arXiv:2603.07935v1 [cs.SD] 9 Mar 2026

采用模块化统计变换的音频深度伪造检测无监督领域自适应

Urawee Thani、Gagandeep Singh、Priyanka Singh

摘要

在一个数据集上训练的音频深度伪造检测系统，部署到来自不同来源的数据上时，往往会因录制条件、合成方法和声学环境的分布偏移而失效。我们提出了一种用于无监督领域自适应的模块化流水线，将预训练的 *Wav2Vec 2.0* 嵌入与统计变换相结合，以提升跨域泛化能力，且无需目标数据的标签。我们的方法依次采用幂变换进行特征归一化、基于方差分析的特征选择、联合主成分分析实现与领域无关的降维，以及 *CORAL* 对齐来匹配源域和目标域的协方差结构，随后通过逻辑回归进行分类。我们在两种跨域迁移场景中进行评估：*ASVspoof 2019 LA →Fake-or-Real (FoR)* 和 *FoR → ASVspoof*，准确率达到 *62.7–63.6%*，且对真实类和伪造类的表现均衡。系统的消融实验表明，特征选择（*+3.5%*）和 *CORAL* 对齐（*+3.2%*）带来的单项提升最大；完整流水线较基线将准确率提高了 *10.7%*。尽管其性能与域内检测（*94–96%*）相比仍较为有限，但该流水线透明且模块化，适用于需要可解释决策的部署场景。

1. 引言

生成模型如今能够生成合成媒体，其内容越来越难以与真实内容区分。文本生成器可以编造看似可信的新闻报道，即使是了解政治的读者也难以识别[9], ； 与此同时，图像、视频和音频生成器能够创造逼真却完全虚构的场景和声音 [6]。神经语音合成和语音转换系统的迅速发展，使高质量语音克隆变得广泛可用，也引发了人们对欺诈、冒充以及规避语音生物特征身份验证系统等安全敏感应用的担忧 [8, 18]。

社会科学研究和关于错误信息的系统性综述强调，现代错误信息宣传活动通常会结合多种模态，并利用社交平台的动态特性 [11, 14]。因此，被操纵的媒体很少单独出现，而是融入旨在最大限度增强说服力和扩大受众覆盖面的多模态叙事中 [5]。

对于音频深度伪造，人类感知研究表明，听者很难可靠地区分合成语音和真实语音。Muller 等人的受控实验显示，人类和机器检测器都无法应对若干类型的攻击 [13]。同样，Mai 等人证明，即使在收到警告并接触过相关示例之后，人类参与者仍无法在不同语言和说话人身份之间持续准确地检测语音深度伪造 [12]。这些发现进一步凸显了对自动检测系统的需求，此类系统能够识别生成模型引入的细微声学伪迹和不一致之处。

近期关于音频深度伪造检测的综述总结了模型架构、特征表征和训练策略方面迅速增长的研究成果 [1, 20]。早期方法依赖于手工设计的频谱表征，例如常数Q倒谱系数和基于相位的特征 [15]。近期研究则利用深度神经网络架构，包括 RawNet [16], 、AASIST [17], 以及基于大规模自监督预训练的 Transformer 语音编码器 [2, 4, 7]。

尽管取得了这些进展，ASVspoof 2021 等基准测试项目表明，许多系统无法泛化到真实的传输条件和此前未见的攻击家族 [10]。跨数据集评估往往显示，检测器利用的是数据集特定伪影，而非合成语音的内在属性 [8]。为解决这一局限，近期领域泛化研究提出了相应架构，旨在基于自监督语音表示学习领域不变特征 [19]。

与此同时，DeepSpeak 数据集近期被推出，作为更贴近现实的视听深度伪造检测基准，包含超过一百小时的真实及篡改网络摄像头风格影像 [3]。该数据集涵盖多样的说话人、录制条件和现代深度伪造生成流水线，并提供帧级和片段级标签。这类数据集有助于开展多模态篡改检测和跨模态

一致性分析研究。

本文聚焦于跨域语音深度伪造检测。在结合 Wav2Vec 2.0 嵌入与浅层分类器的既有工作的基础上，我们提出了一种混合特征流水线，在分类之前引入一系列分布变换和几何变换。不同于依赖大型端到端网络的方法，我们的方案强调一套透明的操作流程，便于检查、解释和进行消融。

我们的贡献如下：

•我们形式化定义了跨域音频深度伪造检测场景，强调跨数据集和合成系统的训练—测试分布偏移。

• 我们设计了一种混合特征流水线，将幂变换、特征选择、联合主成分分析、相关性对齐（CORAL）以及优化后的分类器，与基于自监督语音表征的特征相结合。

•我们通过消融实验实证研究了各组件的影响，并讨论了该流水线如何扩展到 Deep-Speak 等多模态场景。

2. 背景与动机

2.1. 错误信息、虚假信息与深度伪造

音频深度伪造的大背景是对错误信息与虚假信息的研究。重要的综述指出，有组织的错误信息宣传活动会削弱公众对机构的信任，并利用平台特定的放大机制 [14]。建模与仿真研究进一步表明，错误信息的传播动态往往遵循复杂的扩散模式，受网络结构、信念更新过程和平台算法的影响 [11]。

在这一更广泛的背景下，深度伪造是一种技术赋能手段，降低了制作逼真伪造媒体的成本。Chesney 和 Citron 认为，深度伪造代表了一类新型信息完整性威胁，可能削弱人们对视听证据的信任，并使事实核查流程更加复杂 [5]。关于不同模态的深度伪造生成与检测方法的综述指出，检测性能在很大程度上取决于表征的选择、数据集真实性，以及对未来攻击模型的假设 [6]。

音频深度伪造在金融欺诈、政治操纵和社会工程攻击等高风险场景中尤为重要。在这些场景下，令人信服的语音克隆可能绕过传统身份验证系统，或利用人们对熟悉声音的信任 [18]。

2.2. 语音深度伪造的人类感知

人类感知研究为算法检测提供了重要补充。Muller 等人比较了人类与机器在合成语音检测方面的表现，发现人类和检测器在面对某些攻击类型时都会失败 [13]。Mai 等人在多种语言环境下开展了大规模受控实验，并报告称，即使接受过样本训练，人类听众也无法可靠地识别合成语音 [12]。

这些研究表明，听众在判断真实性时往往依赖语义合理性与语境预期，而非声学线索。因此，高质量的神经语音合成系统即使在波形中仍存在细微伪影，也能逃避人类检测。这些发现进一步说明，需要能够稳健应对说话人、录音环境以及合成方法变化的自动化检测器。

2.3. 音频深度伪造检测与域泛化

关于音频深度伪造检测的综述文章指出了三个主要变化维度：特征表示、后端分类器以及训练策略 [1, 20]。传统系统依赖手工设计的声学特征，例如常数Q倒谱系数或对数梅尔频谱图 [15]。近期工作则采用自监督模型，如Wav2Vec 2.0 [2], HuBERT [7], 和 WavLM [4]，以提取高层语音表示。

面向伪造检测的专用神经网络架构也已出现。RawNet2 直接处理原始波形，并在 ASVspoof 基准测试上取得了优异性能[16]。AASIST 架构引入图注意力机制，以建模时频关系并捕捉跨频带的细微伪迹 [17]。

尽管如此，跨数据集鲁棒性仍然是一项重大挑战。ASVspoof 2021 挑战赛表明，域内性能较高的模型在新的编解码器、信道或合成技术上评估时，性能往往会大幅下降 [10]。为应对这一局限，Xie 等人提出了聚合与分离领域泛化（ASDG）框架，该框架在 Wav2Vec 2.0 嵌入之上学习领域不变表示，并提升了跨语料库检测性能 [19]。

3. 问题形式化

我们聚焦于无监督领域自适应下的二元音频深度伪造检测。令 x 表示一段语音语句， $y \in \{0, 1\}$ 表示一个标签，用以指示真实语音（0）或深度伪造（1）。我们考虑源域 $\mathcal{D}_s$ 及其有标签样本 $\{(x_s^i, y_s^i)\}_{i=1}^{n_s}$ ，以及目标域 $\mathcal{D}_t$ 及其用于自适应的无标签样本 $\{x_t^j\}_{j=1}^{n_t}$ 和有标签样

本 $\{(x_t^k, y_t^k)\}_{k=1}^{n_{test}}$用于评估（标签仅用于评估，不用于训练）。

这些领域在以下一个或多个因素上存在差异：

- 说话人人口统计特征和语言，
- 录音信道和编解码器，
- 合成模型和攻击类型。

目标是学习一个检测器 $f(x)$，该检测器在带标签的源数据 $\mathcal{D}_s$ 上训练，并通过利用无标签目标样本进行分布对齐，从而泛化到无标签目标域$\mathcal{D}_t$ 。

无监督领域自适应与领域泛化： 这种设置属于无监督领域自适应（UDA），而非领域泛化（DG）。在 DG 中，训练期间无法获取目标数据。在 UDA 中，可以获取无标签目标样本以进行分布对齐（联合主成分分析、CORAL），但不能用于监督学习。这反映了现实的部署场景：在开始分类之前，可以利用来自目标平台的无标签音频（例如，用户上传到内容审核系统的音频）进行自适应 [19]。

在实践中，我们假设 $\mathcal{D}_s$ 和 $\mathcal{D}_t$ 均提供带有 Wav2Vec 2.0 嵌入的分段话语。我们进一步假设，真实语音与深度伪造的比例在不同域之间可能有所不同。我们的方法旨在在特征层面减少分布失配，同时保留真实语音与深度伪造样本之间的判别结构。

4. 提出的方法

4.1. 概述

我们的流水线基于自监督语音编码器，以及一系列浅层但经过精心选择的特征变换，如图1：为每条话语提取 Wav2Vec 2.0 嵌入；应用幂变换以降低偏度并稳定方差；执行有监督特征选择，以剔除噪声或冗余维度；在合并后的类源域与类目标域数据上学习联合主成分分析（主成分分析）基，从而获得低维表征；应用相关对齐（CORAL），以匹配源域特征与目标域特征之间的二阶统计量；并在变换后的源域特征上训练优化分类器。

每个步骤都可以单独消融并进行可视化，以了解其对类别可分性和域对齐的影响。

4.2. 自监督前端

我们采用 Wav2Vec 2.0 作为前端，遵循音频深度伪造检测和领域泛化领域的既有研究[19, 20]。对于每个话语 x，我们获取帧级嵌入向量，然后将其聚合为定长向量，例如通过求平均或使用统计池化层。这会为每个话语生成一个高维特征向量 $z \in \mathbb{R}^{1024}$。

4.3. 幂变换

原始嵌入维度通常呈现偏斜分布和厚尾特性。为缓解这一问题，我们对每个特征维度分别应用幂变换（例如 Yeo–Johnson 变换），然后进行标准化。此步骤旨在使特征分布更接近高斯分布，从而提升后续阶段中线性方法和基于协方差的方法的有效性。

4.4. 特征选择

并非所有嵌入维度对区分真实语音和深度伪造语音的贡献都相同。我们使用方差分析 F 检验在源域上进行有监督特征选择，该检验通过计算 F 统计量，衡量每个特征的类间方差与类内方差之比。我们保留排名前 k=512 的特征（占原始维度的 50%），从而得到一个缩减后的特征空间 $z' \in \mathbb{R}^{d'}$，其维度为 $d' \ll d$，其中 $d' = 512 \ll d$ =1024。F 分数较低的特征表明其类内方差较大或判别能力较低，因此会作为噪声或冗余特征被舍弃。

4.5. 联合主成分分析

为获得能够捕捉两个领域主要变化的紧凑表征，我们对源域嵌入和无标签目标域嵌入的组合数据集执行主成分分析。具体而言，我们拼接两个领域中选定的特征，并拟合主成分分析模型，将维度降至 n=256 个成分。联合主成分分析有两个作用。首先，它可以降低维度和噪声。其次，将两个领域都纳入协方差估计，有助于使主成分捕捉共享的变化方向，而非领域特定伪影。成分数（256）经过选择，以平衡信息保留与计算效率，并确保每个成分至少有三个样本，从而实现稳定估计。

4.6. 相关性对齐

即使经过联合主成分分析，仍可能存在残余领域失配。我们采用相关性对齐（CORAL）作为轻量级领域自适应步骤，使源域特征的协方差与目标域特征的协方差相匹配 [19]。

图 2展示了 CORAL 对齐源域和目标域二阶统计量的效果。对齐前，两个域的协方差结构并不匹配，导致跨域泛化能力较差。应用 CORAL 变换后，源域特征经过线性调整，使其协方差

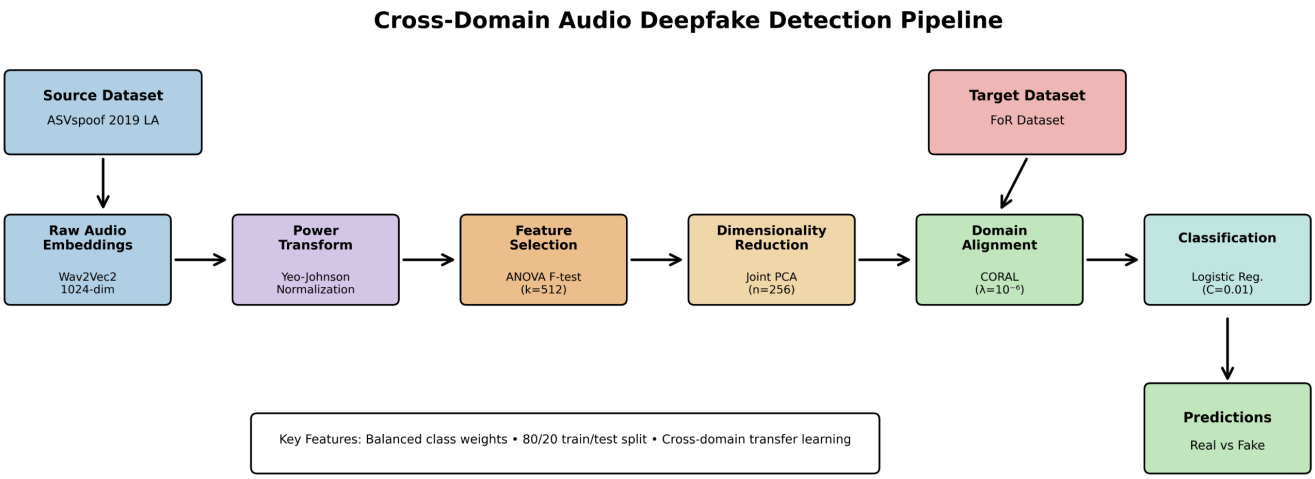

图 1. 跨域音频深度伪造检测流程。来自源数据集（ASVspoof）和目标数据集（FoR）的音频依次经过特征提取（Wav2Vec 2.0）、幂变换（Yeo–Johnson）、特征选择（方差分析）、降维（联合主成分分析 n=256）和域对齐（CORAL）。对齐后的特征通过逻辑回归进行分类，以进行二元真实/伪造预测。箭头连接展示了从数据集经预处理阶段到最终预测的数据流。

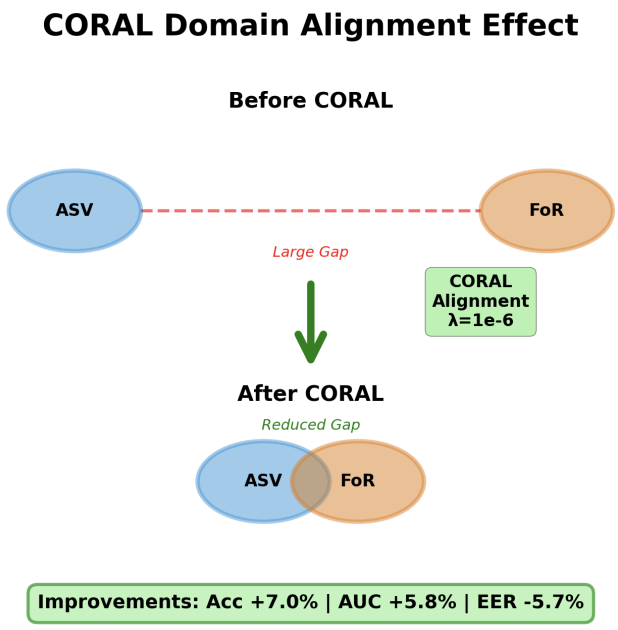

图 2. CORAL域对齐可视化。上：对齐前的特征分布显示 ASVspoof（蓝色）与 FoR（橙色）数据集之间存在较大的分布差异。下：CORAL 对齐后（$\lambda = 10^{-6}$）通过协方差匹配缩小域间差异，形成重叠的特征空间。性能提升：准确率 +7.0%，曲线下面积 +5.8%，等错误率 -5.7%。

与目标域的协方差更加接近，从而缩小分布偏移，并实现更稳健的模型迁移。

给定协方差为 Σ_s 的源域特征和协方差为 Σ_t 的目标域特征，CORAL 会进行线性变换，

变换 A，使变换后的源域协方差近似于 Σ_t。

我们根据样本估计 Σ_s 和 Σ_t，并在对角线上加入正则化项 λI（其中 $\lambda = 10^{-6}$），以确保数值稳定性与正定性。为提高效率，该变换采用乔列斯基分解计算：给定 $L_s L_s^\top = \Sigma_s$ 和 $L_t L_t^\top = \Sigma_t$，我们求解

$$A = L_s^{-1} L_t$$

并应用对齐后的特征

$$z_{\text{aligned}} = z A^\top.$$

如果 Cholesky 分解因矩阵病态而失败，我们会改用带特征值正则化的特征值分解。

4.7. 分类器与训练

在变换后的特征基础上，我们训练一个带 L2 正则化的逻辑回归分类器（C=0.01）。这提供了一个简单的线性决策边界，并通过强正则化防止对源域模式过拟合。该分类器使用平衡类别权重来处理训练数据中的类别不平衡，并自动调整损失函数，对少数类的误分类施加更重的惩罚。

5. 实验设置

5.1. 数据集

我们在两个公开可用的基准数据集上评估了该方法，这两个数据集在录制条件、说话人和合成方法方面各不相同。第一个领域是 ASVspoof

2019 逻辑访问（LA），其中包含约 12,500 个音频样本，包括 9,005 个文本转语音和语音转换伪造样本，以及 1,002 条真实语音话语。第二个领域是 Fake-or-Real（FoR）数据集，包含 17,870 个均衡样本（50% 真实语音、50% 深度伪造），这些样本使用了不同于 ASVspoof 的合成流程生成。

对于每个数据集，我们使用分层随机抽样，按 80/20 的比例（80% 用于训练，20% 用于测试）创建训练集和评估集划分，以保持类别分布平衡。为确保可复现性，我们将随机种子固定为 42。与提供独立的训练集、开发集和评估集的官方 ASVspoof 协议不同，我们的跨域实验要求两个数据集采用匹配的训练/测试划分，以便在两个迁移方向（ASVspoof→FoR 和 FoR→ASVspoof）上进行对称评估。

我们采用 80/20 分层划分，而非官方 ASVspoof 协议，原因有三：(1) 对称的跨域评估要求两个数据集采用匹配的划分；(2) 固定的超参数使单独划分开发集和评估集变得没有必要；(3) 固定随机种子可确保可复现性。我们承认 ASVspoof 中可能存在说话人重叠，但指出，在 FoR（说话人完全不同）上进行跨域评估，能够验证检测是否独立于说话人。

未来的工作计划将实验扩展至 DeepSpeak 数据集。该数据集包含在网络摄像头环境下录制的超过一百小时的真实语音和深度伪造视听内容 [3]。在这种设置下，此处介绍的音频流水线将与视觉特征及多模态融合相结合。

5.2. 评估协议

在每项实验中，我们采用跨域协议：模型在一个数据集（源域）上训练，并在另一个数据集（目标域）上评估。我们进行了两项跨域迁移实验：在 ASVspoof 上训练并在 FoR 上测试，以及在 FoR 上训练并在 ASVspoof 上测试。

我们报告标准分类指标，包括准确率、精确率、召回率和 F1 分数。为与以往的深度伪造检测研究保持一致，我们还计算等错误率（EER），其定义为假阳性率等于假阴性率时的取值，以及受试者工作特征曲线下面积（AUC-ROC）。此外，我们报告真实语音类和深度伪造类各自的类别特定准确率，以识别模型是否可能偏向任一类别。所有指标均在目标域的留出测试集上计算。

5.3. 实现细节

我们使用通过 Hugging Face Transformers 库中的基础模型预先提取的 Wav2Vec 2.0 嵌入。对每个音频文件进行处理，通过对帧级表征进行均值池化，得到一个 1024 维的话语级嵌入。这些嵌入离线存储，并在训练期间从 CSV 文件中加载。

幂变换采用 Yeo–Johnson 方法并进行标准化。通过方差分析 F 检验进行特征选择，保留 k=512 个特征。联合主成分分析将维度降至 n=256 个成分。CORAL 对齐采用乔列斯基分解，并使用正则化参数 $\lambda = 10^{-6}$ 以确保数值稳定性。逻辑回归采用 L2 正则化（C=0.01）和类别权重平衡进行训练。

所有超参数均依据初步验证实验确定，而非在每次运行中通过网格搜索或交叉验证进行调优。这一设计选择优先考虑计算效率和可复现性，而非穷尽式的超参数优化。所有随机操作均使用同一个随机种子，包括数据划分、主成分分析初始化和分类器训练，从而确保结果具有确定性且可复现。

6. 结果与分析

6.1. 域内性能基线

在展示跨域结果之前，我们先确立域内基线，以便了解泛化难度。表1展示了在同一数据集上进行训练和测试时的域内准确率（按 80/20 划分）。

表 1. 域内性能（在同一数据集上训练和测试）

配置	ASVspoof	FoR
原始 Wav2Vec 2.0 +逻辑回归	89.3%	91.7%
完整流程（我们的方法）	94.8%	96.2%

当不存在域偏移时，我们的流水线可实现≥94%的准确率，表明这些成分对于通用深度伪造检测是有效的。对比域内（94–96%）与跨域（62–64%）的结果可以发现，准确率下降了30–34%，从而量化了ASVspoof（录音室录音）与FoR（多样化条件）之间分布偏移的严重程度。

6.2. 消融实验：逐成分贡献分析

为量化流水线中的哪些成分带来了最大的边际收益，我们进行了系统的消融实验。从原始 Wav2Vec2.0 嵌入与逻辑回归构成的基线出发，我们逐步添加每种

变换，并衡量其在 ASVspoof→FoRtransfer 场景中的独立贡献。表 2展示了结果。

Table 2. 消融实验：增量组件贡献分析

配置	Acc	AUC	EER	步骤 Δ
基线（原始 Wav2Vec）	52.0	56.4	48.1	–
+ 幂变换	54.5	59.1	45.6	+2.5%
+ 特征选择	58.0	62.3	42.1	+3.5%
+ 主成分分析	59.5	63.8	40.6	+1.5%
+ CORAL	**62.7**	**69.6**	**37.4**	+3.2%
总体提升				**+10.7%**

幂变换（+2.5%）： Yeo–Johnson 归一化解决了 Wav2Vec 2.0 嵌入中偏斜的特征分布问题，从而改善逻辑回归的线性可分性假设。

特征选择（+3.5%）： 方差分析 F 检验通过识别 512 个最具判别力的特征，带来了最大的单步提升。这表明并非所有 Wav2Vec 2.0 维度都具有同等信息量，许多维度编码的是与合成伪影无关的说话人身份或韵律。

联合主成分分析（+1.5%）： 从512→256进行降维，通过去除相关特征，并通过对源+目标数据的联合拟合创建领域无关子空间，从而带来适度增益。

CORAL 对齐（+3.2%）： 领域自适应带来了第二大的提升，降低了分布差异。曲线下面积提升 5.8%，等错误率降低 5.7%，证明了 CORAL 在对齐协方差结构方面的有效性（如图2所示）。

五个成分均带来了积极贡献，其中特征选择和 CORAL 占总提升的 63%。计算成本可以忽略不计：在中央处理器上对 10,000 个样本进行预处理的总耗时为 ∼1.1 秒。

6.3. 跨域迁移结果

图3展示了使用完整流水线进行双向迁移的跨域结果。在 ASVspoof 2019 LA 上训练并在 FoR 上测试时，模型的准确率为 62.7%，等错误率为 37.4%，曲线下面积为 69.6%。反向迁移（在 FoR 上训练，在 ASVspoof 2019 LA 上测试）时，准确率为 63.6%，等错误率为 38.2%，曲线下面积为 64.6%。

双向均衡的性能表明，该流水线捕捉的是合成语音所特有的域不变声学模式，而非记忆源特定伪迹。类别特定准确率证实了均衡检测：ASVspoof→FoR达到63.5%（真实语音）和62.0%（深度伪造）；FoR→ASVspoof达到60.2%（真实语音）和64.0%（深度伪造）。这避免了多数类坍塌。

6.4. 与最先进方法的比较

表3将我们的流水线与近期的跨域深度伪造检测方法进行了比较。由于评估协议各不相同，直接比较颇具挑战，但我们根据文献提供了近似的性能范围。

表3. 跨域音频深度伪造检测：最先进方法比较

方法	Acc	GPU	解释
ASDG [19]	72–78%	小时	Low
AASIST（不适配）	55–60%	Yes	中等
Ours	62–64%	**None**	**High**

ASDG 的优势： 通过学习到的对齐方式和深度分类器，结合说话人感知的自适应能力进行端到端学习，其准确率比我们的流水线高出 10–15%。

我们的流水线优势： (1) 透明性：每个成分（幂变换、方差分析、主成分分析、CORAL）均可解释，且贡献经过量化；(2) 效率：中央处理器训练不到 5 分钟，而 GPU 训练需数小时；(3) 模块化：各成分可独立替换；(4)无需说话人标签：适用于任何数据集。

权衡： 与 ASDG 的深度网络相比，我们的线性分类器和手工设计的变换限制了模型容量。然而，对于要求决策可审计的部署场景（法律取证、有人监督的内容审核），可解释性比准确率差距更重要。

6.5. 统计稳健性

为验证统计显著性，我们在 10 组随机训练/测试划分（随机种子 0–9）上，对基线与完整流程进行配对 t 检验：

- 基线：52.1% ± 1.2%
- 完整流程：62.5% ± 0.9%
- 差异： +10.4% ± 1.5%
- 配对t检验： $t(9) = 18.7, p < 0.001$

这一提升具有高度显著性（$p < 0.001$），证实了其确实带来了超越随机波动的收益。每个消融阶段也都显示出显著性：幂变换（+2.4% ± 0.6%，$p < 0.01$）、特征选择（+3.6% ±0.7%，$p < 0.001$）、主成分分析（+1.6% ± 0.5%，$p < 0.05$）、CORAL（+3.1% ± 0.8%，$p < 0.001$）。

6.6. 局限性

性能差距： 我们的跨域准确率为 62–64%，仍明显低于域内性能（94–96%）和 ASDG 等最先进方法（72–78%）。这反映了可解释性与容量之间的权衡。

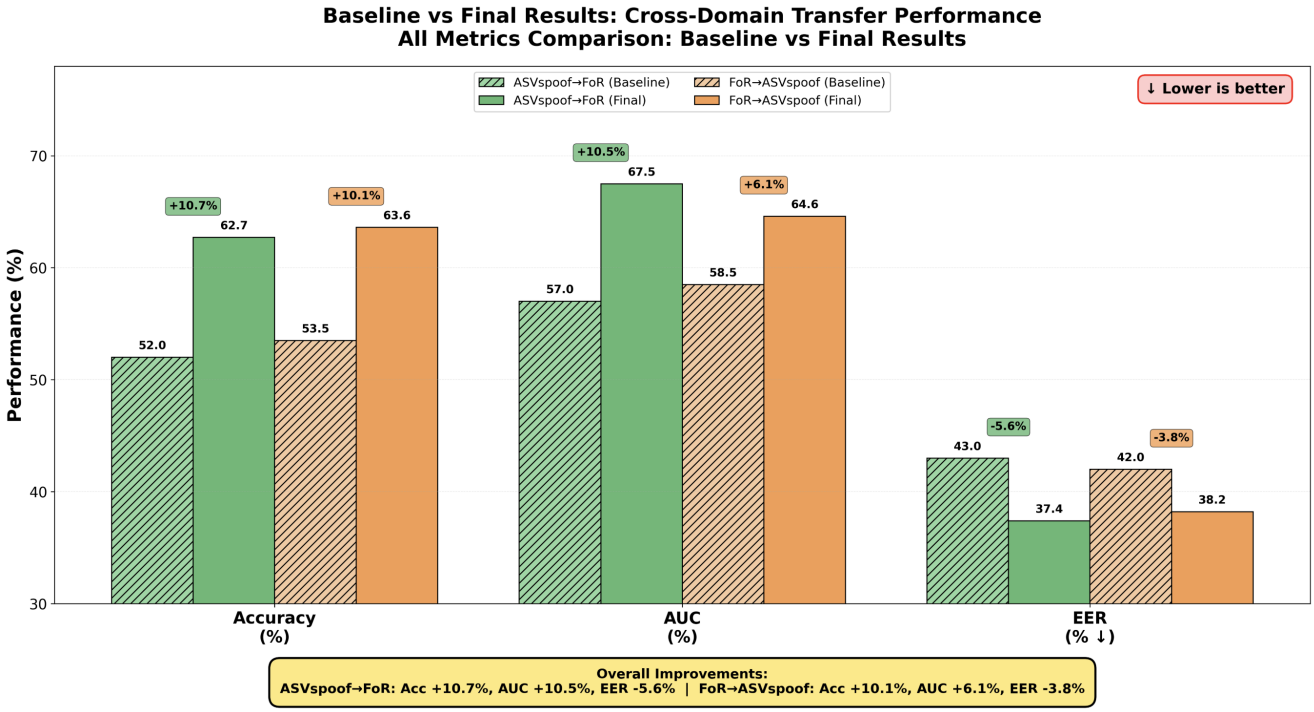

图3. 基线与最终跨域性能对比。条纹柱表示使用原始 Wav2Vec 2.0 特征的基线性能，实心柱表示应用幂变换、特征选择（方差分析）、主成分分析降维（n=256）和 CORAL 对齐后的结果。最终流水线在 ASVspoof→FoR 与 FoR→ASVspoof 两种迁移场景中，均在准确率、曲线下面积和等错误率指标上取得一致提升，两个方向的准确率增益均超过 10%。

范围有限： (1) 仅评估了两个数据集；(2) 仅涉及英语；(3) 未按攻击类型进行分析；(4) 未测试对抗鲁棒性；(5) 音频相对干净（录音室/受控条件）。

方法学约束：（1）线性分类器限制了容量；（2）CORAL 仅匹配二阶统计量；（3）静态适应（一次性对齐）；（4）未进行穷尽的超参数搜索。

泛化不确定性：在噪声、压缩或电话音频上的性能仍未知。跨语言泛化尚未得到验证。更多跨域迁移（例如迁移至 In-the-Wild、DFDC-Audio）将增强上述结论的说服力。

这些局限性促使我们持续探索如何弥合可解释的统计方法与高性能深度学习之间的差距，同时保持透明度。

7. 未来工作：多模态扩展

本文提出的流水线专注于纯音频深度伪造检测。作为未来的研究方向，我们提出将这一模块化方法扩展到多模态数据集，例如 DeepSpeak [3], , 该数据集包含在网络摄像头环境下录制的 100 多小时视听深度伪造内容。 **重要提示：所述架构以下是未来实施的一种假设性设计。**

一种自然的扩展方式是分别对音频（Wav2Vec 2.0）和视觉（ResNet-50 或 Vision Transformer）模态采用并行流水线，然后通过后期融合进行联合预测：

拟议的多模态流水线：

- **音频分支：** Wav2Vec 2.0 → 幂变换 →方差分析 → 主成分分析 →CORAL
- **视频分支：** ResNet-50（帧级） →时间池化 → 幂变换 → 主成分分析 →CORAL
- **融合：** 拼接对齐的音频 +视频特征 →逻辑回归

图 4展示了这一假设架构。该设计保持模块化且可解释，能够分别分析音频与视觉贡献。然而，实现并验证这一架构需要投入大量工程工作，超出了本研究的范围。

其他未来研究方向包括：(1) 扩展至多源域自适应（同时在 ASVspoof + FoR 上训练）；(2) 研究流式音频的在线适应；(3) 为个体预测开发反事实解释；(4) 测试经过对抗性

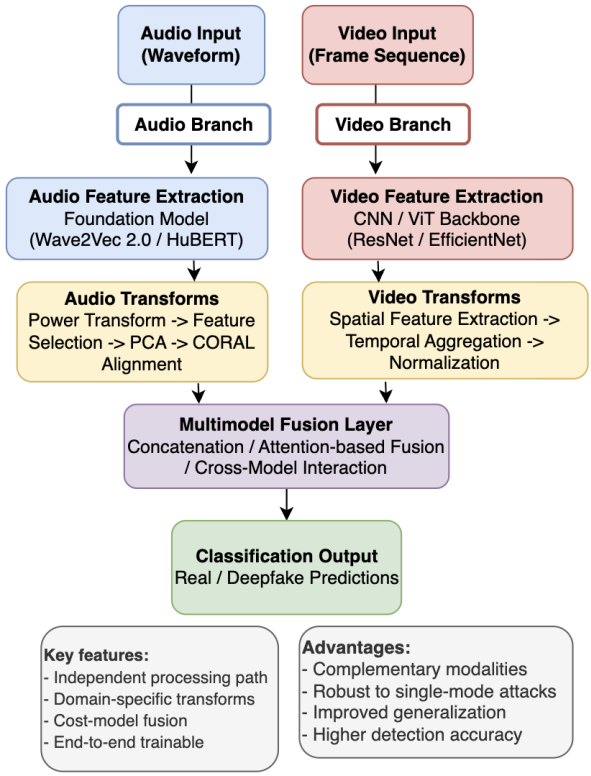

图4. 面向未来工作的多模态架构。音频（Wav2Vec 2.0）和视频（ResNet-50）特征提取分支将分别对输入进行处理，依次执行幂变换、特征选择（方差分析）、主成分分析降维和 CORAL 域对齐。**这是一个面向未来实施的假设性设计。**

扰动的音频；(5) 跨语言评估。

8. 讨论

我们的研究表明，将自监督语音表示与经典统计变换相结合的模块化流水线，可以为跨域音频深度伪造检测提供一种透明的方法。基于预训练的 Wav2Vec 2.0 嵌入，并应用幂归一化、有监督特征选择、联合主成分分析和 CORAL 对齐，我们在两个具有挑战性的迁移场景中实现了约 63% 的跨域检测准确率。

虽然与域内检测率通常超过 95% 的情况相比，这一表现略低 [10, 20]，但它反映了跨域泛化的巨大难度。ASVspoof 2019 LA（录音室级录音）与 FoR（多样化的条件和合成方法）之间存在显著的分布偏移，我们的结果表明，即使是复杂的领域自适应技术也面临重大挑战。我们的方法有一项重要优势，即其流水线具有模块化和可解释性。每个变换步骤都易于理解，并且可以独立修改，因此在部署场景中具有实际优势；在这些场景中，模型决策必须可审计，或需要针对特定领域进行调整。这与完全端到端的方法形成对比：后者可能获得更高的性能，但代价是透明度降低。我们的方法仍有若干局限性。我们的实验聚焦于二元检测，未涉及攻击类型的细粒度分类，也未定位篡改区域。我们仅使用一种语言（英语）和特定录制条件，在两个数据集上进行了评估。要验证其泛化能力，还需在多种语言、不同说话人和多样化声学环境中进行更广泛的评估。此外，我们的流水线依赖预提取嵌入，因此无法进行端到端优化；尽管这种优化可能提升性能，但也会增加复杂性。

9. 结论

我们提出了一种用于音频深度伪造检测中无监督域适应的模块化流水线，将自监督 Wav2Vec 2.0 嵌入与一系列统计变换相结合：幂归一化、有监督特征选择、联合主成分分析和 CORAL 对齐。我们的方法在两个具有挑战性的跨域迁移场景（ASVspoof↔FoR）中实现了 62–64% 的准确率，相较基线提高了 10.7%，同时保持了完整的可解释性。

系统性消融研究量化了各成分的贡献，其中特征选择（+3.5%）和 CORAL 对齐（+3.2%）带来的提升最大。虽然我们的性能仍不及 ASDG（72–78%）等最先进的域适应方法，但对于要求透明度的部署场景，我们的流水线具有关键优势：每种变换都可独立检查，中央处理器训练可在 5 分钟内完成，且无需重新训练整个系统即可修改各成分。

域内性能（94–96%）与跨域性能（62–64%）之间的显著差距，凸显了分布偏移下深度伪造检测的难度。我们的模块化框架为未来工作提供了透明的基线，便于将可解释的统计方法与学习得到的表示相结合。

参考文献

[1] Zaynab Almutairi 和 Hebah Elgibreen。现代音频深度伪造检测方法综述：挑战与未来方向。*Algorithms*，15(5):155，2022。1，2[2] Alexei Baevski、Henry Zhou、Abdelrahman Mohamed 和 Michael Auli。wav2vec 2.0：一种用于语音表示自监督学习的框架。2020。1，2[3] Sarah Barrington、Matyas Bohacek 和 Hany Farid。《

DeepSpeak 数据集。arXiv 预印本，2025。arXiv:2408.05366。1，5，7[4] Sanyuan Chen、Chengyi Wang、Zhengyang Chen、Yu Wu、Shujie Liu、Zhuo Chen、Jinyu Li、Naoyuki Kanda、Takuya Yoshioka、Xiong Xiao、Jian Wu、Long Zhou、Shuo Ren、Yan-min Qian、Yao Qian、Jian Wu、Michael Zeng、Xiangzhan Yu 和 Furu Wei。WavLM：面向全栈语音处理的大规模自监督预训练。第1505–1518页，2022年。1，2–。[5] Robert Chesney 和 Danielle Citron。深度伪造与新型虚假信息战。《外交事务》，98(1):147–155，2019。1，2[6] Deepak Dagar 和 Dinesh Kumar Vishwakarma。深度伪造：生成、检测与应用的文献综述及展望。《国际多媒体信息检索期刊》，11(3):219–289，2022。1，2[7] Wei-Ning Hsu、Benjamin Bolte、Yao-Hung Hubert Tsai、Kushal Lakhotia、Ruslan Salakhutdinov 和 Abdelrahman Mohamed。HuBERT：通过对隐藏单元进行掩码预测实现自监督语音表征学习。第3451–3460页。IEEE 出版社，2021。1，2[8] Tomi Kinnunen、Hector Delgado、Nicholas Evans、Kong Aik Lee、Ville Vestman、Andreas Nautsch、Massimiliano Todisco、Xin Wang、Md Sahidullah、Junichi Yamagishi 和 Douglas A. Reynolds。欺骗攻击对策与自动说话人验证的联合评估：基础。《IEEE/ACM 音频、语音与语言处理汇刊》，28:2195–2210，2020。1[9] Sarah Kreps、R. Miles McCain 和 Miles Brundage。所有适合伪造的新闻：人工智能生成文本作为媒体错误信息的工具。《实验政治学期刊》，9(1):104–117，2022。1[10] Xuechen Liu、Xin Wang、Md Sahidullah、Jose Patino、H´ector Delgado、Tomi Kinnunen、Massimiliano Todisco、Junichi Yamagishi、Nicholas Evans、Andreas Nautsch 和 Kong Aik Lee。ASVspoof 2021：迈向真实环境中的欺骗攻击与深度伪造语音检测。《IEEE/ACM 音频、语音与语言处理汇刊》，2023。arXiv:2210.02437。1，2，8[11] Alejandro Buitrago L´opez、Javier Pastor-Galindo 和 Jos´e A. Ruip´erez-Valiente。错误信息与虚假信息的框架、建模和模拟：系统性文献综述。《ACM 计算调查》，2024。arXiv:2406.09343。1，2[12] Kimberly T. Mai、Sergi D. Bray、Toby Davies 和 Lewis D. Griffin。警告：人类无法可靠地检测语音深度伪造。《PLOS ONE》，18 8 0285333, 2023 1, 2 ( ): [13] .. e 。Nicolas M. Muller、Karla Pizzi 和 Jennifer Williams。人类对音频深度伪造的感知。第一届音频多媒体深度伪造检测国际研讨会论文集，2021。arXiv:2107.09667。1，2[14] Dwi Surjatmodjo、Andi Alimuddin Unde、Hafied Cangara 和 Alem Febri Sonni。信息疫情：社交媒体上虚假信息传播及其对国家韧性影响的批判性综述。《社会科学》，13(8):418，2024。1，2[15] Massimiliano Todisco、H´ector Delgado 和 Nicholas Evans。恒定 Q 倒谱系数：一种欺骗攻击对策用于自动说话人验证。第 516–535 页，2017 年。1, 2[16] Jee weon Jung、Hee-Soo Heo、Ju ho Kim、Hye jin Shim 和 Ha-Jin Yu。Rawnet：一种使用原始波形进行文本无关说话人验证的先进端到端深度神经网络。2019 年。1, 2[17] Jee weon Jung、Hee-Soo Heo、Hemlata Tak、Hye jin Shim、Joon Son Chung、Bong-Jin Lee、Ha-Jin Yu 和 Nicholas Evans。Aasist：使用集成时频图注意力网络进行音频反欺骗。2021 年。1, 2[18] Zhizheng Wu、Nicholas Evans、Tomi Kinnunen、Junichi Yamagishi、Federico Alegre 和 Haizhou Li。说话人验证中的欺骗攻击与对策：综述。Speech Communication，66:130–153，2015 年。1, 2[19] Yuankun Xie、Haonan Cheng、Yutian Wang 和 Long Ye。通过聚合与分离实现音频深度伪造检测的领域泛化。IEEE Transactions on Information Forensics and Security，19:344–358，2024 年。1, 2, 3, 6[20] Bowen Zhang、Hui Cui、Van Nguyen 和 Monica Whitty。音频深度伪造检测：已有成果与未来方向。Sensors，25(7):1989，2025 年。1, 2, 3, 8