arXiv:2611.32593v1 [eess.AS] 24 Nov 2026

满分却错位的解释：音频深度伪造检测可解释性的受控研究

David Combei ,[1] *Dan Oneata,*[2,3] *Francesco Paissan,*[4,5] *Cem Subakan,*[4,5] *Horia Cucu,*[2] *Adriana Stan*[1]

[1]克卢日-纳波卡技术大学，罗马尼亚 [2]布加勒斯特理工大学，罗马尼亚[3]Bitdefender，罗马尼亚 [4]Mila-魁北克人工智能研究所，加拿大 [5]拉瓦尔大学，加拿大

摘要

如果一种解释得分完美，却难以理解，会怎样？我们发现，在解释音频深度伪造检测器时就会出现这种情况。我们使用一种基于优化的方法，解释一个先进的基于 wav2vec2 的检测器。这些解释在标准可解释性指标（即保真度和忠实度）上取得了最优得分。然而，视觉检查却揭示出不直观、难以理解的模式。为了解这一差异，我们开展了一项受控扰动研究，在其中修改语音信号中已知的频率区域。尽管我们提出了一种能够在这些指标上取得近乎最优结果的模型，但只有当被操纵区域在所有样本中固定不变时，解释才能定位该区域；当区域有所变化时，定位结果大多退化至随机水平。我们还表明，用于填补已移除区域的信号在很大程度上决定了解释及其定位结果，但这一影响并未体现在指标中。

索引词— 音频深度伪造检测、反欺骗、可解释性、可解释人工智能、优化掩码、指标

1. 引言

是什么让这段语音样本成为伪造样本？先进的深度伪造检测器可以准确判断文件是生成的，却无法说明是什么导致了这一判断；仅有准确性并不能带来信任 [1]。在政治虚假信息或数字证据完整性等高风险场景中，理解这些判断十分必要。可解释人工智能（xAI）旨在减少这种不透明性。可解释性设计方法 [2]难以构建，因为即使对人类专家观察者 [3]而言，深度伪造检测也很困难。基于人类可理解特征训练的可解释分类器，例如呼吸模式 [4]、基频 [5] 或手工设计的声学特征 [6]，其性能仍落后于高性能检测器。这类检测器基于大型自监督学习（SSL）模型构建，尤其是 wav2vec2 系列 [7]，并已被证明能够很好地泛化到不同数据集和操纵类型 [8, 9]。线性分类器 [8, 10], 图神经网络 [11] 和状态空间模型 [12] 通常基于这些自监督学习特征进行训练。

通讯作者：adriana.stan@com.utcluj .ro

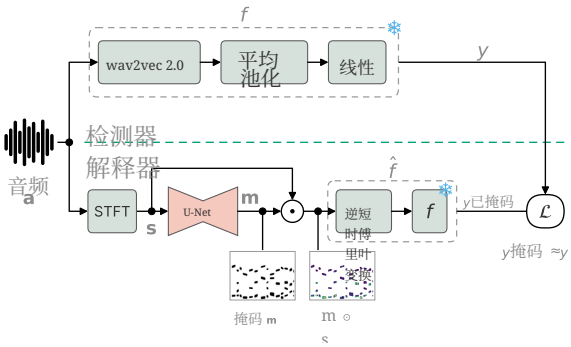


图 1。所提可解释人工智能方法。对于冻结的音频深度伪造检测器 f，我们学习一个解释器，用于预测输入频谱图 $\mathbf{s}$上的软稀疏掩码$\mathbf{m}$。仅保留掩码区域应能保持检测器的预测，而移除这些区域则应使预测发生翻转。

但额外的复杂性使其决策更难解释。迄今为止，可解释的音频深度伪造检测主要依赖于时域中的基于梯度的解释 [13]，、预测真实信号与声码器处理信号之间的频谱图差异 [14]，、伪影专用专家检测器 [15]，，或单类异常检测 [16]。

我们的目标是解释一个复杂且性能卓越的、基于 wav2vec2 的深度伪造检测器，通过定位与其决策相关的时频区域，即显著性图。我们的解释是输入频谱图上的软掩码，通过最大化两个标准可解释性指标获得：保真度，用于检查仅保留掩码区域时预测是否得以保持；以及忠实度，用于检查移除这些区域时预测是否发生变化。这些指标在大量训练样本上进行优化，从而得到一种摊销优化方法 [17]。该工作类似于用于声音 [18]，、图像 [19] 和文本 [20]分类的方法，但这是其首次应用于音频深度伪造检测。

我们发现，我们的方法在两个xAI指标上都取得了近乎完美的分数，超过了基于梯度的解释方法；后者本质上难以实现较高的忠实度。为评估这些指标是否能产生有意义的解释，

我们开展了一项受控评估，将特定频带替换为声码器处理后的对应信号，以此对其进行扰动。尽管深度伪造检测器能够揭示此类操纵，但基于梯度的方法无一能够找回真实的被操纵区域。当不同样本中的频带保持一致时，我们的解释器能够定位该频带；但当频带发生变化时，定位结果大多会降至随机机会水平。最后，用于填充移除区域的信号构成了解释的基线， [21], 改变该信号会使定位结果从远高于随机机会水平降至随机水平，却不会影响这些指标。

总而言之，我们的贡献包括：**（i）**我们构建了一个强大的、基于优化的摊销解释器，为基于 wav2vec2 的深度伪造检测器生成时频掩码；**（ii）**我们表明，该模型在标准可解释性指标上取得了最优分数，但无法提供可解释的解释；**（iii）**通过受控的频率扰动，我们表明，高保真度和忠实度并不自动意味着正确的定位；**（iv）**我们表明，掩码外区域的填充方式决定了解释及其定位，但指标无法反映这一点。

2. PROPOSEDXAI 方法

我们的方法将一个冻结的检测器与一个可训练的解释器配对。解释器预测一个掩码，突出显示与检测器预测最相关的频谱区域（见图 1）。

2.1. 检测器

该检测器是一个二元分类器 f ，接收一维时域音频波形$\mathbf{a}$ 作为输入，并输出两个标签 $y \in$ {真实，伪造}中的一个。它由以下部分组成：wav2vec 2.0 特征提取器 [7]、全局平均池化层和用于分类的线性层。我们使用 wav2vec2 XLS-R 2B 骨干网络，从 48 层中的第 9 层提取特征——研究发现，较早的层效果更好 [22]。音频以 16 千赫兹采样。分类器头是一个低正则化的 scikit-learn 逻辑回归模型（$C = 10^6$）。

2.2. 解释器

由于检测器在波形上运行，直接针对其输入计算的解释只能揭示何时信息很重要，而大多数语音专家会期待时×频域中的解释。因此，我们训练了一个解释器，在输入音频的频谱图$\mathbf{s}$ 上生成软显著性掩码$\mathbf{m}$ 。接近 1 的值标记预测 y 的证据，接近 0 的值标记无关区域。经过掩码处理的频谱图通过可微分的逆短时傅里叶变换（iSTFT）转换回来——从而可以通过优化或梯度获得解释——并输入冻结的ˆ检测器，得到封装检测器 $f(\mathbf{s}) = f$ (iSTFT(s))。封装检测器返回输入为伪造语音的概率。短时傅里叶变换使用与 HiFi-GAN 相同的参数：窗口长度为 1024 个样本，FFT 大小为 1024，跳跃长度为 256。解释器在许多样本上进行训练，这些样本使推理速度更快，并降低其对单个输入中虚假模式的敏感性 [23]。解释器的架构采用 U-Net，包含三个二维卷积层和一个输出 Sigmoid激活函数，将输出解释限制在 $[0, 1]$ 区间内。不同于 [24], ，我们保持分类器冻结，从而防止解释器编码标签信息。尽管输出是线性的，wav2vec 2.0 前端仍使决策边界具有高度非线性的特征。

该**解释** 是一个与频谱图$\mathbf{s}$大小相同的掩码$\mathbf{m}$ $\in [0, 1]^{F\times T}$。被掩码移除的区域将由填充信号$\mathbf{z}$替代。我们将仅保留掩码内区域的输入（$\mathbf{s_m}$）以及保留掩码外区域的输入（$\mathbf{s_{\bar{m}}}$）分别记为：

$$\mathbf{s_m} = \mathbf{m} \odot \mathbf{s} + (1 - \mathbf{m}) \odot \mathbf{z}, \tag{1}$$

$$\mathbf{s_{\bar{m}}} = (1 - \mathbf{m}) \odot \mathbf{s} + \mathbf{m} \odot \mathbf{z}, \tag{2}$$

其中 $\odot$ 表示逐元素乘法。

优化使用以下项：

（a）**保真度** 检查仅保留掩码区域是否能够保持预测：

$$\text{FID} := [\![\, H(\hat{f}(\mathbf{s})) = H(\hat{f}(\mathbf{s_m})) \,]\!], \tag{3}$$

其中 H 将概率以 0.5 进行阈值化。

(b) **忠实度** 衡量移除掩码区域后概率的下降：

$$\text{FTH} := \hat{f}(\mathbf{s}) - \hat{f}(\mathbf{s_{\bar{m}}}). \tag{4}$$

(c) **稀疏性**：掩码应尽可能少地保留频谱图， $\|\mathbf{m}\|_1$,

这些指标组合为训练损失：

$$\mathcal{L} = \alpha\, \text{BCE}\big(\hat{f}(\mathbf{s_m}),\, y\big) + \beta\, \text{BCE}\big(\hat{f}(\mathbf{s_{\bar{m}}}),\, 1 - y\big) + \gamma\, \|\mathbf{m}\|_1, \tag{5}$$

其中， y 是检测器对原始输入的预测，BCE 是二元交叉熵，$\{\alpha, \beta, \gamma\}$ 是权重。

填充值（$\mathbf{z}$）充当解释的零点：通过学习掩码来决定保留哪些区域，以及用$\mathbf{z}$替换哪些区域，因此$\mathbf{z}$ 的选择决定了解释所依据的参照。我们比较四种填充方式：(a) 零填充，其中$\mathbf{z}$全为零；(b) 平均填充，其中$\mathbf{z}$ 是训练数据上的平均频谱图；(c) 真实填充，其中$\mathbf{z}$ 是配对真实录音的频谱图——说话人相同、内容相同，但没有时间对齐；(d) 优化填充，即学习得到的单个频谱图$\mathbf{z}$ ，使检测器对其尽可能不确定， $f(\mathbf{z}) = 0.5$，同时在与真实或伪造输入以各种掩码组合时保留预测结果。优化填充采用自举式训练：使用训练后的解释器生成的掩码重新学习填充值，然后使用新的填充值重新训练解释器。

2.3. 基线方法

我们将结果与两种广泛使用且无需训练的基于梯度的方法进行比较，这些方法通过 Captum 应用于同一个封装检测器 f :

表1。现实数据（MLAAD）以及两种受控设置下的保真度（FID）、忠实度（FTH）、定位AUC和相关性质量准确度（RMA）；在这两种受控设置中，分别对固定或变化的1千赫兹频带进行声码化。第0–4行使用优化填充来计算分数。随机机会水平：AUC为0.5，RMA为0.125。

		现实数据 （EER 7.4%）		**固定频带** （EER 0.0%）				**变化频带** （EER 0.04%）			
	方法	FID↑	FTH↑	FID↑	FTH↑	AUC↑	RMA↑	FID↑	FTH↑	AUC↑	RMA↑
1	随机	0.993	0.005	**1.000**	0.000	0.500	0.125	0.981	0.020	0.500	0.125
2	显著性	0.855	0.057	0.824	0.000	0.497	0.090	0.999	0.004	0.496	0.118
3	Inp×Grad	0.975	0.085	0.755	0.003	0.516	0.137	0.328	0.001	0.509	0.101
4	**我们的方法**，优化填充	**1.000**	**0.999**	**1.000**	**1.000**	0.811	0.289	**1.000**	**1.000**	0.505	0.124
5	**Ours**, 真实填充	0.931	0.886	**1.000**	0.999	**0.897**	**0.436**	**1.000**	**1.000**	**0.728**	**0.358**
6	**我们的方法**，平均填充	**1.000**	**0.999**	**1.000**	0.986	0.448	0.136	**1.000**	0.997	0.495	0.128
7	**Ours**, 零填充	**1.000**	**0.999**	**1.000**	0.994	0.713	0.409	**1.000**	0.999	0.498	0.126

显著性 [25] 和输入 × 梯度 [26]。我们还使用一个随机基线，为 $[0, 1]$ 中的频谱图超像素 [27] 分配均匀分数，以检验这些指标能否将有意义的解释与任意解释区分开来。超像素是由视觉上属于同一整体的相邻时频块组成的区域，例如谐波的一部分或静音区域。

2.4. 结果

数据。我们从 MLAADv9 [28] 中抽取 10% 的数据，针对 298 种模型-语言组合分别选取样本，共计 2.7 万个伪造样本。从 M-AILABS [29] 中选取 2.7 万个真实样本。数据按 80:20 的比例划分为训练集和测试集，用于训练深度伪造检测器。解释器仅在伪造样本上进行训练和评估。这是因为，伪造判定可能依赖于局部异常（例如被篡改的频带），而真实判定可能依赖于不存在此类异常这一事实，掩码无法突出显示这种信息。在所有设置中，语音文件都会裁剪或进行零填充，使其长度达到 5 秒。

定量结果。检测器的等错误率（EER）达到 7.4%，这意味着它基本能够区分伪造样本和真实样本。表1中“Realistic”列展示了所有可解释人工智能方法的保真度和忠实度结果。显著性和输入 × 梯度取得了良好的保真度，但其忠实度接近于 0（第 2–3 行）。即使是随机基线（第 1 行）也取得了 0.99 的保真度，尽管其解释是任意的。与此同时，我们的解释器（第 4–7 行）无论采用何种填充信号，都实现了较高的保真度和忠实度。

定性结果。但这些生成的解释究竟有多可解释？图 2 展示了 MLAAD 中两个样本的解释（第 1–2 行）。随机基线（第 2 列）将第 2 行中的填充静音视为一个整体区块，却获得了近乎完美的保真度。显著性（第 3 列）突出显示宽阔的垂直条带，也将相关性赋予填充静音。输入×梯度（第 4 列）遵循起始音和低频谐波等精细结构，并在静音处消失。相比之下，我们的掩码（第 5–8 列）更加锐利。然而，四种填充方式获得的分数同样高，却为同一输入生成了截然不同的掩码，例如只有部分掩码延伸到了第 2 行的填充静音区域。这类模式很难与任何合理的合成伪影联系起来，也表明较高的可解释人工智能分数并不一定能转化为音频中可识别的操纵线索。下一节将通过受控实验考察这种差异。

3. 理解解释

为了检验解释器是否确实能够定位合成伪影——尽管这些伪影无法被人类理解——我们在已知频谱区域经过操纵的语音上对其进行评估。我们首先从 LJ Speech（单说话人、高质量） [30] 中选取 5,000 个真实音频样本，并将其重采样为 16 千赫兹。随后，使用基于 LibriTTS [32]训练的 HiFi-GAN 声码器 [31] 对每个样本进行声码器处理。接着，我们将一个 1千赫兹频带（8 个可能频带中的一个）替换为其时间对齐的声码器处理版本中的对应频带。替换操作在复频谱域中进行。被替换的频带要么在所有样本中保持相同，要么因样本而异。我们首先在新生成的数据集上训练深度伪造检测器。两种设置均取得了较低的等错误率：固定频带为 0.0%，变化频带为 0.04%。表1还列出了这两种设置下的指标。与真实场景一样，基于梯度的方法忠实度较低，而我们的解释器则实现了近乎完美的保真度和忠实度。

定位与填充值的影响。由于操纵区域已知，我们可以使用掩码值作为评分、将被操纵的频点作为正例，通过 ROC 曲线下面积（AUC）来衡量定位效果。我们还可以报告相关性质量准确率（RMA） [33],即落在被操纵频带内的软掩码总值所占比例。当所有相关性都位于该频带内时，该值为 1；对于随机解释，该值为 0.125，也就是该频带在频谱图中所占的比例（1/8 kHz）。

所有基于梯度的方法仍处于随机水平（AUC≈ 0.50，RMA ≈ 0.125）。采用优化填充后，我们的解释器能够定位固定频带（AUC 0.81），但其相关性中不到三分之一位于该频带内（RMA 0.29）。其掩码覆盖了该频带及其下方的大部分频谱区域（图 2,

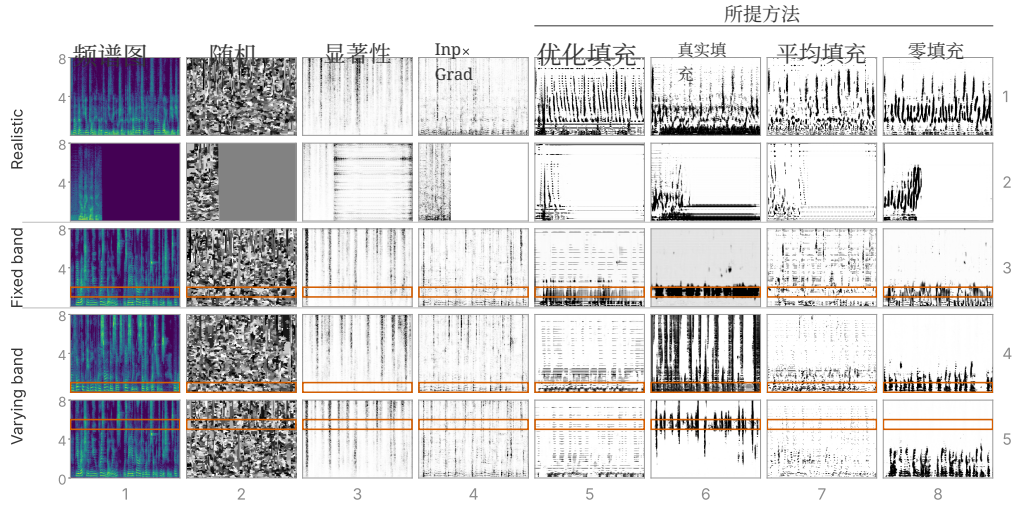


图 2。展示了两个 MLAAD 伪造样本（第 1–2 行）、一个固定频带样本（1–2 kHz，第 3 行）以及两个变化频带样本（0–1 kHz 和 5–6 kHz，第 4–5 行）的解释。第 5–8 列展示了每种填充方式对应的掩码；基线方法的映射不受填充方式影响。橙色方框标出了篡改频带。每个掩码均单独归一化。

第 3 行，第 5 列）。频带变化时，定位能力降至随机水平（AUC 0.51，RMA 0.12），而且无论采用何种操控，掩码都呈现相同的模式（第 4–5 行，第 5 列）。

一个令人意外的结果是，解释器定位篡改的能力很大程度上取决于填充方式（表1第 4–7 行）。真实填充的定位效果最好（AUC 为 0.90 和 0.73），其掩码也紧密贴合篡改频带（图 2 第 3 行和第 5 行的第 6 列）。它是唯一一种在第 5 行将掩码移至 5–6 kHz 频带的填充方式。然而，真实信号与篡改信号仅在被交换的频带上有所不同，因此，在训练过程中，填充方式本身就传达了篡改的位置。平均填充产生的掩码较为分散，定位效果仅达到随机机会水平（第 7 列）。零填充（第 8 列）的表现与优化后的填充方式相似：它覆盖固定频带及其下方区域，在频带变化时仍停留在低频区域，并且在第 5 行将 5–6 kHz 频带留空。

这种行为可能部分是由于将每种填充分别单独输入检测器所致。零填充被三个检测器都判定为伪造；平均填充被两个受控检测器判定为伪造，而逼真检测器则将其判定为真实。在这些填充下，伪造样本的稀疏掩码无论位于何处，都仍会被判定为伪造。即使填充没有偏向性，保真度仍然很高。优化填充经过专门设计，本身不偏向任何一类；真实填充则会将预测推向真实，但任意掩码仍能保留该预测：在 MLAAD 上，使用优化填充时，随机解释的保真度为 0.98–1.00；使用真实填充时为 0.98（表中未展示），而我们的解释器为 0.93。由于软掩码会在各处保留部分输入，足够的伪造信号得以保留，使检测器维持其判断，因此保真度无法将有意义的解释与任意解释区分开来。另一方面，交替训练优化填充与解释器可能会使二者发生协同适应 [24]。但是，零填充并未经过学习，却产生了相同的固定掩码（第 4–5 行、第 5 列和第 8 列），因此不能仅用协同适应来解释这一现象。这也不是容量限制，因为解释器架构能够以 AUC 1.0 定位随机放置的矩形。

简而言之，单凭指标无法判断解释是否指向真实操纵。而在我们的设置中，填充方式似乎决定了很大一部分解释及其得分的表现。

4. 结论

我们提出了一种基于优化的方法，利用时频掩码解释基于 wav2vec2 的深度伪造检测器。该方法达到了近乎完美的保真度和忠实度，但在操纵位置已知的受控操纵实验中发现，其解释往往未能定位被操纵区域。使用优化填充和零填充时，解释器能够定位跨样本固定的频带，但无法定位位置不断变化的频带。真实填充则能够定位这两类频带，但这只是因为其训练过程揭示了该频带。与此同时，各项指标仍接近完美，甚至随机解释也获得了近乎完美的保真度。因此，不应单独使用这些指标来验证解释，而应结合具有已知真实标签的受控操纵进行评估。

这些结果看似在意料之中，因为该方法优化的正是我们所质疑的指标（古德哈特定律：经过优化的指标不再是可靠的衡量标准）。然而，问题并不局限于我们的方法：这些指标可能被故意夸大 [34]，而且我们表明，这种情况也会在无意中发生。设计更稳健的指标留待未来研究。

致谢。 Cem Subakan 得到了加拿大自然科学与工程研究委员会项目 RGPIN-2023-05759 的支持。Francesco Paissan 由魁北克研究基金（FRQ）以及拉瓦尔大学智能与数据研究所（IID）资助

遵守伦理标准。 本研究回顾性地使用了开放获取的人类受试者数据（MLAAD、M-AILABS 和 LJ Speech）。根据这些数据集所附的许可条款，无需进行伦理审批。

LLM 的使用。作者使用 Claude（Anthropic）编辑并精简文本、编写图 1 和图 2 的绘图脚本，以及实现定位指标（AUC 和 RMA）。所有实验、结果和结论均由作者独立完成；作者审阅并核实了所有 AI 辅助生成的文本和代码，并对内容承担全部责任。

5. 参考文献

[1] A. K. Jain、D. Deb 和 J. J. Engelsma，《生物特征识别：信任，但要验证》，*IEEE Trans. Biom. Behav. Identity Sci.*，2022 年。[2] C. Rudin，《在高风险决策中，停止解释黑箱机器学习模型，改用可解释模型》，*Nat. Mach.Intell.*，2019 年。[3] N. M. M¨uller、K. Pizzi 和 J. Williams，《人类对音频深度伪造的感知》，发表于*DDAM@MM*，2022 年。[4] S. Layton、T. D. Andrade、D. Olszewski 等，《你不曾吸入的每一口气：利用呼吸检测深度伪造语音》，*Digit. Threat. Res. Pract.*，2025 年。[5] J. Xue、C. Fan、Z. Lv 等，《基于 F0 信息与实部、虚部频谱图特征组合的音频深度伪造检测》，发表于*DDAM@MM*，2022 年。[6] O. Pascu、D. Oneata、H. Cucu 等，《简单、可解释、有效：openSMILE 用于语音深度伪造检测》，发表于*ICASSP*，2025 年。[7] A. Baevski、Y. Zhou、A. Mohamed 等，《Wav2vec 2.0：一种用于语音自监督学习表示的框架》，发表于*NeurIPS*，2020 年。[8] W. Ge、X. Wang、X. Liu 等，《深度伪造语音检测的后训练》，发表于*ASRU*，2025 年。[9] O. Pascu、A. Stan、D. Oneata 等，《借助自监督表示实现可泛化且校准良好的音频深度伪造检测》，发表于*Interspeech*，2024 年。[10] D. Combei、A. Stan、D. Oneata 等，《揭开真实场景中的音频深度伪造：一种以数据为中心的方法》，发表于*Interspeech*，2025 年。[11] H. Tak、M. Todisco、X. Wang 等，《利用 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》，发表于*Odyssey*，2022 年。[12] H. M. Tran、D. Lolive、D. Guennec 等，《利用 SSL 语音特征和 Mamba 增强深度伪造检测》，发表于*Interspeech*，2025 年。[13] P. Grinberg、A. Kumar、S. Koppisetti 等，《音频深度伪造检测器关注什么？时域研究》，发表于*ICASSP*，2025 年。[14] P. Grinberg、A. Kumar、S. Koppisetti 等，《一种用于音频深度伪造解释的数据驱动扩散式方法》，发表于*Interspeech*，2025 年。[15] V. Negroni、X. Wang、W. Ge 等，《利用伪迹专用专家和校准后的检测分数实现可解释的语音深度伪造检测》，发表于*ACM Deepfake ForensicsWork- shop*，2026 年。[16] E. Coletta、D. Salvi、V. Negroni 等，《基于特征金字塔匹配的语音深度伪造异常检测与定位》，发表于*EUSIPCO*，2025 年。[17] J. Chen、L. Song、M. Wainwright 等，《学习解释：模型解释的信息论视角》，发表于*ICML*，2018 年。[18] F. Paissan、M. Ravanelli 和 C. Subakan，《音频分类器的可聆听映射》，发表于*ICML*，2024 年。[19] R. C. Fong 和 A. Vedaldi，《通过有意义的扰动对黑箱进行可解释的解释》，发表于*ICCV*，2017 年。[20] P. Hase、H. Xie 和 M. Bansal，《可解释性中的分布外问题及特征重要性解释的搜索方法》，发表于*NeurIPS*，2021 年。[21] Pascal Sturmfels、Scott Lundberg 和 Su-In Lee，《可视化特征归因基线的影响》，*Distill*，2020 年。[22] Y. El Kheir、Y. Samih、S. Maharjan 等，《针对音频深度伪造检测的自监督学习模型全面逐层分析》，发表于*Findings NAACL*，2025 年。[23] P. Dabkowski 和 Y. Gal，《黑箱分类器的实时图像显著性》，发表于*NeurIPS*，2017 年。[24] N. Jethani、M. Sudarshan 等，《我们学会解释了吗？可解释性方法如何学会在解释中编码预测》，发表于*AISTATS*，2021 年。[25] K. Simonyan、A. Vedaldi 和 A. Zisserman，《深入卷积网络内部：图像分类模型与显著性图的可视化》，发表于*ICLR* 研讨会，2014 年。[26] A. Shrikumar、P. Greenside、A. Shcherbina 等，《不只是黑箱：通过传播激活差异学习重要特征》，*arXiv preprint arXiv:1605.01713*，2016 年。[27] P. Felzenszwalb 和 D. Huttenlocher，《高效的基于图的图像分割》，《国际计算机视觉杂志》，第 59 卷，第 2 期，2004 年。[28] N. M. Muller、P. Kawa、W. H. Choong 等，《MLAAD：多语言音频反欺骗数据集》，发表于*IJCNN*，2024 年。[29] I. C. A. Solak 和 D. Naumov，《M-AILABS 语音数据集》，https://github.com/i-celeste-aurora/ m-ailabs-dataset，2017 年，访问日期：2026年5月2日。[30] K. Ito 和 L. Johnson，《LJ Speech 数据集》，https: // keithito.com/LJ-Speech-Dataset/，2017 年。[31] J. Kong、J. Kim 和 J. Bae，《HiFi-GAN：用于高效、高保真语音合成的生成对抗网络》，发表于*NeurIPS*，2020 年。[32] H. Zen、V. Dang、R. Clark 等，《LibriTTS：源自 LibriSpeech 的文本转语音语料库》，发表于*Interspeech*，2019 年。[33] L. Arras、A. Osman 和 W. Samek，《CLEVR-XAI：用于神经网络解释真实值评估的基准数据集》，《信息融合》，2022 年。[34] J. Hsia、D. Pruthi、A. Singh 等，《古德哈特定律同样适用于自然语言处理的解释基准》，发表于*FindingsEACL*，2024 年。