多头自注意力中的细粒度帧建模用于语音深度伪造检测

Phuong Tuan Dat[1,*] *Duc-Tuan Truong*[2,*] *Long-Vu Hoang*[1] *Nguyen Thi Thu Trang*[1†]

[1] 信息与通信技术学院，河内科学技术大学，越南[2] 计算与数据科学学院，南洋理工大学，新加坡

arXiv:2602.04702v1 [cs.SD] 4 Feb 2026

摘要

基于 Transformer 的模型在语音深度伪造检测中展现出了优异的性能，这在很大程度上得益于多头自注意力（MHSA）机制的有效性。MHSA 提供帧级注意力分数，这一点尤为重要，因为深度伪造伪迹通常出现在语音时间维度上的小范围局部区域。因此，细粒度帧建模对于准确检测细微的欺骗线索至关重要。在本研究中，我们针对基于 MHSA 的语音深度伪造检测提出了细粒度帧建模（FGFM）方法，首先通过多头投票（MHV）模块选出信息量最大的帧。随后，通过跨层细化（CLR）模块对这些选定的帧进行优化，以增强模型学习细微欺骗线索的能力。实验结果表明，我们的方法优于基线模型，并且在 LA21、DF21 和 ITW 数据集上分别取得了 0.90%、1.88% 和 6.64% 的等错误率（EER）。在多个基准测试中取得的稳定改进凸显了我们提出的细粒度建模方法在实现稳健语音深度伪造检测方面的有效性。

索引词— 深度伪造语音检测、多头自注意力、细粒度

1. 引言

文本转语音（TTS） [1] 和语音转换（VC） [2], 技术的进步推动了深度伪造语音的兴起，对银行业、电信和访问控制领域的生物识别安全系统构成了严峻威胁 [3], ，同时也带来了虚假信息和社会工程攻击等更广泛的风险 [4, 5]。为应对这些挑战，ASVspoof 系列赛事（2015–2021 年） [6] 以及近期的 VSASV 2024 挑战赛[7] 推动了合成语音检测（SSD）方法的发展 [8, 9, 10]。其中，基于 Transformer 的架构，尤其是采用 wav2vec 2.0 [11] 和 WavLM 等预训练自监督学习（SSL）模型的架构，已达到当前最先进水平 [12], 。这主要归功于多头自注意力（MHSA）机制，它能够捕捉帧级依赖关系，并且每个注意力头对声学模式 [13], 表现出不同的敏感度，因此这些模型非常适合检测合成语音中的伪迹。

尽管多头自注意力具有强大的建模能力，但它将注意力输出视为全局聚合特征，忽略了对合成语音检测至关重要的细粒度时间动态。在原始多头自注意力中，注意力分布于整个序列，但由于缺少突出或优先关注局部异常的显式机制，细微的欺骗线索可能会被稀释或忽略。先前研究表明，合成伪迹可能出现在时间上稀疏的短暂区域，例如不自然的过渡 [14] 或特定音素[15]。当帧级建模过于粗略，或在整个序列中隐式地进行平均时，这些局部异常可能无法被检测到。此外，近期研究表明，不同的注意力头专门用于捕捉不同类型的声学模式 [13], ，这表明头级选择或加权机制能够提高对欺骗伪迹的敏感度。

受这些观察结果的启发，我们提出了一种用于语音深度伪造检测的细粒度帧建模（FGFM）方法。所提出的 FGFM 首先应用多头投票（MHV），从每个多头自注意力（MHSA）头中选择高注意力帧，从而识别出用于欺骗检测的最显著细粒度帧。随后，通过跨层细化（CLR）模块在不同层之间聚合这些选定帧，并将细化后的信息整合到分类令牌中。尽管 MHV 和 CLR 模块最初是为细粒度视觉分类任务而提出的， [16], 我们通过展示增强选择性信息建模适合捕捉深度伪造语音所特有的细微且局部化伪影，证明了其在语音深度伪造检测中的更广泛适用性。通过在多个基准测试上的评估，我们提出的方法在 ASVspoof 2021 LA、DF 以及域外 In-the-Wild（ITW）数据集上均优于基线模型，等错误率分别相对降低了 7.2%、27.1% 和 21.1%。

*共同第一作者。 †通讯作者。

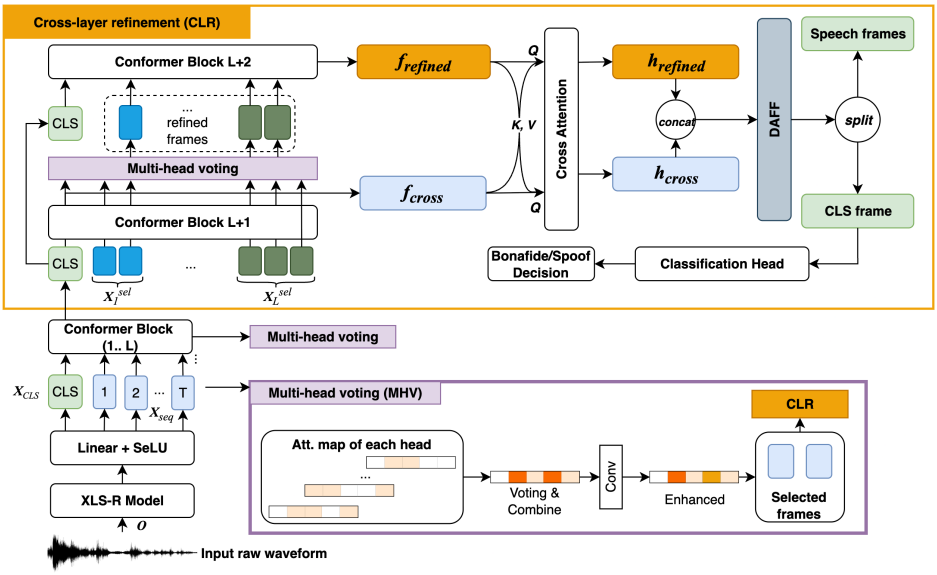


图 1。所提出的细粒度帧建模（FGFM）架构概览。

2. 方法

图 1 概述了我们提出的细粒度帧建模（FGFM）方法。我们首先介绍基线 XLSR-Conformer 模型， [17], 随后介绍所提出方法的组成部分：多头投票（MHV）和跨层细化（CLR）模块。

2.1. 基线模型

我们的基线是 XLSR-Conformer 模型 [18], ，该模型将预训练的 XLS-R 模型 [19] 作为特征提取器（FE），与 Conformer 模型相结合。给定输入信号 O，FE 输出 T 帧序列嵌入，并通过线性投影层将其投影到较低的 D维空间。将一个可学习的分类令牌 X_{CLS} 添加到投影后的输入嵌入 X_{seq}之前，形成编码器输入 X。该输入序列经过 L 基于多头自注意力的 Conformer块，每个块生成输出表示 X^l。最后，将最后一个块输出的更新后的分类令牌输入分类头，以判断输入是真实语音还是伪造语音。该模型构成了我们引入细粒度帧建模的基础。

2.2. 所提出的模型

如图 1 所示，我们提出的 FGFM 架构包含 MHV 和 CLR 模块。本节将介绍各模块的工作机制。

2.2.1. 多头投票（*MHV*）模块

每个 Conformer 块中的多头自注意力在时间维度上将输入划分为K 个注意力头，并将每个组送入每个头以计算自注意力。每个块l 生成 K 注意力图 $A^l = [A^1, ..., A^K]$（分类令牌与其他语音帧之间），$A^k \in \mathbb{R}^T$。然而，每个头的注意力图可以表示对不同区域判别线索的关注 [20]。为了减轻分类性能较差的头的影响，并自适应地学习从每个 Conformer 块中选择有价值的帧，我们提出使用 MHV 模块，将多头自注意力中的每个头视为装袋算法的弱学习器，以选择有价值的帧进行有效判别。

具体而言，对于从 1 到 L的所有 Conformer 块，MHV 根据 A^l中的每个注意力图进行投票，为每个注意力头选出 v 帧。选出的帧将在二值分数图 M^k中标记为 1，其余帧则标记为 0。最后，我们对 K个注意力头的分数图求和，得到$M' = \sum_{k=0}^{K} M^k$。为了进一步优化分数图，我们使用一维类高斯核进行卷积运算，得到优化后的分数图 $M^* = M' * \mathbf{G}$，其中$\mathbf{G}$ 见公式 1。对于每个 Conformer 块 l，我们从 MHV 模块中选出 M^* 中分数最高的 v 帧，其表征记为 $X_l^{sel} = [X_{l,1}^{sel}, ..., X_{l,v}^{sel}]$。

$$\mathbf{G} = \begin{bmatrix} 1 & 2 & 3 & 4 & 3 & 2 & 1 \end{bmatrix} \tag{1}$$

2.2.2. 跨层细化（*CLR*）模块。

虽然 MHV 能够从各个块中提取有价值的帧，但有效的欺骗检测还需要利用跨层信息。为此，我们引入 CLR模块。在第 $(L+1)$ 个 Conformer 块中，输入由分类令牌与所有

先前选取的帧拼接而成： $X_L^{in} = [X_L^{\text{CLS}}; X_1^{sel}; ...; X_L^{sel}]$。输出帧序列记为 $f_{\text{cross}} = [X_{L+1}^{\text{CLS}}; X_{L+1}^{sel}]$，它捕获了跨层聚合的全局判别线索。

为获得精细特征，我们再次对第（$L+1$）个块的输出应用 MHV 模块，并从 X_{L+1}^{sel}中选取帧。这些帧连同来自第 L 个块的分类令牌，被送入额外的第（$L+2$）个 Conformer 块： $X_{L+2}^{in} = [X_L^{\text{CLS}}; X_{L+1}^{sel}]$。该块的输出序列记为精细特征 $f_{\text{refined}} = [X_{L+2}^{\text{CLS}}; X_{L+2}^{sel}]$。为在 f_{cross}与 f_{refined} 之间交换信息，我们采用两个并行的交叉注意力操作，其输入分别为来自跨层特征与精细特征的查询向量（Q_{cross}, Q_{refined}）、键向量（$K_{\text{cross}}, K_{\text{refined}}$）和值向量（$V_{\text{cross}}, V_{\text{refined}}$），分别记为 f_{cross} 和 f_{refined}，并由线性层进行投影：

$$h_{\text{cross}} = \text{softmax}\left(\frac{Q_{\text{cross}} \cdot K_{\text{refined}}^T}{\sqrt{D}}\right) V_{\text{refined}} + f_{\text{cross}}$$

$$h_{\text{refined}} = \text{softmax}\left(\frac{Q_{\text{refined}} \cdot K_{\text{cross}}^T}{\sqrt{D}}\right) V_{\text{cross}} + f_{\text{refined}}$$

拼接后的输出 $[h_{\text{cross}}; h_{\text{refined}}]$ 随后经由轻量级 DAFF 块 [21], 处理，该块作为前馈聚合器，用于增强交叉相关性并强化最终的分类令牌。经过增强的分类令牌融合了层内（MHV）与跨层（CLR+DAFF）处理的信息，最终被送入分类头，以做出真实语音/欺骗的二分类决策。该设计使 FGFM 能够捕捉合成语音中的细粒度伪影。

3. 实验

3.1. 数据集和实验设置

所有模型均在 ASVspoof 2019 LA 训练集上训练。我们在三个基准测试上评估性能：ASVspoof 2021 LA（21LA）、ASVspoof 2021 DF（21DF）和 In-the-Wild（ITW）数据集 [22]。ASVspoof 2021 LA 和 DF 的语音数据受到各种编解码器及压缩变化的影响，而 ITW 数据集则因涵盖此前未见过的设备/信道不匹配情况以及自然环境下的变化而带来不同的挑战。我们使用等错误率（EER）作为主要指标报告性能。

我们采用与基线 Conformer 模型相同的模型配置 [18], ，并在 CLR模块中额外加入两个 Conformer 编码器块。对于 MHV 模块，我们将每个编码器块中每个注意力头选取的细粒度帧数设为 $v = 24$，该值由经验确定。训练设置与基线保持一致。所有实验均在单块 NVIDIA A40 GPU 上进行。

模型	等错误率（%）		
	21LA	**21DF**	**ITW**
XLSR-Conformer+TCM [18]†	1.18	2.25	7.79
XLSR-AASIST [23]	1.00	3.69	10.46
XLSR-AASIST2 [24]	1.61	2.77	-
WavLM-MFA [25]	5.08	2.56	-
XLSR-SLS [26]	5.08	<u>1.92</u>	7.46
XLSR-MoE [27]	2.96	2.54	9.17
XLSR-Mamba [28]	<u>0.93</u>	**1.88**	6.71
XLSR-Transformer	1.96	2.43	6.59
+ FGFM（我们的方法）	1.82	2.37	**6.31**
XLSR-Conformer [17]†	0.97	2.58	8.42
+ FGFM（我们的方法）	**0.90**	**1.88**	<u>6.64</u>

表1。在 ASVspoof 2021 评估集和 ITW 数据集上的最先进系统性能比较（† 表示复现结果，**粗体** 表示最佳结果，下划线表示次佳结果，“-”表示结果不可用）。

3.2. 与最先进模型的比较

表1报告了所提出模型的性能，以及与近期最先进模型的比较。所提出的 XLSR-MHV 在三个基准测试的等错误率（EER）方面均达到最先进水平，优于近期系统。与基线 XLSR-Conformer 相比，该模型在 21LA、21DF 和 ITW数据集上的 EER 分别相对降低了 7.2%、27.1% 和 21.1%。尽管其在 21DF 数据集上的表现与 XLSR-Mamba 不相上下，但所提出的模型在域外ITW数据集上超越了所有其他模型，取得了领先的 6.64% EER，展现出对未见欺骗条件的强泛化能力和鲁棒性。此外，我们将 Conformer块替换为 Transformer模块，以验证所提出模块的鲁棒性。结果表明，所提出的方法在基于 Conformer 和基于 Transformer 的模型上都带来了稳定的提升，平均相对提升 12.2%。这说明我们的方法有益于多种基于多头自注意力的架构，并且具有良好的泛化性。

3.3. 消融研究

为更好地了解所提出框架中各个组件的作用，我们基于 XLSR-Conformer基线开展了一系列消融研究。结果汇总于表2。首先，我们将基线系统中的 Conformer 块数量增加至与所提模型深度相同（$L = 6$）；所提模型包含四个原始 Conformer 块，以及 CLR模块中的两个额外块。结果表明，仅增加基线中的 Conformer 块数量，无法达到与我们的方法相当的性能，这说明我们的改进并非源于模型容量的增加。此

模型	等错误率 (%)		
	21LA	**21DF**	**ITW**
XLSR-Conformer（基线）	0.97	2.58	8.42
XLSR-Conformer（$L = 6$）	2.09	2.68	9.79
XLSR-Conformer-FGFM	**0.90**	**1.88**	**6.64**
-不含 DAFF	0.98	1.98	6.82
-无 MHV 增强	1.04	2.03	6.98

表2。XLSR-MHV 与基线模型在有无 DAFF 和增强模块时的性能。

每个头的投票	等错误率（%）		
	21LA	**21DF**	**ITW**
$v = 16$	1.34	2.04	7.04
$v = 24$	**0.98**	**1.98**	6.82
$v = 32$	1.69	2.27	6.73
$v = 40$	1.71	2.84	**6.66**

表3。MHV 中每个头的投票数的影响 v 。

外，通过系统地移除或修改各个组件，我们评估了它们对整体性能的影响，以凸显每项设计选择的有效性。

3.3.1. 各个所提模块的有效性

首先，我们移除 DAFF 模块，使跨层特征 f_{cross} 直接传递至分类头。此修改使模型在 21LA、21DF 和 ITW 数据集上的相对性能分别下降 8.2%、5.1% 和 2.7%。不过，模型的表现仍与 XLSR-Conformer 基线相当，而且在 ITW 数据集上明显优于该基线。这些结果表明，DAFF 模块能够将细粒度语音帧中的信息有效整合到分类标记中，从而提升性能。其次，我们移除 MHV 模块中采用高斯核卷积的增强操作。这使模型在 21LA、21DF 和 ITW 测试集上的相对性能分别下降 15.6%、8.0% 和 5.1%。这些发现证实，增强步骤有助于优化投票注意力图。

3.3.2. MHV 模块中每个注意力头投票选出的帧数

表3展示了在三个评估数据集上，对 MHV 模块中每个注意力头的投票数进行调整所开展的消融研究。投票数从16增加到24时，性能有所提升，在 LA21、DF21 和 ITW 上分别取得了0.98%、1.98%和6.82%的最佳等错误率。不过，进一步将投票数增加到32和40，会导致所有数据集上的性能下降。这种下降可能是由于纳入了信息量较低的帧，例如代表静音或非语音区域的帧，这些帧会引入噪声并降低判别能力。这些结果表明，对于4秒音频片段，每个注意力头选择24票能取得有效的平衡，既能捕捉关键语音帧，又能排除无关帧。

3.3.3. 模型在语音中关注的关键特征

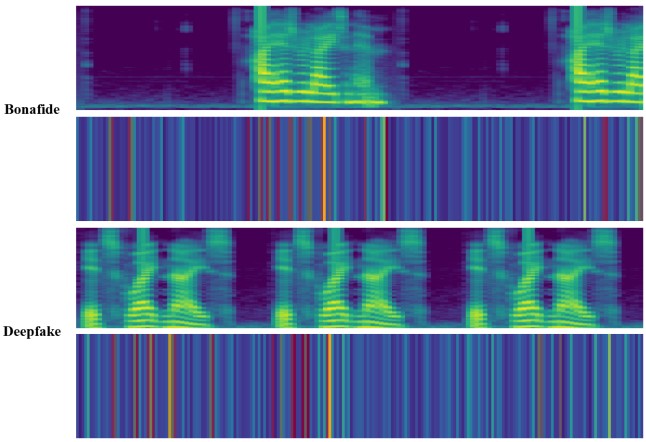


图 2。MHV 中选定帧（红色竖线）与真实语音（上）和深度伪造语音（下）样本频谱图的对照可视化。

图 2 对比了最终 Con-former 模块的注意力图与真实语音和深度伪造语音的频谱图。红线标记了 MHV 模块选出的帧，表明该模块始终关注语音丰富区域，同时避开静音或信息量较低的片段。对于这两种音频，选出的帧均与能量较高的语音丰富区域相对应。这种有针对性的关注至关重要，因为语音信号通常包含静音或信息量不足的区间，若将这些区间一视同仁，可能会削弱分类效果。

4. 结论

本文提出了一种用于语音深度伪造检测的细粒度帧建模方法，增强了基于多头自注意力的模型捕捉细微局部伪迹的能力。我们的框架融合了多头投票机制以筛选显著帧、跨层细化模块以聚合层次信息，以及动态聚合前馈（DAFF）模块，以利用上下文相关的重要特征丰富分类令牌。在 ASVspoof 2021 LA、DF 和 In-the-Wild 数据集上进行的大量实验表明，我们的方法始终优于强大的基线方法和具有竞争力的最先进系统。这些结果凸显了选择性信息建模对于提升语音深度伪造检测模型鲁棒性的有效性。

5. 致谢

本研究由越南教育与培训部资助，项目代码为 CT2025.EA.BKA.04。

6. 参考文献

[1] Y. Kumar、A. Koul 和 C. Singh，“文本转语音系统中的深度学习方法：系统综述与近期研究展望”，《多媒体工具与应用》，第 82 卷，2022年9月。[2] Wen-Chin Huang、Lester Violeta、Songxiang Liu、Jiatong Shi 和 Tomoki Toda，“2023年歌声转换挑战赛”，2023年12月，第 1–8 页。[3] Z. K. Anjum 和 R. K. Swamy，“说话人验证中的欺骗攻击与对策：综述”，载于*2017*年无线通信、信号处理与网络国际会议 *(WiSP-NET)*，2017年，第 467–471 页。[4] Z. Wu、P. L. De Leon、C. Demiroglu、A. Khodabakhsh、S. King、Z.-H. Ling、D. Saito、B. Stewart、T. Toda、M. Wester 和 J. Yamagishi，“文本无关说话人验证的反欺骗：初始数据库、对策比较及人类表现”，*IEEE/ACM*《音频、语音与语言处理汇刊》，第 24 卷，第 4 期，第 768–783 页，2016年。[5] Z. Wu、N. Evans、T. Kinnunen、J. Yamagishi、F. Alegre 和 H. Li，“说话人验证中的欺骗攻击与对策：综述”，《语音通信》，第 66 卷，第 130–153 页，2015年。[6] X. Liu、X. Wang、M. Sahidullah、J. Patino、H. Delgado、T. Kinnunen、M. Todisco、J. Yamagishi、N. Evans、A. Nautsch 和 K.A. Lee，“ASVspoof 2021：面向真实环境中的伪造语音与深度伪造语音检测”，*IEEE/ACM*《音频、语音与语言处理汇刊》，第 31 卷，第 2507–2522 页，2023年。[7] V. Hoang、V.T. Pham、H. N. Xuan、P. Nhi、P. Dat 和 T.T.T. Nguyen，“VSASV：用于防欺骗说话人验证的越南语数据集”，载于*Interspeech 2024*，2024年，第 4288–4292 页。[8] Duc-Tuan Truong、Yikang Wang、Kong Aik Lee、Ming Li、Hiromitsu Nishizaki 和 Eng Siong Chng，“面向深度伪造语音检测的引导式掩蔽数据增强研究”，载于自动说话人验证防欺骗对策研讨会（*ASVspoof 2024*），2024年，第 176–180 页。[9] Tianchi Liu、Duc-Tuan Truong、Rohan Kumar Das、Kong Aik Lee 和 Haizhou Li，“Nes2net：一种由基础模型驱动的轻量级嵌套架构，用于语音防欺骗”，*IEEE*《信息取证与安全汇刊》，第 20 卷，第 12005–12018 页，2025年。[10] Tuan Dat Phuong、Long-Vu Hoang 和 Huy Dat Tran，“借助 Kolmogorov-Arnold 网络和自监督学习模型提升合成语音检测性能”，载于*Interspeech 2025*，2025年，第 5633–5637 页。[11] A. Baevski、H. Zhou、A. Mohamed 和 M. Auli，“wav2vec 2.0：用于语音表征自监督学习的框架”，2020年。[12] 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，“WavLM：面向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》，第 16 卷，第 6 期，第 1505–1518 页，2022年。[13] K. Audhkhasi、Y. Huang、B. Ramabhadran 和 P. J. Moreno，“基于 Conformer 的自动语音识别中自注意力头多样性分析”，载于 *Interspeech 2022*，2022年，第 1026–1030 页。[14] Tianchi Liu、Lin Zhang、Rohan Kumar Das、Yi Ma、Ruijie Tao 和 Haizhou Li，“神经防欺骗对策如何检测部分伪造音频？”，载于*Interspeech2024*，2024年，第 1105–1109 页。[15] Davide Salvi、Viola Negroni、Sara Mandelli、Paolo Bestagini 和 Stefano Tubaro，“目标人物语音深度伪造检测的音素级分析”，2025年。[16] Q. Xu、J. Wang、B. Jiang 和 B. Luo，“通过内部集成学习变换器实现细粒度视觉分类”，*IEEE* 多媒体汇刊，第 25 卷，第 9015–9028 页，2023年。[17] E. Rosello、A. G.-Alanis、A. M. Gomez 和 A. Peinado，“用于反欺骗可变长度话语处理的基于 Conformer 的分类器”，载于*INTERSPEECH2023*，2023年，第 5281–5285 页。[18] D.T. Truong、R. Tao、T. Nguyen、H.T. Luong、K.A. Lee 和 E.S. Chng，“用于合成语音检测的多头自注意力时序-通道建模”，载于*Interspeech 2024*，2024年，第 537–541 页。[19] 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，“XLS-R：大规模自监督跨语言语音表征学习”，载于*Interspeech2022*，2022年，第 2278–2282 页。[20] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、L. Kaiser 和 I. Polosukhin，“注意力就是一切”，神经信息处理系统进展，第 30 卷，2017年。[21] Z. Lu、H. Xie、C. Liu 和 Y. Zhang，“弥合小型数据集上视觉变换器与卷积神经网络之间的差距”，载于神经信息处理系统进展，Alice H. Oh、Alekh Agarwal、Danielle Belgrave 和 Kyunghyun Cho 编，2022年。[22] N. M¨uller、P. Czempin、F. Diekmann、A. Froghyar 和 K. B¨ottinger，“音频深度伪造检测能否泛化？”，载于*Interspeech 2022*，2022年，第 2783–2787 页。[23] H. Tak、M. Todisco、X. Wang、J. Jung、J. Yamagishi 和 N. Evans，“使用 wav2vec 2.0 和数据增强进行自动说话人验证欺骗攻击与深度伪造检测”，2022年。[24] Y. Zhang、J. Lu、Z. Shang、W. Wang 和 P. Zhang，“使用 aasist2 改进短语音反欺骗”，载于*ICASSP 2024 - 2024*年 *IEEE* 声学、语音与信号处理国际会议 *(ICASSP)*，2024年，第 11636–11640 页。[25] Y. Guo、H. Huang、X. Chen、H. Zhao 和 Y. Wang，“结合自监督 WavLM 与多融合注意力分类器的音频深度伪造检测”，载于*ICASSP 2024 - 2024*年 *IEEE* 声学、语音与信号处理国际会议 *(ICASSP)*，2024年，第 12702–12706 页。[26] Q. Zhang、S. Wen 和 T. Hu，“结合自监督 XLS-R 和 SLS 分类器的音频深度伪造检测”，载于*ACM Multimedia 2024*，2024年。[27] Z. Wang、R. Fu、Z. Wen、J. Tao、X. Wang、Y. Xie、X. Qi、S. Shi、Y. Lu、Y. Liu、C. Li、X. Liu 和 G. Li，“使用冻结的 wav2vec 2.0 进行伪造音频检测的专家混合融合”，载于*ICASSP 2025 - 2025*年 *IEEE* 声学、语音与信号处理国际会议 *(ICASSP)*，2025年，第 1–5 页。[28] Y. Xiao 和 R. K. Das，“XLSR-Mamba：用于欺骗攻击检测的双列双向状态空间模型”，*IEEE*《信号处理快报》，第 32 卷，第 1276–1280 页，2025年。