SUPERB 风格的用于音频深度伪造检测的自监督语音模型基准

Hashim Ali、*Nithin Sai Adupa*、*Surya Subramani*、*Hafiz Malik*

密歇根大学电气与计算机工程系，
Evergreen 路 4901 号，迪尔伯恩，密歇根州
48128，美国

arXiv:2603.01482v1 [eess.AS] 2 Mar 2026

摘要

自监督学习已改变语音处理领域，SUPERB 等基准为各种下游任务建立了公平的比较标准。尽管音频深度伪造检测对安全至关重要，却一直未被纳入这些工作。在本文中，我们推出 Spoof-SUPERB，这是一个音频深度伪造检测基准，系统地评估了 20 种涵盖生成式、判别式和基于频谱图的架构的自监督学习模型。我们在多个域内和域外数据集上评估了这些模型。结果表明，XLS-R、UniSpeech-SAT 和 WavLM Large 等大规模判别式模型始终优于其他模型，这得益于多语言预训练、说话人感知目标和模型规模。我们还分析了这些模型在声学退化条件下的稳健性，结果显示生成式方法的性能急剧下降，而判别式模型则能保持韧性。该基准建立了可复现基线，并提供了实用见解，指出哪些自监督学习表征最可靠，有助于保护语音系统免受音频深度伪造的威胁。

索引词— 语音自监督学习，音频深度伪造检测，模型泛化，基准

1. 引言

文本转语音（TTS）和语音转换（VC）技术的飞速发展，使得创建高度逼真的音频深度伪造内容成为可能。尽管这些技术为无障碍服务和内容创作带来了机遇，但其滥用却威胁着安全、信任和取证可靠性。欺骗攻击已被用于破坏自动说话人验证（ASV）系统并传播虚假信息。因此，可靠的音频深度伪造检测已成为亟待研究的问题。

自监督学习（SSL）改变了语音处理的格局。wav2vec 2.0[1], HuBERT [2], WavLM [3], 和 XLS-R [4] 等模型从大型无标注语料库中学习通用表示，并在广泛的任务上取得了最先进的性能。为系统地评估自监督学习模型，人们推出了 SUPERB [5], SUPERB-SG [6], MiniSUPERB [7] 和 ML-SUPERB [8] 等基准，并为各类下游任务制定了标准化协议。然而，这些基准都未考虑音频深度伪造检测。另一系列研究表明，自监督学习模型在反欺骗方面很有潜力。Tak 等人 [9] 展示了 wav2vec 2.0 在 ASVspoof 2021[10] 上的有效性。Zhang 等人 [11]和 Xiao 和 Das [12]在 ASVspoof 21 DF 和 In-the-Wild 数据集上取得了最先进的结果。尽管这些研究展现了自监督学习模型在音频深度伪造检测方面的巨大潜力，但现有研究在模型、数据集和协议方面仍较为零散，导致结果难以比较。

在本研究中，我们通过推出首个用于音频深度伪造检测的 SUPERB 风格基准来填补这一空白。我们在统一的协议下评估了多种自监督学习模型：（i）采用所有 Transformer 层加权和的冻结自监督学习前端；（ii）作为后端的简单全连接分类器；（iii）在多个欺骗数据集上进行评估，包括 ASVspoof 2019（ASV19） [13],ASVspoof 2021（ASV21）LA/DF [10], DeepfakeEval（DFE-val） 2024 [14], In-the-Wild（ITW） [15], Famous Figures [16],和 ASVSpoof 洗白数据库（ASVSpoofLD） [17]。我们的贡献主要有三点：

1. 我们首次提出了一个可复现的音频深度伪造检测自监督学习模型排行榜。2. 我们分析了架构、预训练数据和训练目标如何影响检测性能。3. 我们首次对自监督学习模型在声学劣化条件下（如编解码器、混响和噪声）进行了系统比较。

2. 相关工作

2.1. 语音处理中的自监督学习基准

自监督学习的兴起催生了若干面向通用语音表示-的大规模基准

表示。SUPERB框架 [5] 规范了自动语音识别、说话人识别和情感识别等任务的评估，并随后得到扩展：SUPERB-SG面向语义与生成式任务 [6], , MiniSUPERB用于轻量级评估 [7], , ML-SUPERB用于多语言建模 [8], , TS-SUPERB用于目标说话人场景[18]。这些工作建立了自监督学习模型的可复现比较，但均未涉及音频深度伪造检测。

2.2 音频深度伪造检测的自监督学习

与基准开发同步，越来越多的研究将自监督学习模型应用于音频深度伪造检测。Tak 等人 [9] 首先表明，经过微调和数据增强的 wav2vec 2.0 在 ASVspoof 2021 上表现出色。随后，ASVspoof 5 挑战赛 [19]确立了自监督学习嵌入作为先进系统的骨干：Rohdin 等人 [20] 评估了自监督学习特征的池化策略，Chan 等人 [21]融合了 WavLM-ResNet18-SA，Chen 等人 [22]则将 wav2vec2 嵌入与手工特征相结合。SAFE 挑战赛 [23] 将评估扩展到多语言和对抗性条件，在这些条件下，表现最强的系统仍依赖自监督学习，例如混合专家框架 [24],和 WavLM+MAE-AST-Frame 多语言模型 [25]。

除了这些挑战赛之外，近期研究还开发了专门的自监督学习框架。Zhang 等人提出了一种结合 XLS-R 的敏感层选择（SLS）分类器 [11],，在 ASVspoof 2021 DF 数据集上首次取得低于 2% 的等错误率，并在 In-the-Wild 上表现出色。Xiao 和 Das 提出了 XLSR-Mamba [12], , 将 wav2vec2 与双列双向状态空间模型相结合，以实现高效检测。总体而言，这些研究表明，尽管基于自监督学习的系统代表了音频深度伪造检测领域的先进水平，但由于数据集、后端和评估协议各不相同，相关比较结果仍较为零散。

尽管取得了进展，目前仍没有类似 SUPERB、用于音频深度伪造检测的统一基准。现有研究分别评估自监督学习模型，导致结果难以比较。ASVspoof 5 和 SAFE 等挑战赛，以及 XLSR-SLS 和 XLSR-Mamba 等近期系统，都提升了性能，但评估设置并不一致。目前仍缺少受控分析，以揭示：（i）哪些预训练选择的迁移效果最佳；（ii）模型设计如何影响检测；（iii）哪些自监督学习表征最能抵御声学退化。为此，我们推出一个 SUPERB 风格的音频深度伪造检测基准，依据统一协议，对广泛的自监督学习模型进行可复现的评估。

3. 方法：Spoof-SUPERB

本节将详细介绍**Spoof-SUPERB** 框架，该框架将 SUPERB 基准测试方法 [5–7] 应用于音频深度伪造检测任务。Spoof-SUPERB 统一了在 ASVspoof 2019 [13] 上的训练，并在多个欺骗数据集上评估自监督学习模型，从而能够公平、可复现地比较各类表征。

3.1. 自监督学习（SSL）语音模型

语音自监督学习发展迅速，产生了多种具有不同预训练目标和规模的模型系列。遵循 SUPERB 基准[5], , 我们根据学习目标将自监督学习模型分为三大类：**生成式、判别式**和**混合式** 模型。这种以目标为导向的分类法有助于分析不同的预训练信号如何影响音频深度伪造检测中的表征鲁棒性。

生成式模型：这些模型通过重建输入信号的部分内容来学习表征，例如预测未来帧或还原被遮蔽的片段。早期的例子包括 APC（自回归预测编码） [26]和 VQ-APC（带量化的 APC） [27]。Mockingjay[28]和 TERA [29] 通过遮蔽频谱图中的片段扩展了这一方法；DeCoAR 2.0 [30] 将双向编码与重建相结合；NPC（非自回归预测编码） [31]则对遮蔽卷积进行了改进。这些模型早于当今占主导地位的判别式架构。

判别式模型：这些模型通过区分目标样本（正样本）与干扰样本（负样本）来学习表征，通常采用对比式或预测目标。Wav2vec [32] 和 Modified CPC [33]引入了对比预测编码。wav2vec 2.0 [1]通过遮蔽和量化改进了这一方法，而 HuBERT [2], 、WavLM [3], 、UniSpeech-SAT [34], 、Data2Vec [35], 和 XLS-R [4] 等大规模后继模型，则通过更大的语料库、多语言扩展，以及教师回归或对抗式说话人训练等训练目标，拓展了这一范式。

混合模型： 另一条研究路线是在预训练期间，将掩码重建与判别式学习目标相结合。SSAST [36] 和 MAE-AST [37]等模型采用补丁级掩码和基于 Transformer 的架构，同时利用重建损失和判别式目标。滤波器组特征则用作非自监督学习基线，以供比较。

3.2. 实验设置

遵循 SUPERB [5], , 我们在所有下游训练中固定上游自监督学习模型的参数。针对每个话语，提取所有层的帧级隐状态，并使用可训练的加权求和机制将其聚合为上下文化表示。这些表示被投影到较低维空间（256 个单元），并通过均值池化在话语级别进行池化。池化后的向量输入一个轻量级分类器，该分类器由带有 ReLU激活和丢弃的线性层构成，用于

表1。所研究的自监督学习表征详情。**缩写：**F=未来，M=掩码，G=生成，C=对比，P=词元预测，VQ=向量量化，LS=LibriSpeech，LL=LibriLight，GS=GigaSpeech，VP=VoxPopuli，AS=AudioSet，MLS=多语言LS，CV=CommonVoice，BBL=BABEL，G18=Googlei18n

模型	类别	网络	参数	步幅	输入	语料库	预训练
滤波器组特征	基线	–	–	10毫秒	波形	–	–
APC	生成式	三层门控循环单元	411万	10ms	滤波器组特征	LS-360小时	F-G
VQ-APC	生成式	三层门控循环单元	463万	10ms	滤波器组特征	LS-360小时	F-G + 向量量化
NPC	生成式	4-卷积层，4-掩码卷积	19.38M	10ms	滤波器组特征	LS-360小时	M-G (+向量量化)
Mockingjay	生成式	12层Transformer	∼85M	10ms	滤波器组特征	LS-960h	时间遮蔽
TERA	生成式	3-Transformer层	∼21M	10ms	滤波器组特征	LS-960h	时频 M-G
DeCoAR 2.0	生成式	一维卷积 + 12-Transformer层	–	10ms	滤波器组特征	LS-960h	时间遮蔽（+向量量化）
Wav2vec	判别式	19层卷积	3254万	10毫秒	波形	LS-960h	F-C
wav2vec 2.0 Base	判别式	7层卷积, 12层Transformer	95.04M	20 毫秒	波形	LS-960h	M-C + 向量量化
wav2vec 2.0 Large	判别式	7层卷积，24层Transformer	317.38M	20 毫秒	波形	LL-60kh	M-C + 向量量化
HuBERT Base	判别式	7层卷积，12层Transformer	94.68M	20 毫秒	波形	LS-960h	M-P + 向量量化
HuBERT Large	判别式	7层卷积，24层Transformer	316.61M	20 毫秒	波形	LL-60kh	M-P + 向量量化
MR-HuBERT	判别式	7层卷积，24层Transformer	-	20 毫秒	波形	LS-960h / LL-60kh	M-P + 向量量化（多分辨率）
XLS-R	判别式	7层卷积，24层Transformer	300M	20 毫秒	波形	VP+MLS+CV+BBL	M-C +向量量化
UniSpeech-SAT	判别式	7层卷积，24层Transformer	316M	20 毫秒	波形	LL-60kh+GS+VP	M-P + 说话人对抗训练
Data2Vec	判别式	7层卷积，12层Transformer	95M	20 毫秒	波形	LS-960 小时	EMA 教师回归
WAVLABLM	判别式	7层卷积，24层Transformer	316.61M	20 毫秒	波形	CV+VP+MLS+G18	M-P + 去噪
WavLM Large	判别式	7层卷积，24层Transformer	317M	20 毫秒	波形	LS-960h+GS+VP	M-P + 降噪/混音
SSAST	混合式	视觉 Transformer（12层Transformer）	89M	–	对数梅尔	AS+LS-960h	patch M-G + C
MAE-AST-FRAME	混合式	视觉 Transformer（12层Transformer）	89M	–	对数梅尔	AS+LS-960h	patch M-G + C

生成二元欺骗/真实预测。所有实验均采用固定的随机种子单次运行，以确保不同自监督学习模型之间的可比性。

所有自监督学习模型均在 ASV19 LA 训练集 [13]，上进行训练，该语料库已被反欺骗社区广泛用作参考语料库。Spoof-SUPERB 的目标并非通过训练数据选择来优化检测性能，而是在固定的下游设置下，对自监督学习表征进行公平且受控的比较。评估涵盖多个欺骗数据集，包括 ASV19 LA 评估集 [13], ASV21 LA/DF [10], ITW [38], DFEval 2024 [14], Famous Figures[16], 和 ASVSpoofLD [17]。这种固定训练与跨数据集评估方案能够对自监督学习泛化能力进行受控评估，并使用等错误率（EER）报告性能。尽管 XLS-R 和 WavLM 的预训练数据与 ASVspoof 5（该数据集源自 MLS）可能存在重叠，但这两个模型在多个评估数据集上均表现出稳定的优异性能。

4. 结果与分析

4.1. 整体性能

表2报告了所有自监督学习模型在各评估数据集上的等错误率（EER），并列出FBANK基线以供参考。基于自监督学习的模型始终优于FBANK，后者的平均EER为46.5%。相比之下，多个自监督学习系统将平均EER降至25%以下。表2以粗体突出显示了表现最好的5个模型，其中这些模型分别为XLS-R（17.4%）、UniSpeech-SAT（19.5%）、WavLM Large（20.6%）、HuBERT Large（22.7%）和 MR-HuBERT（23.0%）。这些结果证实，大规模判别式自监督学习模型在音频深度伪造检测方面具有良好的有效性。较早的生成式模型，如APC（31.7%）、TERA（36.1%）和 Mockingjay（35.4%），仍明显落后于判别式方法。

4.2. 模型因素分析

在所有自监督学习模型中，XLS-R 和 UniSpeech-SAT 的整体性能最强（平均等错误率分别为 17.8% 和 19.5%）。我们认为，XLS-R 的成功归功于其大规模多语言预训练：它在数十种语言的超过 40 万小时语音数据上进行训练，因此在不匹配的测试条件下具有更强的鲁棒性。UniSpeech-SAT 采用说话人感知预训练方法，鼓励保留说话人身份，并在 HuBERT 模型的基础上提出了两种方法。首先，它提出了话语级对比学习，通过增强单说话人信息提取来改进说话人验证和说话人识别等下游任务。其次，它提出了话语混合增强，有助于说话人分离等多说话人任务。

表现最佳的模型均为经过掩码预测训练的大规模判别式自监督学习模型。XLS-R、UniSpeech-SAT、WavLM Large、MR-HuBERT 和 HuBERT Large 占据排行榜前列，平均等错误率低于 25%。模型规模发挥着核心作用，大规模版本的表现始终优于对应的基础版本。多语言预训练（XLS-R、UniSpeech-SAT）则能让

表2。自监督学习模型在各评估数据集上的等错误率（EER，%）。每个数据集列中，**表现最佳的模型** 以粗体显示。在平均等错误率 列中，**排名前五的模型** 以粗体突出显示。

模型	ASV19 LA	ASV21 LA	ASV21 DF	ASV5 Eval In-the-Wild		DFEval 2024 Famous Fig.		ASVspoofLD	平均等错误率
滤波器组特征	42.828	43.155	44.789	49.838	48.393	47.113	48.427	47.672	46.527
APC	10.075	16.335	22.276	33.311	36.889	42.662	58.402	34.345	31.787
VQ-APC	12.155	18.872	20.217	30.581	34.860	52.173	58.544	31.799	32.400
NPC	15.243	17.619	25.239	37.868	40.986	49.843	51.979	29.758	33.567
Mockingjay	15.430	19.798	25.312	40.217	35.848	49.800	40.975	56.033	35.427
Mockingjay-960h	13.801	25.525	22.584	37.866	52.387	52.130	49.953	59.283	39.191
TERA	9.112	26.572	17.254	35.656	39.894	54.251	49.282	57.565	36.198
DeCoAR 2.0	7.628	12.352	18.990	29.571	35.029	49.800	54.452	22.126	28.743
wav2vec	8.812	15.500	14.761	30.691	42.239	53.895	51.048	36.263	31.651
wav2vec 2.0 Base	4.661	11.452	10.046	18.698	40.945	56.981	51.921	32.891	28.449
wav2vec 2.0 Large	7.695	18.887	11.617	19.956	40.461	55.764	44.401	30.413	28.649
HuBERT Base	4.867	12.562	13.387	23.990	27.276	53.747	53.749	17.772	25.919
HuBERT Large	2.788	10.049	11.996	21.252	21.039	52.991	48.440	13.146	**22.712**
MR-HuBERT	2.478	9.074	11.635	23.056	23.799	49.696	52.720	11.645	**23.006**
XLS-R	1.985	14.096	**4.314**	**14.394**	20.073	45.392	**29.598**	**9.420**	**17.409**
UniSpeech-SAT	**1.961**	**8.818**	7.443	14.996	**16.791**	49.800	46.601	9.557	**19.496**
Data2Vec	7.695	11.877	16.511	26.773	29.249	50.808	53.092	16.418	26.678
WAVLABLM	3.631	15.380	9.847	21.115	23.402	52.530	52.660	15.500	24.258
WavLM Large	2.273	11.636	11.527	17.549	24.331	49.696	35.367	12.089	**20.558**
SSAST	11.693	24.935	22.909	31.186	47.113	**40.184**	36.885	21.523	29.553
MAE-AST-FRAME	7.685	19.554	17.001	27.295	43.645	47.974	35.214	19.978	27.293

模型接触更广泛的语音和声学变化，从而进一步提升性能。最后，先进的预训练策略也能强化用于音频深度伪造检测的表征，例如 UniSpeech-SAT 中的说话人感知目标，以及 MR-HuBERT 中的掩码区域建模。

4.3. 声学退化条件下的鲁棒性

为评估鲁棒性，我们在 ASVspoofLD 语料库 [17], 的一个子集上评估了所有系统，该子集包含受到多人语音噪声（信噪比 = 10 dB）和混响（RT60 = 6s）破坏的 ASV19 LA 音频文件。表2（ASVSpoofLD 列）中的结果表明，自监督学习表示在处理这些退化时存在显著差异。最具韧性的系统是 XLS-R（9.4%，较 ASV19 LA 相差 +7.4%）、UniSpeech-SAT（9.5%，较 ASV19 LA 相差 +7.6%）、MR-HuBERT（11.6%，+9.2%）、WavLM Large（12.1%， +9.8%）以及 HuBERT Large（13.1%，+10.4%）。相比之下，TERA（57.6%， +31%）和 Mockingjay-960h（59.3%， +34%）等生成式系统在退化条件下性能崩溃。

我们进一步分析了编解码器条件（ASV5）下的性能。表 3 报告了代表性模型在所有编解码器条件（C01–C11）下的平均等错误率。判别式自监督学习模型（XLS-R、UniSpeech-SAT）再次展现出最佳鲁棒性，平均等错误率低于 20%。

5. 结论

在本研究中，我们推出了 Spoof-SUPERB，这是一个采用 SUPERB 风格的音频深度伪造检测基准。通过评估 20 种自监督学习模型，我们深入了解了哪些表征

表 3。所有编解码器条件下的平均等错误率（%）（ASV5 评估）。各类别的代表性模型

模型	平均编解码器等错误率
滤波器组特征（基线）	49.8
APC（生成式）	33.3
XLS-R（判别式）	13.5
UniSpeech-SAT（判别式）	14.0
WavLM Large（判别式）	18.1
SSAST（混合式）	28.8

最适合迁移到欺骗检测任务。结果表明，XLS-R、UniSpeech-SAT 和 WavLM Large 等大规模判别式模型的表现始终优于生成式方法，并且在噪声、混响和编解码器条件下具有更强的鲁棒性。未来的工作将通过扩大噪声退化分析范围，并考察自监督学习模型在不同合成方法下的表现，进一步了解模型对特定生成技术的脆弱性，从而推动开发更稳健的语音对抗措施，以应对不断演变的深度伪造威胁。

参考文献

[1] A. Baevski、Y. Zhou、A. Mohamed 和 M. Auli，《Wav2vec 2.0：语音表示自监督学习框架》，神经信息处理系统进展，第33卷，第12 449–12 460页，2020年。[2] W.-N. Hsu、B. Bolte、Y.-H. H. Tsai、K. Lakhotia、R. Salakhutdinov 和 A. Mohamed，《HuBERT：通过隐藏单元掩码预测进行语音表示自监督学习》，IEEE/ACM 音频、

语音与语言处理汇刊，第29卷，第3451–3460页，2021年。[3] S. Chen 等人，《WavLM：面向全栈语音处理的大规模自监督预训练》，2021年。[4] A. Babu 等人，《XLS-R：大规模自监督跨语言语音表示学习》，arXiv预印本 arXiv:2111.09296，2021年。[5] S. wen Yang 等人，《SUPERB：语音处理通用性能基准》，载于 Interspeech 2021 论文集，2021年，第1194–1198页。[6] H.-S. Tsai 等人，《SUPERB-SG：增强语义与生成能力的语音处理通用性能基准》，arXiv:2203.06849 [cs],2022年3月。[7] Y.-H. Wang、H.-Y. Chen、K.-W. Chang、W. Hsu 和 H.-Y. Lee，《Mini-SUPERB：轻量级语音自监督学习模型基准》，载于2023年 IEEE自动语音识别与理解研讨会（ASRU），2023年12月，第1–8页。[8] J. Shi 等人，《ML-SUPERB：多语言语音通用性能基准》，arXiv:2305.10615 [cs], 2025年2月。[9] H. Tak、M. Todisco、X. Wang、J.-w. 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与深度伪造检测》，载于说话人与语言识别研讨会，2022年。[10] X. Liu 等人，《ASVspoof 2021：迈向真实环境中的欺骗与深度伪造语音检测》，英文，IEEE/ACM 音频、语音与语言处理汇刊，第31卷，2023年，第2507–2522页。[11] Q. Zhang、S. Wen 和 T. Hu，《采用自监督 XLS-R 与 SLS分类器的音频深度伪造检测》，英文，载于第32届 ACM国际多媒体会议论文集，澳大利亚维多利亚州墨尔本：ACM，2024年10月，第6765–6773页。[12] 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。[13] X. Wang 等人，《ASVspoof 2019：合成、转换和重放语音的大规模公共数据库》，《计算机语音与语言》，第64卷，第101 114页，2020年。[14] N. A. Chandra 等人，《Deepfake-Eval-2024：2024年传播的真实环境深度伪造内容多模态基准》，arXiv预印本 arXiv:2503.02857, 2025年。[15] N. M. M¨uller、P. Czempin、F. Dieckmann、A. Froghyar 和 K. B¨ottinger，《音频深度伪造检测能否泛化？》，arXiv:2203.162[cs, eess], 2022年4月。[16] H. Ali、S. Subramani、R. Varahamurthy、N. Adupa、L. Bollinani 和 H. Malik，《为知名人物收集、整理并标注高质量语音深度伪造数据集：流程与挑战》，arXiv预印本 arXiv:2507.00324，2025年。[17] H. Ali、S. Subramani、S. Sudhir、R. Varahamurthy 和 H. Malik，《音频欺骗检测能否抵御洗白攻击？》，载于2024年 ACM 信息隐藏与多媒体安全研讨会论文集，2024年，第283–288页。[18] J. Peng 等人，《TS-SUPERB：面向语音自监督学习模型的目标语音处理基准》，载于ICASSP 2025 - 2025年 IEEE国际声学、语音与信号处理会议（ICASSP），国际标准连续出版物号：2379-190X，2025年4月，第1–5页。[19] X. Wang 等人，《ASVspoof 5：大规模众包语音数据、深度伪造内容与对抗攻击》，载于自动说话人验证欺骗对策研讨会（ASVspoof 2024），2024年，第1–8页。[20] J. Rohdin 等人，《ASVspoof 5 挑战的 BUT系统与分析》，arXiv预印本 arXiv:2408.11152，2024年。[21] P.-C. Chan、W.-Y. Chen 和 J.-C. Wang，《增强 ASVspoof 5 研讨会 2024 中的欺骗检测：融合 WavLM-ResNet18-SA，以实现针对语音深度伪造的最佳性能》，英文，载于《自动说话人验证欺骗对策研讨会（ASVspoof 2024）》，ISCA，2024年8月，第 158–162 页。[22] Y. Chen 等人，《USTC-KXDIGIT 系统对 ASVspoof5 挑战赛的描述》，英文，载于《自动说话人验证欺骗对策研讨会（ASVspoof 2024）》，ISCA，2024年8月，第 109–115 页。[23] T. Kirill 等人，《Safe：合成音频取证评测挑战赛》，载于《ACM 信息隐藏与多媒体安全研讨会论文集》，2025年，第 174–180 页。[24] V. Negroni、D. Salvi、A. I. Mezza、P. Bestagini 和 S. Tubaro，《利用专家混合提升语音深度伪造检测》，载于《ICASSP 2025——2025年 IEEE 国际声学、语音与信号处理会议（ICASSP）》，国际标准连续出版物号：2379-190X，2025年4月，第 1–5 页。[25] H. Ali、S. Subramani、L. Bollinani、N. S. Adupa、S. El-Loh 和 H. Malik，《面向稳健音频深度伪造检测的多语言数据集整合策略：Safe 挑战赛系统》，arXiv 预印本 arXiv:2508.20983，2025年。[26] Y.-A. Chung、W.-N. Hsu、H. Tang 和 J. Glass，《用于语音表示学习的无监督自回归模型》，arXiv 预印本 arXiv:1904.03240，2019年。[27] 《向量量化自回归预测编码》，载于 Interspeech，2020年。[28] D. B. T. Encoders，《Mockingjay：利用……进行无监督语音表示学习》，2020年。[29] A. T. Liu、S.-W. Li 和 H.-y. Lee，《TERA：面向语音的 Transformer 编码器表示自监督学习》，IEEE/ACM《音频、语音与语言处理汇刊》，第29卷，第 2351–2366 页，2021年。[30] S. Ling 和 Y. Liu，《DeCoAR 2.0：结合向量量化的深度上下文化声学表征》，arXiv 预印本 arXiv:2012 06659，2020年。[31] A. H. Liu、Y.-A. Chung 和 J. Glass，《利用局部依赖关系学习语音表征的非自回归预测编码》，arXiv 预印本 arXiv:2011.00406，2020年。[32] S. Schneider、A. Baevski、R. Collobert 和 M. Auli，《Wav2vec：面向语音识别的无监督预训练》，arXiv 预印本 arXiv:1904.05862，2019年。[33] M. Riviere、A. Joulin、P.-E. Mazar´e 和 E. Dupoux，《无监督预训练可有效迁移至不同语言》，载于《ICASSP 2020——2020年 IEEE 国际声学、语音与信号处理会议（ICASSP）》，IEEE，2020年，第 7414–7418 页。[34] S. Chen 等人，《UniSpeech-SAT：结合说话人感知预训练的通用语音表示学习》，载于《ICASSP 2022——2022年 IEEE 国际声学、语音与信号处理会议（ICASSP）》，IEEE，2022年，第 6152–6156 页。[35] A. Baevski、W.-N. Hsu、Q. Xu、A. Babu、J. Gu 和 M. Auli，《Data2Vec：用于语音、视觉与语言自监督学习的通用框架》，arXiv 预印本 arXiv:2202 03555，2022年。[36] Y. Gong、C.-I. Lai、Y.-A. Chung 和 J. Glass，《SSAST：自监督音频频谱图 Transformer》，载于《AAAI 人工智能会议论文集》，第36卷，2022年，第 10 699–10 709 页。[37] A. Baade、P. Peng 和 D. Harwath，《MAE-AST：掩码自编码音频频谱图 Transformer》，arXiv 预印本 arXiv:2203 16691，2022年。[38] N. M. M¨uller、P. Czempin、F. Dieckmann、A. Froghyar 和 K. B¨ottinger，《音频深度伪造检测能否泛化？》，Interspeech，2022年。