arXiv:2303.10539v1 [cs.SD] 19 Mar 2023

无文本 基于情绪相似度的语音到音乐检索 Y

♭SeungHeon Doh, ♮Minz Won, ♯Keunwoo Choi, ♭Juhan Nam

♭韩国科学技术院文化技术研究生院，韩国♮字节跳动，美国♯Gaudio Lab，韩国

摘要

我们介绍了一种根据语音情绪推荐音乐的框架。在内容创作和日常生活中，语音包含有关人类情绪的信息，而音乐能够强化这些情绪。我们的框架聚焦于跨域检索系统，通过情绪标签弥合语音与音乐之间的差距。我们探讨了不同的语音表征，并报告了它们在不同语音类型（包括表演语音和唤醒词）上的影响。我们还提出了一种用于跨域检索任务的情绪相似度正则化项。将该正则化项纳入训练后，情绪空间中相似的语音与音乐配对在联合嵌入空间中的距离会更近。全面的实验结果表明，所提出的模型在无文本语音到音乐检索中效果显著。

索引词— 跨域检索，多模态检索

1. 引言

在多媒体内容中，情绪首先通过语音、面部表情和手势传达。若搭配恰当的音乐，所表达的情绪往往会得到强化。例如，如果电影场景描绘的是一段浪漫对话，浪漫的配乐就会增强其氛围（如图 1 所示）。然而，这种跨域匹配过程，即语音情感识别和音乐搜索，既耗时又需要领域专业知识。

为了解决这类检索问题，以往研究将任务形式化为跨模态检索 [1] ——例如音乐视频到音乐 [2], 、用户视频到音乐 [3, 4], ，以及故事讲述文本到音乐 [5]。跨域检索的一种常见方法是多模态表征学习，它将来自不同模态的数据样本映射到联合嵌入空间中。通过优化联合嵌入空间，使异构模态的数据分布对齐。文献 [6], 中，研究者采用无监督、基于成对数据、基于排序以及监督方法来解决对齐问题，这些方法分别利用共现数据、成对数据、排序列表和标注标签。然而，语音和音乐无法自然共现——它们只有经过刻意安排才会共同出现——因此，我们没有自然配对的语音和音乐数据集。一种可行的解决方案是利用彼此独立的语音数据集和音乐数据集，采用监督方法来利用两个领域标签之间的关系。

然而，不同领域之间不匹配的数据集分类体系（情绪词汇）会立即造成困难。基于情绪的映射可以解决这一问题：将每个标签转换为最相似的情感标签，例如，喜悦（来自语音数据集）→快乐（来自音乐数据）。Won 等人 [5]将

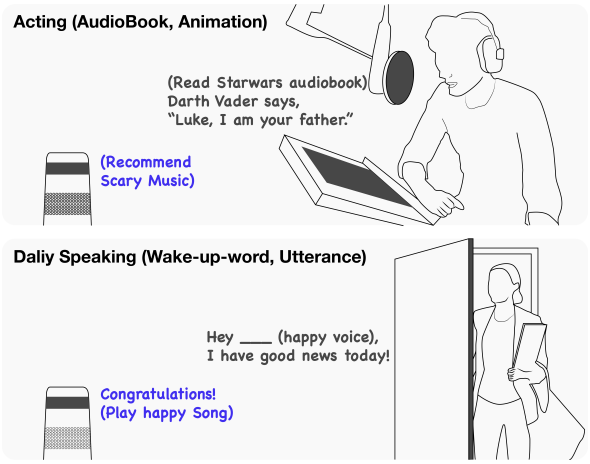


图 1。用户通过智能音箱进行语音到音乐检索的使用场景。

将故事讲述与音乐匹配这一任务形式化为监督式跨模态检索。他们提出了三种可能的基于情绪的映射方法：（1）根据预定义效价-唤醒度空间中情绪标签之间的欧几里得距离进行映射；（2）根据预定义词嵌入空间中情绪标签之间的余弦距离进行映射；或（3）直接手动映射情绪标签。这种情绪映射方法使模型能够弥合模态差距。虽然此类映射方法能够实现数据驱动的对齐，但由于最相似映射具有离散性，因此可能缺少有关连续情绪分布的信息。直接手动映射可通过一对多映射（例如，喜悦→{快乐、兴奋}）解决最相似映射的问题。标注成本会随着标签数量的增加而相应提高。

在本研究中，我们将语音到音乐任务中的匹配形式化为一个有监督的跨域检索问题。[1] 受 [7],启发，我们提出了一种情绪相似度正则化项，利用连续情绪相似度信息。例如，我们引导网络，使快乐语音的特征表示更接近按以下顺序排列的音乐表示：兴奋、有趣、紧张、悲伤、愤怒和恐惧。我们探索了不同的语音表征，并报告它们对不同语音类型的影响，包括表演语音和唤醒词。实验表明，我们能够成功弥合语音与音乐之间的差距，从而促进跨域检索。我们在线发布了代码和演示。[2]

[1]我们使用**“跨域”**而不是**“跨模态”**这一术语，因为语音音频和音乐音频属于相同的模态。[2]https://seungheondoh.github.io/speech-to-music-demo

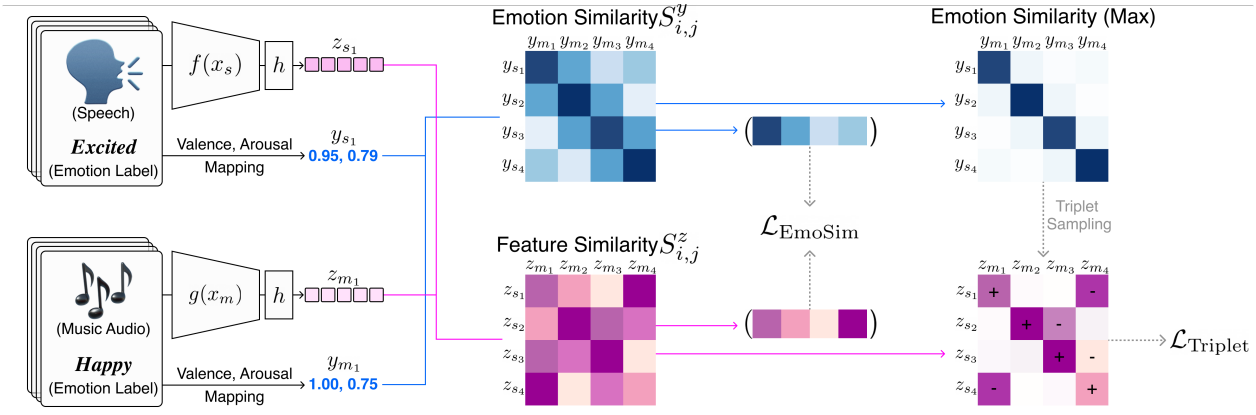


图2。所提出的语音到音乐检索框架。 S_y 和 S_z 分别表示情绪标签和特征的成对相似度编码。损失函数的目标是促使情绪标签相似的项目（语音、音乐）在特征空间中也彼此相似。

2. 语音到音乐检索

与先前的工作 [5], 类似，我们的跨域检索模型由两部分组成：用于特征提取的预训练编码器网络，以及用于弥合领域差距的投影层。在所提出的系统中，语音由音频、文本和融合模态表示。如图 2 所示，该架构接收来自语音和音乐两个不同领域的数据。语音数据包括音频、转录文本和情绪标签。音乐数据包括音频和情绪标签。领域数据由预训练编码器网络和投影多层感知机（MLP）处理。预训练编码器网络提取领域特征，投影 MLP 则将各领域特征的输出映射到一个联合嵌入空间。

2.1. 用于特征提取的预训练模型

大规模预训练模型在情绪识别这一类下游任务中展现出良好的泛化性能，例如基于音乐 [8, 9], 语音 [10, 11, 12], 文本的 [13]。如图 2所示，所提出的架构通过预训练编码器模型，将音乐和语音映射到一个联合空间。对于音乐编码器，我们使用音乐标签转换器 [14], 这是当前最先进的音乐标注模型。对于语音音频编码器，我们使用 Wav2vec 2.0 [15] ，并对其音素级特征进行平均池化，将其汇总为话语级 [11]。对于语音文本编码器，我们使用 DistilBERT [16], ，它是广受欢迎的 BERT Transformer 模型的紧凑变体。对于音乐音频、语音音频和语音文本编码器，我们使用经分类任务微调的预训练网络（见表1）。对于情绪标签，我们使用 VAD 词典 [17]中的效价-唤醒度表示。

2.2. 跨域检索

本节介绍语音到音乐检索框架（如图 2 所示）。在以下描述中，符号x_s， y_s和 x_m， y_m分别表示语音和音乐样例及其对应的情绪标签。符号 f和 g表示语音和音乐的预训练编码器， h_s 和 h_m分别表示语音和音乐的投影多层感知机。每个领域的输入都由相应的预训练编码器 f或 g 以及投影多层感知机 h_s或 h_m处理。我们将两个输出嵌入分别记为 $z_s = h_s(f(x_s))$ 和 $z_m = h_m(g(x_m))$。在跨域检索模型中，我们冻结预训练编码器 f， g，并训练投影多层感知机网络 h_s， h_m。

我们使用带有三元组损失的深度度量学习来训练检索模型。度量学习的目标是学习一个嵌入空间，使来自相同语义的输入彼此之间比来自不同语义的输入更接近。给定一个三元组嵌入 (z,z^+,z^-)，度量学习模型经过优化，以最小化三元组损失。

$$\mathcal{L}_{\text{Triplet}}(z,z^+,z^-) = \max\{0, D(z,z^+) - D(z,z^-) + \delta\} \quad (1)$$

其中， z,z^+,z^- 分别是锚点、正样本和负样本的嵌入。 D 是余弦距离函数， δ是预定义的边距。在我们的跨域检索场景中，锚点是查询语音嵌入 (z_s)，而正样本和负样本是音乐嵌入 (z_m^+, z_m^-)。我们的基线模型是传统的三元组网络，它优化损失函数 $\mathcal{L}$:

$$\mathcal{L}_{\text{Cross}} = \mathcal{L}_{\text{Triplet}}(z_s, z_m^+, z_m^-) \quad (2)$$

与此同时， $\mathcal{L}_{\text{Cross}}$的三元组采样中还存在另一种分类体系问题。为了解决这个问题，我们遵循先前的研究 [5]，使用效价-唤醒度映射连接两个领域的情绪标签 (y_s, y_m)。计算语音和音乐分类体系之间的欧几里得距离后（见图3），我们在情感空间中采样与语音最相似的音乐项。

然而，三元组损失并未考虑模态内部的邻域结构或数据分布。虽然考虑了语音与音乐之间的关系，却没有考虑各项与情感标签之间的关系。先前的工作 [5] 通过结构保持约束 [18, 19] 解决了这一问题。给定真实情感标签，各个损失函数的设计旨在优化情感到语音和情感到音乐的三元组损失，如下所示：

$$\begin{aligned} \mathcal{L}_{\text{SP-Speech}} &= \mathcal{L}_{\text{Triplet}}(z_e, z_s^+, z_s^-) \\ \mathcal{L}_{\text{SP-Music}} &= \mathcal{L}_{\text{Triplet}}(z_e, z_m^+, z_m^-) \end{aligned} \quad (3)$$

其中 z_e 表示情感标签嵌入。我们使用预训练的 GloVe 词嵌入 [20] ，并通过投影多层感知机层获取情感标签嵌入 z_e。该模型学习情感、语音和音乐之间的共享嵌入空间。在这些约束下，我们的最终损失函数为 $\mathcal{L}_{\text{Triplet+SP}} = w_1\mathcal{L}_{\text{Cross}} + w_2\mathcal{L}_{\text{SP-Speech}} + w_3\mathcal{L}_{\text{SP-Music}}$

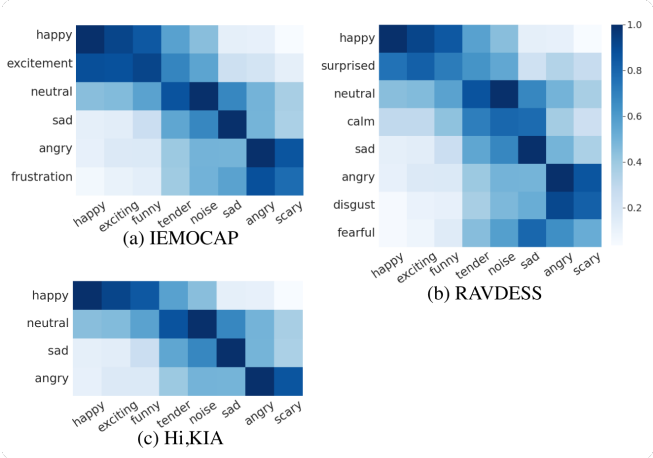


图 3。使用效价-唤醒度映射进行语音到音乐情绪相似度计算。横轴表示 AudioSet 音乐情绪标签，纵轴表示每个语音数据集的情绪标签。颜色表示两个标签之间的效价-唤醒度相似度值。

其中 w 使我们能够控制跨域映射与域内情绪映射之间的权衡。遵循先前的工作 [5], , 我们分别将所有 w_1， w_2， w_3设为 0.4、0.3 和 0.3。

2.3. 情绪相似度正则化

为保留联合嵌入空间中的连续情绪分布，我们采用了 Gong 等人提出的方法，该方法将秩相似度正则化（RankSim） [7] 用于深度不平衡回归任务。针对跨域检索任务，我们提出了情绪相似度正则化（EmoSim），这是 RankSim 的一个改进版本。我们的方法概述如图 2 所示。具体而言，我们的目标是促进情绪空间中邻居的相似度（S^y）与特征空间中邻居的相似度（S^z）之间的一致性。EmoSim 正则化项的公式如下：

$$\mathcal{L}_{\text{EmoSim}} = \frac{1}{N}\sum_{i=1}^{N} \ell(S_i^y, S_i^z) \tag{4}$$

其中， S_i^y 和 S_i^z 分别表示第 i 个语音项和音乐项的相似度分数， ℓ 表示均方误差， N 表示小批量大小。在训练过程中，我们只使用唯一的相似度分数，并增强低频标签的相对表示。我们不计算排名，而是直接计算两个相似度分数之间的差值。通过这种正则化，我们的最终损失为 $\mathcal{L}_{\text{Triplet+Emosim}} = \mathcal{L}_{\text{Cross}} + \lambda\mathcal{L}_{\text{EmoSim}}$

其中， λ 使我们能够控制最相似情绪映射（离散）与整体情绪分布映射（连续）之间的权衡。应用情绪正则化时， λ 的值设为0.5。

3. 实验设计

3.1. 语音数据集

我们在不同类型的语音数据集上验证所提方法的有效性，包括表演语音（IEMOCAP）、

模态	AS 音乐情绪	IEMOCAP	RAVDESS	HI,KIA
Text	-	0.709	0.133	0.250
音频	0.605	0.718	0.733	0.935

表1。音乐与语音情绪识别模型的分类准确率

词汇匹配语音（RAVDESS）和唤醒词语音（HI,KIA）。IEMOCAP [21] 包含约 12 小时、来自 10名说话者的语音，并分为五个场次。我们的实验基于 6 种情绪类别。我们使用前四个场次进行训练，第五个场次用于测试。我们将所提供训练数据的 10% 用作验证集。RAVDESS [22] 是一个包含情绪语音和歌曲的多模态数据库。该数据库收录了 24 位不同的演员（12 位男性和 12 位女性），他们演绎了 2 段陈述语句，并表现出 8 种不同的情绪。我们使用前 20 位演员进行训练，第 20 至 22 位演员用于验证，第 22 至 24 位演员用于测试。Hi,KIA [23]是一个短时唤醒词数据集，标注了四种情绪状态：愤怒、快乐、悲伤和中性。它包含来自 8 名说话者的 488 条带有韩国口音的话语。我们使用官方八折划分之一，即人工验证表现最高的划分（F7）。

3.2. 音乐数据集

我们使用 AudioSet [24]，其中包含从 YouTube 视频中截取、由人工标注的 10 秒音频片段。我们使用其“音乐情绪”本体下的一个子集，该子集包含标注为 7 种音乐情绪类别之一的音乐片段：快乐、有趣、悲伤、温柔、兴奋、愤怒，以及恐惧。由于部分条目已无法获取，实际使用的条目数量少于官方数据集所包含的数量。由于 AudioSet 音乐情绪数据缺少中性标签，我们将随机生成的彩色噪声视为中性语音的正样本对。这个较小的数据集包含一个训练集，其中有 18,643 个片段（13,295 个音乐片段、5,348 个噪声片段），以及一个评估集，其中有 405 个片段（345 个音乐片段、60 个噪声片段）。

3.3. 评估

我们使用准确率评估分类性能。对于检索模型的评估，我们沿用先前的工作 [5],，使用两项评估指标：前5位精确率 (P@5) 和平均倒数排名 (MRR)。相关性分数是一个二元值，根据查询语音和目标音乐 是否具有相同的情绪语义来衡量（如图3所示）。然而，这项指标有两个问题：它不考虑排名顺序，也无法反映连续情绪相似度。例如，温柔 比悲伤 更接近愤怒。为了解决上述问题，我们使用归一化折损累计增益@5 (nDCG@5)，该指标专为具有多个相关性等级的排序任务设计 [25]。我们将预测分数定义为语音与音乐项之间的相似度；将真实值定义为该音乐项情绪标签的效价—唤醒度 (VA) 相似度。

3.4. 训练详情

音乐编码器的输入是采样率为 22,050 Hz、时长为 10 秒的音频片段。该片段会转换为对数刻度的梅尔频谱图，使用 128 个梅尔频带、1024 点快速傅里叶变换、汉宁窗和 512 的帧移。语音音频编码器和文本编码器的输入分别为采样率为 16,000 Hz、最长 16 秒的波形，以及最大长度为 64 的分词文本。对于跨域检索模型，所有模型均使用 AdamW 优化，并

方法	模态	应用于	IEMOCAP [21]			RAVDESS [22]			HIKIA [23]		
			MRR	P@5	NDCG@5	MRR	P@5	NDCG@5	MRR	P@5	NDCG@5
三元组	Text	[5]	0.75±0.03	0.68±0.03	0.87±0.02	0.22±0.07	0.11±0.05	0.48±0.03	0.22±0.11	0.17±0.08	0.51±0.05
三元组 +SP	Text	[5]	0.76±0.01	0.68±0.02	0.88±0.01	0.29±0.03	0.16±0.06	0.54±0.06	0.23±0.14	0.16±0.1	0.52±0.09
三元组 +EmoSim	Text	Ours	0.76±0.01	0.69±0.02	0.88±0.01	0.28±0.1	0.18±0.05	0.62±0.06	0.22±0.11	0.1±0.03	0.46±0.02
三元组	音频	Ours	0.74±0.02	0.67±0.04	0.86±0.03	0.72±0.03	0.65±0.03	0.85±0.03	0.84±0.09	0.73±0.06	0.91±0.03
三元组 +SP	音频	Ours	0.73±0.03	0.65±0.02	0.86±0.02	0.75±0.04	0.65±0.05	0.87±0.02	0.83±0.10	0.73±0.10	0.88±0.06
三元组 +EmoSim	音频	Ours	0.76±0.03	0.68±0.03	0.88±0.02	0.75±0.03	**0.67±0.02**	**0.88±0.01**	**0.9±0.04**	**0.79±0.04**	**0.93±0.02**
三元组	融合	Ours	0.80±0.03	0.73±0.05	0.89±0.02	**0.76±0.02**	0.65±0.02	0.86±0.01	0.87±0.06	0.74±0.10	0.92±0.04
三元组 + SP	融合	Ours	0.82±0.01	0.74±0.02	0.90±0.02	0.73±0.07	0.66±0.02	0.87±0.01	0.82±0.16	0.75±0.07	0.91±0.08
三元组 + EmoSim	融合	Ours	**0.83±0.01**	**0.75±0.02**	**0.90±0.01**	0.74±0.04	0.63±0.05	0.86±0.02	0.82±0.07	0.73±0.04	0.92±0.02

表2。在三个不同语音数据集上的语音到音乐检索结果。**SP** 代表结构保持，**EmoSim** 代表情绪相似度正则化。

采用 64 的批大小。模型以 1e-4 的学习率和 0.4 的间隔 δ 进行训练。

4. 结果

4.1. 定量结果

表1比较了分类任务上的基线性能，以评估特征提取性能。显而易见，音频模态在词汇匹配语音（“RAVDESS”、“Hi,KIA”）上的表现更好。值得注意的是，即使在包含语言信息的 IEMOCAP 数据集中，音频模态也表现出色。这表明，无需语音转文本模型，仅凭音频模态就足以进行情绪识别。

表2比较了采用文本、音频和融合语音模态的不同检索模型。每个跨域检索模型都使用不同的随机种子训练了5次，并报告其平均分数和标准差。对于融合方法，我们通过后期融合利用模态间动态。 [3] 比较单一模态时，音频模态在词汇匹配语音数据集（“RAVDESS”、“Hi,KIA”）上表现出色。在包含语言信息的 IEMOCAP 数据集中，文本模态的结果与音频模态相当。一个可能的原因是语言结构中包含丰富的情绪线索。与其他单一模态相比，融合模型的性能有明显提升。

比较各方法时，我们提出的情绪相似度（EmoSim）正则化方法在三个数据集和多种模态上均表现出较高的性能和较低的标准差。特别是在三个数据集的音频模态中，三元组 + *EmoSim* 方法的 NDCG@5 表现出色。这是因为在失败案例中，模型会检索与正确答案尽可能接近的音乐项目。

4.2. 定性结果

度量学习嵌入空间使用统一流形近似与投影（UMAP） [26]投影到二维空间。我们使用音乐嵌入和语音嵌入拟合 UMAP，然后将它们投影到二维空间。图4展示了多模态嵌入空间的可视化结果。受篇幅限制，我们仅展示四个模型。所有嵌入向量均由测试集计算得出。

首先，我们观察到所有模型都能成功区分情绪语义。此外，我们还可以看到情绪呈连续分布

[3]在初步实验中，我们发现后期融合的表现优于早期融合

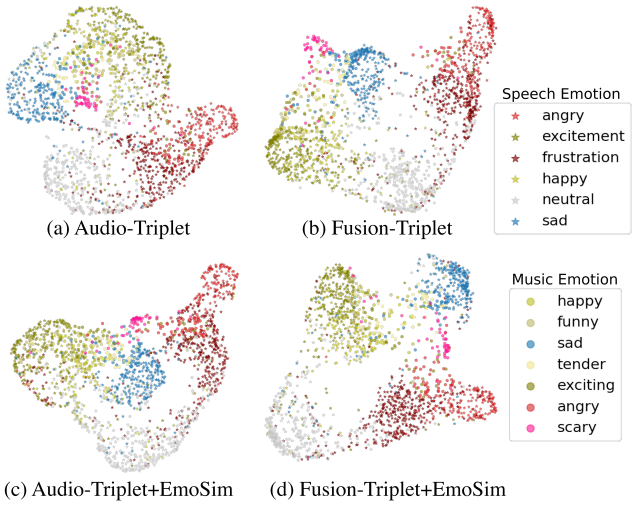


图4。语音—音乐联合嵌入空间的 UMAP 可视化。星形标记表示语音嵌入，圆形标记表示音乐嵌入。

，从快乐→兴奋→温柔→悲伤 延伸至所有模型中。然而，在三元组模型中，恐惧 音乐嵌入位于悲伤和快乐语音嵌入簇之间（图4 (a)、(b)）。根据图3，在效价-唤醒度空间中，恐惧情绪与沮丧和愤怒情绪相似。三元组 +EmoSim 模型缓解了这一问题。靠近愤怒和沮丧簇的恐惧音乐样本相对较少（图4 (c)、(d)）。我们认为，联合嵌入空间从连续情绪相似度中学习到了跨模态邻域结构。

5. 结论

我们提出了一种语音到音乐跨域检索框架，用于找出与语音情绪相匹配的音乐。我们探索了多种语音模态表示，并在跨域检索任务中提出了情绪相似度正则化项。实验结果表明，在没有语言信息（词汇匹配语音）的情况下，音频模态能够检索出与说话者情绪相匹配的音乐。特别是，我们提出的正则化项有助于联合嵌入空间理解情绪的连续性。未来，我们的方法可以扩展到纳入其他模态（例如视觉），并通过更好的层次化融合方法加以改进。

6. 参考文献

[1]Bokun Wang、Yang Yang、Xing Xu、Alan Hanjalic 和 Heng Tao Shen，“对抗性跨模态检索”，载于美国计算机协会多媒体国际会议论文集，2017年。[2] Sungeun Hong、Woobin Im 和 Hyun Seung Yang，“Cbvmr：基于内容的视频音乐检索及软性模态内结构约束”，载于美国计算机协会国际多媒体检索会议论文集，2018年。[3] Bochen Li 和 Aparna Kumar，“以视频查询：跨模态音乐检索”，载于国际音乐信息检索学会会议论文集，*ISMIR*，2019年。[4] D´ıdac Sur´ıs、Carl Vondrick、Bryan Russell 和 Justin Salamon，“音乐与视频中的艺术对应关系正当其时”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2022年，第10564–10574页。[5] Minz. Won、Justin. Salamon、Nicholas J. Bryan、Gautham J. Mysore 和 Xavier. Serra，“用于音乐与故事匹配的情感嵌入空间”，载于国际音乐信息检索学会会议论文集，*ISMIR*，2021年。[6] Kaiye Wang、Qiyue Yin、Wei Wang、Shu Wu 和 Liang Wang，“跨模态检索综合调查”，*arXiv*预印本 *arXiv:1607.06215*，2016年。[7] Yu Gong、Greg Mori 和 Frederick Tung，“RankSim：深度不平衡回归的排序相似度正则化”，载于国际机器学习会议（*ICML*），2022年。[8]Keunwoo Choi、Gy¨orgy Fazekas、Mark Sandler 和 Kyunghyun Cho，“音乐分类与回归任务的迁移学习”，载于国际音乐信息检索学会会议论文集，*ISMIR*，2017年。[9] Rodrigo Castellon、Chris Donahue 和 Percy Liang，“编码式音频语言建模学习对音乐信息检索有用的表征”，载于国际音乐信息检索学会会议论文集，*ISMIR*，2021年。[10] Li-Wei Chen 和 Alexander Rudnicky，“探索 Wav2vec 2.0 微调以改进语音情感识别”，*arXiv*预印本 *arXiv:2110.06309*，2021年。[11] Xingyu Cai、Jiahong Yuan、Renjie Zheng、Liang Huang 和 Kenneth Church，“采用多任务学习进行语音情感识别”，载于*Interspeech*论文集，2021年。[12] Leonardo Pepino、Pablo Riera 和 Luciana Ferrer，“使用 Wav2vec 2.0 嵌入进行语音情感识别”，载于*Interspeech*论文集，2021年。[13] Acheampong Francisca Adoma、Nunoo-Mensah Henry 和 Wenyu Chen，“BERT、RoBERTa、DistilBERT 和 XLNet 在基于文本的情感识别中的比较分析”，载于*2020*年第*17*届国际小波、主动媒体技术与信息处理计算机会议，*ICCWAMTIP*，2020年。[14]Minz Won、Keunwoo Choi 和 Xavier Serra，“半监督音乐标签转换器”，载于国际音乐信息检索学会会议论文集，*ISMIR*，2021年。[15] Alexei Baevski、Yuhao Zhou、Abdelrahman Mohamed 和 Michael Auli，“Wav2vec 2.0：用于语音表征自监督学习的框架”，*NeurIPS*神经信息处理系统进展，2020年。[16] Victor Sanh、Lysandre Debut、Julien Chaumond 和 Thomas Wolf，“DistilBERT：BERT 的精简版——更小、更快、更便宜、更轻量”，神经信息处理系统能源高效机器学习与认知计算研讨会，2019年。[17] Saif Mohammad，“为 20,000 个英语单词获取可靠的效价、唤醒度和支配度人工评分”，载于计算语言学协会年会论文集，*ACL*，2018年。[18] Li wei Wang、Yin Li 和 Svetlana Lazebnik，“学习保持深层结构的图文嵌入”，载于*IEEE*计算机视觉与模式识别会议论文集，*CVPR*，2016年。[19] Keunhyoung Kim、Jongpil Lee、Sangeun Kum 和 Juhan Nam，“利用结构保持三元组损失学习人声与混合音频的跨域嵌入空间”，载于国际音乐信息检索学会会议论文集，*ISMIR*，2021年。[20] Jeffrey Pennington、Richard Socher 和 Christopher D Manning，“GloVe：词表示全局向量”，载于自然语言处理实证方法会议论文集，*EMNLP*，2014年。[21]Carlos Busso、Murtaza Bulut、Chi-Chun Lee、Abe Kazemzadeh、Emily Mower、Samuel Kim、Jeannette N Chang、Sungbok Lee 和 Shrikanth S Narayanan，“IEMOCAP：交互式情绪双人动作捕捉数据库”，《语言资源与评估》，2008年。[22] Steven R Livingstone 和 Frank A Russo，“Ryerson情绪语音与歌曲音视频数据库（RAVDESS）：一套动态多模态的北美英语面部与声音表达数据”，《*PLOS ONE*》，2018年。[23] Taesu Kim、SeungHeon Doh、Gyunpyo Lee、Hyung seok Jun、Juhan Nam 和 Hyeon-Jeong Suk，“Hi, Kia：面向唤醒词的语音情绪识别数据集”，载于第*14*届亚太信号与信息处理协会年度峰会暨会议论文集（*APSIPA*），2022年。[24] Jort F Gemmeke、Daniel PW Ellis、Dylan Freedman、Aren Jansen、Wade Lawrence、R Channing Moore、Manoj Plakal 和 Marvin Ritter，“Audio Set：用于音频事件的本体与人工标注数据集”，载于国际声学、语音与信号处理会议论文集，*IEEE*，*ICASSP*，2017年。[25] Kalervo J¨arvelin 和 Jaana Kek¨al¨ainen，“基于累积增益的检索技术评估”，美国计算机协会信息系统汇刊（*TOIS*），2002年。[26] Leland McInnes、John Healy 和 James Melville，“UMAP：用于降维的统一流形近似与投影”，*arXiv*预印本 *arXiv:1802.03426*，2018年。