech*。5618–5622。[13] Q. Zhang、S. Wen 和 T. Hu。2024。基于自监督XLS-R和SLS分类器的音频深度伪造检测。在 *ACM MM*。6765–6773。