arXiv:2601.06560v1 [eess.AS] 10 Jan 2026

亮点

基于跨尺度注意力与一致性学习的轻量级分辨率感知音频深度伪造检测

K. A. Shahriar

- 提出了一种轻量级分辨率感知框架，用于音频深度伪造检测
- 跨尺度注意力可自适应融合多分辨率频谱特征
- 一致性学习可促使模型学习分辨率不变表示，从而增强鲁棒性
- 在 ASVspoof、FoR 以及野外基准测试中展现出强大的泛化能力
- 可解释性分析揭示了有意义且与分辨率一致的检测线索

基于跨尺度注意力与一致性学习的轻量级分辨率感知音频深度伪造检测

K. A.Shahriar[a],1,*

[a] 孟加拉国达卡孟加拉国工程技术大学（*BUET*）电气与电子工程系（*EEE*）

摘要

由于语音合成和语音转换技术的迅速发展，音频深度伪造检测变得愈发困难，尤其是在信道失真、重放攻击和真实世界录音条件下。本文提出了一种具备分辨率感知能力的音频深度伪造检测框架，通过跨尺度注意力和一致性学习，对多分辨率频谱表征进行显式建模与对齐。与传统的单分辨率方法或隐式特征融合方法不同，所提出的方法在互补的时频尺度之间强制实现一致性。所提出的框架在三个具有代表性的基准上进行了评估：ASVspoof 2019（逻辑访问（LA）和物理访问（PA））、Fake-or-Real（FoR）数据集以及野外音频深度伪造数据集，并采用说话人互斥协议。该方法在 ASVspoof LA 上取得近乎完美的性能（等错误率 0.16%），在 ASVspoof PA（等错误率 5.09%）、FoR 重录音频（等错误率 4.54%）和野外深度伪造音频（曲线下面积 0.98，等错误率 4.81%）上展现出强大的鲁棒性，在具有挑战性的条件下显著优于单分辨率和非注意力基线。所提出的模型轻量且高效，仅需 159k 个参数，每次推理的计算量不到 1 GFLOP，适合实际部署。全面的消融研究证实了跨尺度注意力和一致性学习的关键作用，而基于梯度的可解释性分析表明，该模型能够在多种欺骗条件下学习到分辨率一致且语义明确的频谱线索。这些结果表明，显式跨分辨率建模为新一代音频深度伪造检测系统提供了有原则、鲁棒且可扩展的基础。

*通讯作者。电子邮箱：kh.ashikshahriar@gmail.com[1]ORCID：0009-0008-2571-7485

关键词：音频深度伪造检测、音频欺骗、多分辨率频谱图、注意力机制、
一致性学习

1. 引言

音频深度伪造是指由机器学习模型生成的合成或篡改语音，这些模型会模仿真实说话人的声学特征、语言模式和声音身份 [1, 2, 3, 4]。与传统的文本转语音（TTS）系统不同，现代音频深度伪造采用神经声码器 [5, 6], 、自回归生成器 [7, 8], 和基于扩散的模型 [9], 等深度神经网络架构生成，能够实现高度自然且具有说话人特征的语音生成 [10]。攻击者只需使用有限的语音样本对这些模型进行条件训练，就能生成任意话语，使其听起来与目标人物的声音极为相似，往往令仅凭人类感知难以可靠地区分 [11, 12, 13]。

语音合成和语音转换技术的快速发展，大幅降低了生成逼真音频深度伪造的门槛[14]。尽管这些进步推动了无障碍服务、语音修复和人机交互等有益应用 [15], ，但同时也带来了严重的安全和社会风险 [16, 17]。音频深度伪造可被用于绕过说话人验证系统、冒充公众人物、伪造证据，或实施社会工程和金融欺诈[18, 19, 20]。开源合成工具包和商业语音克隆服务日益普及，进一步凸显了开发可靠且具有泛化能力的检测机制的紧迫性。早期关于音频深度伪造和欺骗攻击检测的研究，主要聚焦于识别合成流程引入的伪影 [21, 22, 23, 24]。这些方法依赖手工设计的声学特征（例如相位不一致性、频谱统计特征或倒谱系数），并结合传统分类器 [25, 26, 27]。这类方法对早期合成语音颇为有效，但本质上受限于对人工设计特征以及特定合成伪影假设的依赖。随着合成模型不断改进，许多此类手工设计的线索变得越来越细微，甚至不复存在 [2, 28]。

深度学习的兴起标志着音频深度伪造检测领域发生了范式转变。卷积神经网络和基于注意力的神经网络使模型能够直接从频谱图和梅尔表示等时频输入中学习判别性表示 [29, 30, 31, 32, 33]。大规模基准测试和社区驱动的挑战赛 [34] 通过统一评估协议和数据集，进一步加速了进展。在

受控基准测试中，现代检测器如今已达到接近极限的性能，准确率往往超过人类水平。然而，这种表面上的成功掩盖了一项关键弱点：在干净或经过整理的数据集上训练的检测系统，在遇到重放攻击、信道失真、压缩伪影或真实世界的录音条件时，性能往往会下降 [35, 36]。实际上，音频深度伪造很少以未经处理的形式出现；它们会经由麦克风、扬声器、电信信道和在线平台传播。由此产生的分布偏移表明，许多现有检测器实际上依赖于数据集特定伪影或狭窄的频谱线索，而这些特征无法泛化到训练条件之外。这一问题的根本局限在于对时频表示的建模方式。大多数现有方法采用单一频谱分辨率，或在使用多种分辨率时采用简单的融合策略。然而，不同分辨率所捕捉的信息本质上各不相同：细粒度表示突出短时频谱异常，而较粗的分辨率则编码更长时间范围的时间动态和韵律结构 [37, 38]。独立处理这些表示，或在缺乏明确对齐的情况下将其合并，可能导致决策边界不一致或不稳定，尤其是在信道失真和重放失真下。

本文指出，稳健的音频深度伪造检测需要显式的跨分辨率建模。为此，我们提出了一种分辨率感知检测框架，该框架在多个频谱分辨率上联合学习，同时确保表示对齐的一致性。所提出的架构引入了一种跨尺度注意力机制，能够动态融合互补的分辨率特定特征，使模型可以根据声学条件自适应地突出重要线索。此外，我们采用一致性学习目标来约束不同分辨率下的预测，促使模型关注分辨率不变特征，而非无关伪影。我们在多种基准数据集上评估了该框架，这些数据集涵盖受控合成攻击、基于重放的失真以及真实环境音频深度伪造，并包括说话人互斥评估协议。实验结果表明，该方法在标准基准测试中取得了近乎顶尖的性能，同时显著提升了其在严苛真实世界条件下的鲁棒性。广泛的消融研究验证了各架构组件的作用，基于梯度的可解释性分析则揭示了模型在不同分辨率下利用的频谱线索。

总而言之，本研究作出了三项重要贡献。首先，提出了一种分辨率感知音频深度伪造检测框架，能够显式地对多分辨率频谱表示进行建模和对齐。其次，研究表明，交叉尺度注意力和一致性学习能够显著提

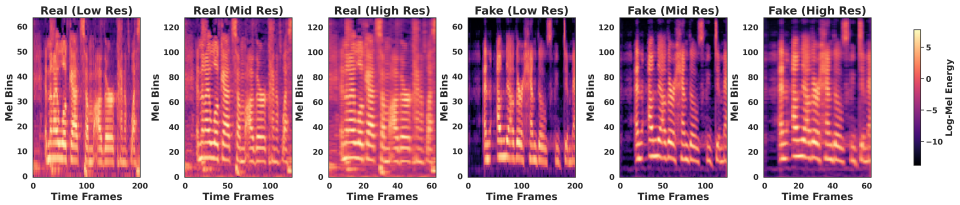

图 1：In-the-Wild 数据集中的真实与伪造音频样本的多分辨率对数梅尔频谱图。低、中、高分辨率表征呈现了互补的时间与频谱伪影。

升回放及非受限条件下的鲁棒性。第三，研究通过实证证据和基于可解释性的证据表明，显式的跨分辨率建模是构建具备泛化能力且值得信赖的音频深度伪造检测系统的关键因素。

2. 提出的方法

2.1. 理论动机

音频深度伪造检测可表述为一个二元分类问题：将输入语音信号映射到一个标签，以指示其属于真实样本还是伪造样本。设 $x(t)$ 表示原始音频波形，$y \in \{0, 1\}$ 表示其对应标签，其中 0 和 1 分别表示真实样本和伪造语音。大多数基于深度学习的检测系统首先将 $x(t)$ 转换为时频表示，随后通过神经网络学习具有判别性的映射。

本研究的一个关键动机是，单一时频分辨率不足以捕捉现代语音合成和语音转换系统引入的所有伪影。精细分辨率的频谱图着重呈现短期频谱异常和高频伪影，而较粗的分辨率则编码较长时间的时间结构和韵律不一致性。图 1 展示了这一现象，其中给出了 In-the-Wild 数据集中具有代表性的真实与伪造音频样本的多分辨率对数梅尔频谱图。某些合成伪影在较高分辨率下更为明显，而另一些则只有在较粗的时间尺度下才能观察到，这凸显了不同分辨率之间的互补性。

形式化地，给定音频信号 $x(t)$，可以在不同分辨率下提取一组 K 频谱表示，记作

$$\mathcal{S}(x) = \{S_1(x), S_2(x), \ldots, S_K(x)\}, \tag{1}$$

其中，每个 $S_k(\cdot)$ 对应一种时频变换，具有不同的时间和频谱粒度。所提出的框架对这些表征进行显式建模和对齐，以利用互补分辨率线索，实现稳健的音频深度伪造检测。设 $f_\theta(\cdot)$ 表示一个由 θ 参数化的共享编码器网络，它将频谱表示映射到潜在嵌入空间，

$$\mathbf{z}_k = f_\theta(S_k(x)) \in \mathbb{R}^d, \quad k = 1, \ldots, K. \tag{2}$$

在传统的多特征方法中，这些嵌入通常会在分类之前进行拼接或平均。然而，这类隐式融合策略并未显式建模不同分辨率之间的关系，并且可能使分辨率特定伪影主导决策函数。这可能导致脆弱表示，使其在信道失真或重新录制条件下难以实现良好的泛化 [39]。

为解决这一局限，所提出的框架引入了一种显式的跨尺度交互机制，使不同分辨率的嵌入能够彼此进行注意力交互。令 $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}2, \ldots, \mathbf{z}_K]^\top \in \mathbb{R}^{K\times d}$ 表示堆叠后的多分辨率嵌入。跨尺度注意力定义为

$$\tilde{\mathbf{Z}} = \mathrm{Attn}(\mathbf{Z}, \mathbf{Z}, \mathbf{Z}), \tag{3}$$

其中，Attn(·)表示多头自注意力算子。该公式使每个分辨率特定嵌入能够动态地重新加权来自其他分辨率的信息，从而捕获互补的频谱线索。尽管跨尺度注意力有助于实现特征融合，但它并未显式强制不同分辨率之间保持一致性。实际上，从不同分辨率中提取的嵌入可能编码相互冲突或依赖于分辨率的线索，尤其是在信道失真条件下。为减轻这一影响，引入了一致性学习目标，以促进对应于真实语音的嵌入之间实现对齐。具体而言，鼓励归一化嵌入 $\hat{}\mathbf{z}_k = \mathbf{z}_k/\|\mathbf{z}_k\|_2$ 在不同分辨率之间保持相似，

$$\mathcal{L}_{\mathrm{cons}} = \sum_{i<j} \mathbb{E}_{x\sim\mathcal{D}_{\mathrm{real}}} \left[\|\hat{\mathbf{z}}_i - \hat{\mathbf{z}}_j\|_2^2\right], \tag{4}$$

其中 $\mathcal{D}_{\mathrm{real}}$ 表示真实语音的分布。该目标反映了这样一个假设：真实语音应表现出分辨率不变特征，而伪造语音通常会在不同频谱尺度之间引入不一致性。

最终的分类分数通过聚合注意力嵌入并应用线性分类器获得，

$$\hat{y} = \sigma\left(g\left(\frac{1}{K}\sum_{k=1}^{K}\tilde{\mathbf{z}}_k\right)\right), \tag{5}$$

其中， $g(\cdot)$表示可学习仿射变换， $\sigma(\cdot)$表示 Sigmoid 函数。整体训练目标将二元交叉熵损失 $\mathcal{L}_{\text{cls}}$ 与一致性正则化项相结合，

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda\mathcal{L}_{\text{cons}}, \tag{6}$$

其中， λ 控制一致性学习的贡献。该公式明确编码了跨分辨率关系，并强制学习分辨率一致的表征，为提升音频深度伪造检测的鲁棒性和泛化能力提供了一种有理论依据的方法。

2.2. 数据集

所提出的方法在三个公开可用的音频深度伪造检测数据集上进行评估，这些数据集共同涵盖受控、重新录制和真实世界条件。这些数据集在相关文献中得到广泛使用，并提供了互补的评估场景，用于评估鲁棒性和泛化能力。

2.2.1. ASVspoof 2019 数据集

ASVspoof 2019 数据集 [40]是音频欺骗和深度伪造检测领域广泛使用的基准数据集，包含两种互补的评估场景：逻辑访问（LA）和物理访问（PA）。两个子集均采用 16 kHz 采样率，并遵循标准化的实验方案。LA 子集包含真实语音录音，以及使用多种文本转语音和语音转换系统生成的伪造语句。该子集代表一种受控实验环境，其中欺骗伪迹主要源于算法，而非环境因素。相比之下，PA 子集侧重于基于重放的欺骗攻击：真实语音和伪造语音在不同声学环境和信道条件下，通过物理录制设备和播放设备采集，从而引入真实的环境失真。

数据集按照预定义协议划分为训练集、开发集和评估子集。在本研究中，我们在逻辑访问（LA）和物理访问（PA）子集上均进行了实验，以评估所提出的方法在

表 1：本研究使用的 ASVspoof 2019 LA 和 PA 数据集统计概览

子集	划分	伪造	真实	总计
LA	训练集	22,800	2,580	25,380
	Dev	22,296	2,548	24,844
PA	训练集	48,600	5,400	54,000
	Dev	24,300	5,400	29,700

算法和物理欺骗攻击场景下的表现。按照文献中既定的评估惯例，训练集和开发集分别用于模型优化和验证。本研究所用的数据集统计汇总见表 1。

2.2.2. Fake-or-Real（*FoR*）数据集

Fake-or-Real（FoR）数据集 [41] 是一个面向合成语音和音频深度伪造检测的大规模基准数据集。该数据集包含超过 198,000 条话语，涵盖真实人类语音和由多种现代文本转语音系统生成的伪造语音。与基于经典合成技术的早期数据集不同，FoR 侧重于基于深度学习的语音合成方法，涵盖开源系统和商业系统。数据集中的伪造部分包含使用多种先进神经网络文本转语音模型生成的话语，例如 DeepVoice 3、Google Cloud TTS、WaveNet、Amazon AWS Polly、Microsoft Azure TTS 和百度云文本转语音，共涉及 33 种合成语音。真实语音样本收集自不同来源，包括开放语音语料库和真实场景录音，以确保说话人、口音、麦克风和录音环境的多样性。

FoR 数据集以多个版本发布。*for-original* 版本包含原始采集语句，而 *for-norm* 版本则进行归一化处理，包括重采样至 16 kHz、音量归一化、单声道转换和静音裁剪。*for-2sec* 版本进一步将话语截断为固定时长的两秒，以消除与长度相关的偏差。最后，*for-rerecorded* 版本通过使用消费级播放和录音设备回放并重新录制话语来模拟重放攻击，从而引入真实的信道失真。本文使用 FoR 数据集的归一化、两秒和重新录制版本，在受控、归一化和重放类条件下评估所提出的分辨率感知框架。这些子集可用于系统地

表2：本研究所使用的 FoR 数据集子集的数据分布（16千赫兹采样率）。

子集	伪造	真实	总计
for-norm	34,700	34,700	69,400
for-2sec	8,935	8,935	17,870
for-rerecorded	8,935	8,935	17,870

评估模型对录音可变性和信道效应的鲁棒性。

2.2.3. 野外音频深度伪造数据集

为评估现实世界泛化能力，我们在野外音频深度伪造数据集 [42]上开展实验。该数据集包含从公开在线来源收集的真实语音和伪造语音，其中包括社交媒体平台和视频流媒体网站，因此说话人、录音环境、压缩伪影和背景噪声存在显著差异。该数据集包含来自 58 位公众人物的音频录音，其中包括政治人物及其他知名人士。每位说话人均提供真实和伪造音频样本，从而支持说话人感知评估和说话人互斥评估协议。该数据集总计包含约 20.8 小时的真实语音和 17.2 小时的伪造语音，相当于每位说话人平均有 23 分钟真实音频和 18 分钟伪造音频。所有音频样本的采样率均为 16 kHz。

本研究采用说话人不相交划分协议，以确保训练集、验证集和测试集中的说话人互不重叠。该协议可避免身份泄漏，并严格评估模型对未见说话人及录音条件的泛化能力。由于该数据集具有多样且不受约束的特性，其评估场景比精选基准更具挑战性，也可作为部署条件的现实替代。本文使用的数据集统计汇总见表 3。

2.3. 预处理与特征提取

所有实验均采用统一的预处理和特征提取流程，以确保在具有不同录音条件和时长的数据集之间进行一致且公平的评估。该流程旨在保留相关频谱伪影，同时尽量减少采样率变化和话语长度差异等混杂因素。

表 3：本研究使用的野外音频深度伪造数据集概览。

特征	数值
样本总数	31,779
真实样本数	19,963
伪造样本	11,816
说话人数量	58
真实音频时长总计	20.8 小时
伪造音频时长总计	17.2 小时
采样率	16 kHz

2.3.1. 音频预处理

所有音频信号均重新采样至 16 kHz 的统一采样率，并转换为单通道（单声道）格式。对于包含可变长度话语的数据集，提取固定时长片段，以减少话语长度相关偏差。对于 FoR 数据集，在预处理期间通过随机裁剪获取 2 秒片段。对于野外音频深度伪造数据集，在训练期间通过随机裁剪提取更长的 4 秒片段；对于较短的录音，则在必要时进行填充。随机裁剪仅应用于训练阶段，以引入时间变化性并提升泛化能力 [43], ；而在验证阶段则使用确定性裁剪或完整片段。

2.3.2. 多分辨率频谱特征提取

从每个音频片段中提取多个对数梅尔频谱图表示，以捕获互补的频谱表示。具体而言，使用不同的短时傅里叶变换（STFT）配置计算三种梅尔频谱图：

- 一种精细分辨率表征，使用 $n_{\text{fft}} = 400$，跳跃长度为 160 个样本，包含 64 个梅尔频率箱。
- 一种中等分辨率表征，使用 $n_{\text{fft}} = 1024$，跳跃长度为 256 个样本，包含 128 个梅尔频率箱。
- 一种粗略分辨率表征，使用 $n_{\text{fft}} = 2048$，跳跃长度为 512 个样本，包含 128 个梅尔频率箱。

对每个梅尔频谱图应用对数振幅缩放，以稳定优化过程并突出与感知相关的频谱差异。

由此得到的表征提供了对短期频谱不规则性和长期时间模式的互补视角，所提出的分辨率感知框架会联合利用这些信息。

2.4. 模型架构

所提出的分辨率感知音频深度伪造检测模型旨在联合利用互补的频谱分辨率，同时保持计算效率。该架构由三个主要部分组成：共享卷积编码器、跨尺度注意力模块以及分辨率无关分类头。表4总结了模型复杂度概况。

2.4.1. 共享编码器

每个输入音频片段均由三个在不同时间–频率分辨率下计算得到的对数梅尔频谱图表示。对每种分辨率分别应用共享卷积编码器，从而确保参数效率以及跨尺度的一致特征提取。编码器由三个带有 ReLU 激活的卷积层构成，随后通过自适应平均池化生成固定维度的嵌入。给定输入频谱图 S_k，编码器将其映射为潜在表示 $\mathbf{z}_k \in \mathbb{R}^{128}$。

对于涉及数据集感知训练（FoR）的实验，会对潜在嵌入应用可学习的缩放和平移参数（γ 和 β），以实现轻量级的数据集条件调制。此操作保留共享编码器结构，同时支持针对特定分布进行小幅调整。

2.4.2. 跨尺度注意力

将三个潜在向量堆叠后，输入多头自注意力模块。该跨尺度注意力机制使每种分辨率都能关注其他分辨率中的互补信息，从而让模型动态突出与分辨率无关且具有判别性的频谱线索。随后，对注意力输出沿分辨率维度进行平均池化，得到融合表示，整合多尺度信息。

2.4.3. 分类头

融合后的嵌入被送入轻量级线性分类器，该分类器输出一个对应于伪造概率的标量逻辑值。在多数据集实验中，各数据集使用独立的分类头，同时共享编码器和注意力模块。这种设计能够隔离各数据集特定的决策

表4：所提出的模型的参数量和计算复杂度。

模块	**# 参数**
编码器卷积层1	320
编码器卷积层2	18,496
编码器卷积层3	73,856
数据集调制（γ）	384
数据集调制（β）	384
多头注意力	49,536
注意力输出投影	16,512
分类器头（每个数据集）	129
可训练参数总数	159,875

边界，而不会增加共享特征提取器的表征能力。

2.4.4. 模型复杂度

所提出的架构依然紧凑且计算效率高。可训练参数总数约为 1.6×10^5，采用单精度格式时，模型大小不到 1 MB（0.62 MB）。

所提出的模型包含 159,875 个可训练参数，采用单精度格式时内存占用约为 0.62 MB，采用半精度格式时约为 0.30 MB。每次前向传播的浮点运算总数约为 936.66 MFLOPs，而训练期间的峰值激活内存使用量为 42.07 MB。这些特性表明，该模型在保持良好准确率与计算效率平衡的同时，实现了出色的检测性能。

2.5. 模型训练与评估

本节介绍用于评估所提出的分辨率感知音频深度伪造检测框架的训练流程和评估方法。所有实验均采用统一的训练策略，以确保在不同数据集和实验条件下进行公平比较。

2.5.1. 训练流程

所提出的模型作为二分类器进行训练，以区分真实样本和伪造语音。给定一个由多个频谱分辨率表示的输入音频片段，模型会输出一个标量逻辑值

该逻辑值表示遭受欺骗攻击的可能性。训练采用二元交叉熵（BCE）损失。对于涉及一致性学习的实验，还会加入一个额外的正则化项，以促进不同分辨率特定嵌入之间的对齐；总损失则由分类损失和一致性损失加权求和得到。

所有模型均使用 Adam优化器进行优化，学习率固定为 1×10^{-4}。采用小批量训练，批量大小为 32。训练共进行 15 个轮次，并根据验证表现选择性能最佳的模型。

2.5.2. 模型选择

模型选择基于各数据集对应的验证集进行。等错误率（EER）作为主要选择标准，因为它不依赖阈值，且在音频欺骗攻击文献中被广泛采用。对于每项实验，选择验证集 EER 最低的模型并保存，以供后续评估。

2.5.3. 评估指标

采用多种互补指标评估性能。报告分类准确率，以直观衡量分类正确性。使用 EER 作为主要指标，量化错误接受率与错误拒绝率之间的权衡。此外，在选定的实验中报告受试者工作特征曲线下面积（ROC-AUC），以刻画模型不受决策阈值影响的整体判别能力。

对于 ASVspoof 2019 LA 和 PA 数据集，评估遵循官方协议，在训练子集上进行训练，并在开发子集上评估性能。对于 FoR 数据集，分别在标准化版本、两秒版本和重新录制版本上进行评估，以分析信道失真程度不断增加时的鲁棒性。对于野外音频深度伪造数据集，采用说话人互斥协议，确保训练集、验证集和测试集中的说话人互不重叠。

如图2所示，所提出的框架通过共享编码器、跨尺度注意力和一致性正则化，联合利用多个频谱分辨率进行学习。

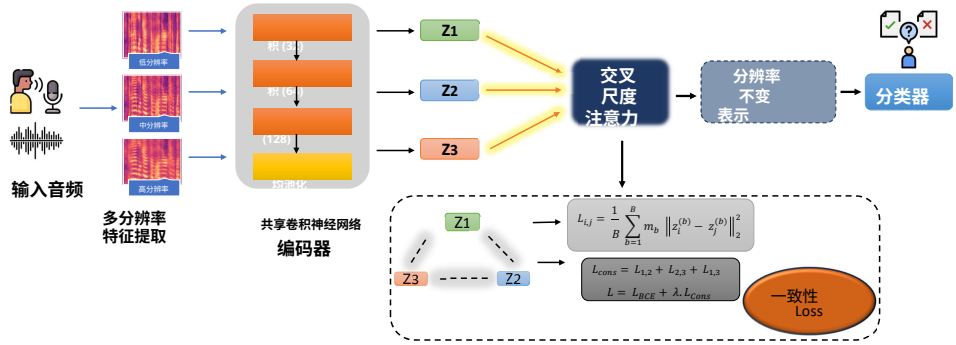

图2：所提出的分辨率感知音频深度伪造检测框架概览。输入音频信号被转换为低、中、高分辨率的梅尔频谱图，再由共享卷积编码器处理，提取分辨率特定嵌入 z_1、z_2 和 z_3。跨尺度注意力模块动态融合这些嵌入，形成用于分类的分辨率不变表示。在训练期间，对真实语音应用跨分辨率一致性损失，以确保不同分辨率之间保持对齐，从而提升模型在回放及真实世界条件下的鲁棒性。

3. 结果与分析

3.1. ASVspoof 2019 上的结果

3.1.1. 逻辑访问（*LA*）结果

表 5 总结了在 ASVspoof 2019 LA 开发集上的分类性能。所提出的方法取得了近乎完美的性能，准确率达到 99.90%，曲线下面积为 1.000，等错误率为 0.0016。真实样本和伪造类均以极高的精确率和召回率被检测出来，表明该模型能够有效捕捉由文本转语音和语音转换系统引入的算法伪影。

图3展示了 LA 子集的混淆矩阵、ROC 曲线和 t-SNE 嵌入。混淆矩阵显示误分类数量微乎其微，而 ROC 曲线表明类别之间几乎可以理想地区分

表 5：ASVspoof 2019 LA（开发集）上的性能。

类别	精确率	召回率	F1 分数
真实样本	0.9933	0.9965	0.9949
伪造	0.9996	0.9992	0.9994
准确率		0.9990	
AUC		1.0000	
EER		0.0016	

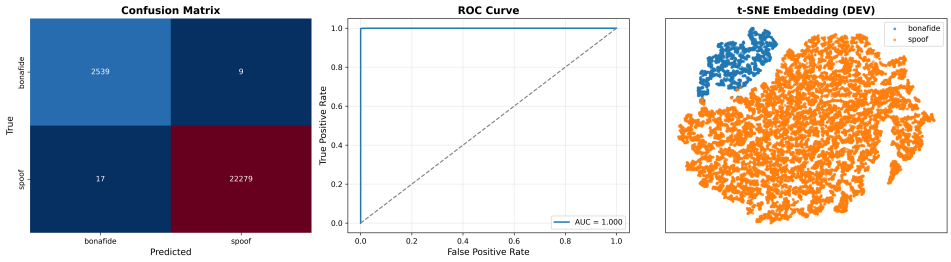

图3：ASVspoof 2019 LA 结果：混淆矩阵（左）、ROC 曲线（中）以及开发集表征的 t-SNE 嵌入（右）。

能力。t-SNE 可视化进一步揭示了清晰的聚类结构，真实样本和伪造样本在学习到的嵌入空间中形成了彼此分离良好的流形。

3.1.2. 物理访问（*PA*）结果

表 6 报告了模型在 ASVspoof 2019 PA 开发集上的性能。在这一更具挑战性的基于重放的场景中，所提出的模型准确率达到 95.32%，曲线下面积为 0.9884，等错误率为 0.0509。尽管性能依然出色，但与 LA 子集相比，错误率有所上升，这反映出声学环境、播放设备和录音通道带来了更大的变化。

图 4 展示了 PA 子集对应的混淆矩阵、ROC 曲线和 t-SNE 嵌入。与 LA 情况相比，混淆矩阵显示，被误分类为伪造的真实样本比例更高；这与重放攻击会引入信道失真、使其部分类似于欺骗伪迹这一情况相符。尽管如此，ROC 曲线仍保持陡峭上升，表明模型具有很强的判别能力。t-SNE 可视化显示，类别间的重叠有所增加，这反映了重放攻击检测固有的难度，但仍保留了有意义的分离程度

表 6：ASVspoof 2019 PA（开发集）上的性能。

类别	精确率	召回率	F1 分数
真实样本	0.8647	0.8804	0.8725
伪造	0.9733	0.9694	0.9713
准确率		0.9532	
AUC		0.9884	
EER		0.0509	

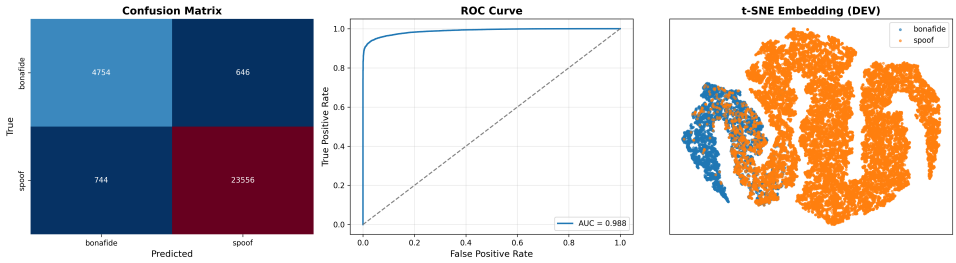

图 4：ASVspoof 2019 PA 结果：混淆矩阵（左）、ROC 曲线（中）以及开发集表征的 t-SNE 嵌入（右）。

，在学习到的表征空间中。

3.2. FoR 数据集上的结果

表7总结了所有 FoR子集上的分类性能。所提出的方法取得了始终如一的优异结果，在归一化数据和时长控制数据上接近完美，并且在重新录音条件下仅出现平稳的性能下降。特别是，*for-2sec* 子集取得了最高的总体准确率，为 99.12%，等错误率为 0.0057。

3.2.1. for-2sec 上的结果

图5展示了*for-2sec* 子集的混淆矩阵、ROC 曲线和 t-SNE 嵌入。混淆矩阵显示，误分类数量很少，且各类别的平衡错误率相近。ROC 曲线表现近乎理想，曲线下面积达到 0.9993。t-SNE 可视化显示，真实样本语音与伪造语音形成了两个分离良好的簇，这表明即使将话语限制为固定的两秒时长，学习得到的嵌入仍具有很强的判别能力。

表7：在不同录音条件下 FoR 数据集上的性能。

子集	准确率	EER	ROC 曲线下面积
for-2sec	0.9912	0.0057	0.9993
for-norm	0.9810	0.0046	0.9996
for-rerec	0.9563	0.0454	0.9900

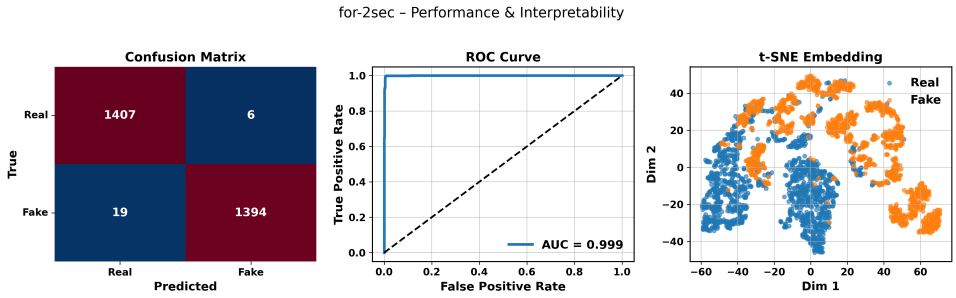

图5：FoR for-2sec 结果：混淆矩阵（左）、ROC 曲线（中）和 t-SNE 嵌入（右）。

3.2.2. for-norm 上的结果

归一化 FoR 子集上的结果如图6所示。该设置代表一种干净且受控的场景，其中录音条件经过标准化。所提出的方法取得了 98.10% 的准确率和 0.9996 的曲线下面积。混淆矩阵显示，假阳性率和假阴性率都极低；而 t-SNE 嵌入则显示，真实与伪造语音流形之间存在明显分离。

3.2.3. for-rerec 上的结果

图7报告了在重新录制的 FoR 子集上的结果，该子集模拟了重放攻击和信道失真。与干净子集相比，性能有所下降，但所提出的方法仍取得了 95.63% 的较高准确率，曲线下面积为 0.9900。混淆矩阵显示，真实样本的误分类有所增加，反映了重放引起的伪影所带来的影响。t-SNE 可视化显示类别之间的重叠有所增加，但仍保留了可辨别的分离结构，说明即使在不利的录音条件下，学习到的表征仍保留了有意义的欺骗相关信息。

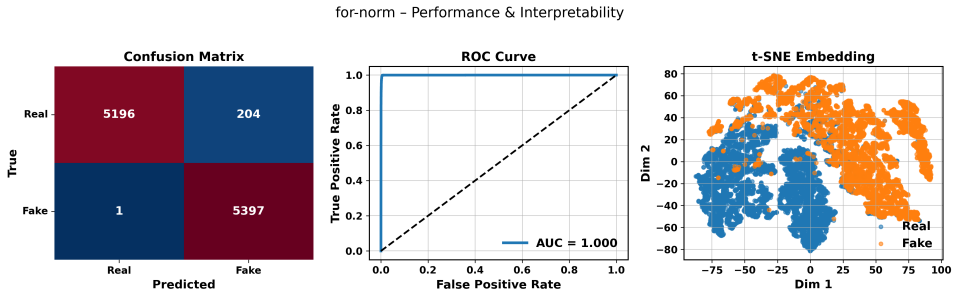

图6：FoR for-norm 结果：混淆矩阵（左）、ROC 曲线（中）和 t-SNE 嵌入（右）。

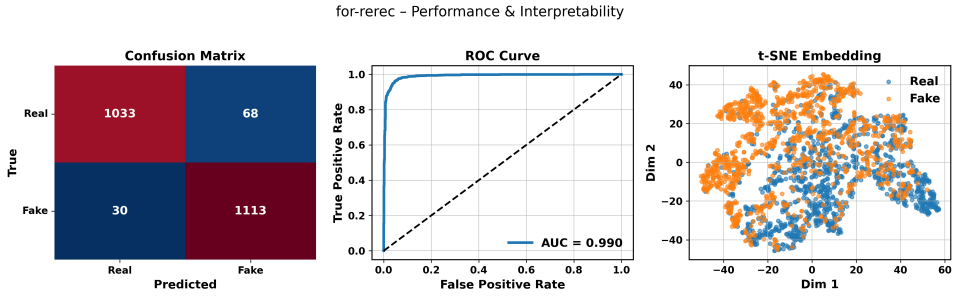

图7：FoR for-rerec 结果：混淆矩阵（左）、ROC 曲线（中）和 t-SNE 嵌入（右）。

3.3. 真实环境音频深度伪造的结果

表8总结了所提出的模型在测试集上的性能。该方法的总体准确率为 95.70%，ROC 曲线下面积为 0.9800，等错误率为 0.0481。

图8展示了测试集的混淆矩阵和 ROC 曲线。混淆矩阵显示，绝大多数真实语音都被正确分类，仅出现两次假阳性错误，因此真实类别的召回率接近完美，达到 1.00。伪造样本的检测精确率很高，但也观察到数量适中的假阴性。ROC 曲线在原点附近迅速上升，表明模型在广泛的阈值范围内具有很强的判别能力。曲线下面积达到 0.98，证实学习到的表征在真实世界声学条件和说话人变异性下仍然稳健。

表8：野外音频深度伪造数据集（测试集）上的性能。

指标	数值
准确率	0.9570
ROC 曲线下面积	0.9800
EER	0.0481

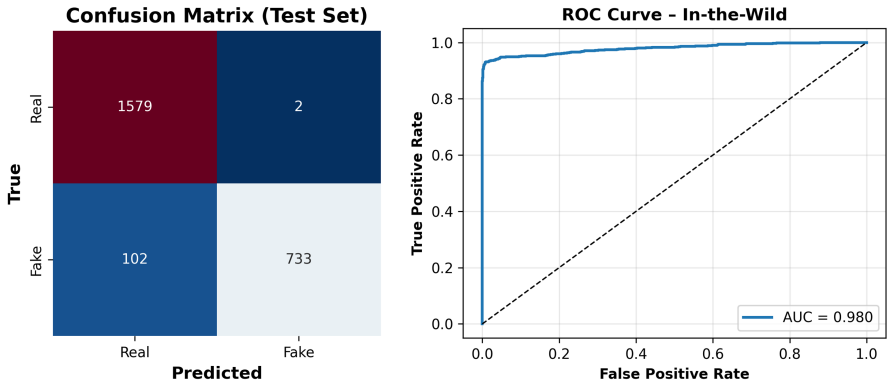

图8：野外音频深度伪造在测试集上的结果：混淆矩阵（左）和 ROC 曲线（右）。

3.4. 可解释性分析

采用多分辨率时频可视化和基于 Grad-CAM 的激活映射进行可解释性分析。分析聚焦于三个具有代表性的评估场景：（i）ASVspoof 2019（PA）；（ii）FoR 数据集的重录子集；以及（iii）In-the-Wild 音频深度伪造数据集。

3.4.1. ASVspoof PA 上的 *Grad-CAM* 分析

图 9 展示了 ASVspoof 2019 PA 数据集中低、中、高分辨率梅尔频谱图表示的显著性分布。低分辨率表示主要捕捉长期时域能量分布和粗略的频谱包络，从而提供与语音节奏和韵律相关的上下文线索。中分辨率特征突出共振峰过渡和时间调制模式，而高分辨率表示则凸显精细的频谱不规则性，尤其是高频区域中的不规则性。

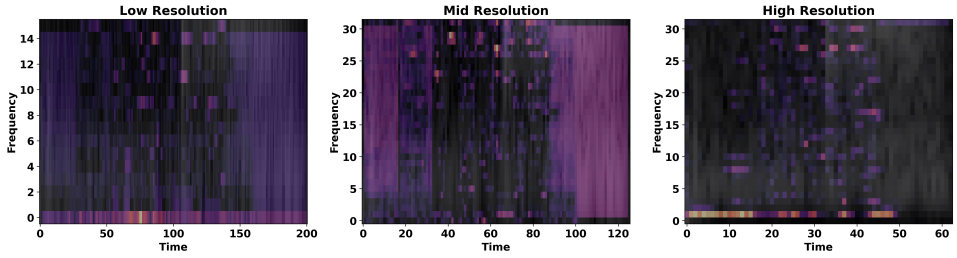

图 9：ASVspoof 2019 PA 数据集上的分辨率级显著性可视化。从左到右依次为低分辨率、中分辨率和高分辨率梅尔频谱图表示。

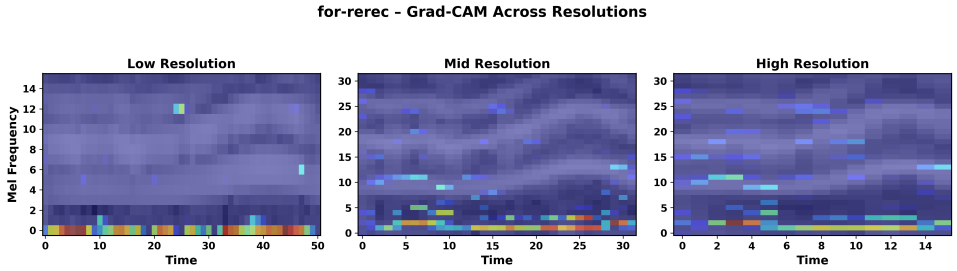

图 10：FoR 重录数据集跨分辨率的 Grad-CAM 可视化，展示了回放和信道效应引起的注意力转移。

3.4.2. 重录音频（*FoR*）上的 *Grad-CAM* 分析

图 10 展示了 FoR 重录子集在不同分辨率下的 Grad-CAM 激活图，该子集模拟了重放攻击和信道失真。与受控的 PA 条件相比，激活模式在空间上分布得更加分散，在瞬态片段中对中频和高频区域的关注也有所增加。这表明模型会调整其注意力，以适应播放—录制链路引入的伪影，包括麦克风染色、混响和频谱模糊。值得注意的是，低分辨率路径仍能提供稳定的上下文信息，而更高分辨率则捕捉信道引起的不一致性。

3.4.3. 真实世界条件下的可解释性（野外）

野外音频深度伪造数据集的可解释性结果如图 11 所示。与精选基准相比，真实世界样本在说话人、背景噪声、压缩伪影和录音设备方面表现出显著差异。

值得注意的是，该模型并不依赖孤立的谱峰，而是

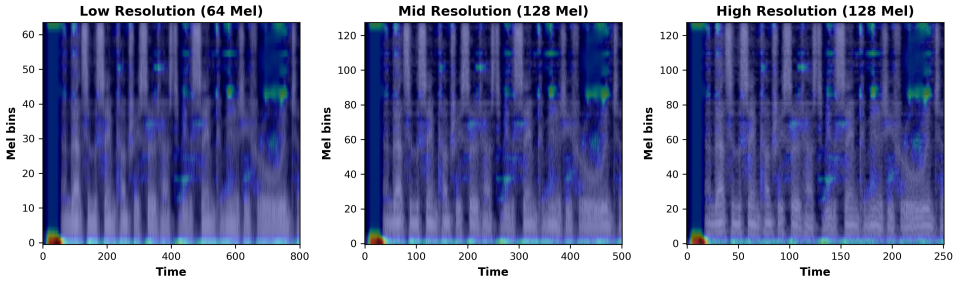

图 11：基于 Grad-CAM 的野外音频深度伪造数据集可解释性分析，展示了在时间和频率维度上分布式的注意力。

在较长时间跨度内汇总证据。高分辨率激活对于检测合成伪影仍然有用，而中分辨率表征则捕捉了在各种录音条件下持续存在的时间动态。在所有三个数据集中，都呈现出一致的趋势：没有任何单一分辨率主导决策过程。相反，所提出的框架会动态整合粗粒度时间上下文与精细频谱细节，从而能够在基于合成、基于回放以及真实场景深度伪造场景中实现稳健检测。

3.5. FoR 数据集上的消融分析

为分析所提出的分辨率感知框架中各个组件的贡献，以 FoR 数据集为例开展了消融研究。实验在三个 FoR 变体上进行——*for-2seconds*、*for-norm* 和 *for-rerecorded*——这些变体的难度逐步增加，从时间归一化信号到带有信道失真的重放音频。所有消融模型均遵循相同的训练和评估流程，并采用等错误率和 ROC 曲线下面积报告性能。

评估了四种模型变体：

- **完整模型**：所提出的完整框架，包含多分辨率输入、跨尺度注意力和跨分辨率一致性学习。
- **无一致性**：移除不同分辨率之间的一致性正则化。
- **单分辨率**：仅使用单一梅尔频谱图分辨率，不采用多尺度输入。
- **无注意力**：将跨尺度注意力替换为对不同分辨率进行简单平均。

表 9：FoR 数据集上的消融结果。较低的等错误率和较高的曲线下面积表示更好的性能。

数据集	模型变体	EER	AUC
for-2sec	完整模型	0.0057	0.993
	无一致性	0.0086	0.974
	单分辨率	0.0114	0.962
	无注意力	0.0427	0.953
for-norm	完整模型	0.0046	0.999
	无一致性	0.0071	0.999
	单分辨率	0.0094	0.998
	无注意力	0.0124	0.985
for-rerecorded	完整模型	0.0454	0.9900
	无一致性	0.0615	0.9814
	单分辨率	0.0846	0.8970
	无注意力	0.261	0.811

表 9 总结了三个 FoR子集上的消融结果。

在所有 FoR子集中都能观察到若干一致的趋势。首先，移除跨尺度注意力机制会导致性能下降最为严重，尤其是在回放的 *for-rerecorded* 条件下。这凸显了当欺骗攻击线索受到信道效应扭曲时，自适应分辨率加权的重要性。其次，将模型限制在单一分辨率会导致等错误率明显上升，这证实了欺骗伪迹分布在互补的时域和频域尺度上。随着声学条件变得不那么可控，性能下降会更加显著。最后，移除跨分辨率一致性学习会导致所有子集的性能出现中等程度但稳定的下降。这表明，一致性学习通过防止模型过度依赖分辨率特定伪影，并鼓励其在不同尺度上做出连贯的决策，从而提升鲁棒性。

4. 讨论

本研究探讨了显式跨分辨率建模在提升音频深度伪造检测系统鲁棒性与泛化能力方面的作用。在受控数据集、基于回放的数据集和真实世界数据集上的实验结果一致表明，对多个时频分辨率之间的交互进行建模，相较于传统的单分辨率方法或隐式融合方法具有显著优势。

实验评估得出了几项重要发现。首先，所提出的分辨率感知框架在 ASVspoof 2019 LA 等受控基准测试中取得了近乎满分的表现，证实了其检测现代文本转语音和语音转换系统所引入算法伪影的有效性。更重要的是，该方法在具有挑战性的基于回放的场景和真实世界条件下仍保持了出色的性能，ASVspoof 2019 PA、FoR 重录数据和野外音频深度伪造数据集上的结果也证实了这一点。这些场景的特点是存在信道失真、背景噪声以及未知的欺骗攻击方法，这些因素往往会导致现有检测器的性能严重下降。其次，对 FoR 数据集进行的消融分析表明，没有任何单一架构组件能够独自解释所观察到的性能提升。跨尺度注意力在动态加权分辨率特定特征方面发挥着关键作用，尤其是在存在重放引起的失真时；一致性学习则有助于稳定不同分辨率下的表征。移除这些组件后，性能下降会随着录音条件变得不那么受控而愈发明显，这凸显了它们对于提升鲁棒性的重要性，而不仅仅是改善干净数据上的表现。第三，可解释性分析为定量结果提供了定性证据。Grad-CAM 可视化结果表明，所提出的模型并不依赖孤立的谱峰或狭窄的频带，而是整合多个分辨率下的时频信息，并根据声学上下文调整其注意力。在回放场景和野外场景中，这一行为尤为明显，因为伪影在空间上分散，且与分辨率有关。

这项工作的主要创新在于，将音频深度伪造检测明确表述为一个跨分辨率学习问题。以往研究虽然探索过多特征或多尺度表示，但通常依赖拼接或池化等隐式融合策略。相比之下，本研究提出了一种有原则的分辨率感知框架：（i）通过跨尺度注意力显式建模多个频谱分辨率之间的交互；（ii）通过专门的正则化目标强制实现分辨率一致性。这种组合促使模型关注分辨率不变线索，使其对信道变化和重放效应更具鲁棒性。此外，所提出的架构在实现这些改进的同时，仍保持紧凑且计算高效。模型的可训练参数少于 2×10^5，复杂度低于十亿次浮点运算，表明在音频深度伪造检测中，鲁棒性与效率并非不可兼得。所提方法的一大优势是，在范围广泛的欺骗攻击条件下都能保持稳定的性能

这些条件涵盖了从干净的合成式攻击到不受约束的真实世界深度伪造。共享编码器既确保了参数效率，也缓解了过拟合；跨尺度注意力则能够自适应地整合特征，而不会带来过高的复杂度。学习到的表征具有可解释性，也是该方法的另一项显著优势。定量性能的提升与定性显著性模式相吻合，这让人相信模型学到的是有意义且可泛化的线索，而非利用数据集特定伪影。这一特性对于取证和安全应用尤为重要，因为这些领域要求结果可信且过程透明。

尽管所提出的框架具有诸多优点，但仍存在若干局限。首先，评估仅限于基于梅尔谱图的表征。尽管梅尔特征应用广泛且具有感知学依据，但其他表征（例如原始波形 [44], 基于相位的特征 [45], 或学习型滤波器组 [46], ）可能包含本研究尚未充分利用的互补信息。其次，一致性学习目标仅应用于真实语音，其前提是假设真实语音具有分辨率不变特征。尽管这一假设在实证中效果良好，但它未必适用于所有录音条件，尤其是在严重退化的环境中。

最后，尽管该模型在所评估的数据集上表现出良好的泛化能力，但生成式语音模型快速演进的特性可能会引入训练期间未曾出现的新伪迹 [47]。因此，持续进行适应和评估仍十分必要 [48, 49]。本研究揭示了若干值得探索的未来研究方向。将分辨率感知框架扩展至纳入其他特征模态（如相位信息或自监督语音表征），有望进一步提升鲁棒性。探索自适应或数据驱动的分辨率选择机制，也可能在保持性能的同时降低计算开销。另一个重要方向是整合域适应或持续学习策略，以应对不断演进的合成技术和录音条件 [50, 51]。最后，将所提出的框架应用于多模态深度伪造检测——联合分析音频与视觉或文本线索——是一个自然的拓展方向，具有重要的实际意义。

5. 结论

本文提出了一种具备分辨率感知能力的音频深度伪造检测框架，通过跨尺度注意力和一致性学习，显式利用互补的时频表征

。该方法通过联合建模多种频谱分辨率，解决了现有检测系统的一项关键局限：这些系统依赖单分辨率特征或隐式融合策略，因此往往难以适应信道失真和真实世界录音条件。在 ASVspoof 2019、Fake-or-Real（FoR）数据集和野外音频深度伪造基准上的全面评估表明，该方法在算法生成、回放式和不受约束的欺骗攻击场景中，均能实现出色且稳定的性能。该框架在受控基准测试中达到接近极限的准确率，同时在回放和真实世界条件下表现出更强的鲁棒性，且模型设计紧凑、计算效率高。

实验和可解释性分析进一步凸显了跨分辨率注意力和一致性正则化在捕捉分辨率不变的欺骗线索方面的重要性。这些组件使模型能够动态融合粗粒度时间上下文与细粒度频谱伪影，从而获得稳定且具有泛化能力的表征。总而言之，本研究提出了一种有原则且实用的稳健音频深度伪造检测方案，强调了显式跨分辨率建模对于在现实且不断变化的声学环境中部署的重要价值。

致谢

作者感谢开放研究社区公开提供 ASVspoof 2019、Fake-or-Real（FoR）和野外音频深度伪造数据集，使本研究得以开展。作者还感谢 Grammarly 和 OpenAI 的 ChatGPT 对语言润色提供的帮助。

竞争利益声明

作者声明，不存在可能影响本文所报告工作的已知竞争性财务利益或个人关系。

数据和代码可用性

本研究使用的数据集可从各自来源公开获取：ASVspoof 2019 数据集、Fake-or-Real（FoR）数据集以及野外音频深度伪造数据集。访问这些数据集

须遵守其原始提供方规定的条款与条件。为支持可复现性和未来研究，实现所提出的分辨率感知框架的代码（包括预处理、训练和评估脚本）将在论文获接收后公开发布。

参考文献

[1] O. A. Shaaban, R. Yildirim, A. A. Alguttar，音频深度伪造方法，《IEEE Access》11（2023）132652–132682。[2] A. Dixit, N. Kaur, S. Kingra，音频深度伪造检测技术综述：问题与前景，《Expert Systems》40（8）（2023）e13322。[3] T. Wang, X. Liao, K. P. Chow, X. Lin, Y. Wang，深度伪造检测：可靠性视角下的全面综述，《ACM Computing Surveys》57（3）（2024）1–35。[4] A. Chadha, V. Kumar, S. Kashyap, M. Gupta，深度伪造：概述，载于：第二届计算、通信与网络安全国际会议论文集：IC4S 2020，Springer，2021，第557–566页。[5] C. Sun, S. Jia, S. Hou, S. Lyu，利用神经声码器伪影检测人工智能合成语音，载于：IEEE/CVF计算机视觉与模式识别会议论文集，2023，第904–912页。[6]Y. Patel, S. Tanwar, R. Gupta, P. Bhattacharya, I. E. Davidson, R. Nyameko, S. Aluvala, V. Vimal，深度伪造生成与检测：案例研究与挑战，《IEEE Access》11（2023）143296–143323。[7]M. Morrison, R. Kumar, K. Kumar, P. Seetharaman, A. Courville, Y. Bengio，用于条件波形合成的分块自回归生成对抗网络，arXiv预印本 arXiv:2110.10139（2021）。[8] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, H. Malik，深度伪造生成与检测：最新进展、开放挑战、对策与未来方向，《Applied Intelligence》53（4）（2023）3974–4026。[9] K. Bhagtani, A. K. S. Yadav, P. Bestagini, E. J. Delp，扩散模型深度伪造语音生成器的归因，载于：2024年IEEE信息取证与安全国际研讨会（WIFS），IEEE，2024，第1–6页。

[10] B. Liu, B. Liu, T. Zhu, M. Ding，深度伪造及其检测综述：从生成对抗网络到扩散模型，《International Journal of Intelligent Systems》2025（1）（2025）9987535。[11] M. Rabhi, S. Bakiras, R. Di Pietro，音频深度伪造检测：对抗攻击与对策，《Expert Systems with Applications》250（2024）123941。[12] N. M. Müller, F. Dieckmann, J. Williams，音频深度伪造的攻击者归因，arXiv预印本 arXiv:2203.15563（2022）。[13] D. Bilika, N. Michopoulou, E. Alepis, C. Patsakis，你好，我的真实自我：针对语音助手的音频深度伪造攻击，arXiv预印本 arXiv:2302.10328（2023）。[14]R. Mubarak, T. Alsboui, O. Alshaikh, I. Inuwa-Dutse, S. Khan, S. Parkinson，关于视觉、音频和文本格式中深度伪造检测及其影响的综述，《IEEE Access》11（2023）144497–144529。[15] V. Danry, J. Leong, P. Pataranutaporn, P. Tandon, Y. Liu, R. Shilkrot, P. Punpongsanon, T. Weissman, P. Maes, M. Sra，人工智能生成角色：让深度伪造发挥积极作用，载于：CHI人机交互系统会议扩展摘要，2022，第1–5页。[16] H. Lee, C. Lee, K. Farhat, L. Qiu, S. Geluso, A. Kim, O. Etzioni，深度伪造生成与检测之间的拉锯战，arXiv预印本 arXiv:2407.06174（2024）。[17] N. Amezaga Vélez，语音深度伪造技术的普及及其带来的善恶影响（2021）。[18] M. Sareen，深度伪造技术带来的威胁与挑战，载于：《深度伪造》，CRC Press，2022，第99–113页。[19] A. Hery, O. Joseph, O. Femi, H. Luz，音频深度伪造：对语音助手和语音激活系统的威胁（2024）。[20] M. Mcuba, A. Singh, R. A. Ikuesan, H. Venter，深度学习方法对数字调查中深度伪造音频检测的影响，《Procedia Computer Science》219（2023）211–219。[21] M. Li, Y. Ahmadiadli, X.-P. Zhang，深度伪造音频检测中物理特征与感知特征的比较研究，载于：第一届音频多媒体深度伪造检测国际研讨会论文集，2022，第35–41页。

[22] N. Chakravarty，M. Dua，深度伪造音频检测的轻量级特征提取技术，Multimedia Tools and Applications 83 (26) (2024) 67443–67467。[23] P. Grinberg，A. Kumar，S. Koppisetti，G. Bharaj，音频深度伪造检测器关注什么？时域研究，载于：ICASSP 2025-2025 IEEE国际声学、语音与信号处理会议（ICASSP），IEEE，2025，第1–5页。[24] A. Jellali，I. Ben Fredj，K. Ouni，采用梅尔频率倒谱系数和频谱对比度混合方法，拓展音频深度伪造检测的边界，Multimedia Tools and Applications 84 (18) (2025) 20249–20268。[25] C. Bisogni，V. Loia，M. Nappi，C. Pero，面向可解释机器学习的音频欺骗检测声学特征分析，Computer Vision and Image Understanding 249 (2024) 104145。[26] N. Sandotra，B. Arora，基于特征的人工智能技术在深度伪造检测中的综合评估，Neural Computing and Applications 36 (8) (2024) 3859–3887。[27] M. Gohari，D. Salvi，P. Bestagini，N. Adami，歌声深度伪造检测的音频特征研究，载于：ICASSP 2025-2025 IEEE国际声学、语音与信号处理会议（ICASSP），IEEE，2025，第1–5页。[28] Z. Khanjani，G. Watson，V. P. Janeja，深度伪造有多深？聚焦音频深度伪造：综述，arXiv预印本 arXiv:2111.14203 (2021)。[29] K. Verma，D. Mittal，S. Samanta，K. Gulati，O. Kulkarni，M. A. Dar，C. Biji，音频深度伪造检测：先进深度学习模型的比较研究，IEEE Access (2025)。[30] B. Kumar，S. R. Alraisi，使用深度卷积神经网络的深度伪造音频检测技术，载于：2022年国际机器学习、大数据、云计算与并行计算会议（COM-IT-CON），第1卷，IEEE，2022，第463–468页。[31] A. Heidari，N. Jafari Navimipour，H. Dag，M. Unal，使用深度学习方法进行深度伪造检测：系统性全面综述，Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 14 (2) (2024) e1520。

[32] L. A. Passos，D. Jodas，K. A. Costa，L. A. Souza Júnior，D. Rodrigues，J. Del Ser，D. Camacho，J. P. Papa，面向深度伪造内容检测的基于深度学习的方法综述，Expert Systems 41 (8) (2024) e13570。[33] V. K. Sharma，R. Garg，Q. Caudron，深度伪造检测技术的系统性文献综述，Multimedia Tools and Applications 84 (20) (2025) 22187–22229。[34] P. Edwards，J.-C. Nebel，D. Greenhill，X. Liang，深度伪造技术综述：架构、检测与数据集，IEEE Access (2024)。[35] N. M. Müller，N. Evans，H. Tak，P. Sperl，K. Böttinger，更难还是不同？理解音频深度伪造检测的泛化能力，arXiv预印本 arXiv:2406.03512 (2024)。[36] X. Li，P.-Y. Chen，W. Wei，衡量音频深度伪造检测器的鲁棒性，arXiv预印本 arXiv:2503.17577 (2025)。[37] D. L. Jones，T. W. Parks，若干时频表征的分辨率比较，载于：国际声学、语音与信号处理会议，IEEE，1989，第2222–2225页。[38] K. Umapathy，B. Ghoraani，S. Krishnan，使用时频方法进行音频信号处理：编码、分类、指纹识别与水印，EURASIP 信号处理进展期刊 2010 (1) (2010) 451695。[39]A. Firc，K. Malinka，P. Hanáček，深度伪造语音检测评估框架：最先进深度伪造语音检测器的比较研究，Cybersecurity 8 (1) (2025) 50。[40] X. Wang，J. Yamagishi，M. Todisco，H. Delgado，A. Nautsch，N. Evans，M. Sahidullah，V. Vestman，T. Kinnunen，K. A. Lee，等，ASVspoof 2019：合成、转换和重放语音的大规模公共数据库，Computer Speech & Language 64 (2020) 101114。[41] R. Reimao，V. Tzerpos，FoR：用于合成语音检测的数据集，载于：2019年语音技术与人机对话国际会议（SpeD），2019，第1–10页。doi:10.1109/SPED.2019. 8906599。[42] N. M. Müller，P. Czempin，F. Dieckmann，A. Froghyar，K. Böttinger，音频深度伪造检测能否泛化？，Interspeech (2022)。

[43] W. Li, X. Xue，一种能够抵御随机裁剪的音频水印技术，Computer Music Journal 27 (4) (2003) 58–68。[44] G. Lee, J. Lee, M. Jung, J. Lee, K. Hong, S. Jung, Y. Han，双通道深度伪造音频检测：利用直达波形与混响波形，IEEE Access (2025)。[45] E. Prashnani, M. Goebel, B. Manjunath，基于相位的运动分析实现可泛化的深度伪造检测，IEEE Transactions on Image Processing (2024)。[46] L. Pfister, Y. Bresler，学习滤波器组稀疏变换，IEEE Transactions on Signal Processing 67 (2) (2018) 504–519。[47] S. Lyu，深度伪造检测：当前挑战与后续方向，arXiv 预印本 arXiv:2003.09234 (2020)。[48] B. Zhang, H. Cui, V. Nguyen, M. Whitty，音频深度伪造检测：已取得的成果与未来方向，Sensors（巴塞尔，瑞士）25 (7) (2025) 1989。[49] H. Dinh-Xuan, T.-P. Doan, S. Han, K. Hong, S. Jung，迈向资源受限环境中的实时音频深度伪造检测，载于：国际神经信息处理会议，Springer，2024，第 409–422 页。[50] X. Chen, W. Lu, R. Zhang, J. Xu, X. Lu, L. Zhang, J. Wei，面向音频深度伪造检测的持续无监督领域自适应，载于：ICASSP 2025-2025 IEEE 国际声学、语音与信号处理会议，IEEE，2025，第 1–5 页。[51] M. H. Kabir, A. A. Noman, A. A. Afiq, H. Jaki, M. N. Hasan, Ahmad, M. S. Islam, H. Hasan, I. A. Mojumder，深度伪造检测机器学习技术综述，Applied Computational Intelligence and Soft Computing 2025 (1) (2025) 8883527。