arXiv:2303.03717v1 [cs.SD] 7 Mar 2023

通过特征多样性与去相关改进音频表征的自监督学习

Bac Nguyen、*Stefan Uhlich*、*Fabien Cardinaux*

Sony Europe B.V.，研发中心，斯图加特实验室1，德国

摘要

自监督学习（SSL）近期在缩小监督学习与无监督学习之间的差距方面取得了显著成果。其核心思想是学习对输入数据失真保持不变的鲁棒特征。尽管这一思想取得了成功，但仍可能出现坍塌问题，即网络生成恒定的表征。为此，我们提出SELFIE，这是一种通过特征多样性与去相关实现音频表示的新型**自**监督**学**习方法**特**征多**样**性与去**相**关方法。SELFIE 通过确保表征（i）在嵌入之间保持较高的多样性，以及（ii）消除各维度之间的相关性，来避免坍塌问题。SELFIE 在大规模 AudioSet 数据集上进行预训练，其嵌入在九项音频下游任务上得到验证，包括语音、音乐和声音事件识别。实验结果表明，SELFIE 在多项任务中优于现有的自监督学习方法。

关键词— 自监督学习，表征学习，无监督学习，音频表示。

1. 引言

深度学习已在许多应用领域取得了巨大成功，包括计算机视觉、音频处理、文本处理等 [1]。然而，其训练通常需要大量标注数据。为降低大规模数据标注的高昂成本，自监督学习（SSL）旨在利用输入数据本身的监督信号来学习强大的特征表征。学习通常通过解决一个人工设计的前置任务来完成，而无需使用任何人工标注标签。针对自监督学习，人们提出了各种前置任务，包括未来帧预测 [2], 掩码特征预测 [3, 4], 对比学习 [5–8], 以及预测编码 [9]。一旦网络被训练以解决该前置任务，便可从预训练模型中提取特征表征，以解决新的下游任务。强大且通用的表征能够惠及下游任务，尤其是标注数据有限的任务。

自监督学习的一般框架是在潜在空间中最大化同一对象不同视图之间的相关性 [10]。通过学习对输入失真具有不变性的特征，自监督学习模型旨在提取高层语义信息。根据是否使用负样本（即 来自其他对象的视图），自监督学习方法可分为两类：对比学习和预测学习。在对比学习中，正样本的嵌入会被映射到彼此接近的位置，而负样本的嵌入则会被映射到彼此远离的位置。正样本对可以是来自同一片段的音频片段，负样本对则可以是来自不同片段的音频片段 [5, 11]。负样本用于促进表征的多样性，从而防止出现平凡恒定嵌入。另一方面，预测模型只最大化正样本之间的相似度，而不考虑任何负样本。为了避免模型坍塌，人们提出了若干种技术。例如，BYOL [12] 根据一个视图预测另一个视图的输出，并采用动量编码器来维持一致的表征。SimSiam [13] 在目标分支上引入停止梯度操作。VICReg [14] 则通过在一个批次内为每个嵌入维度维持较大的方差，来确保多样性。

音频和语音表征的学习近年来备受关注，因为它在各种具有挑战性的任务中展现出出色的性能。例如，HuBERT [4] 和 wav2vec 2.0 [3]在训练数据有限的情况下，实现了自动语音识别（ASR）的最先进结果。其他自监督学习方法在非语义语音任务上也取得了显著进步，包括 TRILL [15], COLA [5],基于 Conformer 的模型 [16], 以及 BYOL-A [17]。在最成功的自监督学习方法中，对比学习已广泛应用于语音和声音事件领域 [5, 11]。例如， [18] 采用度量学习中的三元组损失来学习音频表征。三元组仅通过应用音频增强来构造。另一种方法是，CPC [9] 使用自回归模型，根据过去的全局上下文预测未来信息。尽管广受欢迎，对比学习方法往往计算开销较大，并且需要较大的小批量大小才能收敛。此外，这些方法还受到维度坍塌问题的影响：嵌入只位于一个低维子空间中，而非整个嵌入空间 [19]。也有少数方法尝试不使用对比学习。HuBERT [4] 采用类似 BERT 的预训练方式，并使用 k-均值聚类生成的伪标签。BYOL-A [17]借鉴了 BYOL [12], , 后者在不使用负样本的情况下学习表征。尽管实证研究显示这些方法取得了令人鼓舞的结果，但它们如何避免坍塌问题仍未得到充分探讨。

提取良好表征的关键要求包括三个方面。（R1）首先，学习到的表征应对数据增强造成的输入失真具有不变性。这有助于表征捕捉高层语义信息。（R2）其次，表征应通过从数据中提取有用信息，促进学习系统完成下游任务。（R3）第三，表征不应包含冗余信息（信息坍塌） [14, 20] ，即嵌入的各个维度高度相关。最后一项要求有助于形成解耦的表征。遵循这些指导原则，我们提出了一种新颖的自监督学习方法，以学习良好的特征表征。

总而言之，我们工作的贡献如下。(i) 我们提出了 SELFIE，这是一种学习通用声音表征的简单方法。我们的方法通过最大化音频梅尔频谱图两个增强视图之间的相似度来进行学习。我们采用了多种数据增强技术来生成不同的视图。为避免表征坍塌，我们通过最大化同一小批量内各嵌入之间的距离，迫使它们彼此不同。此外，我们通过对所有维度对进行去相关，防止信息坍塌。(ii) 我们展示了该方法在九项下游音频任务上的表现

，并使用线性评估进行评估。具体而言，我们在预训练嵌入之上训练一个线性分类器，以评估下游任务。结果表明，SELFIE 能够学习到有用的音频表征。

2. 提出的方法

本节将介绍我们提出的自监督学习方法及其训练目标函数，随后说明实现细节。

2.1. SELFIE

设 $\boldsymbol{v}_i = t(\boldsymbol{x}_i)$ 和 $\boldsymbol{v}'_i = t'(\boldsymbol{x}_i)$ 是对输入 $\boldsymbol{x}_i$ 进行不同随机数据增强后得到的两个视图，其中 t 和 t' 是分别从两个数据增强分布 $t \sim \mathcal{T}$ 和 $t' \sim \mathcal{T}'$ 中采样得到的变换。SELFIE 与近期提出的 BYOL 方法 [12] 使用相同的训练方案。图 1 展示了 SELFIE 的整体网络架构和损失函数。首先，两个视图 $\boldsymbol{v}_i$ 和 $\boldsymbol{v}'_i$ 分别由在线网络 f_θ 和目标网络 f_ξ 编码。随后，它们被送入投影器网络 p_θ 和 p_ξ，得到投影 $\boldsymbol{z}_i = p_\theta(f_\theta(\boldsymbol{v}))$ 和 $\boldsymbol{z}'_i = p_\xi(f_\xi(\boldsymbol{v}'))$。与目标网络不同，在线网络还额外包含一个预测器网络 q_θ，随后该网络生成预测 $q_\theta(\boldsymbol{z}_i)$。最后，我们对在线网络和目标网络的输出进行归一化，即 $\boldsymbol{z}'_i = \boldsymbol{z}'_i/\|\boldsymbol{z}'\|_2$ 和 $q_\theta(\boldsymbol{z}_i) = q_\theta(\boldsymbol{z}_i)/\|q_\theta(\boldsymbol{z}_i)\|_2$。请注意，这两个网络具有相同的架构，但使用不同的权重。具体而言，在线网络通过梯度下降进行训练，而目标网络则通过缓慢变化的指数平均进行更新，

$$\xi \leftarrow \tau\xi + (1-\tau)\theta\,,$$

其中，$\tau \in [0,1]$ 是所选的衰减率。接下来，. 我们将介绍用于训练在线网络的训练目标。

对齐损失。 我们的第一个要求（R1）是确保对数据增强的不变性。通过最小化对齐损失即可满足这一要求，该损失定义为归一化预测与目标投影之间的欧几里得距离平方，

$$\mathcal{L}_{\text{align}} = \mathbb{E}_{\boldsymbol{z}_i,\boldsymbol{z}'_i}\left[\left\|\overline{q_\theta}(\boldsymbol{z}_i) - \overline{\boldsymbol{z}'_i}\right\|_2^2\right]. \tag{1}$$

通过最小化对齐损失，学习到的表示应提取正样本之间共享的最相关信息，并对噪声保持不变。正如 Grill等人 [12], 通过实证所示，预测器 q_θ 和动量编码器有助于缓解坍塌问题并稳定训练，尤其是在小批量较小时。

多样性损失。 我们的第二个要求（R2）是确保学习到的表示能够为下游任务提供有用信息。我们引入多样性损失 $\mathcal{L}_{\text{diversity}}$，通过最大化在线网络所有归一化预测之间的距离来实现这一点，

$$\mathcal{L}_{\text{diversity}} = \mathbb{E}_{\boldsymbol{z}_i,\boldsymbol{z}_j}\left[-\left\|\overline{q_\theta}(\boldsymbol{z}_i) - \overline{q_\theta}(\boldsymbol{z}_j)\right\|_2^2\right]. \tag{2}$$

通过提高嵌入的多样性，有助于捕捉数据中不同的结构信息，这可能对不同的下游任务有所帮助。多样性损失的另一个好处是，它能防止嵌入坍缩为相同的表示。我们的多样性损失与 Wang和Isola [21]提出的均匀性概念密切相关。 [22]中给出了多样性的理论论证。令 n 表示小批量大小，$\boldsymbol{V} \in \mathbb{R}^{d\times n}$ 表示包含 n 归一化预测向量的矩阵，其中每个向量的维度为 d。式（2）的一种朴素实现的复杂度为 $\mathcal{O}(n^2d)$，因为它需要计算所有可能向量对之间的距离

由于投影已经归一化，我们可以将 $\mathcal{L}_{\text{diversity}}$ 简化如下

$$\begin{aligned}\mathcal{L}_{\text{diversity}} &= \frac{1}{n^2}\left[\mathbf{e}^{1\times n}(\mathbf{e}^{n\times n} - 2\boldsymbol{V}^\top\boldsymbol{V} + \mathbf{e}^{n\times n})\mathbf{e}^{n\times 1}\right] \\ &= 2 - \frac{2}{n^2}(\boldsymbol{V}\mathbf{e}^{n\times 1})^\top(\boldsymbol{V}\mathbf{e}^{n\times 1}) = 2 - \frac{2}{n^2}\left\|\boldsymbol{V}\mathbf{e}^{n\times 1}\right\|_2^2,\end{aligned}$$

其中，$\mathbf{e}^{n\times m} \in \mathbb{R}^{n\times m}$ 是一个所有元素均为 1 的矩阵。因此，我们可以将计算复杂度降低至 $\mathcal{O}(nd)$。

去相关损失。我们的第三项要求（R3）是确保每个特征维度编码不同的信息，以避免信息坍塌问题。受 VICReg [14], 启发，我们采用如下定义的去相关损失

$$\mathcal{L}_{\text{decorrelation}} = \sum_{i\neq j}\left[\boldsymbol{C}\big(\overline{q_\theta}(\boldsymbol{z})\big)_{ij}\right]^2, \tag{3}$$

其中，$\boldsymbol{C}(\overline{q}_\theta(\boldsymbol{z}))$ 是归一化预测的协方差矩阵

$$\boldsymbol{C}\big(\overline{q_\theta}(\boldsymbol{z})\big) = \frac{1}{n-1}\sum_{\boldsymbol{z}_i}\big(\overline{q_\theta}(\boldsymbol{z}_i) - \widehat{q_\theta}(\boldsymbol{z})\big)\big(\overline{q_\theta}(\boldsymbol{z}_i) - \widehat{q_\theta}(\boldsymbol{z})\big)^\top$$

而 $\widehat{q_\theta}(\boldsymbol{z})$ 表示均值向量

$$\widehat{q_\theta}(\boldsymbol{z}) = \frac{1}{n}\sum_{\boldsymbol{z}_i}\overline{q_\theta}(\boldsymbol{z}_i)\,.$$

去相关损失旨在最小化 $\boldsymbol{C}(q_\theta(\boldsymbol{z}))$ 的非对角系数，从而避免特征维度编码相似的信息。

利用公式 (1) 至 (3)，我们的最终损失函数定义如下 $\mathcal{L} = \mathcal{L}_{\text{align}} + \lambda_{\text{diversity}}\mathcal{L}_{\text{diversity}} + \lambda_{\text{decorrelation}}\mathcal{L}_{\text{decorrelation}}$，其中 $\lambda_{\text{diversity}} \geq 0$ 和 $\lambda_{\text{decorrelation}} \geq 0$ 是权重项。在我们的实验中，这些项设为 $\lambda_{\text{diversity}} = 1$ 和 $\lambda_{\text{decorrelation}} = 1$。

2.2. 实现细节

虽然可以将原始音频波形输入编码器网络，但由于时频特征在学习表征方面相比原始音频信号有显著优势，我们使用时频特征来学习表征 [7, 23]。给定一个音频片段，我们使用 64 ms 的窗长、10 ms 的帧移和 64 个梅尔滤波器组计算对数梅尔频谱图。这些对数梅尔频谱图作为网络的输入。下文将详细介绍我们在实现中使用的编码器网络架构和数据增强策略。

网络架构。 我们使用 BYOL-A [17] 提出的简单轻量级编码器网络来提取特征表征。该网络架构基于音频嵌入模块 [24],, 由两个卷积神经网络（CNN）模块和两个全连接层（FC）模块组成。所有 CNN 层均使用卷积核大小为 3×3的卷积核，后接批归一化层和卷积核大小为 2×2、步幅为 2 的最大池化层。我们使用 ReLU 作为激活函数。CNN 模块的输出被视为局部特征。这些输出经过重塑，将频率维度和通道维度展平，然后输入 FC 模块以学习全局特征。随后，我们拼接 CNN 模块和 FC 模块的输出，以结合局部特征和全局特征。最后，应用均值-最大池化操作，得到维度为 3072 的嵌入。投影器和预测器网络均为多层感知机，包含一个大小为 4096 的隐藏层，后接批归一化层、作为激活函数的 ReLU，以及用于输出维度为 256 的嵌入的线性层。指数移动平均的衰减率设为 $\tau = 0.995$。在我们的实现中，我们将 [17] 中的 Pre-Norm 模块替换为批归一化层，并且不使用任何 Post-Norm。

Data 数据增强。 我们使用三种数据增强策略 men-

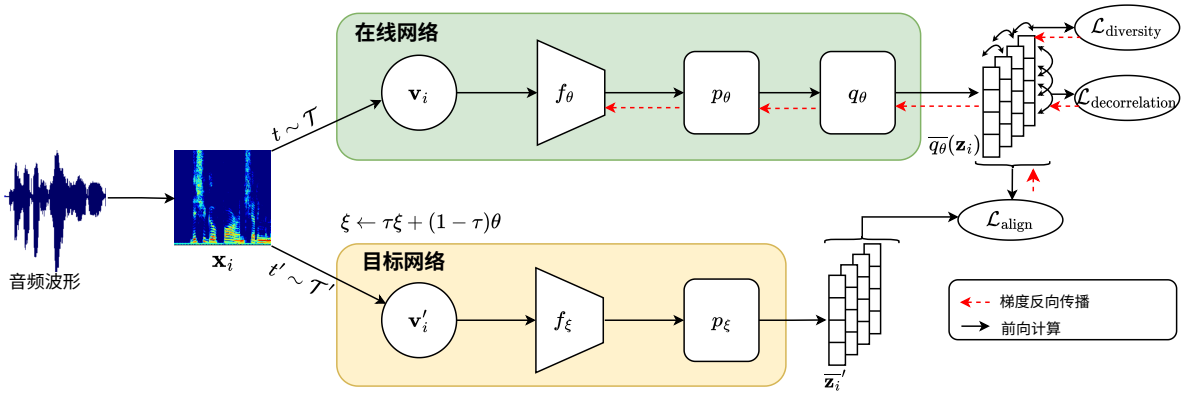


图 1。SELFIE：通过特征多样性和去相关实现的自监督学习。它由一个在线网络和一个目标网络组成。两个网络采用相同的架构，但使用不同的权重。在线网络通过随机梯度下降进行训练，而目标网络则通过缓慢移动的指数平均进行更新。

。首先，对梅尔频谱图输入应用随机缩放裁剪（RRC），随机裁剪区域的尺度范围为 0.6 到 1.5，裁剪区域的宽高比范围为 0.75 到 1.33。尽管 RRC 主要用于计算机视觉，但也可以将其视为音高/时间偏移和拉伸的近似。换言之，对 RRC 具有不变性的表征，对音高和时间扰动也具有不变性 [17]。其次，我们通过混合随机配对的输入来应用随机背景噪声（RBN） [25]。给定梅尔频谱图 $\boldsymbol{x}$，我们将其与同一小批量中的另一个输入 $\boldsymbol{x}'$ 进行线性组合，混合比例较小 $\lambda \sim \mathcal{U}$（0，0。2），即 $(1-\lambda)\boldsymbol{x} + \lambda\boldsymbol{x}'$。RBN 能使表征对噪声更加鲁棒。最后，随机线性淡变器 [7, 17]（RLF）对音频输入进行线性音量变化，以模拟淡入/淡出效果。

3. 实验

本节通过多种不同的下游任务验证 SELFIE 的有效性。首先，在大规模音频数据集上预训练 SELFIE，然后将其嵌入迁移到下游任务中。在线性分类器中使用冻结嵌入进行训练。报告各项任务的分类准确率作为性能指标。

3.1. 数据集和下游任务

我们在常用的 AudioSet [26] 数据集上训练 SELFIE。该数据集包含从 YouTube 视频中提取的两百万个音频片段，类别十分多样，涵盖范围广泛的人类、动物、音乐和环境声音。尽管 AudioSet 为每个音频片段提供多标签注释，但我们的训练并未使用这些信息。所有音频片段均以 16 kHz 的采样率采样。输入 SELFIE 的音频片段长度为 0.95 秒，从音频输入中随机裁剪得到。

表 1 总结了我们实验中使用的下游任务。这些数据集常被以往研究用于评估音频领域中的自监督学习方法 [5, 17]。该基准包含两个声音事件识别（SER）任务、四个非语义语音（NOSS）任务和三个音乐任务。下面简要介绍这些任务。ESC-50 [27] 是一个声音分类问题，包含 50 个环境声音类别。UrbanSound8K [28]（US8K）是另一个声音分类问题，包含 10 个城市声音类别。这两个任务均使用官方划分进行留一交叉验证，并报告准确率。Speech Commands

表 1。用于下游任务的数据集描述。

数据集	Task	类别数	样本数	平均长度
ESC-50 [27]	SER	50	2,000	5.0 秒
US8K [28]	SER	10	8,732	4.0 秒
SPCV2 [29]	NOSS	35	105,829	1.0 秒
VC1 [30]	NOSS	1,251	153,516	8.2 秒
VF [31]	NOSS	6	176,428	5.8 秒
CRM-D [32]	NOSS	6	7,438	2.5 秒
GTZAN [33]	音乐	10	930	30.0 秒
NSynth [34]	音乐	11	305,979	4.0 秒
Surge [35]	音乐	88	183,392	4.0 秒

V2 [29]（SPCV2）是一个单词分类任务，包含从一秒音频中识别出的 35 个口语指令（类别）。VoxCeleb1 [30]（VC1）是一个说话人识别任务，包含从名人访谈视频中提取的 1,251 位说话人。VoxForge [31]（VF）是一项根据音频输入识别所说语言的任务。口语音频涵盖六种语言（英语、法语、德语、西班牙语、俄语和意大利语）。CREMA-D [32]（CRM-D）是一个语音情感识别任务，包含 6 个类别（愤怒、厌恶、恐惧、快乐、中性和悲伤）。GTZAN [33] 是一个音乐流派识别任务，包含 10 个音乐流派类别。NSynth [34] 是一个乐器家族分类任务，包含 11 个类别。Surge [35] 是一个音高音频分类任务，包含 88 个 MIDI音符类别。训练、验证和测试均遵循 [17] 中的相同方案。在线性评估中，下游任务的音频输入会随机裁剪至平均时长。所有结果均为使用不同随机种子训练 SELFIE 的五次运行的平均值。

3.2. 下游任务的结果

我们将 SELFIE 与几种基线音频自监督学习方法进行比较，包括 TRILL [15],COLA [5], Wav2Vec2-C [3],Wav2Vec2-F [3], 和 BYOL-A [17]。此外，我们还报告了使用对数梅尔频谱特征时的性能。Wav2Vec2-C 和 Wav2Vec2-F 分别是 wav2vec 2.0 [3] 的 CNN编码器输出的 512 维嵌入，以及 Transformer网络输出的 1,024 维嵌入。我们使用 Hugging Face[1]，发布的 wav2vec 2.0 权重，这些权重是在 LibriSpeech [37] 语料库上预训练的。除 Wav2Vec2-C 和 Wav2Vec2-F 外，其他参赛方法均在 AudioSet 上预训练。SELFIE 使用 Adam优化器训练 100 个训练轮次，学习率为 0.0001，小批量大小为 512。

[1]https://huggingface.co/facebook/wav2vec2-large-960h-l

表 2。下游任务上的线性评估准确率（%）及 95% 置信区间（CI）。SELFIE 在 AudioSet [26] 上进行预训练。

方法	SER 任务		NOSS 任务				音乐任务			平均值
	ESC-50	US8K	SPCV2	VC1	VF	CRM-D	GTZAN	NSynth	Surge	
COLA [5]	N/A	N/A	62.4	29.9	71.3	N/A	N/A	63.4	N/A	N/A
梅尔频谱图	16.9 ± 1.9	43.7± 0.3	24.8 ± 0.1	9.2 ± 0.1	66.1 ± 0.1	40.4 ± 1.0	32.4 ± 4.1	33.4 ± 0.8	32.9 ± 0.1	33.2
TRILL [15]	75.4 ± 0.7	75.2 ± 1.3	78.4 ± 0.8	40.1 ± 1.1	88.8 ± 0.3	58.8 ± 2.3	64.4 ± 1.8	**74.3 ± 1.8**	28.7 ± 1.0	64.9
Wav2Vec2-F [3]	65.6 ± 1.7	67.8 ± 0.3	85.8 ± 0.2	32.0 ± 0.3	81.7 ± 0.1	56.4 ± 0.5	62。3 ± 1.0	62.2 ± 0.8	30.0 ± 0.4	60.4
Wav2Vec2-C [3]	59.3 ± 0.4	64.7 ± 0.6	**96.7 ± 0.1**	14.3 ± 0.2	**98.5 ± 0.0**	59.4 ± 1.3	54.7 ± 1.3	56.3 ± 1.2	13.4 ± 0.1	57.5
BYOL-A [17]	**83.2 ± 0.6**	**79.7 ± 0.5**	93.1 ± 0.4	**57.6 ± 0.2**	93。3 ± 0.3	63.8 ± 1.0	70.1 ± 3.6	73.1 ± 0.8	**37.6 ± 0.3**	72.4
SELFIE	82.5 ± 0.7	78.7 ± 0.2	94.0± 0.1	55.9± 0.2	93.8 ± 0.2	**65.4 ± 0.8**	**73.6 ± 1.1**	74.0 ± 0.5	36.4 ± 0.1	**72.7**

Ta表 3. 下游任务上的线性评估准确率（%）。所有模型均在 FSD50K 上进行预训练 [36] .

方法	SER 任务		NOSS 任务				音乐任务			平均值
	ESC-50	US8K	SPCV2	VC1	VF	CRM-D	GTZAN	NSynth	Surge	
BYOL-A [17]	82.5	**78.8**	91.5	51.4	91.4	58.5	65.1	**75.5**	**38.3**	70.3
SELFIE	**83.1**	78.0	**92.7**	**54.1**	**92.5**	**63.9**	**69.2**	74.4	36.6	**71.6**
$\mathcal{L}_{\text{align}}$ 和 $\mathcal{L}_{\text{diversity}}$	82.4	78.3	92.5	51.1	92.2	63.0	64.8	75.5	37.0	70.8
$\mathcal{L}_{\text{align}}$ 和 $\mathcal{L}_{\text{decorrelation}}$	82.9	78.1	92.9	53.2	92.6	62.7	66.2	75.0	36.8	71.1
$\mathcal{L}_{\text{align}}$	80.4	77.4	91.8	50.1	91.5	59.3	69.8	75.6	38.3	70.5

表 2 报告了九项下游任务的性能。总体而言，所有自监督学习方法的表现均优于使用梅尔频谱图特征的基线。SELFIE 在所有任务上的平均准确率最高，达到 72.7 %。具体而言，它在 CRM-D 和 GTZAN 数据集上取得了最佳准确率，且优势显著。在语音情感识别任务上，SELFIE 的表现与 BYOL-A 不相上下。Wav2Vec2-F 和 Wav2Vec2-C 在这些任务上的结果较差。由于它们仅在语音数据上预训练，因此语音情感识别下游任务对它们而言可能极具挑战性。另一方面，它们在 SPCV2 和 VF 等口语任务上的表现相当出色。有趣的是，与 BYOL-A 和 SELFIE 相比，COLA 和 TRILL 等对比学习方法表现不佳。

3.3. 消融研究

我们通过实证消融研究评估了所提出的方法。SELFIE 在 FSD50K [36], 声音事件数据集上进行预训练。该数据集涵盖了 200 个类别的各种声音事件。

损失的重要性。 我们展示了添加不同损失对自监督学习框架的影响。结果见表3。我们将 [17] 中报告的 BYOL-A 结果作为参考。仅使用 $\mathcal{L}_{\text{align}}$ 时，BYOL-A 与 SELFIE 等效（见表3最后一行）。我们用粗体标出了 BYOL-A 和 SELFIE 中的最佳结果。如表所示，添加 $\mathcal{L}_{\text{decorrelation}}$ 和 $\mathcal{L}_{\text{diversity}}$ 可提升 BYOL-A 的性能。我们发现，对于音乐任务，仅使用对齐损失效果更好。不过，对于其他任务，添加多样性损失或去相关损失都能显著提升性能。

在线网络与目标网络。 SELFIE 预训练完成后，可使用目标网络而非在线网络计算得到的嵌入来执行下游任务。为便于说明，图 2 展示了在线网络和目标网络在 ESC-50 数据集上的表现。二者的表现相当相似。这一结果符合预期，因为在训练过程中，在线网络的预测会逐渐收敛到目标网络的输出。

小批次大小的影响。我们在不同的小批量大小下运行 SELFIE，范围从 256 到 2048。表4报告了改变小批量大小时的性能—

表4。SELFIE 在不同小批量大小下的平均准确率（%）。

小批量大小	平均准确率
256	71.6
512	71.6
1,024	71.7
2,048	70.6

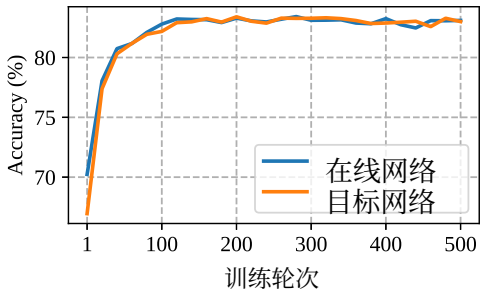


图 2。ESC-50 [27] 数据集上在线网络和目标网络在不同训练轮次下的分类准确率。

表现。我们观察到，SELFIE 的性能非常稳定（几乎保持不变）。

4. 结论

本文介绍了 SELFIE，这是一种简单的音频表征自监督学习方法，通过最大化同一对数梅尔频谱图不同视图之间的相似度来进行学习。为避免表征坍塌问题，学习到的表征被要求具有多样性且不包含冗余信息。我们在九个基准数据集上进行了广泛实验，结果表明 SELFIE 能够学习到强大的音频表征。未来的工作包括探索更多数据增强技术，以进一步改进 SELFIE。

5. 参考文献

[1] I. Goodfellow、Y. Bengio 和 A. Courville，深度学习，MIT出版社，2016年。[2] Y.-A. Chung、W.-N. Hsu、H. Tang 和 J. Glass，“一种用于语音表征学习的无监督自回归模型”，载于*INTERSPEECH*，2019年，第146–150页。[3] A. Baevski、Y. Zhou、A. Mohamed 和 M. Auli，“wav2vec 2.0：一种用于语音自监督学习的框架”，载于*NeurIPS*，2020年，第12449–12460页。[4] W.-N. Hsu、B. Bolte、Y.-H. H. Tsai、K. Lakhotia、R. Salakhutdinov 和 A. Mohamed，“HuBERT：通过隐藏单元掩码预测进行自监督语音表征学习”，*IEEE/ACM TASLP*，第3451–3460页，2021年。[5] A. Saeed、D. Grangier 和 N. Zeghidour，“通用音频表征的对比学习”，载于*ICASSP*，2021年，第3875–3879页。[6] L. Wang、P. Luc、Y. Wu、A. Recasens、L. Smaira、A. Brock、A. Jaegle、J.-B. Alayrac、S. Dieleman、J. Carreira 和 A. van den Oord，“迈向通用音频表征学习”，载于*ICASSP*，2022年，第4593–4597页。[7] H. Al-Tahan 和 Y. Mohsenzadeh，“CLAR：听觉表征的对比学习”，载于*AISTATS*，2021年，第2530–2538页。[8] X. Favory、K. Drossos、T. Virtanen 和 X. Serra，“COALA：用于学习语义丰富音频表征的协同对齐自编码器”，载于*ICML*，2020年。[9] A. v. d. Oord、Y. Li 和 O. Vinyals，“基于对比预测编码的表征学习”，*arXiv*预印本 *arXiv:1807.03748*，2018年。[10] S. Liu、A. Mallol-Ragolta、E. Parada-Cabeleiro、K. Qian、X. Jing、A. Kathan、B. Hu 和 B. W. Schuller，“音频自监督学习：综述”，*Patterns*，2022年。[11] E. Fonseca、D. Ortego、K. McGuinness、N. E. O’Connor 和 X. Serra，“声音事件表征的无监督对比学习”，载于*ICASSP*，2021年，第371–375页。[12] J.-B. Grill、F. Strub、F. Altch´e、C. Tallec、P. H. Richemond、E. Buchatskaya、C. Doersch、B. A. Pires、Z. D. Guo、M. G. Azar、B. Piot、K. Kavukcuoglu、R. Munos 和 M. Valko，“Bootstrap your own latent：一种自监督学习的新方法”，载于*NeurIPS*，2020年，第21271–21284页。[13] X. Chen 和 K. He，“探索简单孪生表征学习”，载于*CVPR*，2021年，第15750–15758页。[14] A. Bardes、J. Ponce 和 Y. LeCun，“VICReg：用于自监督学习的方差、不变性与协方差正则化”，载于*ICLR*，2021年。[15] J. Shor、A. Jansen、R. Maor、O. Lang、O. Tuval、F. d. C. Quitry、M. Tagliasacchi、I. Shavitt、D. Emanuel 和 Y. Haviv，“迈向语音的通用非语义表征学习”，载于*INTERSPEECH*，2020年，第140–144页。[16] J. Shor、A. Jansen、W. Han、D. Park 和 Y. Zhang，“ 使用自监督 Conformer 学习通用副语言语音表征 ”，载于*ICASSP*，2022年，第3169–3173页。[17] D. Niizumi、D. Takeuchi、Y. Ohishi、N. Harada 和 K. Kashino，“面向音频的 BYOL：探索预训练通用音频表征”，*IEEE/ACM T ASLP*，第137–151页，2023年。[18] A. Jansen、M. Plakal、R. Pandya、D. P. W. Ellis、S. Hershey、J. Liu、R. C. Moore 和 R. A. Saurous，“无监督学习语义音频表示”，载于*ICASSP*，2018年，页码126– 130。[19] L. Jing、P. Vincent、Y. LeCun和Y. Tian，“理解对比自监督学习中的维度坍塌”，载于*ICLR*，2022年。[20] M. Cogswell、F. Ahmed、R. B. Girshick、L. Zitnick和D. Batra，“通过去相关表示减少深度网络中的过拟合”，载于*ICLR*，2016年。[21] T. Wang和P. Isola，“通过超球面上的对齐与均匀性理解对比表示学习”，载于*ICML*，2020年，页码9929–9939。[22] F. Laakom、J. Raitoharju、A. Iosifidis和M. Gabbouj，“能量模型中的特征多样性”，载于*W-ICLR*，2021年。[23] Y. Tokozume和T. Harada，“使用端到端卷积神经网络学习环境声音”，载于*ICASSP*，2017年，页码2721–2725。[24] Y. Koizumi、D. Takeuchi、Y. Ohishi、N. Harada和K. Kashino，“NTT DCASE2020 挑战赛任务6系统：基于关键词和句长估计的自动音频描述”，技术报告，DCASE2020 挑战赛，2020年。[25] H. Zhang、M. Cisse、Y. N. Dauphin和D. Lopez-Paz，“mixup：超越经验风险最小化”，载于*ICLR*，2018年。[26] J. F. Gemmeke、D. P. W. Ellis、D. Freedman、A. Jansen、W. Lawrence、R. C. Moore、M. Plakal和M. Ritter，“Audio Set：音频事件本体与人工标注数据集”，载于*ICASSP*，2017年，页码776–780。[27] K. J. Piczak，“ESC：环境声音分类数据集”，载于*ACM-MM*，2015年，页码1015–1018。[28] J. Salamon、C. Jacoby和J. P. Bello，“城市声音研究的数据集与分类体系”，载于*ACM-MM*，2014年，页码1041–1044。[29]P. Warden，“Speech Commands：有限词汇语音识别数据集”，*arXiv*预印本 *arXiv::1804.03209*，2018年。[30] A. Nagrani、J. S. Chung和A. Zisserman，“VoxCeleb：大规模说话人识别数据集”，载于*INTERSPEECH*，2017年，页码2616–2620。[31] Voxforge.org，“Voxforge”，2018年。[32] H. Cao、D. G. Cooper、M. K. Keutmann、R. C. Gur、A. Nenkova和R. Verma，“CREMA-D：众包多模态演员数据集”，*IEEE*情感计算汇刊，第5卷，页码377–390，2014年。[33] G. Tzanetakis和P. Cook，“音频信号的音乐流派分类”，*IEEE*语音与音频处理，第10卷第5期，页码293–302，2002年。[34] J. Engel、C. Resnick、A. Roberts、S. Dieleman、M. Norouzi、D. Eck和K. Simonyan，“使用WaveNet自编码器合成音乐音符的神经音频合成”，载于*ICML*，2017年，页码1068–1077。[35] J. Turian、J. Shier、G. Tzanetakis、K. McNally和M. Henry，“来自GPU驱动模块化合成的十亿个音频声音”，载于*DAFx*，2021年，页码222–229。[36] E. Fonseca、X. Favory、J. Pons、F. Font和X. Serra，“FSD50K：人工标注声音事件的开放数据集”，*IEEE/ACM TASLP*，第30卷，页码829–852，2022年。[37] V. Panayotov、G. Chen、D. Povey和S. Khudanpur，“LibriSpeech：基于公版有声读物的自动语音识别语料库”，载于*ICASSP*，2015年，页码5206–5210。