通过图增强实例学习连接合成语音的归因与开放集检测

MohdMujtabaAkhtar[1]*、Girish[2]*、Farhan Sheth[3]* 和 Muskaan Singh[4] †
[1]印度 Veer Bahadur Singh Purvanchal University[2]印度 UPES[3]印度 Manipal University Jaipur[4]英国阿尔斯特大学

arXiv:2601.07064v1 [eess.AS] 11 Jan 2026

摘要

我们提出一种统一框架，不仅能够将合成语音归因于其来源，还能检测由训练期间未曾遇到的合成器生成的语音。这需要超越简单检测的方法，以支持细致的取证分析和开放集泛化。为此，我们提出**SIGNAL**，这是一种混合框架，将语音基础模型（语音基础模型）与基于图的建模和开放集感知推理相结合。我们的框架整合了图神经网络（GNN）和K近邻（K近邻）分类器，使其能够捕捉话语之间有意义的关系，并识别不属于任何已知生成器的语音。该框架在生成器类别原型上构建查询条件图，使GNN能够推理候选生成器之间的关系，同时K近邻分支通过基于置信度的阈值判定支持开放集检测。我们使用DiffSSD数据集评估**SIGNAL**；该数据集涵盖了来自开源和商业基于扩散的文本转语音系统的真实语音与合成音频，类型丰富。为进一步评估泛化能力，我们还在SingFake基准上进行测试。结果表明，**SIGNAL** 在两项任务中均持续提升性能，其中基于Mamba的嵌入表现尤为突出。据我们所知，这是首项将基于图的学习与开放集检测相结合、以追溯合成语音来源的研究。

1 引言与背景

合成语音检测（SSD）在保障数字通信安全方面发挥着关键作用，使系统能够识别并降低高度逼真的机器生成语音所带来的风险（Todisco等，2019年；Wu等，2015年）。随着先进的文本转语音（TTS）和语音转换（VC）模型的发展，合成语音的逼真度已达到近乎模仿自然人类韵律和音色的水平（Ren等，2021年；Kong等，2021年）。尽管此类进步推动了无障碍和个性化的发展（Cooper等，2020年），但也带来了音频冒充、欺诈和虚假信息等形式的新漏洞（Yi等，2023年）。因此，检测和分析合成语音的能力不仅对安全至关重要，对于维护人机交互中的信任同样不可或缺。过去几年中，神经语音合成取得了显著进展，这些进展由基于扩散的模型、富表现力的文本转语音系统以及多语言语音转换技术推动。VALL-E（Ju等，2024年）、NaturalSpeech 3（Ju等，2024年）和Voicebox（Le等，2023年）等最先进的模型已经展现出生成语音的能力：不仅能够模仿说话人身份，还能捕捉情感、韵律和表现力等细粒度声学属性。这些模型通常建立在大规模语音语言预训练的基础上，并采用包括Transformer（Li等，2019年）、状态空间模型（Gu和Dao，2024年）以及去噪扩散过程（Ren等，2021年）在内的强大骨干网络。通过开源实现和商业API，这些模型日益普及，使合成语音生成比以往任何时候都更加无处不在。然而，这种快速发展也凸显了检测任务日益增加的复杂性，尤其是在推理时生成模型未知或未见的情况下。

尽管合成语音检测日益受到关注，但现有的大多数方法都将合成语音检测视为二分类问题，只专注于区分真实音频与合成音频，而不识别其生成来源（Shin etal.，2024年；Huang 和 Pun，2024年）。然而，面对以往未见的文本转语音或语音转换模型时，现有大多数方法都无法良好泛化

*作为第一作者贡献相同。
†通讯作者：m.singh@ulster.uk.in

，在开放集条件下性能显著下降（Guo et al.，2024年；Stan et al.，2025年）。这一局限制约了它们在实际法证场景中的应用，而在这些场景中，来源归因以及对分布外语音的稳健检测至关重要。尽管近期研究推动了合成语音检测的发展，但仍存在一个关键缺口：既能识别语音的来源文本转语音模型，又能检测由未见生成器合成的语音。来源归因日益重要，在法证和监管场景中，了解合成音频的来源至关重要。同时，现实世界的系统必须能够应对开放集场景，即测试时可能出现新的或未知的模型。弥合这两项挑战对于构建更可靠、泛化能力更强的检测框架至关重要。作为我们在DiffSSD 研究中的核心重点，我们探索了多种语音基础模型（语音基础模型），并提出假设：将图神经网络（*GNN*）进行的关系建模——在生成器类别原型（节点）上构建由查询样本条件化的图，并通过基于注意力的消息传递进行优化——与基于 *K*近邻的开放集推理相结合，是同时解决来源归因和未见生成器检测的有效策略。 这一方法的出发点是两个组成部分优势互补：图神经网络对嵌入节点间的类别级交互进行建模，捕捉用于生成器判别的细微关系线索；而 K近邻分支则通过基于置信度的阈值判定，为处理开放集场景提供轻量且稳健的机制。为验证这一假设，我们在已见和未见生成器条件下，使用多种语音基础模型嵌入开展了广泛评估，包括 Whisper、Unispeech、ECAPA 和 Mamba。我们从可解释来源归因的先前研究（Mishra et al.，2025年）、基于图结构的深度伪造检测研究（Febrinantoetal.，2025年）以及用于欺骗检测的图注意力网络研究（Tan et al.，2025年）中汲取灵感。为解决这一问题，我们提出了 **SIGNAL：** **S**peech **I**nferenceviaGraph**N**etworks and **A**ugmented **L**earning——一种统一框架，通过基于图神经网络的节点分类和事后近邻过滤，同时解决这两项任务。据我们所知，这是首项探索混合图神经网络–近邻方法，用于合成语音来源追踪和开放集检测的研究。

我们的主要贡献如下：

- 我们提出了**SIGNAL**，这是一种新颖的混合框架，将用于关系建模的图神经网络（GNN）与用于开放集推理的K近邻（KNN）相结合，专为合成语音中的联合来源归因和未见生成器检测而设计。
- 我们对多种语音基础模型（SFM）进行了大规模基准测试，包括Whisper、Unispeech、ECAPA和Mamba，并在已见和未见生成器设置下评估其表现，揭示了它们在不同任务中的互补特性。
- 据我们所知，这是首项研究基于图的方法如何对语音基础模型嵌入进行事后推理，以实现开放集合成语音取证，并在DiffSSD和SingFake基准数据集上均取得了出色表现。

本研究的相关资源可在以下地址获取：
https://github.com/Helixometry/SIGNAL.git

2 表征

本节详细介绍用于提取话语级嵌入以进行下游处理的语音基础模型（SFMs）。

我们采用x向量[1] （Snyder等，2020年）和ECAPA-TDNN[2] （Desplanques等，2020年）作为说话人识别模型。二者均基于时延神经网络，并在VoxCeleb1+2 数据集上进行训练。ECAPA通过引入Res2Net模块和SE模块扩展了标准x向量架构，其性能优于x向量，而后者本身也显著超越了传统的i-vector基线。x向量包含约420万个参数。对于单语预训练模型，我们采用WavLM[3] （Chen等，2022年）、UniSpeech-SAT[4] （Chen等，2021年）和wav2vec 2.0[5] （Baevski等，2020年），它们均为基础版本，并在来自LibriSpeech的960小时英语音频上进行训练。WavLM和UniSpeech-SAT是SUPERB基准上表现最佳的预训练模型，

[1] https://huggingface.co/speechbrain/spkrec-xvect-voxceleb
[2] https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb
[3] https://huggingface.co/microsoft/wavlm-base
[4] https://huggingface.co/microsoft/ unispeech-sat-base
[5] https://huggingface.co/facebook/ wav2vec2-base

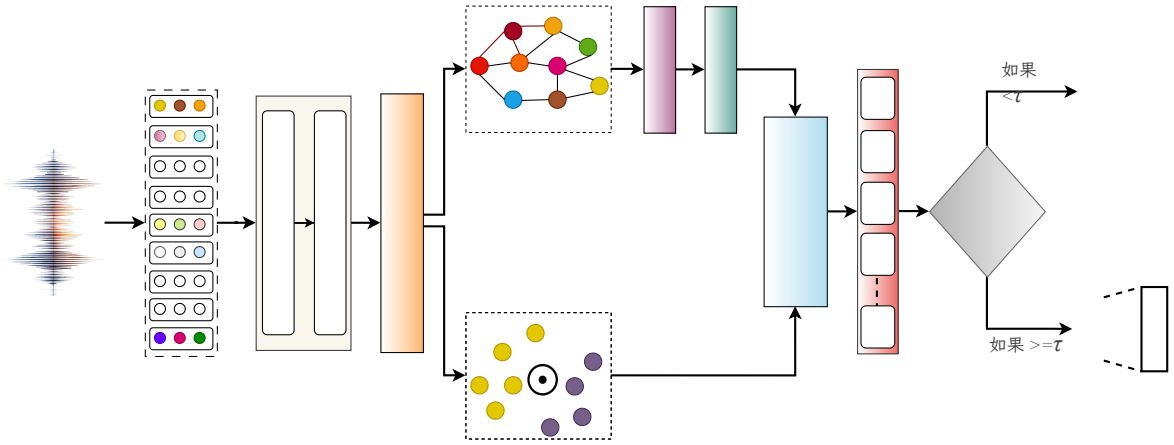


图1：所提出的框架：**SIGNAL**。模型提取表征，随后通过图注意力网络（GAT）和K近邻（KNN）模块进行并行推理。各模块的输出由集成头融合。最终的路由决策基于置信度阈值（$\tau = 0.5$），将样本分配至已见类（$S_1 \dots S_n$）或未见类预测。

其中，WavLM采用掩码语音建模和去噪目标进行训练，而UniSpeech-SAT采用融合说话人感知对比学习的多任务框架。wav2vec 2.0通过对比损失进行优化，以区分真实码本条目和干扰码本条目。这三个模型的参数规模相近：WavLM（9470万）、UniSpeech-SAT（9468万）和wav2vec 2.0（9504万）。我们进一步纳入Whisper[6]（Radford等，2023年），这是一种在68万小时多样化音频上训练的多语言自动语音识别模型。Whisper采用Transformer解码器架构，在零样本设置下能够跨语言、跨任务稳定运行。我们使用其包含7400万个参数的基础版本，并从编码器层提取表示。此外，我们还考虑Audio-MAMBA[7]（Yadav和Tan，2024年），这是一种在AudioSet上训练、用于重建掩码频谱图块的状态空间模型。其SSM骨干网络能够有效捕捉时间依赖关系和频谱细节。我们使用其tiny（480万）、small（1790万）和base（6930万）变体，以研究可扩展性影响。

将音频样本重采样至 16 KHz 后，再输入基础模型（FM）。对每个冻结模型的最终隐藏状态应用平均池化，得到固定长度的嵌入。所得嵌入维度如下：Whisper 和 x向量为 512；WavLM、UniSpeech-SAT 和 wav2vec 2.0 为 768；Audio-MAMBA（基础版）为 3840。ECAPA-TDNN 生成 192 维嵌入。

[6]https://huggingface.co/openai/whisper-base[7]https://github.com/SarthakYadav/ audio-mamba-official

3 建模流程

本节将详细介绍个体表征的建模流程，以及用于来源追踪和未见生成器检测的混合框架。我们采用全连接网络（FCN）和卷积神经网络（CNN）作为基于个体表征的建模下游模型。此外，我们提出**SIGNAL**，将图神经网络（GNN）与 K近邻（KNN）模块相结合，从而能够同时利用全局关系结构和局部实例级相似性。图1 展示了整体架构。

3.1 个体表征模型 g

为建立强有力的基线，我们首先使用全卷积网络和卷积神经网络骨干网络，评估采用单一预训练语音表征的标准分类器。卷积神经网络由两个一维卷积层组成，滤波器数量分别为 64 和 128（卷积核大小为 =3），后接 ReLU 激活和最大池化（池化大小为 =2）。输出展平后输入含 128 个神经元的全连接层，再接 Softmax 输出层。全卷积网络模型采用与卷积神经网络相同的全连接块，但不包含卷积层。

所提出的框架：SIGNAL

我们提出了**SIGNAL**，用于合成语音中来源归因与未见生成器检测的联合任务。**SIGNAL** 的架构如图1所示。

表征编码： 设$\mathbf{x} \in \mathbb{R}^{T \times F}$表示输入音频信号，其中 T 为

PTMs	FCN									CNN								
	DEV			TEST			分布内/分布外			DEV			TEST			分布内/分布外		
	准确率 ↑	F1 ↑	等错误率 ↓	准确率 ↑	F1 ↑	等错误率 ↓	准确率 ↑	F1 ↑	等错误率 ↓	准确率 ↑	F1 ↑	等错误率 ↓	准确率 ↑	F1 ↑	等错误率 ↓	准确率 ↑	F1 ↑	等错误率 ↓
基线																		
Whisper	80.75	79.3	7.78	74.52	72.66	**9.77**	52.35	51.17	41.09	**83.10**	**81.40**	5.71	79.45	78.13	7.81	54.12	52.85	30.11
Unispeech	70.58	69.12	27.20	72.00	70.26	29.70	37.54	36.77	53.99	78.35	76.86	20.34	68.96	67.35	23.45	38.81	38.05	40.32
x向量	66.91	65.42	19.86	69.47	67.52	22.67	43.83	42.80	52.20	74.15	72.90	16.42	67.82	65.58	21.39	45.50	44.39	37.78
wav2vec 2.0	76.82	75.45	8.73	69.46	67.31	12.88	41.90	41.08	44.33	78.13	76.42	6.49	79.19	77.44	9.40	43.43	42.62	32.66
wavLM	78.24	76.81	12.48	73.16	71.07	15.53	44.74	43.65	52.49	80.60	78.25	9.06	76.44	74.21	11.37	46.38	45.34	39.11
ECAPA	63.77	62.55	17.28	62.31	60.09	25.29	32.33	31.63	57.80	71.92	69.87	13.13	64.15	63.29	20.13	33.56	32.84	41.45
MAMBA-T	69.35	68.12	15.87	76.01	**75.40**	19.08	59.23	58.13	46.59	80.15	78.17	11.37	72.23	71.04	11.79	60.94	59.92	35.00
MAMBA-S	72.01	70.66	10.86	71.04	69.16	13.19	57.33	56.21	40.56	81.22	79.56	7.82	74.51	73.18	9.80	59.32	58.11	29.64
MAMBA-B	**81.68**	**80.29**	**7.63**	**76.10**	74.35	10.74	**62.91**	**61.50**	**36.14**	82.90	80.11	**5.59**	**80.08**	**78.47**	**7.78**	**64.62**	**63.35**	**26.27**

表1：不同 PTM 在 FCN 和 CNN 骨干网络下的性能指标（准确率、F1分数、等错误率）。性能最佳的分数以**粗体**显示，次优分数以 下划线标示。研究中的所有表格均采用相同的格式。

帧数， F 为特征维度。该信号经过一个冻结的语音基础模型（SFM），得到一个固定长度的话语嵌入：

$$\mathbf{z}_0 = \mathrm{SFM}(\mathbf{x}) \in \mathbb{R}^{d_0}$$

然后，我们使用卷积神经网络编码器 f_{cnn}将$\mathbf{z}_0$投影到低维潜在空间中：

$$\mathbf{z} = f_{\text{cnn}}(\mathbf{z}_0) \in \mathbb{R}^d, \quad \text{where } d = 64$$

编码器 f_{cnn} 由两个带有 ReLU 激活和最大池化的一维卷积层组成，随后连接一个全连接投影层。

图神经网络头： 为捕捉类别级关系并增强来源归因，我们设计了一个以输入查询和已知生成器类别为中心的基于图的注意力模块。该图由以下部分组成：

- 一个**查询节点**，表示编码后的话语嵌入$\mathbf{z}$。
- **类别节点**由 N可学习的原型向量 $\{\mathbf{e}_1, \ldots, \mathbf{e}_N\}$ 表示，每个向量对应一个已见过的文本转语音生成器。
- **边**连接查询样本与每个类别节点，使模型能够评估相似度并对它们进行推理。

为何使用图神经网络进行归因： 我们采用图神经网络进行归因，因为来源预测不仅受益于各类别相似度分数，还受益于对生成器类别之间关系的推理。通过在类别节点间传播信息，图神经网络能够捕捉相对结构，而这种结构难以用独立分类器表示。

查询嵌入 $\mathbf{z}$ 首先通过一个可学习的变换投影到潜在空间：

$$\mathbf{s} = \mathbf{W}_s \mathbf{z}, \quad \mathbf{s} \in \mathbb{R}^d$$

其中$\mathbf{W}_s \in \mathbb{R}^{d \times d}$ 是一个可训练权重矩阵。随后，将投影后的向量$\mathbf{s}$ 加到每个类别原型上，以生成查询感知节点特征：

$$\tilde{\mathbf{e}}_i = \mathbf{e}_i + \mathbf{s}, \quad \forall i \in \{1, \ldots, N\}$$

这些 N 修改后的节点嵌入会传入**多头自注意力**机制，每个节点都会关注其他节点，并完善自身表示：

$$\tilde{\mathbf{e}}'_i = \mathrm{MultiHeadAttn}(\tilde{\mathbf{e}}_i, \{\tilde{\mathbf{e}}_j\}_{j=1}^N)$$

为计算类别分数，将每个更新后的节点投影为一个标量 logit：

$$\ell_i = \mathbf{w}^\top \tilde{\mathbf{e}}'_i, \quad \forall i \in \{1, \ldots, N\}$$

最终的类别概率通过对 logits 进行 softmax 得到：

$$\mathbf{p}_{\text{GNN}} = \mathrm{softmax}([\ell_1, \ldots, \ell_N])$$

为了估计模型在归因方面的不确定性，我们计算预测分布的注意力熵：

$$\mathcal{H}_{\text{attn}} = -\sum_{i=1}^{N} p_{\text{GNN},i} \log p_{\text{GNN},i}$$

直观而言，熵较低意味着模型确信样本归属于某个特定生成器；而熵较高则表明模型不确定，注意力在各个类别之间的分布也更均匀。

KNNBranch： 为应对开放集场景，我们引入了一个 K 近邻模块，在嵌入空间中执行基于实例的推理。之后

训练 f_{cnn}后，我们收集所有训练嵌入$\{\mathbf{z}_j\}$ 并拟合一个距离加权 KNN 分类器。
在测试时，对于查询样本 $\mathbf{z}$，K近邻的输出计算如下：

$$\mathbf{p}_{\text{KNN}} = \frac{\sum_{k=1}^{K} w_k \cdot \mathbf{y}_k}{\sum_{k=1}^{K} w_k}, \quad w_k = \frac{1}{\|\mathbf{z} - \mathbf{z}_k\|_2^2 + \epsilon}$$

其中，$\mathbf{z}_k$ 是具有相应标签 $\mathbf{y}_k$的 K 个最近邻，ϵ 是一个用于确保数值稳定性的小常数。

集成融合与开放集检测：我们通过凸组合融合图神经网络和K近邻的预测：

$$\mathbf{p}_{\text{ens}} = \alpha \cdot \mathbf{p}_{\text{GNN}} + (1 - \alpha) \cdot \mathbf{p}_{\text{KNN}}, \quad \alpha \in [0, 1]$$

为判断样本属于已知还是未知生成器，我们使用基于置信度的路由，阈值为 τ，并可选择使用基于熵的不确定性信号：

- **置信度阈值判定（主要方法）：**如果 $\max(\mathbf{p}_{\text{ens}}) < \tau$，则将该样本标记为未见。
- **熵阈值判定（可选）：**如果 $\mathcal{H}_{\text{attn}} > \tau_e$，则将该样本标记为未见。

除非另有说明，我们使用置信度阈值作为默认决策规则，并在 τ 第4.4节（图3）中进行消融；熵标准作为辅助不确定性信号纳入考量，但不用于主要模型选择。
训练目标：图神经网络头使用交叉熵损失在已见类别上进行训练：

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{N} y_i \log\left(\mathbf{p}_{\text{GNN},i}\right)$$

K 近邻模块是非参数的，无需训练。

4 实验

4.1 基准数据集

我们在基于扩散模型的合成语音数据集（DiffSSD）（Bhagtani 等人，2025年）上开展研究。该数据集包含高质量的合成语音和真实语音样本，专为评估模型在合成语音来源追踪和未见生成器检测任务中的表现而设计。DiffSSD 总计包含约 200 小时带标签音频，其中包括 70,000 个合成语音样本和 24,226 个真实语音样本。DiffSSD 中的真实语音来自 LibriSpeech（提供 11,126 个样本）和 LJ Speech（提供 13,100 个样本）。合成语音部分由十种文本转语音系统生成器生成：八种开源生成器（GradTTS、OpenVoiceV2、ProDiff、WaveGrad2、Xttsv2、YourTTS、DiffGAN-TTS 和 UnitSpeech）以及两种商业工具（ElevenLabs 和 PlayHT）。每个合成语音样本都使用一组 5,000 条英文文本行生成，内容涵盖对话、天气、引言和一般描述等日常主题。这些文本行由 ChatGPT-3.5 生成，并经过仔细筛选以避免重复。数据集分为三个部分：训练集包含 31,690 个样本，验证集包含 7,423 个样本，测试集包含 54,613 个样本。此划分支持两种评估方式：闭集评估，即训练期间涵盖所有生成器类型；开放集评估，即测试集中包含训练期间未见的生成器所生成的合成语音，例如商业工具 PlayHT 和 ElevenLabs。

训练详情：我们遵循 DiffSSD 提供的预定义训练集/开发集/测试集划分。模型在训练集划分上训练，并使用开发集划分进行选择；所有最终结果均在留出测试集划分上报告（包括分布内/分布外设置）。我们使用 Adam 训练至多 50 个轮次，学习率为 1×10^{-3}，批量大小为 32，并根据开发集表现应用早停。

4.2 评估指标

为了评估所提出框架的性能，我们采用了广泛使用的指标：准确率（准确率）、F1分数（F1）和等错误率（EER）。准确率直观地衡量了分类正确的程度。F1分数是精确率和召回率的调和平均值，能够综合反映模型在类别不平衡场景中的表现。EER表示误接受率与误拒绝率相等时的点，因此对开放集评估尤为有参考价值。这些指标已被用于Phukan 等人（2025年）的深度伪造检测和来源归因研究，确保了不同领域间方法的一致性，并进一步证明了我们评估框架的泛化性。

PTM	DEV			TEST			分布内/分布外		
	准确率 ↑	F1 ↑	等错误率 ↓	准确率 ↑	F1 ↑	等错误率 ↓	准确率 ↑	F1 ↑	等错误率 ↓
KNN									
Whisper	92.59	90.15	5.52	85.41	84.50	7.83	60.36	58.15	25.56
Unispeech	86.14	85.06	16.05	75.13	73.32	19.99	43.14	41.37	30.14
x向量	82.63	80.27	14.32	76.39	74.54	19.47	49.22	47.61	32.91
wav2vec 2.0	88.41	86.38	7.18	87.76	85.91	8.22	50.37	48.04	27.33
WavLM	90.43	88.26	8.94	82.43	80.83	10.52	49.19	47.21	30.36
ECAPA	78.37	76.15	12.39	72.88	70.43	18.19	38.01	36.95	33.57
Mamba-T	89.32	87.19	11.05	80.91	79.30	10.45	67.91	65.36	28.39
Mamba-S	91.26	89.07	7.11	80.66	78.81	8.44	69.95	68.09	23.68
Mamba-B	**94.05**	**93.28**	**5.32**	**89.41**	**87.19**	**7.62**	**72.56**	**70.24**	**22.83**
GNN									
Whisper	**97.29**	96.11	**4.37**	**93.64**	93.21	5.78	64.13	62.88	22.46
Unispeech	91.26	89.26	15.89	81.38	75.18	17.54	47.28	45.86	28.85
x向量	86.95	82.94	12.33	80.23	79.25	17.48	52.34	50.58	27.61
wav2vec 2.0	92.05	85.20	5.74	93.40	90.79	6.92	54.01	52.06	24.38
WavLM	93.62	91.67	6.86	89.37	86.33	8.25	53.27	51.52	28.46
ECAPA	82.36	78.25	10.39	75.03	72.71	17.09	42.69	40.54	30.57
MAMBA-T	94.29	91.22	8.96	85.51	82.64	9.23	71.92	69.66	26.29
MAMBA-S	96.41	92.64	5.18	87.26	85.42	7.34	**72.75**	70.12	21.41
Mamba-B	96.35	**96.15**	5.29	93.40	**93.32**	**5.62**	71.22	**71.02**	**19.28**
K近邻 + 图神经网络									
Whisper	97.56	96.29	3.38	94.19	92.22	4.33	76.67	74.14	17.10
Unispeech	92.31	91.25	9.03	82.63	80.36	13.72	63.15	61.90	22.91
X向量	88.19	86.99	7.79	81.59	80.41	13.20	64.66	62.35	20.96
Wav2vec 2.0	93.46	90.34	3.95	94.38	92.37	5.39	72.05	70.26	19.65
WavLM	95.09	93.34	5.15	91.21	89.84	6.71	74.28	71.47	21.35
ECAPA	83.27	81.06	6.55	77.61	75.29	15.07	62.36	60.32	23.91
Mamba-T	95.64	93.12	4.16	87.37	85.28	7.94	80.17	78.66	20.31
Mamba-S	96.10	94.37	3.91	89.24	87.32	5.90	86.54	84.12	17.36
Mamba-B	**98.11**	**96.86**	**2.33**	**95.52**	**94.21**	**4.32**	**88.91**	**86.53**	**14.78**

表2：使用图神经网络、K近邻及其组合对不同预训练模型（预训练模型）进行的性能比较。蓝色渐变表示性能从高到低的变化。表 4 采用相同的方案。

4.3 实验结果

表1 展示了在单个预训练表示上使用全卷积网络和卷积神经网络分类器的性能。在各基线模型中，Mamba-B 在所有数据划分上的表现均稳定优于其他预训练模型。在卷积神经网络设置下，它在开发集上的准确率为 83.27%、F1分数为 82.63%、等错误率为 4.26%；同时，在分布内/分布外（分布内/分布外）设置下仍保持了较好的泛化能力（准确率：69.01%，等错误率：17.97%）。Whisper 和 wav2vec 2.0 等其他表示在已见场景中表现具有竞争力，但在存在未见生成器时性能显著下降。ECAPA 和 UniSpeech 的结果相对较低，反映出以说话人为中心的表示和单语嵌入在泛化至多样化生成伪迹方面存在局限。这些发现表明，尽管基于卷积神经网络的分类器能够利用强大的预训练表示进行归因，但在开放集场景中仍受到限制。

基线说明：我们评估三类方法：(i) 仅归因模型，假设所有测试样本均来自已知生成器，并输出闭集类别预测（例如，仅使用图神经网络）；(ii) 仅开放集方法等方法侧重于检测未见生成器，而不对已见类别进行细粒度归因（例如，仅使用 KNN）；(iii) 统一模型则同时执行归因和开放集检测。**SIGNAL** 属于第三类，并明确结合了基于 GNN 的归因与基于 KNN 的开放集推理。表2 汇报了这三种设置，便于直接比较归因准确率、开放集检测及其联合优化之间的权衡。单独的 KNN 分支在分布内/分布外检测方面优于 CNN/FCN 基线——例如，Mamba-B 的分布内/分布外等错误率为 22.83%，而 CNN 为 26.27%。图神经网络头通过对类别间关系建模，进一步提升了归因性能，并改善了包括 Whisper 和 Wav2Vec 2.0 在内的若干预训练模型的 F1分数和等错误率。值得注意的是，在开发集上，Whisper 的等错误率从 KNN 的 5.52% 降至 GNN 的 4.37%。组合式 GNN+KNN 配置——即我们提出的**SIGNAL**框架——在所有设置中均取得了最佳性能。使用 Mamba-B 时，**SIGNAL** 在开发集上达到准确率：98.11%、F1分数：96.86% 和等错误率：2.33%，同时对未见生成器展现出强大的泛化能力，其分布内/分布外准确率为 88.91%，并且

等错误率：14.78%。即使是Mamba-T这样更小的模型，在这种混合设置下也取得了显著提升，验证了图推理与基于局部邻域的推理所具有的互补优势。

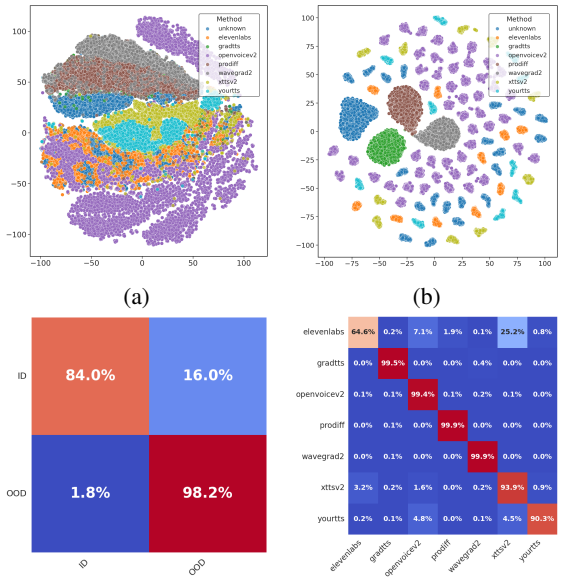


图2：子图a和b展示了t-SNE结果：(a)显示DiffSSD测试集上的原始嵌入空间，(b)则展示了经过基于图神经网络的细化后增强的类别分离效果。子图c和d展示了混淆矩阵：(c)显示了在DiffSSD上使用图神经网络+K近邻集成实现的分布内/分布外分离，(d)显示了整个测试集上的归因性能。

PTM	DEV			分布内/分布外		
	ACC	F1	EER	ACC	F1	EER
FCN						
Whisper	69.62	67.42	10.64	58.16	56.67	26.79
Unispeech	48.59	47.11	20.35	56.31	54.08	33.89
x向量	71.12	69.73	12.16	64.26	62.36	32.61
wav2vec 2.0	66.25	64.56	11.28	57.45	54.34	28.50
wavLM	42.81	41.19	23.30	59.25	57.13	34.11
ECAPA	70.85	68.34	13.98	62.58	60.24	36.22
MAMBA-T	70.92	69.46	9.04	60.62	59.37	30.52
MAMBA-S	79.56	78.02	6.22	63.76	61.52	25.22
MAMBA-B	**80.39**	**79.14**	**5.49**	**65.77**	**63.28**	**23.14**
CNN						
Whisper	71.64	68.91	8.66	61.03	60.17	21.26
Unispeech	50.05	49.88	18.05	59.16	58.28	27.91
x向量	73.04	71.33	10.23	67.39	66.58	25.80
wav2vec 2.0	67.71	66.17	9.38	60.74	59.95	24.52
wavLM	45.67	44.05	20.67	62.65	62.02	26.13
ECAPA	72.34	70.12	11.28	65.89	64.90	29.89
MAMBA-T	72.55	70.99	8.29	63.74	62.97	25.12
MAMBA-S	81.64	80.16	5.01	67.05	66.12	21.34
MAMBA-B	**83.27**	**82.63**	**4.26**	**69.01**	**68.12**	**17.97**

表3：使用全卷积网络和卷积神经网络分类器时，各种预训练模型在SingFake数据集上的性能。模型在DiffSSD上训练，并直接在SingFake上评估（零样本设置）。

这些结果再次证明，在合成语音检测不仅取决于模型规模或架构，还需要恰当结合表征与推理。 **SIGNAL** 通过有效结合结构化图学习与基于局部相似性的推理，体现了这一原则。此外，图2对 **SIGNAL** 的行为进行了全面的可视化分析。子图 2a和 2b展示了学习到的嵌入空间的 t-SNE 投影。在2a中，DiffSSD测试集的原始嵌入在不同生成器类别间呈现重叠区域，表明其可分性有限。经过2b中的基于图神经网络的细化后，各簇变得更加紧凑且彼此分离，凸显了图神经网络在建模类间关系、改善归因方面的有效性。子图 2c和 2d分别展示了分布内/分布外设置和完整测试设置下的混淆矩阵。

补充实验： 为评估我们的框架在合成语音之外的泛化能力，我们在 SingFake（Zang 等人，2024年）上进行了实验。SingFake 是一个用于歌声深度伪造检测（SVDD）的基准数据集。该语料库包含 28.93 小时的真实语音和 29.40 小时的深度伪造片段，涵盖五种语言和 40 位歌手，其中包括多种生成模型和音乐场景。我们在零样本设置下评估 **SIGNAL** ——直接在 SingFake 上测试使用 DiffSSD 训练的模型，不进行任何微调。从表 3可以看出，基线卷积神经网络分类器取得了中等水平的性能，其中 Mamba-B 表现最佳（F1分数：**82.63%**，等错误率：**17.97%**，在分布内/分布外划分上）。然而，表 4显示，我们的混合式 **SIGNAL** 架构显著提升了所有预训练模型的性能。Mamba-B 再次拔得头筹，在域内/域外划分下取得最高 F1分数 **90.99%**和最低等错误率 **7.92%** ，展现出对域外歌唱数据的强大泛化能力。即使是 Mamba-S 这样更小的模型，也从中获益良多；采用混合架构后，等错误率降至 **10.38%** 。这些发现证实了 **SIGNAL** 在声学结构不同的领域（如音乐和多语言内容）中的稳健性与可迁移性。

4.4 消融研究

为评估阈值选择对我们的图神经网络+K近邻混合架构的影响，我们在 DiffSSD 和 SingFake 数据集上，使用性能最佳的表征 Mamba-B 进行消融分析。如前所述，我们的模型采用基于置信度的路由机制：预测置信度高于某一阈值的样本

PTM	DEV			分布内/分布外			DEV			分布内/分布外			DEV			分布内/分布外		
	准确率 ↑	F1 ↑	等错误率 ↓	准确率 ↑	F1 ↑	等错误率 ↓	准确率 ↑	F1 ↑	等错误率 ↓	准确率 ↑	F1 ↑	等错误率 ↓	准确率 ↑	F1 ↑	等错误率 ↓	准确率 ↑	F1 ↑	等错误率 ↓
	KNN						GNN						K近邻 + 图神经网络					
Whisper	75.51	74.31	7.57	62.97	61.04	18.26	77.45	76.44	6.17	64.98	64.07	16.19	85.52	84.41	4.92	70.46	69.61	14.13
Unispeech	70.27	69.76	17.71	61.19	60.26	19.17	72.42	71.34	15.48	63.01	62.26	17.08	80.38	79.17	14.73	68.32	67.36	15.67
x向量	76.39	75.76	9.54	67.31	66.34	16.35	78.85	77.91	8.73	71.77	70.26	15.36	87.34	86.29	7.18	77.64	76.86	14.88
wav2vec 2.0	71.77	70.13	9.01	61.72	60.19	15.61	73.45	71.97	8.31	64.65	63.81	14.34	80.89	79.76	6.64	70.06	69.15	13.83
WavLM	74.57	73.01	17.83	63.44	61.52	18.87	76.54	75.07	15.66	66.79	66.12	16.27	85.17	84.32	14.39	72.41	71.33	16.02
ECAPA	73.89	72.11	10.32	68.28	65.22	21.36	78.22	77.18	9.74	70.05	69.27	19.56	86.61	85.48	8.75	76.04	75.20	19.07
MAMBA-T	76.65	75.32	7.99	64.15	62.78	1	78.22	77.44	7.38	67.97	67.15		87.14	86.19	6.76	73.96	72.29	13.92
MAMBA-S	84.11	83.28	4.87	67.64	66.91		86.12	85.16	4.58	71.42	70.42	11.96	89.29	87.95	4.29	77.26	76.49	10.38
Mamba-B	**86.79**	8	**4.19**	**70.14**			**87.20**	8	**4.18**	**73.47**	7	**9.52**	**92.28**	9	**3.81**	**79.66**	7	**7.92**

表4：使用 K近邻、图神经网络以及所提出的 GNN+KNN，在 SingFake数据集上对不同预训练模型的性能进行评估混合式 **SIGNAL**。

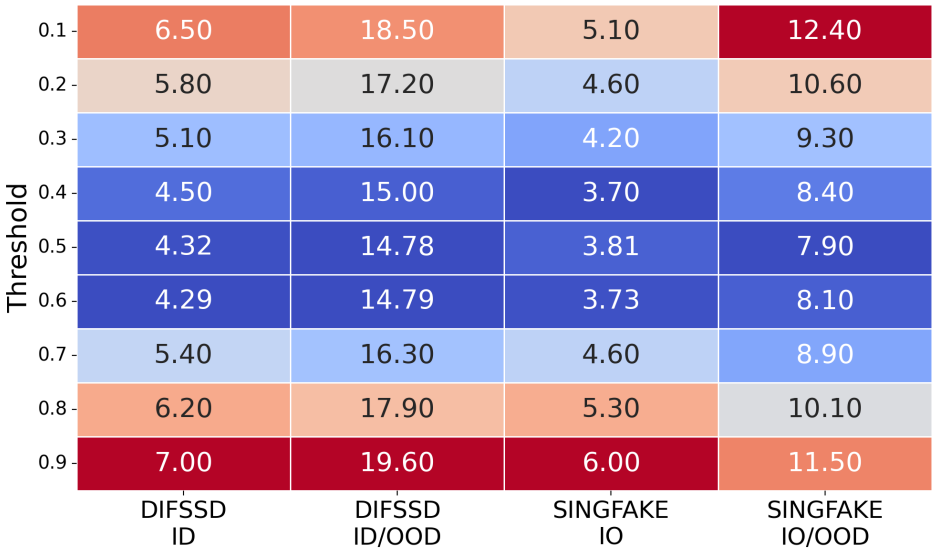


图3：Mamba-B 下 **SIGNAL** (K近邻+图神经网络) 的阈值 τ 敏感性。

τ 将被判定为来自已见生成器，而低于 τ 的样本则标记为未见。图3展示了不同阈值（$\tau \in [0.1, 0.9]$）下的等错误率（EER），评估涵盖四个数据划分：DiffSSD 分布内、DiffSSD 分布内/分布外、SingFake 分布内，以及 SingFake 分布内/分布外。结果清楚地表明，$\tau = 0.5$ 能够实现均衡的折衷，在分布内和开放集划分上均可最大限度地降低等错误率。这项分析验证了我们的设计选择：将 $\tau = 0.5$ 设为**SIGNAL**中的开放集决策边界。

5 结论

本研究应对合成语音归因与未见生成器检测这一双重挑战。我们发现，基于 Mamba 的表征能够高效捕捉生成器特定特征，包括韵律模式和合成伪影，这得益于其先进的时序建模能力。基于这一发现，我们提出了一种混合框架 **SIGNAL**，将图结构关系建模与实例级 K近邻推断相结合，以利用全局结构和局部相似性。本研究凸显了图增强建模在稳健合成语音检测方面被忽视的潜力，为推进生成语音取证中的归因与泛化奠定了坚实基础。本研究为结构化且具备泛化能力的合成语音分析未来研究打下了坚实的基线。

6 局限性与未来工作

首先，我们在两个公开可用的基准数据集（DiffSSD 和 SingFake）上进行了评估。尽管这些数据集具有多样性，但并未涵盖所有可能的合成条件或生成器。其次，SIGNAL 依赖于用于开放集检测的决策阈值；尽管该阈值在实证上表现稳定，但在完全不受约束的部署场景中，阈值的选择仍然是一项挑战。第三，我们当前的框架

能够识别样本是否源自未见生成器，但无法在多个未见来源之间进行细粒度归因。未来，一个有前景的研究方向是将 SIGNAL 扩展至对未见生成器进行分层建模或基于聚类的建模。

可解释性与可解释性： 尽管该框架取得了出色的性能，但其决策过程仍有些不透明。目前的框架缺少相关机制，无法解释为什么某个语音样本会被归因于特定生成器，或被标记为未见。这限制了它在透明度至关重要的高风险场景中的实用性。通过注意力可视化或原型级解释来增强可解释性，是未来工作的重要方向。

阈值敏感性： SIGNAL 依赖决策阈值来平衡归因与开放集检测。尽管图3 显示，在较宽的取值范围内性能稳定，但我们也承认，在所有现实部署场景中，不存在普遍适用的固定阈值。自适应或数据驱动的阈值选择仍是未来工作的重要方向。

7 伦理声明

本研究针对合成语音和歌声生成日益引发的关注，提出了检测并归因人工生成内容的方法。我们仅使用公开可用的数据集（DiffSSD 和 SingFake），不涉及任何个人或敏感数据。我们的框架纯粹用于检测和研究，而非生成或滥用合成内容。

参考文献

Alexei Baevski、Yuhao Zhou、Abdelrahman Mohamed 和 Michael Auli。2020年。wav2vec 2.0：一种用于语音表征自监督学习的框架。神经信息处理系统进展，33：12449–12460。

Kratika Bhagtani、Amit Kumar Singh Yadav、Paolo Bestagini 和 Edward J Delp。2025年。Diffssd：一个基于扩散模型的语音取证数据集。载于*ICASSP 2025-2025 IEEE* 国际声学、语音与信号处理会议 *(ICASSP)*，第 1–5 页。IEEE。

Sanyuan Chen、Chengyi Wang、Zhengyang Chen、Yu Wu、Shujie Liu、Zhuo Chen、Jinyu Li、Naoyuki Kanda、Takuya Yoshioka、Xiong Xiao 和另外 1 位作者。2022年。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，16(6):1505–1518。

Sanyuan Chen、Yu Wu、Chengyi Wang、Zhengyang Chen、Zhuo Chen、Shujie Liu、Jian Wu、Yao Qian、Furu Wei、Jinyu Li 和 Xiangzhan Yu。2021年。Unispeech-sat：通过具备说话人感知能力的预训练进行通用语音表征学习。*ICASSP 2022 - 2022 IEEE* 国际声学、语音与信号处理会议 *(ICASSP)*，第 6152–6156 页。

Erica Cooper、Cheng-I Lai、Yusuke Yasuda、Fuming Fang、Xin Wang、Nanxin Chen 和 J unichi Ya- magishi。2020年。采用最先进神经说话人嵌入的零样本多说话人文本转语音。在*2020*年国际电气与电子工程师协会声学、语音与信号处理国际会议（*ICASSP*），第 6184–6188 页。IEEE。

Brecht Desplanques、Jenthe Thienpondt 和 Kris De- muynck。2020年。Ecapa-tdnn：在基于 tdnn 的说话人验证中强调通道注意力、传播与聚合。发表于*2020*年国际语音通信协会年会，第 3830–3834 页。

Falih Gozi Febrinanto、Kristen Moore、Chandra Thapa、Jiangang Ma、Vidya Saikrishna 和 Feng Xia。2025年。Visiongraph：通过有限标签进行音频深度伪造检测的非对比学习。*arXiv* 预印本 *arXiv:2501.04942*。

Albert Gu 和 Tri Dao。2024年。Mamba：基于选择性状态空间的线性时间序列建模。收录于第一届语言建模会议。

Yinlin Guo、Haofan Huang、Xi Chen、He Zhao 和 Yuehai Wang。2024年。采用自监督WavLM和多重融合注意力分类器的音频深度伪造检测。收录于*ICASSP 2024-2024 IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第12702–12706页。IEEE。

Lian Huang 和 Chi-Man Pun。2024年。用于重放和深度伪造音频检测的自注意力与混合特征。*ArXiv*预印本 *arXiv:2401.05614*。

Zeqian Ju、Yuancheng Wang、Kai Shen、Xu Tan、Detai Xin、Dongchao Yang、Yanqing Liu、Yichong Leng、Kaitao Song、Siliang Tang、Zhizheng Wu、Tao Qin、Xiang-Yang Li、Wei Ye、Shikun Zhang、Jiang Bian、Lei He、Jinyu Li 和 Sheng Zhao。2024年。Natural-speech 3：采用因式分解编解码器和扩散模型的零样本语音合成。收录于第*41*届国际机器学习会议论文集，ICML'24。JMLR.org。

Zhifeng Kong、Wei Ping、Jiaji Huang、Kexin Zhao 和 Bryan Catanzaro。2021年。Diffwave：一种用于音频合成的多功能扩散模型。发表于国际学习表征会议。

Matthew Le、Apoorv Vyas、Bowen Shi、Brian Karrer、LedaSari、Rashel Moritz、Mary Williamson、Vimal Manohar、Yossi Adi、Jay Mahadeokar，以及其他 1 位作者。

2023。Voicebox：大规模文本引导的多语言通用语音生成。神经信息处理系统进展，36：14005–14034。

Naihan Li、Shujie Liu、Yanqing Liu、Sheng Zhao 和 Ming Liu。2019年。使用 Transformer 网络进行神经语音合成。载于《*AAAI* 人工智能会议论文集》，第 33 卷，第 6706–6713 页。

Jagabandhu Mishra、Manasi Chhibber、Hye-jin Shim 和 Tomi H Kinnunen。2025年。迈向可解释的伪造语音归因与检测：一种用于表征语音合成器组件的概率方法。《计算机语音与语言》，第 101840 页。

Orchid Chetia Phukan、Girish、Mohd Mujtaba Akhtar、Swarup Ranjan Behera、Priyabrata Mallick、Pailla Balakrishna Reddy、Arun Balaji Buduru 和 Rajesh Sharma。2025年。使用多模态基础模型进行歌声深度伪造来源归因。载于 *Interspeech2025*，第 1673–1677 页。

Alec Radford、Jong Wook Kim、Tao Xu、Greg Brock- man、Christine McLeavey 和 Ilya Sutskever。2023年。通过大规模弱监督实现稳健的语音识别。载于国际机器学习会议，第 28492–28518 页。PMLR。Yi Ren、Chenxu Hu、Xu Tan、Tao Qin、Sheng Zhao、Zhou Zhao 和 Tie-Yan Liu。2021年。FastSpeech 2：快速、高质量的端到端文本转语音。载于国际学习表征会议 Hyun-seo Shin、Jungwoo Heo、Ju-ho Kim、Chan-yeong Lim、Wonbin Kim 和 Ha-Jin Yu。2024年。HM-Conformer：一种基于 Conformer 的音频深度伪造检测系统，采用层次池化和多层级分类令牌聚合方法。载于*ICASSP 2024—2024*年 *IEEE* 声学、语音与信号处理国际会议（*ICASSP*），第 10581–10585 页。IEEE。David Snyder、Daniel Garcia-Romero、Gregory Sell、Daniel Povey 和 Sanjeev Khudanpur。2018年。X 向量：用于说话人识别的稳健深度神经网络嵌入。*2018*年 *IEEE* 声学、语音与信号处理国际会议（*ICASSP*），第 5329–5333 页。Adriana Stan、David Co mbei、Dan Oneata 和 Horia Cucu。2025年。TADA：音频深度伪造的免训练归因与域外检测。载于*Interspeech2025*，第 1543–1547 页。Yun Tan、Xiaoqian Weng 和 Jiangzhang Zhu。2025年。基于图注意力网络的双通道伪造语音检测。《*Symmetry*》，17(5):641。Massimiliano Todisco、Xin Wang、Ville Vestman、Md. Sahidullah、Héctor Delgado、Andreas Nautsch、Junichi Yamagishi、Nicholas Evans、Tomi H. Kinnunen 和 Kong Aik Lee。2019年。ASVspoof 2019：伪造及合成音频检测的未来展望。载于 *Interspeech2019*，第 1008–1012 页。

Zhizheng Wu、Tomi Kinnunen、Nicholas Evans、Junichi Yamagishi、Cemal Hanilçi、Md Sahidullah 和 Aleksandr Sizov。2015年。ASVspoof 2015：首届自动说话人验证欺骗攻击及对策挑战赛。载于*INTERSPEECH 2015* 第 *16* 届国际语音通信协会年会，第 2037–2041 页。国际语音通信协会。

Sarthak Yadav 和 Zheng-Hua Tan。2024年。Audio Mamba：用于自监督音频表示的选择性状态空间。载于*Interspeech 2024*，第 552–556 页。

Jiangyan Yi、Chenglong Wang、Jianhua Tao、Xi-aohui Zhang、Chu Yuan Zhang 和 Yan Zhao。2023年。音频深度伪造检测：综述。*ArXiv*，abs/2308.14970。

Yongyi Zang、You Zhang、Mojtaba Heydari 和 Zhiyao Duan。2024年。Singfake：歌声深度伪造检测。载于*ICASSP 2024-2024 IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第 12156–12160 页。IEEE。