深度伪造语音检测的后训练

Wanying Ge, Xin Wang, Xuechen Liu, Junichi Yamagishi
National Institute of Informatics, Tokyo, Japan
{gewanying, wangxin, xuecliu, jyamagis}@nii.ac.jp

摘要—我们介绍了一种后训练方法，通过弥合通用预训练和领域特定微调之间的差距，使自监督学习（SSL）模型适应于深度伪造语音检测。我们提出了 AntiDeepfake 模型系列，这些模型是使用一个包含超过 56,000 小时真实语音和 18,000 小时具有各种伪影的语音的大规模多语言语音数据集开发的后训练模型。实验结果显示，这些后训练模型已经表现出对未见过的深度伪造语音的强大鲁棒性和泛化能力。当它们进一步在 Deepfake-Eval-2024 数据集上进行微调时，这些模型始终超越不使用后训练的现有最先进的检测器。模型检查点 [1] 和源代码 [2] 在线可用。

***Index Terms*—训练后，深度伪造检测，语音**

I. 介绍

自监督学习（SSL）显著推动了深度伪造语音对策（CMs）的发展和性能 [1]–[3] 。在深度伪造检测中，SSL 模型用作特征提取模块，并通常通过自我监督目标在数千小时的语音数据上进行预训练，以使其能够生成泛化能力强、有力的潜在表示，这些表示已被证明在解决下游任务方面非常有效 [4]–[8] 。

然而，自我监督的目标与深度伪造检测中的目标不同，在深度伪造检测中，主要目标是区分真实的语音（即由真人发声的语音）和带有伪影的语音。因此，从预训练 SSL 模型中提取的表示通常会在每次针对每个深度伪造数据集 [1], [3], [9], [10] 进行微调阶段使用特定领域的数据进一步优化。然而，为许多深度伪造数据集中的每一个学习这样的有意义的表示是资源消耗的。还有各种新的深度伪造检测任务，如部分伪造 [11], [12] 和源追踪 [13], [14]，因此单独为这些任务之一学习有意义的表示是低效的。是否有可能从现有的 *SSL* 模型创建一个新的基础模型，以提取更适合深度伪造检测的判别性表示？

[1]Zenodo:https://doi.org/10.5281/zenodo.15580542
Hugging Face:https://huggingface.co/nii-yamagishilab
[2]GitHub:https://github.com/nii-yamagishilab/AntiDeepfake

为了实现这一目标，我们采用**训练后**，在初始自监督学习和最终优化 [15]–[17] 之间的关键步骤。这个额外的训练阶段帮助我们更新 SSL 模型，使其更容易获得各种深度伪造数据集及相关任务中有用的表现形式。虽然预训练和后训练都使用了大量语音数据，但在数据类型和目标上存在显著差异（图 1）。建议的后训练使用真实的语音，并且更重要的是，多种带有瑕疵的语音类型（合成语音、转换语音、编码语音、恢复语音、编解码器语音等）。我们还引入了一个判别性目标，以便使表现形式更适合检测带有瑕疵的语音。虽然后训练和微调都旨在适应特定任务的模型，但在范围和目的上存在显著差异。后训练在大规模多样化的数据集上进行，并旨在为模型配备与领域无关的表现形式 [16]。另一方面，微调则在一个狭窄的小目标数据集上进行，专注于仅优化该特定领域的性能。

在这项工作中，我们开发了 AntiDeepfake 模型，一系列用于深度伪造检测的后训练基础模型，使用了一个包含超过 56,000 小时的真实语音和 18,000 小时带有各种伪影语音的大规模语音数据集，涵盖了一百多种语言。我们将 74,000 小时的语音数据应用于基于 wav2vec 2.0 [4], [18], [19] 和 HuBERT [5] 的 SSL 模型，并进行了不同大小的后训练。我们开展了两种类型的实验并发现，这些经过后训练的模型在没有微调的情况下已经表现出强大的稳健性和对未见深度伪造语音的泛化能力。此外，当它们进一步在 Deepfake-Eval-2024 数据集 [20] 上进行微调时，这些模型始终超越了那些不使用后训练的现有最先进的检测器。

II. 相关工作

A. 语音自监督学习模型

流行的语音 SSL 模型包括 wav2vec 2.0 [4], [18], [19] 和 HuBERT [5]。wav2vec 2.0 架构利用多层卷积

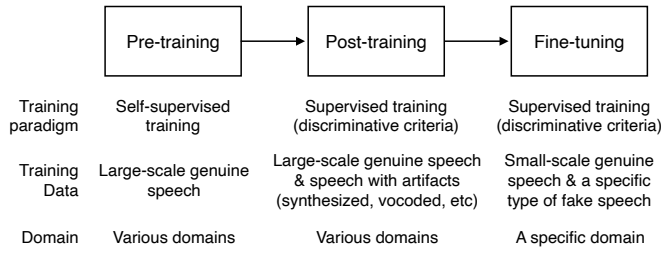


图 1: 提出的深度伪造相关任务的后训练流程。

编码器将原始语音波形处理成深度特征表示，并使用 Transformer [21] 捕捉整个序列的上下文信息。该模型通过预测随机掩码输入中的潜在语音表示进行自监督预训练。HuBERT 也利用了整个序列的上下文信息，并预测来自被掩码语音区域对应语音特征的 k-means 聚类索引。然而，这些自监督目标并没有设计用来区分真实的语音和深度伪造的语音，因此从根本上与深度伪造检测的目标不同。

B. 使用 vocoded 语音训练 SSL 模型

之前的尝试是为深度伪造检测训练 SSL 模型。例如，[22] 提出了使用在几种类型的编码语音上训练的 SSL 模型，并利用原始 SSL 表示和在编码语音上训练的 SSL 表示之间的差异进行深度伪造检测。虽然该方法的有效性得到了确认，但使用两个 SSL 模型进行特征提取或提炼两个 SSL 模型之间差异到学生模型的方法要么是低效的，要么是复杂的 [22]。

III. 深度伪造检测的后训练

A. 训练集

训练后阶段是一种监督学习，旨在通过让模型接触各种伪影类型来适应深度伪造检测的 SSL 模型。为此，我们为训练后阶段整理了一个多样化的训练集。如表 I 所示，我们的训练集结合了来自 27 个公开来源和两个自生成数据集的语音文件。这些来源根据伪影类型进行了分类：

1) **包含生成的深度伪造语音的数据集***:* 我们收集了深度伪造检测社区常用的多种数据集，从 ASVspoof 挑战赛 [23]–[25] 到最近的多样化合成器用于深度伪造语音检测语料库（DSD-corpus）[31]。这些数据集提供了广泛使用文本到语音 (TTS) 和语音转换 (VC) 算法生成的深度伪造语音，以及相应的真正语音样本。尽管大多数深度伪造数据集主要以英语和普通话为主导，我们也包括了最近的一些多语言数据集，如 HABLA [32]（拉丁美洲西班牙语）、多种语言音频防欺骗数据集 [33]（MLAAD，38 种语言）以及 SpoofCeleb [34]（多语言，源自 VoxCeleb1 [49]）[3]。

表 I: 用于训练后 SSL 模型的数据集统计（上部分）及其性能评估（下部分）。如果确切的语言数量未知，则将数据集标记为多语言。

Dataset	Language	Genuine (hrs)	Fake (hrs)	Attack
文本转语音和语音转换				
ASVspoof2019-LA [23]	en	11.85	97.80	TTS, VC
ASVspoof2021-LA [24]	en	16.40	116.10	TTS, VC
ASVspoof2021-DF [24]	en	20.73	487.00	TTS, VC
ASVspoof5 [25]	en	413.49	1808.48	TTS, VC
CFAD [26]	zh	171.25	224.55	TTS
DECRO [27]	en, zh	35.18	102.44	TTS, VC
DFADD [28]	en	41.62	66.01	TTS
Diffuse or Confuse [29]	en	0	231.66	TTS
DiffSSD [30]	en	0	139.73	TTS
DSD [31]	en, ja, ko	100.98	60.23	TTS, VC
HABLA [32]	es	35.56	87.83	TTS, TTS-VC
MLAAD [33]	38 languages	0	377.96	TTS
SpoofCeleb [34]	Multilingual	173.00	1916.2	TTS
VoiceMOS [35]	en	0	448.44	TTS
***vocoded* 语音**				
CVoiceFake [36]	en, fr, de, it, zh	315.14	1561.16	Vocoded
LibriTTS [37]	en	585.83	0	–
LibriTTS-Vocoded	en	0	2345.14	Vocoded
LJSpeech [38]	en	23.92	0	–
VoxCeleb2 [39]	Multilingual	1179.62	0	–
VoxCeleb2-Vocoded	Multilingual	0	4721.46	Vocoded
WaveFake [40]	en, ja	0	198.65	Vocoded
恢复的语音				
FLEURS [41]	102 languages	1388.97	0	–
FLEURS-R [42]	102 languages	0	1238.83	Restored & vocoded
LibriTTS-R [43]	en	0	583.15	Restored & vocoded
神经编码器语音				
Codecfake [44]	en, zh	129.66	808.32	Neural codec
CodecFake [45]	en	0	660.92	Neural codec
额外的真实言语				
AISHELL3 [46]	zh	85.62	0	–
CNCeleb2 [47]	zh	1084.34	0	–
MLS [48]	8 languages	50558.11	0	–
训练集	Over 100 languages	56.37 k	18.28 k	–

2) **带有 *vocoded* 语音的数据集***:* 为了增强模型检测语音生成过程中引入的波形级伪影的能力，我们将编码语音及其原始版本作为第二类。尽管这些编码语音文件的内容与其原始版本相同，但它们的波形是使用带有特定编码器伪影的编码器生成的。这一类别包括 CVoiceFake [36] 和 WaveFake [40]，这是 LJSpeech [38] 的编码版本。我们还通过处理来自 LibriTTS [37] 和 VoxCeleb2 [39] 的真实语音，并使用几种神经网络和 DSP 编码器 [50]，生成了额外的编码数据。

3) **恢复了语音的数据集***:* 我们将神经修复语音作为第三类包括在内。具体来说，我们在微调后加入了来

[3]虽然它最初是为了预测合成语音的平均意见得分而设计的，我们也把 VoiceMOS 数据集的合成部分 [35] 添加到了我们的训练集中。数据集中的真人语音部分被排除在外。

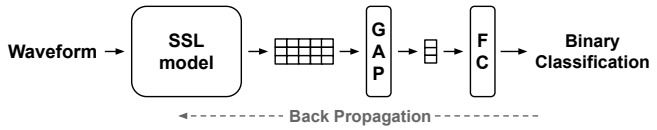


图 2: 用于训练后模型的架构。它包含一个自监督前端、一个全局平均池化层（GAP）和一个全连接层（FC）。

自 FLEURS-R [42] 和 LibriTTS-R [43] 以及它们的原始低质量版本 [41] [37] 的语音文件。语音内容和源说话人与原件相同，但通过高质量扩散声码器 [51] 提高了音频质量。

4) **具有神经编码器语音的数据集***:* 我们还在后训练集的第四类中包含了两个基于神经编解码器的数据集：Codecfake [44] 和 CodecFake [45]。这两个数据集都包含通过应用最近的神经编解码算法重新生成的真实语音文件。需要注意的是，编解码语音不应被视为深度伪造语音，而应视为带有伪影的语音后训练任务之一，而后训练的目标是获得能够检测细微差异的表现形式。

5) **包含额外真实语音的数据集***:* 为了进一步多样化训练集，我们纳入了仅包含真实数据的多语言数据集，涵盖了各种说话人、语言和声学条件。具体来说，我们包含了 AISHELL3 [46], CNCeleb2 [47]，它们通常用于说话人识别，并提供了广泛的说话人和录音条件，以及 Multilingual LibriSpeech (MLS) [48]。

B. 数据预处理和训练后处理过程

本研究中使用的所有语音文件均被重新采样至 16 千赫兹，转换为单声道，并进行了归一化处理。本文中报告的所有后训练过程都与图 2 中所示的方式相同。使用预训练的自监督学习（SSL）模型作为初始权重，从输入语音数据中提取序列级表示。这些表示直接被送入全局平均池化层，该层聚合了时间信息，随后是一个全连接（FC）层，生成一个输出分数，指示输入话语为真实语音的可能性，并且不包含属于上述任何一类的人工痕迹。然后使用基于交叉熵损失函数的反向传播进行后训练并更新 SSL 模型的权重。在这里，我们计算每个类别交叉熵之和，没有任何加权。在后训练阶段，SSL 模型和 FC 层被联合更新。后训练后的模型可以以零样本的方式用于提取深度伪造检测特征，或者进一步针对特定数据集或任务进行微调和优化。

IV. 实验

A. 实验条件

我们使用了在表 II 中所示的各种数据上预训练的不同大小的 wav2vec 2.0 和 HuBERT 进行后训练和评估。训练数据通过将相似时长的文件分组并零填充以形成小批量来动态采样。训练期间，超过 13 秒的文件被随机裁剪到 10 至 13 秒之间的时长。我们还使用了来自 [61] 的最佳配置 RawBoost 作为数据增强方法。对于测试，所有文件都保持其原始时长而不进行增强。

所有模型的最大批处理时长设置为 100 秒，除了最大的一个模型，`XLS-R-2B`，其最大批处理时长为 50 秒。我们每 100k 小批次（步数）进行一次验证。所有实验均使用了 AdamW 优化器 [62]，权重衰减值设置为 0.01。除了 `HuBERT-XL`（使用 5e-6 的学习率），所有模型均使用线性增加到 1e-7 的学习率，在前 80k 步进行优化，然后在 800k 步中线性衰减至 0，此时训练停止。这使得 `XLS-R-2B` 至少被训练数据更新一次超过 90%，而其他模型则看到了大部分训练数据两次。所有实验均在八个 NVIDIA H100 GPU 上使用相同的随机种子进行。

B. 多个测试集

我们的测试集包括 FakeOrReal [63] 和其第二个版本 FakeOrReal-norm, In-the-Wild [64], DEEP-VOICE [65], 音频深度伪造检测挑战赛（ADD 2023）赛道 1.2 评估集 [66] 和 Deepfake-Eval-2024 数据集 [20]。它们均未用于预训练和后训练。

FakeOrReal 和 In-the-Wild 是广泛用于深度伪造检测社区的基于英语的数据集，用于评估 CM 的泛化性能。ADD 2023 是一个基于汉语的挑战数据集，其深度伪造语音样本来自同一挑战的任务生成。DEEP-VOICE 包含八位名人的真实演示录音及其转换版本。最后，Deepfake-Eval-2024 数据集包含在 2024 年从社交媒体和深度伪造检测平台收集的各种真实和深度伪造多媒体内容。在我们的研究中，我们仅使用了其音频部分，其中包括超过 50 种语言的样本。

我们进行了两种类型的实验，零样本和微调性能评估。对于第 IV-C 节中描述的零样本性能评估，我们仅使用了所选测试数据集的测试部分；如果存在训练部分，则完全排除在模型训练和测试之外。请注意，DEEP-VOICE 和 Deepfake-Eval-2024 中的原始语音未分段并且可能持续数分钟，这需要大量的 GPU 内存不仅用于

表 II: 不同模型在各种测试数据集上的零样本评估等错误率（EER）结果，所有系统均未进行微调。对于训练后的情况，每个单元格以“使用 RawBoost / 不使用 RawBoost”的格式呈现结果。**粗体**中的值表示训练后的 SSL 模型中的最佳结果，而 下划线的 的值则突出显示每个测试数据集基线系统中的最佳结果。DEEP-VOICE 作为未见过的测试数据集，尚未公开报告其 EER 结果。

	模型 ID	参数数量。	ADD 2023	DEEP-VOICE	真假		In-the-Wild	Deepfake-Eval-2024
			Track-1.2-R2-Test	Segmented Full Set	original-Test	norm-Test	Full Set	Audio Test
Pre-training + Post-training	HuBERT-XL	964 M	18.90 / 35.34	5.67 / 14.87	2.49 / 3.67	3.17 / 15.52	5.23 / 17.99	34.08 / 47.72
	W2V-小	95 M	13.02 / 19.41	9.80 / 16.22	21.94 / 1.05	17.85 / 6.47	4.24 / 4.65	33.33 / 31.97
	W2V-大型	317 M	13.25 / 12.67	4.53 / 5.01	0.63 / 0.80	**0.97** / 1.44	1.91 / 2.25	33.38 / 30.05
	MMS-300M	317 M	7.93 / 11.22	**2.27** / 3.04	1.35 / **0.46**	5.92 / 2.71	2.90 / 2.00	32.80 / 31.38
	MMS-1B	965 M	9.06 / 9.46	2.56 / **2.27**	1.22 / 0.89	1.73 / 1.10	1.82 / 1.86	27.70 / 27.55
	XLS-R-1B	965 M	5.39 / 6.58	2.52 / 2.96	5.74 / 3.16	12.14 / 10.91	1.35 / 1.36	26.76 / 26.17
	XLS-R-2B	2.2 B	**4.67** / 6.84	2.30 / 2.63	2.62 / 1.18	1.65 / 1.73	**1.23** / 1.31	27.77 / **25.78**
Zero-shot evaluation results in the literature	XLSR-Mamba [52]	319 M	19.36	-	6.71	-	6.70	-
	Resemble AI [53]	2.1 B	6.11	-	1.36	-	3.94	-
	SpeechFake [2]	317 M	-	-	4.88	-	2.01	-
	Wav2Vec + VIB [31]	-	-	-	-	3.93	1.99	-
	UniSpeech-SAT [53], [54]	96 M	28.21	-	1.06	-	15.05	-
	XLS-R + SLS [55]	340 M	21.10	-	5.08	-	7.45	-
	XLSR-Conformer + TCM [56]	319 M	22.74	-	10.69	-	7.79	-
	AdaLAM & f-InfoED [57]	-	-	-	-	-	8.36	-
	P3 [20], [58]	317 M	-	-	-	-	-	43.00
	AASIST [20], [59]	0.3 M	32.47	-	21.64	-	43.01	55.22
	RawNet2 [20], [60]	18 M	64.55	-	65.68	-	49.19	48.20

基于 SSL 的方法，还用于其他基于深度神经网络的深度伪造 CMs。因此，我们在将其添加到我们的测试集之前手动将它们分成较短的片段。对于 DEEP-VOICE，我们发布的代码提供了在发布代码中使用的分段时间戳。对于 Deepfake-Eval-2024，我们使用了 [20] 中的方法将测试集中的每个音频文件分割成非重叠的 4 秒段落，使用 CMs 独立评分每一段，并测量这些段落上的等错误率（EER）。我们也多次进行了评估，在此期间，音频文件被分别分成 10 秒、13 秒、30 秒和 50 秒的片段。包含 13 秒时长是因为它对应于后训练和微调过程中使用的最大输入长度。来自 Deepfake-Eval-2024 的训练文件在第 IV-D 节描述的微调阶段中保持其原始长度。

C. 零样本评估结果

首先，我们分析了零样本评估结果。后训练模型在没有任何微调过程的情况下直接用于测试集。表 II 展示了使用数据增强和不使用数据增强的后训练模型的 EER 评估结果。模型按架构（HuBERT 和 wav2vec 2.0）分组，并根据其参数数量排序。

我们可以看到，尽管这些后训练模型并未针对任何测试集进行微调，它们仍然表现出对未见过的深度伪造语音的强大鲁棒性和泛化能力。结果总体上随着模型大小的增加而提高，除了 HuBERT-XL 以外。对于 ADD 2023 (4.67%) 和 In-the-Wild (1.23%) 的最佳 EER 结果由我们的最大模型 XLS-R-2B 实现，而对于 DEEP-VOICE (2.27%) 的最佳结果则来自我们第二大模型 MMS-1B。另一方面，在 FakeOrReal 测试集上，较小的模型往往表现更好，这些测试集的最佳结果来自于带有 317 M 参数的 SSL 模型（对于 MMS-300M 为 0.46%，对于 W2V-大型为 0.97%）。值得注意的是，W2V-大型在 FakeOrReal-norm (0.97%) 中取得了最佳结果，并且在原始的 FakeOrReal 数据集中排名第二 (0.63%)。此外，所有模型在 Deepfake-Eval-2024 数据集上的表现都不理想，尽管较大的模型仍然倾向于提供略好的性能。

关于基于 RawBoost 的数据增强的使用，表 II 中的结果并不表明它总是有益的。事实上，许多 CM 在用 RawBoost 训练后，在 FakeOrReal 数据集上的表现更差。然而，对于在 ADD 2023、DEEP-VOICE 和 In-the-Wild 数据集上测试的模型来说，其使用几乎总是有益的。

D. 微调和评估在 Deepfake-Eval-2024 上

在测试数据集中，Deepfake-Eval-2024 是最具挑战性的。有几个原因：它是在 2024 年从多个社交媒体平台手动收集并发布的，使其成为最新和最及时的数据集；它包含广泛的真实世界录音，具有复杂的声学背景条件。这些特点使得 Deepfake-Eval-2024 比其他测试集更难。因此，它作为微调和评估我们的 AntiDeepfake 模型泛化能力的理想基准。

为了比较，我们还包含了一个开源的 CM 系统 P3，该系统是通过类似的预训练过程构建的，在大规模声码器数据上进行了另一步预训练，并进行微调 [20]。图 3 说明了这两个模型的设置，具体细节如下所述。这个 P3 系统与提出的后训练框架高度相关，但使用不同的

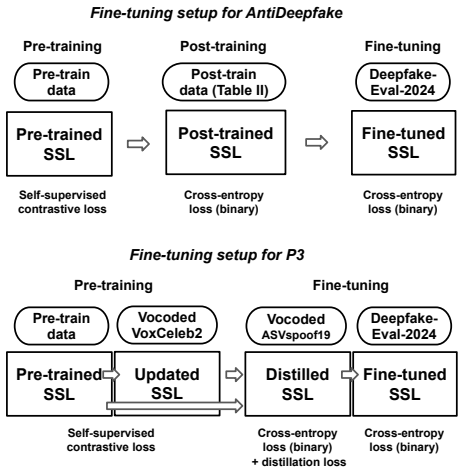


图 3: AntiDeepfake 模型（上）和 P3 模型 [20]（下）在 Deepfake-Eval-2024 数据集上的训练设置示例。

训练标准和多个微调阶段。比较预计将为可能改进提出后训练框架的设计选择提供线索。此外，我们还包含了非 SSL 系统作为参考。

1) ***AntiDeepfake* 模型的微调设置***:* 我们对最佳表现的后训练 AntiDeepfake 模型（报告于第 IV-C 节）进行了微调，使用了 Deepfake-Eval-2024 音频数据集的训练部分。我们随机选择了 90%的训练文件进行微调，并将剩余的 10%用于最佳模型检查点的选择。使用 AdamW 优化器以 1e-6 的学习率更新了模型参数共 6k 步，并应用了 RawBoost 增强。

2) ***P3* 模型的微调设置***:* P3 模型从一个在多语言人类语音数据上预训练的 wav2vec 2.0 模型（XLSR-53 [67]）开始。预训练的 SSL 模型权重使用 VoxCeleb2 及其 vocoded 版本进行了更新，采用与初始预训练阶段完全相同的自监督训练标准。在微调阶段，预训练的 SSL 和更新后的 SSL 模型被蒸馏成一个同样大小的 wav2vec 2.0 模型，并附加了一个二元分类交叉熵损失以及 ASVspoof 2019 数据集的 vocoded 版本 [23]。在微调过程中，全局平均池化和全连接层被添加到了学生 wav2vec 2.0 模型上。最后一步中，微调后的模型进一步在 Deepfake-Eval-2024 训练集上进行微调。两个微调步骤都使用了 Adam 优化器，学习率为 1e-6，并采用了 RawBoost 数据增强。

3) **非 *SSL* 模型的微调设置***:* 作为参考，流行的非 SSL 模型 AASIST [59] 和 Rawnet2 [60] 被包含在实验中。这两个系统首先使用标准交叉熵损失在 ASVspoof 2019 训练集上进行二值深度伪造检测的训练，然后在同一损失下在 Deepfake-Eval-2024 训练集上进行了微调。

表 III: AntiDeepfake 模型在 Deepfake-Eval-2024 测试集上不同输入时长的 EER 结果。我们比较了带有和不带后训练先验微调的模型。**粗体**字体表示每列中的最佳结果。

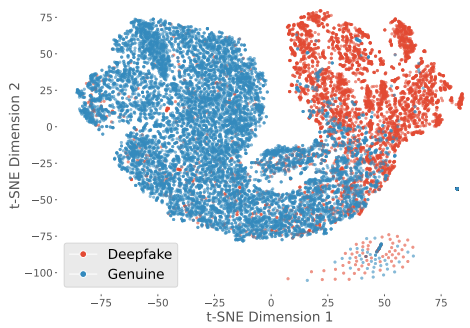
	预训练+后训练+微调					预训练 + 微调				
Model ID	4s	10s	13s	30s	50s	4s	10s	13s	30s	50s
W2V-大型	19.56	12.10	10.94	10.52	11.37	24.42	22.46	22.14	21.15	21.51
MMS-300M	17.15	13.37	12.31	11.05	10.75	19.77	13.29	12.77	12.01	12.29
MMS-1B	12.11	10.36	10.03	8.61	9.37	19.86	**10.32**	**11.55**	11.05	11.52
XLS-R-1B	**11.85**	10.00	**9.27**	**8.50**	**8.29**	19.95	17.18	16.31	10.63	11.21
XLS-R-2B	12.14	**9.80**	9.98	9.46	9.68	**12.88**	10.75	10.39	**9.67**	**9.98**
P3 [20]	-	-	-	-	-	15.38	-	-	-	-
AASIST [20]	-	-	-	-	-	16.99	-	-	-	-
RawNet2 [20]	-	-	-	-	-	20.91	-	-	-	-

4) **评估结果***:* 表 III 提供了在微调后选定的 AntiDeepfake 模型在 Deepfake-Eval-2024 测试集上的 EER 结果。为了评估后期训练的效果，我们在微调前比较了每个模型有和没有后期训练时的表现。如前所述，评估进行了多次，其中输入数据的持续时间为 4、10、13、30 或 50 秒。

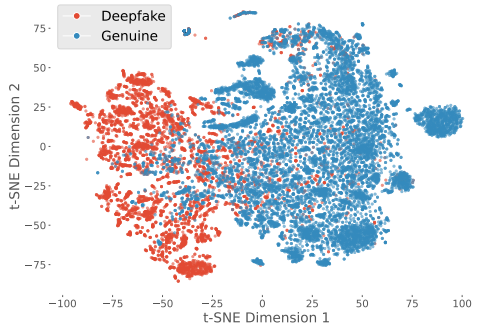
首先，我们可以看到持续一致地应用后训练明显提高了所有模型的 EER 性能。我们还可以观察到较大的模型表现通常更好，并且几乎所有模型都从后训练中受益。其中，XLS-R-1B 取得了最佳结果，其 EER 从 4 秒时的 11.85%降低到了 50 秒时的 8.29%——这是表格中的最低 EER。

我们还观察到，较长的输入持续时间通常会导致较低的 EER，因为较长的部分更有可能包含语音而不是静默区域。值得注意的是，后训练的好处在较短的持续时间内最为明显（4 秒和 10 秒）。例如，在 4 秒钟时，后训练将 XLS-R-1B 的 EER 降低了 7.1%，将 MMS-1B 的 EER 降低了 7.8%，这表明在低信息场景中改进了泛化能力。虽然 13 秒是训练期间使用的最大持续时间，但将输入长度延长到 30 秒和 50 秒继续降低 EER——尽管回报有限。

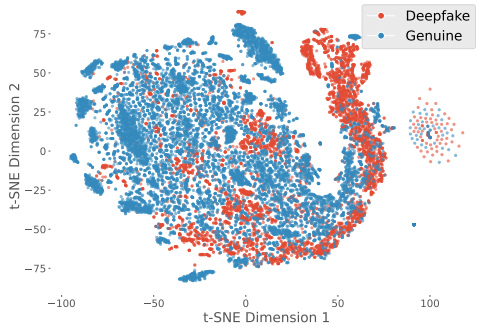
在使用 4 秒段进行评估的条件下，P3 模型的表现强于具有相似模型大小的 MMS-300M。尽管不同的 SSL 前端等因素可能影响了结果，但这种比较至少暗示了一些可以纳入所提出的后训练框架中的潜在技术。特别地，对比损失可能比交叉熵损失更适合用于后训练。这也提出了一个问题：合成数据是否足以进行后训练？如果是这样，由于可以在短时间内创建更大规模的合成数据，后训练的成本将显著降低。P3 模型与 AntiDeepfake 模型之间的比较为未来的工作提供了线索。



(a) 预训练+后训练+微调。



(b) 预训练+微调。



(c) 预训练+后训练。

图 4: t-SNE 可视化了整个 Deepfake-Eval-2024 测试集中经过 GAP 层后由 `XLS-R-2B` 模型提取的嵌入表示。蓝色点代表真实文件，红色点代表深度伪造文件。

E. 嵌入表示的可视化

为了分析微调后训练的效果，我们使用了 t-SNE 来可视化来自我们的最大 AntiDeepfake 模型 `XLS-R-2B` 的嵌入，如图 4 所示。图 4a、4b 和 4c 分别显示了具有（a）预训练+后训练+微调、（b）预训练+微调以及（c）预训练+后训练的模型的嵌入。

所有三个模型都在一定程度上展现了深度伪造（红色）和真实样本（蓝色）之间的分离。然而，仅使用预训练和微调的模型（图 4b）与经过后期训练后进行微调的模型（图 4a）相比，显示出更混杂且重叠的分布。

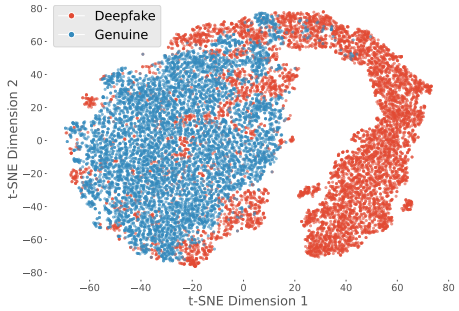


图 5: t-SNE 可视化了由 `XLS-R-2B` 模型提取的嵌入表示使用**预训练和后训练**的结果。**部分欺骗测试集**的一部分被使用。蓝色点代表真实文件，红色点代表部分深度伪造文件。

这一可视化表明后期训练提高了模型学习更具判别性的表示的能力。

最后，为了展示所提出的后训练生成的嵌入在其他不同深伪相关任务中的实用性，我们可视化了 PartialSpoof 数据集 [11] 中的一部分测试集，其中只有一部分片段而非整个话语是由 TTS 或 VC 系统合成的。请注意，测试集是平衡的，使得真实与伪造的比例为一对一。尽管 PartialSpoof 是一个不同的任务，但可视化结果显示真实的和伪造的类别通常是分开的。换句话说，从后训练模型获得的特征表示预计对其他与深伪音频相关的任务有用。

V. 结论与限制

我们介绍了一种新的深度伪造语音检测的后训练方法，以弥合通用自监督预训练与在特定目标域数据集上的微调之间的差距。在后训练阶段，我们将预先训练好的自监督模型暴露于真实的语音样本以及包含各种语言和领域中不同伪影的样本，并对其进行后训练以区分带有和不带伪影的样本。这些样本不仅包括使用文本到语音和语音转换系统的合成或转换语音，还包括超过一百种语言的编码语音、恢复语音和编解码器语音。实验结果表明，这种后训练使模型能够在未见过的深度伪造音频数据上实现高鲁棒性和泛化性能，即使在没有进行任何微调的情况下以零样本方式使用该模型也是如此。此外，我们还展示了后训练的模型为微调提供了更好的表示，并且与仅对预先训练好的自监督模型进行微调相比，通过微调后训练的模型在最具挑战性的测试集上的性能得到了提升。

尽管本文展示了后训练的有效性，但仍有一些方面需要进一步研究。例如，在本研究中使用了 RawBoost

作为后训练的数据增强方法，但在 Deepfake-Eval-2024 实验（表 II ）中显示，使用 RawBoost 进行数据增强对于处理来自现实世界的音频是不够的，并且可能需要使用如 MUSAN [68] 等方法来模拟在不同信噪比下的复杂背景噪声和背景音乐。此外，本研究在后训练时使用的训练标准为交叉熵损失，但这未必是最优的选择。根据第 IV-D 节的结果表明，基于其他标准如监督对比学习 [69] 进行的后训练也是可行的。此外，在后训练过程中并未对多种语音伪迹类别之间的损失权重进行调整，但定义优先级类别并在后训练期间改变权重是可能的。对于构建和生成耗时的文本到语音和/或基于声码转换合成语音是否在后训练中至关重要尚不清楚。从大量人类语音中轻松创建的编码、恢复和带编码的语音可能足以用于后训练。最后，由于篇幅限制，我们没有进行除深度伪造检测外的对完全伪造音频的其他实验。有必要扩展我们的研究范围，使用所提出的后训练模型的强大特征来进行来源追踪和部分伪造定位任务。这些任务是我们未来的工作。

致谢

本论文基于由新能源与工业技术开发组织（NEDO）委托的项目 JPNP22007 的研究成果。本研究部分得到了 JST AIP Acceleration Research (JPMJCR24U3) 和 JST PRESTO (JPMJPR23P9) 的支持。本研究使用了位于东京理科大学的 TSUBAME4.0 超级计算机进行。

参考文献

[1] H. Tak, M. Todisco, X. Wang, J. weon Jung, J. Yamagishi, and N. Evans, "Automatic speaker verification spoofing and deepfake detection using Wav2vec 2.0 and data augmentation," in *Proc. Odyssey*, 2022, pp. 112–119.

[2] W. Huang, Y. Gu, Z. Wang, H. Zhu, and Y. Qian, "SpeechFake: A large-scale multilingual speech deepfake dataset toward cutting-edge speech generation methods," 2025.

[3] T. Liu, D.-T. Truong, R. K. Das, K. A. Lee, and H. Li, "Nes2Net: A lightweight nested architecture for foundation model driven speech anti-spoofing," *arXiv preprint arXiv:2504.05657*, 2025.

[4] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Proc. NeurIPS 2020*, vol. 33, pp. 12 449–12 460, 2020.

[5] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[6] S.-w. Yang, H.-J. Chang, Z. Huang, A. T. Liu, C.-I. Lai, H. Wu, J. Shi, X. Chang, H.-S. Tsai, W.-C. Huang, T.-h. Feng, P.-H. Chi, Y. Y. Lin, Y.-S. Chuang, T.-H. Huang, W.-C. Tseng, K. Lakhotia, S.-W. Li, A. Mohamed, S. Watanabe, and H.-y. Lee, "A large-scale evaluation of speech foundation models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2884–2899, 2024.

[7] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2018, pp. 353–355.

[8] S. Zaiem, Y. Kemiche, T. Parcollet, S. Essid, and M. Ravanelli, "Speech self-supervised representation benchmarking: Are we doing it right?" in *Proc. Interspeech 2023*, 2023, pp. 2873–2877.

[9] X. Wang and J. Yamagishi, "Investigating self-supervised front ends for speech spoofing countermeasures," in *Proc. Odyssey*, 2022, pp. 100–106.

[10] D. Combei, A. Stan, D. Oneata, and H. Cucu, "WavLM model ensemble for audio deepfake detection," in *The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)*, 2024, pp. 170–175.

[11] L. Zhang, X. Wang, E. Cooper, N. Evans, and J. Yamagishi, "The PartialSpoof database and countermeasures for the detection of short fake speech segments embedded in an utterance," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 813–825, 2022.

[12] H.-T. Luong, H. Li, L. Zhang, K. A. Lee, and E. S. Chng, "Llama-PartialSpoof: An LLM-driven fake speech dataset simulating disinformation generation," in *Proc. ICASSP 2025*, 2025, pp. 1–5.

[13] N. Klein, T. Chen, H. Tak, R. Casal, and E. Khoury, "Source tracing of audio deepfake systems," in *Proc. Interspeech 2024*, 2024, pp. 1100–1104.

[14] Y. Xie, R. Fu, Z. Wen, Z. Wang, X. Wang, H. Cheng, L. Ye, and J. Tao, "Generalized source tracing: Detecting novel audio deepfake algorithm with real emphasis and fake dispersion strategy," in *Proc. Interspeech 2024*, 2024, pp. 4833–4837.

[15] H. Xu, B. Liu, L. Shu, and P. Yu, "BERT post-training for review reading comprehension and aspect-based sentiment analysis," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, 2019, pp. 2324–2335.

[16] G. Tie, Z. Zhao, D. Song, F. Wei, R. Zhou, Y. Dai, W. Yin, Z. Yang, J. Yan, Y. Su *et al.*, "A survey on post-training of large language models," *arXiv preprint arXiv:2503.06072*, 2025.

[17] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang *et al.*, "Qwen2.5 technical report," *arXiv preprint arXiv:2412.15115*, 2024.

[18] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi *et al.*, "Scaling speech technology to 1,000+ languages," *Journal of Machine Learning Research*, vol. 25, no. 97, pp. 1–52, 2024.

[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,

and M. Auli, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," in *Proc. Interspeech 2022*, 2022, pp. 2278–2282.

[20] N. A. Chandra, R. Murtfeldt, L. Qiu, A. Karmakar, H. Lee, E. Tanumihardja, K. Farhat, B. Caffee, S. Paik, C. Lee *et al.*, "Deepfake-Eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024," *arXiv preprint arXiv:2503.02857*, 2025.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[22] X. Wang and J. Yamagishi, "Can large-scale vocoded spoofed data improve speech spoofing countermeasure with a self-supervised front end?" in *Proc. ICASSP 2024*, 2024, pp. 10 311–10 315.

[23] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee *et al.*, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, pp. 101–114, 2020.

[24] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans *et al.*, "ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection," in *Proc. ASVspoof 2021 Workshop*, 2021, pp. 47–54.

[25] X. Wang, H. Delgado, H. Tak, J.-w. Jung, H.-j. Shim, M. Todisco, I. Kukanov, X. Liu, M. Sahidullah, T. Kinnunen *et al.*, "ASVspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale," *arXiv preprint arXiv:2408.08739*, 2024.

[26] H. Ma, J. Yi, C. Wang, X. Yan, J. Tao, T. Wang, S. Wang, and R. Fu, "CFAD: A Chinese dataset for fake audio detection," *Speech Communication*, vol. 164, pp. 103–122, 2024.

[27] Z. Ba, Q. Wen, P. Cheng, Y. Wang, F. Lin, L. Lu, and Z. Liu, "Transferring audio deepfake detection capability across languages," in *Proceedings of the ACM Web Conference 2023*, 2023, pp. 2033–2044.

[28] J. Du, I.-M. Lin, I.-H. Chiu, X. Chen, H. Wu, W. Ren, Y. Tsao, H.-y. Lee, and J.-S. R. Jang, "DFADD: The diffusion and flow-matching based audio deepfake dataset," in *2024 IEEE Spoken Language Technology Workshop*, 2024, pp. 921–928.

[29] A. Firc, K. Malinka, and P. Hanáček, "Diffuse or Confuse: A diffusion deepfake speech dataset," in *2024 International Conference of the Biometrics Special Interest Group*, 2024, pp. 1–7.

[30] K. Bhagtani, A. K. S. Yadav, P. Bestagini, and E. J. Delp, "Diff-SSD: A diffusion-based dataset for speech forensics," *arXiv preprint arXiv:2409.13049*, 2024.

[31] T.-P. Doan, H. Dinh-Xuan, T. Ryu, I. Kim, W. Lee, K. Hong, and S. Jung, "Trident of Poseidon: A generalized approach for detecting deepfake voices," in *Proceedings of ACM CCS 2024 Conference*, 2024.

[32] P. T. Flórez, R. Manrique, and B. P. Nunes, "HABLA: A dataset of Latin American Spanish accents for voice anti-spoofing," in *Proc. Interspeech 2023*, 2023, pp. 1963–1967.

[33] N. M. Müller, P. Kawa, W. H. Choong, E. Casanova, E. Gölge, T. Müller, P. Syga, P. Sperl, and K. Böttinger, "MLAAD: The multi-language audio anti-spoofing dataset," in *2024 International Joint Conference on Neural Networks*, 2024, pp. 1–7.

[34] J.-w. Jung, Y. Wu, X. Wang, J.-H. Kim, S. Maiti, Y. Matsunaga, H.-j. Shim, J. Tian, N. Evans, J. S. Chung *et al.*, "SpoofCeleb: Speech deepfake detection and SASV in the wild," *IEEE Open Journal of Signal Processing*, 2025.

[35] W. C. Huang, E. Cooper, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "The VoiceMOS Challenge 2022," in *Proc. Interspeech 2022*, 2022, pp. 4536–4540.

[36] X. Li, K. Li, Y. Zheng, C. Yan, X. Ji, and W. Xu, "SafeEar: Content privacy-preserving audio deepfake detection," in *Proceedings of ACM CCS 2024 Conference*, 2024, pp. 3585–3599.

[37] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," in *Proc. Interspeech 2019*, 2019, pp. 1526–1530.

[38] K. Ito and L. Johnson, "The LJ Speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[39] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech 2018*, 2018, pp. 1086–1090.

[40] J. Frank and L. Schönherr, "WaveFake: A data set to facilitate audio deepfake detection," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[41] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "FLEURS: Few-shot learning evaluation of universal representations of speech," in *2022 IEEE Spoken Language Technology Workshop*, 2023, pp. 798–805.

[42] M. Ma, Y. Koizumi, S. Karita, H. Zen, J. Riesa, H. Ishikawa, and M. Bacchiani, "FLEURS-R: A restored multilingual speech corpus for generation tasks," in *Proc. Interspeech 2024*, 2024, pp. 1835–1839.

[43] Y. Koizumi, H. Zen, S. Karita, Y. Ding, K. Yatabe, N. Morioka, M. Bacchiani, Y. Zhang, W. Han, and A. Bapna, "LibriTTS-R: A restored multi-speaker text-to-speech corpus," in *Proc. Interspeech 2023*, 2023, pp. 5496–5500.

[44] Y. Xie, Y. Lu, R. Fu, Z. Wen, Z. Wang, J. Tao, X. Qi, X. Wang, Y. Liu, H. Cheng *et al.*, "The Codecfake dataset and countermeasures for the universally detection of deepfake audio," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 386–400, 2025.

[45] H. Wu, Y. Tseng, and H. yi Lee, "CodecFake: Enhancing anti-spoofing models against deepfake audios from codec-based speech synthesis systems," in *Proc. Interspeech 2024*, 2024, pp. 1770–1774.

[46] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "AISHELL-3: A multi-speaker mandarin TTS corpus," in *Proc. Interspeech 2021*, 2021, pp. 2756–2760.

[47] L. Li, R. Liu, J. Kang, Y. Fan, H. Cui, Y. Cai, R. Vipperla, T. F. Zheng, and D. Wang, "CN-Celeb: Multi-genre speaker recognition," *Speech Communication*, vol. 137, pp. 77–91, 2022.

[48] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A large-scale multilingual dataset for speech research," in *Proc. Interspeech 2020*, 2020, pp. 2757–2761.

[49] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Proc. Interspeech 2017*, 2017, pp. 2616–2620.

[50] X. Wang and J. Yamagishi, "Spoofed training data for speech spoofing countermeasure can be efficiently created using neural vocoders," in *Proc. ICASSP 2023*, 2023, pp. 1–5.

[51] Y. Koizumi, K. Yatabe, H. Zen, and M. Bacchiani, "Wavefit: An iterative and non-autoregressive neural vocoder based on fixed-point iteration," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 884–891.

[52] Y. Xiao and R. K. Das, "XLSR-Mamba: A dual-column bidirectional state space model for spoofing attack detection," *arXiv preprint arXiv:2411.10027*, 2024.

[53] Speech Arena, "Speech Arena: Speech deepfake leaderboard," https://huggingface.co/spaces/Speech-Arena-2025/Speech-DF-Arena, 2025.

[54] S. Chen, Y. Wu, C. Wang, Z. Chen, Z. Chen, S. Liu, J. Wu, Y. Qian, F. Wei, J. Li, and X. Yu, "UniSpeech-SAT: Universal speech representation learning with speaker aware pre-training," in *Proc. ICASSP 2022*, 2022, pp. 6152–6156.

[55] Q. Zhang, S. Wen, and T. Hu, "Audio deepfake detection with self-supervised XLS-R and SLS classifier," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, p. 6765 – 6773.

[56] D.-T. Truong, R. Tao, T. Nguyen, H.-T. Luong, K. A. Lee, and E. S. Chng, "Temporal-channel modeling in multi-head self-attention for synthetic speech detection," in *Proc. Interspeech 2024*, 2024, pp. 537–541.

[57] B. Zhao, Z. Kang, Y. He, X. Qu, J. Peng, J. Xiao, and J. Wang, "Generalized audio deepfake detection using frame-level latent information entropy," *arXiv preprint arXiv:2504.10819*, 2025.

[58] X. Wang and J. Yamagishi, "Can large-scale vocoded spoofed data improve speech spoofing countermeasure with a self-supervised front end?" in *Proc. ICASSP 2024*, 2024, pp. 10 311–10 315.

[59] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *Proc. ICASSP 2022*, 2022, pp. 6367–6371.

[60] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with RawNet2," in *Proc. ICASSP 2022*, 2021, pp. 6369–6373.

[61] H. Tak, M. Kamble, J. Patino, M. Todisco, and N. Evans, "RawBoost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing," in *Proc. ICASSP 2022*, 2022, pp. 6382–6386.

[62] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *Proc. ICLR*, 2019.

[63] R. Reimao and V. Tzerpos, "FoR: A dataset for synthetic speech detection," in *2019 International Conference on Speech Technology and Human-Computer Dialogue*, 2019, pp. 1–10.

[64] N. Müller, P. Czempin, F. Diekmann, A. Froghyar, and K. Böttinger, "Does audio deepfake detection generalize?" in *Proc. Interspeech 2022*, 2022, pp. 2783–2787.

[65] J. J. Bird and A. Lotfi, "Real-time detection of AI-generated speech for deepFake voice conversion," *arXiv preprint arXiv:2308.12734*, 2023.

[66] J. Yi, J. Tao, R. Fu, X. Yan, C. Wang, T. Wang, C. Y. Zhang, X. Zhang, Y. Zhao, Y. Ren *et al.*, "ADD 2023: The second audio deepfake detection challenge," in *IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis*, 2023, pp. 125–130.

[67] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-Lingual representation learning for speech recognition," in *Proc. Interspeech 2021*, 2021, pp. 2426–2430.

[68] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[69] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 18 661–18 673.