arXiv:2303.01211v1 [cs.SD] 2 Mar 2023

向自己学习：一种用于伪造语音检测的自蒸馏方法

Jun Xue[1],*Cunhang Fa n*[1,*], *Jiangyan Yi*[2],*Chenglong Wang*[2],*Zhengqi Wen*[3],张丹[4],*Zhao Lv*[1,*]

[1]安徽省多模态认知计算重点实验室、计算机科学与技术学院，安徽大学，中国合肥230601[2]模式识别国家重点实验室，中国科学院自动化研究所，中国北京100190[3]启元实验室，中国北京100190[4]清华大学心理学系，中国北京100190e21201068@stu.ahu.edu.cn, {cunhang.fan, kjlz}@ahu.edu.cn, {jiangyan.yi,chenglong.wang}@nlpr.ia.ac.cn, wenzhengqi@qiyuanlab.com, dzhang@tsinghua.edu.cn

摘要

本文提出一种用于伪造语音检测（FSD）的新型自蒸馏方法，无需增加模型复杂度即可显著提升 FSD 的性能。对于 FSD，一些细粒度信息至关重要，例如声谱图缺陷、静音片段等，这些信息通常由浅层网络感知。然而，浅层网络中的噪声较多，难以很好地捕捉这些信息。为解决这一问题，我们提出利用最深层网络指导浅层网络，以增强浅层网络的性能。具体而言，我们将 FSD 网络划分为多个部分，将最深层网络用作教师模型，并通过添加分类器，使所有浅层网络成为多个学生模型。同时，利用最深层网络特征与浅层网络特征之间的蒸馏路径，缩小特征差异。在 ASVspoof 2019 LA 和 PA 数据集上的一系列实验结果表明，所提方法行之有效，相较于基线取得了显著提升。

索引词— 伪造语音检测、自蒸馏、自动说话人验证、ASVspoof

1. 引言

随着生物特征识别的发展，自动说话人验证（ASV）也开始得到广泛应用。然而，合成语音技术的发展严重威胁着自动说话人验证系统的安全。自动说话人验证系统面临的主要攻击类型包括音频重放、文本转语音（TTS）和语音转换（VC）。因此，人们开展了一系列伪造语音检测（FSD）挑战，以提高自动说话人验证系统的安全性。

FSD 的研究围绕寻找具有判别性的特征和设计稳健的网络展开。前端特征主要包括原始波形 [1, 2] , 线性频率倒谱系数（LFCC）、对数功率谱图（LPS）等。针对不同的前端特征，研究人员提出了一系列卷积神经网络建模方法。大多数网络都基于残差网络、LCNN[3], 和图网络。为了进一步提升模型性能，研究人员将注意力机制引入不同的网络中。Ling 等人 [4] 提出了一种基于注意力的卷积神经网络，用于 FSD 任务；作者设计了频率注意力模块和通道注意力模块，以关注特征中的判别性信息。在 [5], 中，他们首次将图注意力（GAT）网络应用于 FSD 任务，以进一步分析频谱特征中相邻子带之间的关系。Jung 等人 [6]提出了一种基于 GAT 的架构，采用异构注意力机制并堆叠节点，以整合时域和频域信息。Zhang 等人 [7] 将压缩与激励（SE）网络用于 FSD，其中 SE 部分用于计算全局通道注意力权重。此外，还提出了高效通道注意力（ECA） [8] ，以解决 SE 模块降维导致的信息损失问题；该方法已广泛应用于其他领域。

此外，自蒸馏作为一种流行的方法，在许多领域都取得了良好的效果。在 [9], 中，作者提出了一种基于注意力的特征自蒸馏方法，可利用更多高层信息来提升自动语音识别的性能。Liu等人[10] 为说话人验证设计了一种特征增强自蒸馏方法，并取得了良好的效果。Ge等人 [11] 提出了一种集成的知识传播形式来设计自蒸馏框架，该框架对图像分类任务非常有效。

本文提出了一种用于 FSD 的新型自蒸馏方法。许多研究表明，语音的浅层特征对 FSD 至关重要。在 [7], 中，作者发现语音中的静音片段会严重影响 FSD 的性能。此外，Deng等人 [12] 利用浅层特征信息，通过

* 通讯作者。

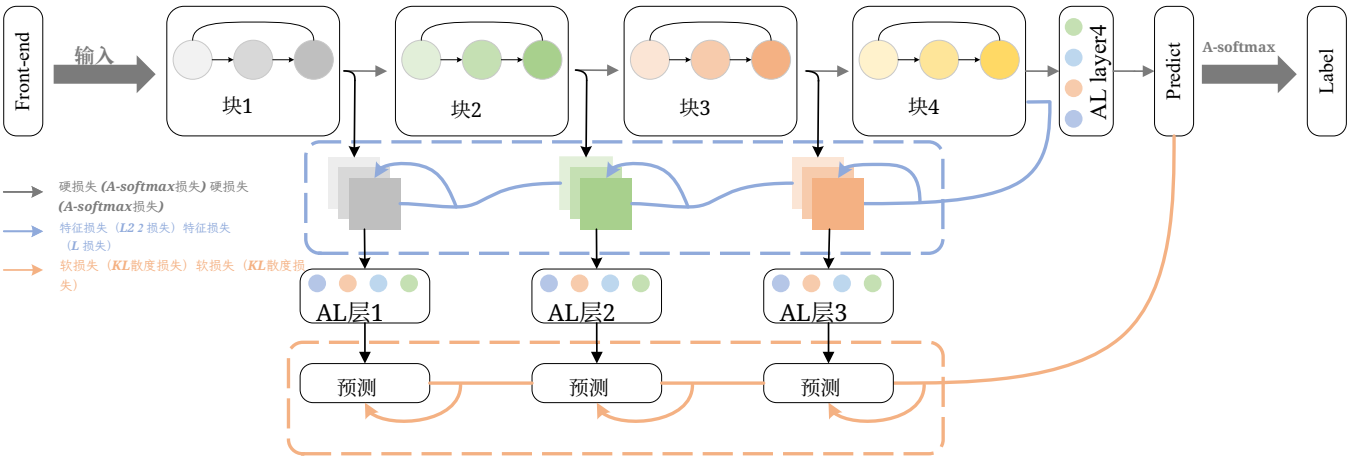


图1。所提出的自蒸馏框架基于 SENet 和 ECANet 网络。该框架分为四个模块，每个模块均设置了一个分类器（AngleLinear，AL）。注意：在评估过程中可以移除所有额外的分类器，因此不会增加模型负载。

设计频谱缺陷、静音片段等简单分类器。虽然浅层网络对这类信息较为敏感，但其捕获能力不如深层网络。为解决这一问题，我们提出了一种自蒸馏框架，由最深层网络指导浅层网络，从而进一步增强浅层网络。具体而言，在训练期间，我们为所有浅层网络添加分类器，将其作为多个学生模型，并将最深层网络设为教师模型。教师模型依据预测结果指导学生模型，将最深层网络中的知识迁移至所有浅层网络。同时，为缩小浅层网络与最深层网络之间的特征差异，我们还在特征维度上构建了自蒸馏模型。值得注意的是，推理期间可以移除所有浅层中的分类器，因此不会增加额外的计算开销或模型复杂度。为验证所提方法的有效性，我们采用 ECANet 和 SENet 网络作为基线，并使用 ASVspoof 2019 挑战赛中的 LA 和 PA 作为数据集。与基线相比，使用自蒸馏后，系统性能显著提升。

2. 所提出的自蒸馏方法

2.1. 自蒸馏

本节提出一种自蒸馏方法，如图 1 所示。具体而言，受 [13]，启发，我们根据模型架构将其划分为四个部分。本文将模型划分为四个模块，每个模块为 SE 模块或 ECA模块。随后，我们在每个模块后设置一个分类器，并使用 AngleLinear 获取每一层的预测结果。需要注意的是，这些额外的分类器仅用于训练，不会增加推理时的负担。在训练阶段，我们将网络的第四层用作教师模型，并将前三层用作学生模型。深层网络的知识会在特征和预测维度上迁移到浅层。

为充分利用网络信息，在训练过程中引入了三种损失:

- 硬损失：使用 A-Softmax函数计算标签与第四层分类器之间的损失。该损失根据训练数据集的标签和 AngleLinear 分类器的网络输出进行计算，从而充分提取训练数据集中的隐含知识。

- 特征损失：使用 L2函数计算第四层网络与各浅层网络之间的特征映射。这可以将最深层特征中的非显式知识引入浅层特征，使浅层网络在进行预测时能够更好地匹配深层网络。

- 软损失：在教师-学生模型中使用 KL 散度函数计算软损失。将最深层网络输出作为教师模型，并将若干浅层网络输出作为学生模型。计算两种输出的分布之间的差异，从而引导浅层网络学习更多知识。

2.2. 训练方法

使用自蒸馏方法进行训练时，损失由三个部分组成。首先，我们计算最深层网络与标签之间的硬损失:

$$\mathcal{L}_{hard} = A_softmax\left(p^n, L\right) \tag{1}$$

其中 p^n 是网络最深层的输出。在本文中，我们将 n 设为 4。 L 是训练数据集的标签，

表 1。基于 ASVspoof 2019 LA 和 PA 数据集的不同架构的等错误率和 t-DCF 结果。SD 表示使用了自蒸馏方法。

	模型		ECANet9	ECANet18	ECANet34	ECANet50	SENet9	SENet18	SENet34	SENet50
LA 数据集	EER	基线	1.48	1.18	1.44	1.87	1.97	1.52	1.23	1.83
		SD	1.22	**0.88**	1.14	1.09	1.23	1.37	**1.08**	**1.00**
	t-DCF	基线	0.0493	0.0378	0.0460	0.0605	0.0610	0.0497	0.0358	0.0536
		SD	0.0376	**0.0295**	0.0334	0.0318	0.0388	0.0417	**0.0347**	**0.0309**
	模型		ECANet9	ECANet18	ECANet34	ECANet50	SENet9	SENet18	SENet34	SENet50
PA 数据集	EER	基线	0.83	0.95	0.88	0.90	0.97	0.93	1.14	0.82
		SD	0.74	0.83	**0.70**	0.82	0.85	0.87	**0.65**	0.79
	t-DCF	基线	0.0221	0.0284	0.0255	0.0262	0.0269	0.0266	0.0334	0.0228
		SD	0.0199	0.0219	**0.0208**	0.0222	0.0232	0.0239	**0.0174**	0.0219

表2。ECANet 和 SENet 的模型架构与配置。维度分别指（通道、频率和时间）。批归一化（BN）和线性整流单元（ReLU）。ECA 和 SE 分别是高效通道注意力模块和压缩与激励模块。

层	输入：27000 个样本	输出形状
前端	F0 子带	(45,600)(F,T)
后处理	添加通道	(1,45,600)
	二维卷积 1 批归一化 & ReLU	(16,45,600)
模块1	$c_1 \times$ { 二维卷积 3 二维卷积 3 高效通道注意力或压缩与激励 }	(32,45,600)
模块2	$c_2 \times$ { 二维卷积 3 二维卷积 3 高效通道注意力或压缩与激励 }	(64,23,300)
模块3	$c_3 \times$ { 二维卷积 3 二维卷积 3 高效通道注意力或压缩与激励 }	(128,12,150)
模块4	$c_4 \times$ { 二维卷积 3 二维卷积 3 高效通道注意力或压缩与激励 }	(256,6,75)
输出	二维平均池化(1,1) AngleLinear	(256,1,1) 2

而 $A_softmax$ 表示 A-softmax 函数。 $\mathcal{L}_{hard}$ 是硬损失。

软损失用于浅层和深层网络的知识蒸馏，我们使用每个浅层网络和最深层网络计算 KL 散度。

$$\mathcal{L}_{soft} = \sum_{i}^{n-1} KL\left(p^i, p^n\right) \tag{2}$$

KL 表示 KL 散度函数， p^i 表示网络中每一层经过 AngleLinear 分类器后的输出。 $\mathcal{L}_{soft}$ 是最终的软损失。

特征损失用于平衡浅层网络与最深层网络之间的差异，并可反馈至浅层网络的分类输出，以促进软损失拟合。

$$\mathcal{L}_{feature} = \sum_{i}^{n-1} L2\left(\mathcal{F}^i, \mathcal{F}^n\right) \tag{3}$$

$\mathcal{F}^i$ 是每一层的输出特征， $\mathcal{F}^n$ 是最深层的输出特征。 $L2$ 是 L2 损失函数， $\mathcal{L}_{feature}$ 是最终的特征损失。

$$\mathcal{L} = \alpha * \mathcal{L}_{hard} + (1-\alpha) * \mathcal{L}_{soft} + \beta * \mathcal{L}_{feature} \tag{4}$$

训练损失由以下三个部分组成， α 和 β 是用于平衡这三个损失来源的超参数。 $\mathcal{L}$ 是最终损失。

3. 实验与结果

3.1. 数据集

我们在 ASVspoof 2019 LA 和 PA 数据集上训练并评估模型。LA 集包含三种类型的欺骗攻击（TTS、VS 和音频重放），分为 19 种攻击算法（A01–A19）。PA 集仅包含重放攻击，涵盖 27 种不同声学环境下的重放攻击。本文采用等错误率（EER）和最小归一化串联检测代价函数（min t-DCF）作为评估指标，以评估不同系统的性能。

3.2. 训练设置与基线

前端特征：受 [20], 启发，我们使用 F0 子带作为输入特征。首先，提取 LPS 的完整频带，并采用布莱克曼短时傅里叶变换（STFT）作为窗函数，将窗长和帧移分别设为 1728 和 130。我们将帧数固定为 600，并将 LPS 特征转换为 865×600。最后，取频率维度的前 45 维，最终得到大小为 45×600的前端 F0 子带特征。

后端分类器：如表2所示，我们使用SENet和ECANet作为深度神经网络分类器。其中，$c_1 - c_4$ 向量分别对应9、18、34层网络，取值为 (1，1，1，1)、(2，2，2，2) 和 (3，4，6，3)。50层网络的块中设置了三个卷积，

表 3。我们提出的自蒸馏系统与其他已知单系统的比较。

LA 数据集				PA 数据集			
系统	前端	等错误率(%)	t-DCF	系统	前端	等错误率(%)	t-DCF
AASIST [6]	原始波形	0.83	0.0275	T28 [14]	-	0.52	0.1470
ECANet18(SD) **我们的方法**	F0 子带	**0.88**	**0.0295**	SENet34(SD) **我们的方法**	F0 子带	**0.65**	**0.0174**
SENet50(SD) **我们的方法**	F0 子带	**1.00**	**0.0309**	ECANet34(SD) **我们的方法**	F0 子带	**0.70**	**0.0208**
RawGAT-ST [15]	原始波形	1.06	0.0340	SE-Res2Net50 [16]	Spec	0.74	0.0207
SENet34(SD) **我们的方法**	F0 子带	**1.08**	**0.0347**	T10 [14]	**-**	1.08	0.1598
FFT-L-SENet [7]	LPS	1.14	0.0368	T45 [14]	**-**	1.23	0.1610
MCG-Res2Net50 [17]	CQT	1.78	0.0520	T44 [14]	**-**	1.29	0.1666
Resnet18-OC-softmax [18]	LFCC	2.19	0.0590	T53 [14]	-	1.66	0.1729
ResNet18-GAT-T [5]	LFB	4.71	0.0894	胶囊 [19]	LFCC	2.76	0.0730

卷积核大小为1、3、1。其余设置与34层网络相同。训练时，我们使用Adam作为优化器，参数为 $\beta_1 = 0.9$、$\beta_2 = 0.98$、 $\epsilon = 10^{-9}$，权重衰减为 10^{-4}。训练轮次为32。两个超参数 α 和 β 分别设为0.7和0.3。

3.3. LA数据集上的实验结果

表1展示了 ASVspoof 2019 LA数据集上基线系统和自蒸馏系统的等错误率（EER）与 t-DCF。“SD”表示自蒸馏。根据表1可以看出，我们的自蒸馏方法显著优于基线系统。此外，我们还观察到了一些有趣的现象。首先，基线系统“ECANet18”的性能最佳，EER为1.18%。即便如此，自蒸馏方法仍能使其性能提升25%。其次，随着网络深度增加，不同网络架构的性能显著下降。例如，基线系统“SENet50”的EER为1.83%，而其自蒸馏系统的EER为1.00%，性能提升了45%。自蒸馏有效减轻了FSD因网络深度增加而导致的性能下降，并使其在不同网络深度下的性能更加平稳。表3展示了目前已知的最先进水平（SOTA）单系统的性能，而我们的最佳系统排名第二。总体而言，自蒸馏方法能够充分利用网络不同层级的信息。此外，该方法具有很强的通用性，能够提升不同深度下不同架构的性能。

3.4. PA 数据集上的实验结果

表 1 展示了 ASVspoof 2019 PA 数据集上基线和自蒸馏系统的等错误率和 t-DCF。根据表 1 可以看出，自蒸馏系统优于基线。基线“SENet34”的等错误率为 1.14%，这是所有基线中表现最差的。然而，“SENet34(SD)”系统的等错误率可达到 0.65%，成为表现最佳的系统。这可能是因为浅层网络包含更多尚未探索的信息，因此自蒸馏系统有教师提供指导，使浅层网络也能够挖掘更细粒度的信息。此外，表 3 对比了 PA 数据集上的顶级系统，我们的方法也取得了第二名。这表明自蒸馏方法非常有效。此外，我们的方法还可以适配不同的数据集。

4. 结论

本文提出了一种用于 FSD 任务的新型自蒸馏方法。该方法无需增加负载即可进一步提升 FSD 的性能，并且适用于不同架构的网络，具有良好的通用性。具体而言，我们在浅层网络后添加分类器，使其在特征维度和预测维度上与最深层网络进行交互，从而增强浅层网络捕获细粒度判别信息的能力。特征蒸馏旨在减小深层特征与浅层特征之间的差异，而预测维度上的蒸馏则旨在充分利用网络各层的信息，进一步优化网络。我们采用了不同架构的 ECANet 和 SENet，并在 ASVspoof 2019 LA 和 PA 数据集上的实验结果验证了所提方法的有效性和通用性，显著提升了基线的性能。未来，我们将致力于构建参数量更少且鲁棒性更高的 FSD 系统。

5. 致谢

本研究得到国家重点研发计划（编号：2020AAA0140003）、国家自然科学基金（NSFC）（编号：62201002、61972437）、安徽省杰出青年科学基金（编号：2208085J05）、安徽省科技重点项目专项资金（编号：202203a07020008）、之江实验室开放研究项目（编号：2021KH0AB06）以及国家模式识别重点实验室开放项目计划（编号：202200014）的资助。

6. 参考文献

[1]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”，*Interspeech 2019* 论文集，第1268–1272页，2019年。

[2] Hemlata Tak、Jose Patino、Massimiliano Todisco、An- dreas Nautsch、Nicholas Evans 和 Anthony Larcher，“基于 RawNet2 的端到端反欺骗”，载于*ICASSP2021*。IEEE，2021 年，第 6369–6373 页。

[3] Chenglong Wang、Jiangyan Yi、Jianhua Tao、Haiyang Sun、Xun Chen、Zhengkun Tian、Haoxin Ma、Cunhang Fan 和 Ruibo Fu，“全自动端到端伪造音频检测”，载于*DDAM 2022*，2022 年，第 27–33 页。

[4]Hefei Ling、Leichao Huang、Junrui Huang、Baiyan Zhang 和 Ping Li，“用于自动说话人验证欺骗检测的基于注意力的卷积神经网络”，载于*Interspeech* 论文集，2021 年，第 4289–4293 页。

[5] Hemlata Tak、Jee-weon Jung、Jose Patino、Madhu Kam- ble、Massimiliano Todisco 和 Nicholas Evans，“用于说话人验证反欺骗和语音深度伪造检测的端到端谱时图注意力网络”，载于 *ASVSpoof* 挑战赛论文集，2021年，第1–8页。

[6] Jee-weon Jung、Hee-Soo Heo、Hemlata Tak、Hye-jin Shim、Joon Son Chung、Bong-Jin Lee、Ha-Jin Yu 和 Nicholas Evans，“AASIST：利用集成谱时图注意力网络进行音频反欺骗”，载于*ICASSP 2022*。IEEE，2022年，第6367–6371页。

[7] Yuxiang Zhang、WenchaoWang 和 Pengyuan Zhang，“静音和双频带融合对反欺骗系统的影响”，载于*Interspeech* 论文集，2021年，第4279–4283页。

[8] Qilong Wang、Banggu Wu、Pengfei Zhu、P. Li、Wang-meng Zuo 和 Qinghua Hu，《Eca-net：面向深度卷积神经网络的高效通道注意力》，*CVPR2020*，第 11531–11539 页，2020 年。

[9] Qiang Xu、Tongtong Song、Longbiao Wang、Hao Shi、Yuqin Lin、Yongjie Lv、Meng Ge、Qiang Yu 和 Jianwu Dang，《基于高层信息监督的自蒸馏，用于压缩端到端 ASR 模型》，*2022* 年 *Interspeech* 会议论文集，第 1716–1720 页，2022 年。

[10] Bei Liu、Haoyu Wang、Zhengyang Chen、Shuai Wang 和 Yanmin Qian，《通过特征增强进行说话人验证的自知识蒸馏》，载于*ICASSP2022*。IEEE，2022 年，第 7542–7546 页。

[11] Yixiao Ge、Xiao Zhang、Ching Lam Choi、Ka Chun Cheung、Peipei Zhao、Feng Zhu、Xiaogang Wang、Rui Zhao 和 Hongsheng Li，《通过批次知识集成提升 ImageNet 分类效果的自蒸馏》，*arXiv*，2021 年。

[12] JiaCheng Deng、Terui Mao、Diqun Yan、Li Dong 和 Mingyu Dong，《基于频谱缺陷的合成语音检测》，载于*DDAM 2022*，2022 年，第 3–8 页。[13] Linfeng Zhang、Jiebo Song、Anni Gao、Jingwei Chen、Chenglong Bao 和 Kaisheng Ma，《做自己的教师：通过自蒸馏提升卷积神经网络的性能》，载于*ICCV 2019*。IEEE 计算机学会，2019 年，第 3712–3721 页。

[14] Andreas Nautsch、Xin Wang、Nicholas Evans、Tomi H Kinnunen、Ville Vestman、Massimiliano Todisco、H´ector Delgado、Md Sahidullah、Junichi Yamagishi 和 Kong Aik Lee，《ASVspoof 2019：用于检测合成、转换和重放语音的欺骗对策》，*IEEE* 生物特征识别、行为与身份科学汇刊，第 3 卷，第 2 期，第 252–265 页，2021 年。

[15] Hemlata Tak、Jee-Weon Jung、Jose Patino、Madhu Kam- ble、Massimiliano Todisco 和 Nicholas Evans，《用于说话人验证反欺骗和语音深度伪造检测的端到端谱时图注意力网络》，载于*ASVSPOOF 2021*，自动说话人验证与欺骗对策挑战赛。ISCA，2021年，第1–8页。

[16] Xu Li、Na Li、Chao Weng、Xunying Liu、Dan Su、Dong Yu 和 Helen Meng，《使用 Res2Net 架构进行重放和合成语音检测》，载于*ICASSP 2021*。IEEE，2021年，第6354–6358页。

[17] Xu Li、Xixin Wu、Hui Lu、Xunying Liu 和 Helen Meng，《通道门控 Res2Net：迈向对合成语音攻击的稳健检测》，载于*Interspeech 2021* 论文集，2021年，第4314–4318页。

[18] You Zhang、Fei Jiang 和 Zhiyao Duan，《面向合成语音欺骗检测的单类学习》，*IEEE* 信号处理快报，第 28 卷，第 937–941 页，2021 年。

[19] AnweiLuo、Enlei Li、YongliangLiu、Xiangui Kang 和 Z Jane Wang，《一种基于胶囊网络的音频欺骗攻击检测方法》，载于 *ICASSP 2021*。IEEE，2021 年，第 6359–6363 页。

[20] Jun Xue、Cunhang Fan、Zhao Lv、Jianhua Tao、Jiangyan Yi、Chengshi Zheng、Zhengqi Wen、Minmin Yuan 和 Shegang Shao，《基于基频信息与实部加虚部频谱特征组合的音频深度伪造检测》，载于 *DDAM2022*，2022 年，第 19–26 页。