arXiv:2304.07449v1 [cs.SD] 15 Apr 2023

用于基于音乐相似度的检索和自动标签的度量学习自监督辅助损失

Taketo Ak ama[1,*]、Hiroaki Kitano[1], Katsuhiro Takematsu[2], Yasushi Miyajima[2], 和 Natalia Polouliakh [1]

[1]Sony ComputerScienceLaboratories, Inc, 东京，日本[2]Koozyt 公司,
东京，日本[*]通讯作者：Taketo Akama，taketo.akama@sony.com

摘要

在音乐信息检索领域，基于相似度的检索和自动标签是必不可少的组成部分。鉴于人工监督信号存在局限性且难以扩展，模型必须从其他来源学习，以提升性能。自监督学习仅依赖于从音乐音频数据中提取的学习信号，在自动标签任务中已展现出有效性。本研究提出一种基于自监督学习的方法，通过引入结合自监督辅助损失的度量学习方法，解决基于相似度的检索难题。此外，与传统的自监督学习方法不同，我们发现，在不冻结预训练模型的情况下，同时使用自监督信号和监督信号训练模型具有优势。我们还发现，微调阶段不采用数据增强能够取得更好的结果。实验结果证实，所提出的方法提升了两种不同场景下的检索和标签性能指标：一种场景中，所有音乐曲目均持续提供人工标注标签；另一种场景中，此类标签仅适用于部分音轨。

1 引言

正如网络搜索引擎彻底改变了人们获取信息的方式，音乐搜索系统的进步也有望成为向听众和创作者定制化提供音乐曲目的关键力量，从而推动个性化音乐的发展。随着 YouTube 和 TikTok 等社交媒体平台上的视频内容不断增多，以及婚礼庆典等活动的兴起，背景音乐已成为至关重要的组成部分。因此，能够搜索与视频内容相配或符合特定场合的音乐，需求日益增长。Shazam（由苹果公司收购）、SoundHound 和 Sony TrackID 等商业应用的广泛使用，充分体现了音乐检索技术日益增长的普及程度。这些应用使用户能够通过设备麦克风采集的短片段识别歌曲。然而，正如网络搜索引擎的用途不仅限于识别网页，音乐搜索也不应局限于歌曲识别，而应帮助用户发现与视频内容、创意用途的广告相契合的音乐，或符合个人偏好和情绪、适合聆听体验的音乐。此类音乐搜索系统的两项基本功能包括：通过自动标签实现基于标签的搜索，以及通过基于相似度的音乐检索实现相似音乐搜索。

为了提供卓越的用户体验，搜索系统必须具备出色的准确性，确保其输出体现出对音乐以及人类音乐感知或识别的深刻理解。基本方法是利用音乐的人工标注标签。在自动标签中，依据给定音乐曲目的标签进行分类学习，并通过度量学习，使具有相同标签的音乐曲目嵌入表示彼此相似。然而，由于人工监督信号存在局限性且难以扩展，模型必须纳入其他信号以提升性能。我们通过从自监督信号中学习来应对这一挑战，这些信号源自音乐音频数据本身。

传统的基于相似度的音乐检索技术依赖监督学习，其学习信号源自人工标注标签 [8]。相比之下，自监督学习已被用于自动标签 [12]。在本研究中，我们提出了一种融合度量学习与自监督学习的模型。我们证明，自监督学习不仅有利于自动标签，也有利于基于相似度的检索任务。此外，我们还介绍了改进传统自监督学习方法的精炼技术。

如何直观地理解我们的自监督信号？音乐曲目之间的相似性通常由其全局相似度来定义，这取决于它们的全局属性有多接近[8]。自动标签的效果根据从每首音乐曲目推断全局标签的能力来评估[8, 13, 14]。我们的神经网络旨在提取这类全局属性特征，而不只依赖人工标注的标签。我们在以下假设下构建学习信号：与来自不同音轨的片段相比，来自同一音轨的片段更有可能具有相似的全局属性特征。此外，我们假设，即使经过带通滤波或移调等数据增强变换，音轨的全局属性特征也基本保持不变。由于这种学习方法所依据的监督无需人工标注，而是依赖从音频数据本身提取的标注，因此被称为自监督学习。

为了利用自监督信号，我们需要仔细设计用于自监督和度量学习的层。鉴于全局属性特征与度量学习嵌入的关联比与分类概率更直接，我们经过审慎考量，确定将度量学习嵌入放在其输出特征由自监督信号学习的那一层之后。我们还仔细考虑了归一化操作的应用位置，并将其置于分支之后、监督损失函数头之前，以免影响自监督损失函数头。

我们的自监督损失在几个方面不同于传统的自监督损失。自监督学习通常是在表征学习的背景下引入的，此时获得的表征或特征是固定的（训练好的神经网络被冻结），并在所谓的微调阶段用于其他任务 [4, 12]。在本文中，我们利用自监督学习提升任务性能，并提出了相应的学习技术。具体而言：1）在微调阶段，神经网络不会被冻结，因此可以训练整个网络，充分发挥其表达能力。2）即使在微调阶段，也会使用自监督学习信号。3）在微调阶段的自监督学习中不使用数据增强，使神经网络能够使用质量更高的数据进行训练。总体而言，我们将自监督信号视为相对于主要度量学习损失的辅助损失；与采用标准自监督方法相比，这种做法能够提升性能。

为了进一步利用自监督信号，我们通过实验证明，我们的方法在解决半监督场景问题时同样有效：在这些场景中，获取音乐曲目的人工标注标签成本高昂，而且用于训练模型的曲目也可能并非都能获得标签。

我们的主要贡献可总结如下：

- 我们提出了一种模型，通过采用自监督学习，在监督和半监督场景中提升基于音乐相似度的检索性能。
- 我们为基于音乐相似度的检索和音乐自动标注引入了一种自监督辅助损失。在监督场景中，与传统自监督方法相比，该损失有助于进一步提升效果。

2 方法

2.1 问题设定

我们来考虑一个数据集

$$\mathcal{D} = \{(\boldsymbol{x}_k, \boldsymbol{y}_k)\}_{k=1}^{N_{\text{label}}} \bigcup \{\boldsymbol{x}_k\}_{k=1}^{N_{\text{unlabel}}},$$

一组由 N_{label} 音乐曲目 $\boldsymbol{x}_k \in \mathcal{X}$ 及其多标签 $\boldsymbol{y}_k \in \mathcal{Y}$ 组成的配对，以及一组 N_{unlabel} 音乐曲目 $\boldsymbol{x}_k \in \mathcal{X}$。我们的目标是学习一个映射 F_{sim}：给定 $\mathcal{X} \to \mathcal{Z}$ $\mathcal{D}$，其中 $F_{\text{sim}}(\boldsymbol{x}_k) \in \mathcal{Z} \subset \mathbb{R}^D$ 是一个嵌入向量，而潜在空间 $\mathcal{Z}$ 中的某种距离能够反映数据点 $\boldsymbol{x}_k \in \mathcal{X}$之间的相似性。这适用于基于相似度的

检索任务。我们的目标还包括学习一个映射 F_{tag}：给定 $\mathcal{X} \to \mathcal{Y}$ $\mathcal{D}$，其中 $F_{\mathrm{tag}}(\boldsymbol{x}_k) \in \mathcal{Y} \subset [0,1]^T$是一个概率向量，其第 t个元素表示将第 t个标签分配给 $\boldsymbol{x}_k$的概率。这适用于自动打标签任务。

2.2 概述

图1展示了我们模型的概览。不同于直接学习 F_{sim} 或 F_{tag}，我们的模型学习以 $\boldsymbol{x}^{\mathrm{exc}} \in \mathcal{X}^{\mathrm{exc}}$ 为输入的映射；这些片段从音乐曲目中截取，沿用先前研究 [8]的方法。形式上，我们的模型学习映射 f_{sim}：$\mathcal{X}^{\mathrm{exc}} \to \mathcal{Z}$，并定义

$$F_{\mathrm{sim}}(\boldsymbol{x}_k) = \mathrm{Aggregate}_{\mathrm{sim}}\left(\left(f_{\mathrm{sim}}(\boldsymbol{x}_{k,e}^{\mathrm{exc}})\right)_{e=1}^{E}\right), \tag{1}$$

其中，$\left(\boldsymbol{x}_{k,e}^{\mathrm{exc}}\right)_{e=1}^{E}$ 是从音轨 $\boldsymbol{x}_k$ 中截取的一系列片段，而 $\mathrm{Aggregate}_{\mathrm{sim}}(\cdot)$ 表示先执行算术平均值操作，再除以 ℓ^2-范数。类似地，我们的模型还学习映射 f_{tag}：$\mathcal{X}^{\mathrm{exc}} \to \mathcal{Y}$，并定义

$$F_{\mathrm{tag}}(\boldsymbol{x}_k) = \mathrm{Aggregate}_{\mathrm{tag}}\left(\left(f_{\mathrm{tag}}(\boldsymbol{x}_{k,e}^{\mathrm{exc}})\right)_{e=1}^{E}\right), \tag{2}$$

其中，$\mathrm{Aggregate}^{\mathrm{tag}}(\cdot)$ 表示先执行算术平均值操作，再执行 softmax 操作。在实验中，每条音轨中的片段均为互不重叠的滑动窗口。正如先前研究 [15, 11]所揭示的，相似度学习（度量学习）通过基于标签的分类方法实现。因此，我们的模型学习f_{tag}，使得

$$f_{\mathrm{tag}}(\boldsymbol{x}_{k,e}^{\mathrm{exc}}) = \sigma\left(W f_{\mathrm{sim}}(\boldsymbol{x}_{k,e}^{\mathrm{exc}})\right), \tag{3}$$

其中，$W \in \mathbb{R}^{T \times D}$ 和 σ 表示 Sigmoid 激活。在这种设定下，用于基于相似度的检索和自动标签的模型架构大多是共享的，因此在训练和推理阶段，尤其是在同时使用基于相似度的检索和自动标签两项功能时，这种做法在时间、内存和存储方面都很有优势。在第 2.3 节和第 2.4 节中，我们将详细说明如何训练 f_{sim} 和 W（即 f_{tag}），其中 f_{sim} 定义为

$$f_{\mathrm{sim}}(\cdot) = \frac{\mathrm{LN}(f(\cdot))}{\|\mathrm{LN}(f(\cdot))\|_2}, \tag{4}$$

其中，LN 表示层归一化 [1]。因此，第 2.3 节和第 2.4 节的目标归结为学习f 和 W，我们选择使用 SampleCNN 架构来实现 f [9]。 f 使用自监督学习损失和度量学习损失进行训练，而 W 仅使用度量学习损失进行训练。由于 W 和 f_{sim}的每一行之间的距离度量是内积（$\boldsymbol{x}_{k\,e,}^{\mathrm{exc}}$），因此在进行基于相似度的检索时，我们在相似度空间中使用内积作为距离度量。

2.3 自监督学习

考虑从数据集 $\mathcal{D}$ 中取出一个小批量 $\{\boldsymbol{x}_k\}_{k=1}^{B}$，以及一组数据增强操作 $\mathcal{A}$（实验中 $\mathcal{A}$ 的选择见第 3.2 节）。我们遵循音乐表征对比学习（CLMR）[12],，该方法使用 SimCLR 框架进行自监督学习 [4]。对于每次小批量训练，我们采样两种数据增强操作 $a,\ a' \sim \mathcal{A}$ 并计算以下变换。

$$\tilde{\boldsymbol{x}}_{2k-1}^{\mathrm{exc}} = a\left(\mathrm{RandCrop}\left(\boldsymbol{x}_k\right)\right) \tag{5}$$

$$\boldsymbol{h}_{2k-1} = f\left(\tilde{\boldsymbol{x}}_{2k-1}^{\mathrm{exc}}\right) \tag{6}$$

$$\boldsymbol{o}_{2k-1} = g\left(\boldsymbol{h}_{2k-1}\right) \tag{7}$$

$$\tilde{\boldsymbol{x}}_{2k}^{\mathrm{exc}} = a'\left(\mathrm{RandCrop}\left(\boldsymbol{x}_k\right)\right) \tag{8}$$

$$\boldsymbol{h}_{2k} = f\left(\tilde{\boldsymbol{x}}_{2k}^{\mathrm{exc}}\right) \tag{9}$$

$$\boldsymbol{o}_{2k} = g\left(\boldsymbol{h}_{2k}\right) \tag{10}$$

其中，$(\tilde{\boldsymbol{x}}_{2k-1}^{\mathrm{exc}},\ \tilde{\boldsymbol{x}}_{2k}^{\mathrm{exc}})$ 这一对称为正样本对，并假设随机裁剪（记作 $\mathrm{RandCrop}(\cdot)$）和数据增强操作能够保留全局属性。随机裁剪是指裁剪

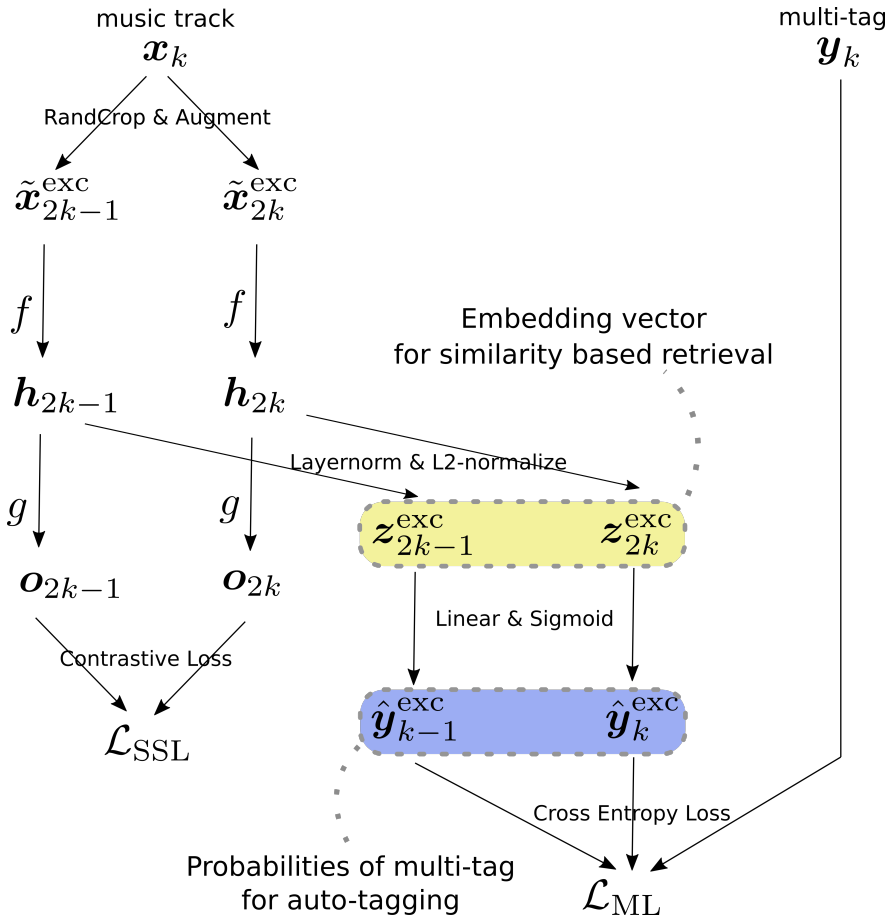


图 1：**模型概览**。对于每个由音乐曲目 $\boldsymbol{x}$ 及其对应的多标签 $\boldsymbol{y}$组成的批次，我们会对音乐曲目进行变换（以箭头表示），以计算自监督学习损失 $\mathcal{L}_{\text{SSL}}$ 和度量学习损失 $\mathcal{L}_{\text{ML}}$。这些损失用于定义总损失函数 $\mathcal{L}_{\text{SSML}} = \lambda\mathcal{L}_{\text{SSL}} + \mathcal{L}_{\text{ML}}$ （公式（17）），以训练我们提出的模型。模型训练完成后，给定一条音乐曲目$\boldsymbol{x}$, 嵌入向量 $\boldsymbol{z}^{\text{exc}}$ 和多标签的估计概率 $\hat{\boldsymbol{y}}^{\text{exc}}$ 分别用于基于相似度的检索和自动标签。

音乐曲目中的一个片段，其中片段在音乐曲目中的位置是从所有可能的位置中均匀抽取的。对于 g 的架构，我们使用一个线性层，后接一个 ReLU 层，再接一个线性层；这些线性层均不使用偏置项。

给定一个集合 $\{\tilde{\boldsymbol{x}}^{\text{exc}}_l\}_{l=1}^{2B}$，其中包含一对正样本 $\tilde{\boldsymbol{x}}^{\text{exc}}_i$和 $\tilde{\boldsymbol{x}}^{\text{exc}}_j$，对比预测任务旨在针对给定的 $\tilde{\boldsymbol{x}}^{\text{exc}}_i$，从 $\{\tilde{\boldsymbol{x}}^{\text{exc}}_l\}_{l\neq i}$ 中识别出 $\tilde{\boldsymbol{x}}^{\text{exc}}_j$。形式上，令 $\text{sim}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top\boldsymbol{v}/\|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2$，则对比预测任务的对比损失函数可定义为

$$L_{\text{SSL}}(i,j) = -\log\frac{\exp\left(\text{sim}\left(\boldsymbol{o}_i, \boldsymbol{o}_j\right)/\tau\right)}{\sum_{l=1}^{2B}\mathbb{1}_{[l\neq i]}\exp\left(\text{sim}\left(\boldsymbol{o}_i, \boldsymbol{o}_l\right)/\tau\right)}, \tag{11}$$

其中，τ 是温度参数，其设定为 SimCLR [4]提出的默认值。所有增强样本对都会计算 $L_{\text{SSL}}(i,j)$，即 $(i,j) \in \{(2k-1,2k)\}_{k=1}^{B}\bigcup\{(2k,2k-1)\}_{k=1}^{B}$，然后取平均，得到总损失函数

$$\mathcal{L}_{\text{SSL}} = \frac{1}{2B}\sum_{k=1}^{B}[L_{\text{SSL}}(2k-1,2k) + L_{\text{SSL}}(2k,2k-1)]. \tag{12}$$

2.4 基于自监督辅助损失的度量学习

我们提出将基于分类的度量学习与自监督学习相结合。对 $\boldsymbol{h}_i$应用层归一化（记作 LN(·)），然后使用 ℓ^2-范数进行归一化，得到用于基于相似度的检索的嵌入向量$\boldsymbol{z}_i^{\text{exc}} \in \mathbb{R}^D$ 。形式化地,

$$\boldsymbol{z}_i^{\text{exc}} = \frac{\text{LN}(\boldsymbol{h}_i)}{\|\text{LN}(\boldsymbol{h}_i)\|_2}. \tag{13}$$

$\boldsymbol{z}_i^{\text{exc}}$随后与 W相乘，再逐元素应用 Sigmoid 激活，得到分类输出 $\hat{\boldsymbol{y}}_i^{\text{exc}}$ ，即:

$$\hat{\boldsymbol{y}}_i^{\text{exc}} = \sigma\left(W\boldsymbol{z}_i^{\text{exc}}\right). \tag{14}$$

我们对每个标签使用二元交叉熵损失，并计算其平均值:

$$L_{\text{ML}}(i) = \frac{1}{T}\sum_{t}^{T}\left[-\boldsymbol{y}_i[t]\log\left(\hat{\boldsymbol{y}}_i^{\text{exc}}[t]\right) - \left(1-\boldsymbol{y}_i[t]\right)\log\left(1-\hat{\boldsymbol{y}}_i^{\text{exc}}[t]\right)\right]. \tag{15}$$

设 $\mathcal{K}_{\text{label}}$为一个索引集合，使得 $\{\boldsymbol{x}_k:\ k \in \mathcal{K}_{\text{label}} \subseteq \{1,2,...,B\}\}$ 是$\{\boldsymbol{x}_k\}_{k=1}^{B}$中所有已标注样本的集合。对已标注子集中的样本计算 $L_{\text{ML}}(i)$并取平均，得到损失函数

$$\mathcal{L}_{\text{ML}} = \frac{1}{|\mathcal{K}_{\text{label}}|}\sum_{k\in\mathcal{K}_{\text{label}}}[L_{\text{ML}}(2k-1) + L_{\text{ML}}(2k)]. \tag{16}$$

最后，我们提出的模型的损失函数为

$$\mathcal{L}_{\text{SSML}} = \lambda\mathcal{L}_{\text{SSL}} + \mathcal{L}_{\text{ML}}. \tag{17}$$

其中， $\lambda \in \mathbb{R}$ 是两个损失 $\mathcal{L}_{\text{SSL}}$ 和 $\mathcal{L}_{\text{ML}}$之间的平衡因子。

在实际应用中，自监督学习需要更长的训练时间，因此我们首先仅使用 $\mathcal{L}_{\text{SSL}}$训练模型，这一阶段称为预训练阶段。然后，我们使用 $\mathcal{L}_{\text{SSML}}$进行训练，这一阶段称为微调阶段。

3 实验设置

3.1 数据集

3.1.1 MagnaTagATune 数据集

MagnaTagATune 数据集包含来自 6,622 首不同歌曲的 25,000 条音乐曲目 [7]。我们使用排名前 50 的标签，并采用与以往工作相同的训练-测试划分 [12]。我们通过以下代码获取 MagnaTagATune 数据集: https://github.com/Spijkervet/CLMR/blob/master/clmr/datasets/magnatagatune.py；数据集本身则从 https://github.com/minzwon/sota-music-tagging-models/ tree/master/split/mtat 下载。

3.1.2 MTG-Jamendo 数据集

MTG-Jamendo 包含 55,000 首完整音频曲目（320kbps，MP3），涵盖流派、乐器和情绪/主题的 195 个标签 [2]。该数据集根据目标任务预先划分。我们使用预先划分的数据集和排名前 50 的标签进行训练和评估。我们从 https://github.com/MTG/mtg-jamendo-dataset 获取了 MTG-Jamendo 数据集。

3.2 模型配置

为确保公平比较，数据增强操作集合 $\mathcal{A}$ 遵循 CLMR [12] 。具体而言，以下操作以概率 p 依次应用，以生成 $\mathcal{A}$ 的一个元素。

- 极性反转（$p = 0.8$）
- 加性高斯噪声的分贝值从 $[80, 40]$ ($p = 0.01$) 中均匀采样
- 增益的分贝值从 $[-6, 0]$ ($p = 0.3$) 中均匀采样
- 以相同概率选择低通滤波或高通滤波，截止频率分别从 $[2200, 4000]$ Hz 和 $[200, 1200]$ Hz 中均匀采样（$p = 0.8$）
- 将延迟信号以 0.5 的音量因子叠加到原始信号上，延迟时间从 $\{200, 250, 300, ..., 500\}$ ms 中随机采样（$p = 0.3$）
- 移调的半音数从 $[-7, 7]$ ($p = 0.6$) 中均匀采样
- 混响脉冲响应的房间大小、混响和阻尼因子均从 $[0, 100]$ ($p = 0.6$) 中均匀采样

为确保公平比较，我们按照 CLMR [12] 将片段长度设为 59049，音频采样率设为 22.05 kHz。

为确定公式（17）中 λ 的取值，我们首先引入两个项 $\mathcal{L}_{\mathrm{ML}}$ 和 $\mathcal{L}_{\mathrm{SSL}}$ 的基础平衡因子 r。 r 定义为 $r = \mathcal{L}^{\mathrm{only\ ML}} / \mathcal{L}^{\mathrm{only\ SSL}}$，其中 $\mathcal{L}^{\mathrm{only\ ML}}$ 和 $\mathcal{L}^{\mathrm{only\ SSL}}$ 分别表示模型仅使用 $\mathcal{L}_{\mathrm{ML}}$ 或 $\mathcal{L}_{\mathrm{SSL}}$ 进行训练时的收敛损失值；使用 $\mathcal{L}_{\mathrm{ML}}$ 训练时则使用所有可用标签。MagnaTagATune 数据集和 MTG-Jamendo 数据集的 r 值分别为 22.00 和 18.95。接着，公式（17）中 λ 的候选值设为 $\{\alpha/r : \alpha \in \{0.05, 0.1, 1, 10\}\}$。为简洁起见，表 1 和表 2 分别列出了 MagnaTagATune 数据集的 $\{\alpha/r : \alpha \in \{0.1, 1, 10\}\}$ 和 MTG-Jamendo 数据集的 $\{\alpha/r : \alpha \in \{0.05, 0.1, 1\}\}$。

在我们的模型预训练阶段，仅使用 $\mathcal{L}_{\mathrm{SSL}}$，批量大小设为 48；我们采用 Adam 优化器，学习率为 0.0003，并使用 $\beta_1, \beta_2 =$ (0.9,0.999)。对于 MagnaTagATune 和 MTG-Jamendo，模型分别训练 10,000 和 1,000 轮。

在我们的模型精调阶段，使用总损失 $\mathcal{L}_{\mathrm{SSML}}$，批量大小设为 48。我们采用 Adam 优化器，学习率为 0.001，并使用 $\beta_1, \beta_2 =$ (0.9,0.999)；当验证损失连续 5 个训练轮次未改善时，学习率乘以 0.1。我们使用权重为 $1.0e-6$ 的权重衰减，模型最多训练 200 轮。当验证损失连续 10 个训练轮次未改善时，训练将停止，这称为提前停止。

3.3 评估指标

3.3.1 基于相似度的检索

为了评估基于相似度的检索，我们采用 K 值召回率（R@K）指标来衡量检索质量，遵循图像检索 [15, 11] 和基于音乐相似度的检索模型 [8]中的标准评估设置。该指标适用于评估搜索方法，因为它衡量排名靠前的 K 个检索结果的质量；与排名较低的结果相比，这些结果更为重要，也更有可能被用户看到。

3.3.2 自动标签

音乐自动标注已得到广泛研究，并且已经开发出多种模型架构 [8, 13,14]。我们遵循标准的基准测试和评估标准，并报告逐标签受试者工作特征曲线下面积（ROC 曲线下面积）和平均精度（精确率-召回率曲线下面积）分数，以衡量基于标签的检索性能。

3.4 基线方法

我们将模型与称为 Inception 模型的模型进行比较。后者是用于基于相似度的检索和自动标签的最先进模型 [8]。我们还将模型与 CLMR [12], 进行比较，后者是一种用于自动标签的模型，使用 SimCLR 作为自监督学习方法进行预训练 [4]。

3.5 学习技术的变体

本节讨论三种学习技术，它们决定了所提方法及基线方法的不同变体。

3.5.1 微调增强

微调增强是在微调阶段应用数据增强操作（详见第 3.2 节）。请注意，Inception 模型和 CLMR 不采用此技术。

3.5.2 微调对比学习

微调对比学习是在微调阶段进行对比自监督学习，其损失函数由公式（12）给出。值得注意的是，Inception 模型和 CLMR 均未采用此技术。

3.5.3 加载预训练

加载预训练是指在微调阶段加载预训练模型权重。预训练采用式 (12) 所定义的对比式自监督损失进行。需要指出的是，CLMR 使用了这一技术，而 Inception 模型则没有。此外，在我们提出的方法中，即使加载了预训练权重，我们也不会冻结模型。

4 结果

4.1 监　　督式：标签始终适用于音乐曲目的场景

表 1 展示了 MagnaTagATune 数据集在监督场景下的结果，其中“微调增强”“微调对比学习”和“加载预训练”是用于体现我们提出的方法（见第 3.5 节）不同变体的学习技术。我们的方法 G 在基于相似度的检索和自动标注任务上的表现均优于先前的方法 Inception 和 CLMR。我们的方法 A 使用与 Inception 相同的学习算法，但输入表示和网络架构除外；结果表明，这些改变并不总能带来更高的性能。我们的方法 B 是在我们的方法 A 上加入“微调增强”，这使部分指标略有提升，而另一些指标略有下降，尽管数据增强通常是一种有效策略。在我们的方法 A 上加入“加载预训练”后，得到我们的方法 C，其性能有了相当可观的提升。“加载预训练”与 CLMR 采用相同的策略，但我们的方法 C 表现更佳，可能是因为我们的方法不会冻结预训练网络，并能利用其表达能力。我们发现，在微调过程中进行自监督学习能够提升性能（我们的方法 F、G）；尤其是在微调时不进行数据增强的情况下（我们的方法 G），提升更为明显。

表2展示了 MTG-Jamendo 数据集在监督场景下的结果。我们的方法 M 在基于相似度的检索方面效果最佳，在自动标签方面的表现则与 Inception 相当。请注意，我们的方法 M 和 G 使用相同的方法（我们的方法采用“微调对比学习”和“加载预训练”），并且这种方法在不同数据集上的表现始终最佳。

4.2 半监督：音乐曲目的标签并非总是可用的场景

我们通过降低可用标签的比例来模拟半监督设置。图2展示了 MagnaTagATune 数据集在半监督场景下的结果。与 Inception 模型相比，随着标注数据量减少，我们模型的性能增益会变得更大。在基于相似度的检索（a-d）中，即使标注数据减少了 99%（也就是说，仅使用 1% 的标注数据），我们模型的性能也只略有下降。

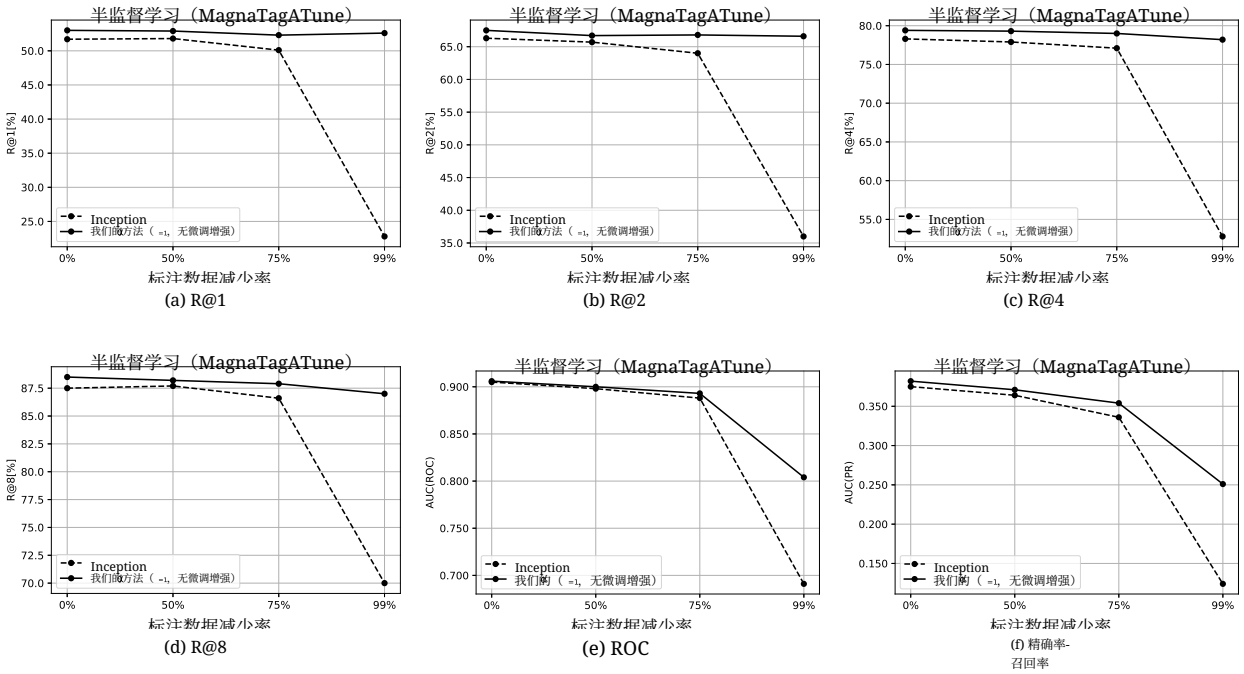
图3展示了 MTG-Jamendo 数据集在半监督场景下的结果。与 MagnaTagATune 数据集的情况类似，相比 Inception 模型，随着标注数据量减少，我们模型的性能增益往往会变得更大。

表 1：**MagnaTagATune 数据集监督场景的结果**。 我们的方法 A-I 与基线方法 Inception 和 CLMR 进行了比较。“微调增强”、“微调对比学习”和“加载预训练”是学习技术，用于描述我们所提方法的不同变体（见第 3.5 节）。我们的方法 G 通常在两项任务中都取得最高分。

模型	微调增强	微调对比式	Load 预训练	基于相似度的检索 R@1	R@2	R@4	R@8	自动标注 AUC ROC	PR
Inception				51.7	66.3	78.3	87.5	0.905	0.375
CLMR			✓					0.894	0.368
我们的方法 A				52.1	66.4	78.7	87.6	0.901	0.371
我们的方法 B	✓			51.0	66.1	78.8	87.8	0.900	0.373
我们的方法 C			✓	52.4	66.8	79.3	**88.6**	0.904	0.377
我们的方法 D（α=0.1）	✓	✓	✓	52.2	66.7	78.8	88.2	0.905	0.381
我们的方法 E（α=0.1）		✓	✓	**53.0**	67.1	78.8	88.1	**0.906**	0.381
我们的方法 F（α=1）	✓	✓	✓	**53.0**	66.7	79.2	88.3	0.905	0.381
我们的方法 G（α=1）		✓	✓	**53.0**	**67.5**	**79.4**	88.5	**0.906**	**0.382**
我们的方法 H（α=10）	✓	✓	✓	52.3	66.6	78.5	87.7	0.891	0.352
我们的方法 I (α=10)		✓	✓	52.8	66.6	78.6	87.7	0.897	0.361

表 2：**MTG-Jamendo 数据集监督场景的结果**。我们的方法 J–O 与基线方法 Inception 和 CLMR 进行了比较。“微调增强”、“微调对比学习”和“加载预训练”是用于描述所提方法（尤其是我们的方法）不同变体的学习技术（见第 3.5 节）。对于基于相似度的检索任务，我们的方法 M 通常得分最高；对于自动打标签任务，Inception 得分最高。请注意，我们的方法 M 和 G（见表 1）采用相同的方法（我们的方法结合“微调对比学习”和“加载预训练”），且在我们的方法中，这种方法在不同数据集上始终最为有效。

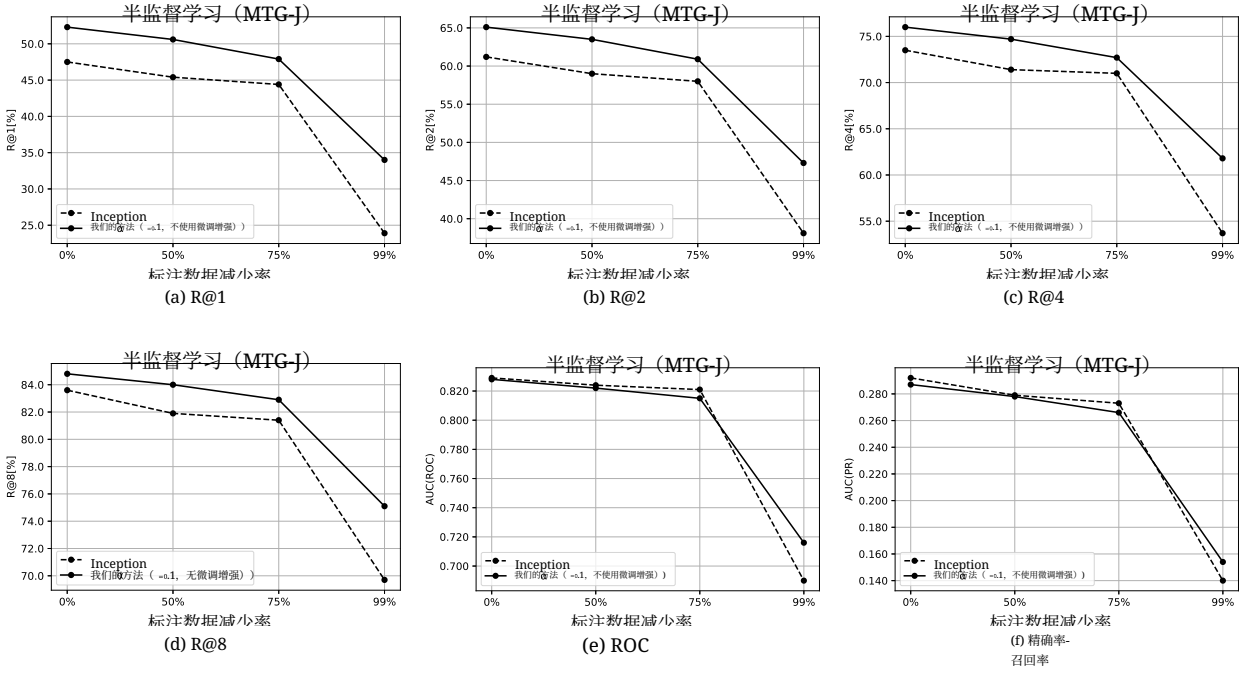
模型	微调增强	微调对比学习	Load 预训练	基于相似度的检索 R@1	R@2	R@4	R@8	自动标注 AUC ROC	PR
Inception				47.5	61.2	73.5	83.6	**0.829**	**0.292**
我们的方法 J α=0.05	✓	✓	✓	49.3	62.3	73.7	83.5	0.825	0.285
我们的方法 K α=0.05		✓	✓	52.1	64.5	75.7	84.6	0.826	0.286
我们的方法 L α=0.1	✓	✓	✓	49.7	62.5	74.2	83.8	0.826	0.288
我们的方法 M α=0.1		✓	✓	**52.3**	**65.1**	**76.0**	**84.8**	0.828	0.287
我们的方法 N α=1	✓	✓	✓	47.6	60.2	72.2	82.4	0.822	0.278
我们的方法 O α=1		✓	✓	50.0	62.2	73.5	82.8	0.825	0.285



(a) R@1 (b) R@2 (c) R@4

(d) R@8 (e) ROC (f) 精确率-召回率

图 2：**MagnaTagATune 数据集半监督场景的结果**。(a-d) 基于相似度的检索 R@K 结果。**(e-f)** 自动标注 AUC 结果。



(a) R@1 (b) R@2 (c) R@4

(d) R@8 (e) ROC (f) 精确率-召回率

图3：**MTG-Jamendo 数据集半监督场景的结果**。（a-d）基于相似度的检索 R@K 结果。（e-f）自动标注 AUC 结果。

5 相关工作

Spijkervet和Burgoyne展示了基于SimCLR的自监督学习在音乐自动标注中的有效性 [12]。我们已经证明，自监督学习不仅对自动标注有效，对基于相似度的音乐检索同样有效。此外，我们的目标是提升实际性能，而不仅仅是评估表征质量。为此，我们提出了一种自监督辅助损失，并配合一个简单的改进流程，其性能优于他们的自监督方法。

Thom´e等人提出了四种用于学习音乐相似度的三元组学习项，其中包括经过变换的片段、来自同一音轨的片段，以及流派和情绪归属 [3]。相比之下，我们的模型采用基于SimCLR的对比学习进行自监督学习，通过基于分类的度量学习处理一般的多标签设置，解决自动打标签任务，并证明了其在半监督设置中的有效性。

Manocha等人利用SimCLR进行预训练，在JND数据上训练损失网络，并采用三元组比较进行学习 [10]。他们的方法通过精心设计的语音领域数据集，专注于语音相似度；而我们的方法则利用广泛可用的标签注释，针对音乐领域中的全局音频相似度。

Duan等人采用自监督学习训练教师网络 [5]。随后，他们利用教师网络生成伪标签，再将其用于结合排序损失的度量学习。我们的方法则直接对“学生”网络应用自监督，无需教师网络。此外，他们的方法针对的是图像领域，而非音乐领域。

Fu等人以自监督方式引入了类内排序损失，并结合度量学习来处理类间方差 [6]。然而，他们的自监督采用类内排序损失，这与我们的对比式自监督损失不同；而且他们的方法针对的是图像领域，而非音乐领域。

6 结论

本文提出了一种模型，可提升基于音乐相似度的检索和音乐自动标签的质量。我们探讨了自监督在度量学习中的作用，并提出将自监督用作度量学习的辅助损失。我们的模型优于基线方法，并且在人工提供的音乐标签有限时仍能发挥有效作用。音乐产业常常需要处理规模庞大、异构且具有长尾属性的音乐数据库。人工标注标签可能无法获取、不够干净，或在不同数据库分区之间不一致。我们预计，我们的方法无需人工标注即可生成学习信号，因此在此类真实场景中能够发挥作用。

原则上，我们的方法可以扩展到其他类型的信号数据，例如生物信号（脑电图、心电图、肌电图等）和科学测量。我们计划将这一方法应用于生物信号，并研究跨模态检索，将其作为连接生物信号与音乐的途径。

7 数据可用性

当前研究中生成和/或分析的数据集可在 sota-music-tagging-models 存储库中获取：
https://github.com/minzwon/sota-music-tagging-models/tree/master/split/mtat。

当前研究中生成和/或分析的数据集可在 mtg-jamendo 数据集存储库中获取：
https://github.com/MTG/mtg-jamendo-dataset。

参考文献

[1] Lei Jimmy Ba、Jamie Ryan Kiros 和 Geoffrey E. Hinton。层归一化。*CoRR*，abs/1607.06450，2016。

[2] Dmitry Bogdanov、Minz Won、Philip Tovstogan、Alastair Porter 和 Xavier Serra。用于自动音乐标注的 MTG-Jamendo 数据集。发表于音乐发现机器学习研讨会、国际机器学习会议（*ICML 2019*），2019年。

[3] Sebastian Piwell Carl Thom´e 和 Oscar Utterb¨ack。利用自监督卷积神经网络实现音乐音频相似性。*CoRR*，abs/2202.02112，2022年。[4] Ting Chen、Simon Kornblith、Mohammad Norouzi 和 Geoffrey E. Hinton。用于视觉表征对比学习的简单框架。发表于第*37*届国际机器学习会议论文集（*ICML 2020*），*2020*年*7*月*13*日至*18*日，线上活动，第119卷机器学习研究论文集，第1597–1607页。PMLR，2020年。[5] Jiali Duan、Yen-Liang Lin、Son Dinh Tran、Larry S. Davis 和 C.-C. Jay Kuo。SLADE：用于距离度量学习的自训练框架。发表于*CVPR*，第9644–9653页。计算机视觉基金会 / IEEE，2021年。[6] Zheren Fu、Yan Li、Zhendong Mao、Quan Wang 和 Yongdong Zhang。利用自监督排序进行深度度量学习。发表于*AAAI*，第1370–1378页。AAAI出版社，2021年。[7] Edith Law、Kris West、Michael I. Mandel、Mert Bay 和 J. Stephen Downie。利用游戏评估算法：音乐标签的案例。发表于*ISMIR*，第387–392页。国际音乐信息检索学会，2009年。[8] Jongpil Lee、Nicholas J. Bryan、Justin Salamon、Zeyu Jin 和 Juhan Nam。用于解耦音乐表征学习的度量学习与分类。编者为 Julie Cumming、Jin Ha Lee、Brian McFee、Markus Schedl、Johanna Devaney、Cory McKay、Eva Zangerle 和 Timothy de Reuse，发表于第*21*届国际音乐信息检索学会会议论文集（*ISMIR 2020*），蒙特利尔（加拿大），*2020*年*10*月*11*日至*16*日，第439–445页，2020年。[9] Jongpil Lee、Jiyoung Park、Keunhyoung Luke Kim 和 Juhan Nam。SampleCNN：采用极小滤波器进行音乐分类的端到端深度卷积神经网络。《应用科学》，8(1):150，2018年。[10] Pranay Manocha、Zeyu Jin、Richard Zhang 和 Adam Finkelstein。CDPAM：用于感知音频相似性的对比学习。发表于*IEEE*国际声学、语音与信号处理会议（*ICASSP 2021*）。[11] Qi Qian、Lei Shang、Baigui Sun、Juhua Hu、Tacoma Tacoma、Hao Li 和 Rong Jin。SoftTriple损失：无需三元组采样的深度度量学习。发表于*ICCV*，第6449–6457页。IEEE，2019年。[12] Janne Spijkervet 和 John Ashley Burgoyne。音乐表征的对比学习。*CoRR*，abs/2103.09410，2021年。[13] Minz Won、Keunwoo Choi 和 Xavier Serra。半监督音乐标签Transformer。发表于*ISMIR*，第769–776页，2021年。[14] Minz Won、Andres Ferraro、Dmitry Bogdanov 和 Xavier Serra。基于CNN的自动音乐标签模型评估。*CoRR*，abs/2006.00751，2020年。[15] Andrew Zhai 和 Hao-Yu Wu。分类是深度度量学习的强基线。发表于*BMVC*，第91页。BMVA出版社，2019年。