arXiv:2302.14757v1 [cs.MM] 28 Feb 2023

多模态设计文档音频检索：新数据集与算法

Prachi Singh[1,2] , *Srikrishna Karanam*[1],*Sumit Shekhar*[1]

[1]Adobe Research，印度班加罗尔[2]印度科学理工学院，印度班加罗尔prachisingh@iisc.ac.in

摘要

我们研究并提出一个新问题：如何根据多模态设计文档输入检索相关音频文件。这类输入包含文本元素和视觉图像，例如生日贺卡。除了提升用户体验，整合与这些输入的主题或风格相匹配的音频，也有助于提高此类文档的可访问性（例如，视障人士可以改为聆听音频）。

尽管近期已有音频检索相关研究，但这些方法和数据集明确面向自然图像。然而，我们研究的问题针对的是多模态设计文档（由用户使用创意软件创建），这类文档与自然拍摄的照片有着显著差异。为此，我们的首项贡献是收集并整理了一个名为 Melodic-Design（简称 MELON）的大规模新数据集，其中包含代表各种风格、主题、模板、插画等的设计文档，并为其配对音乐音频。基于这一图像-文本-音频配对数据集，我们的另一项贡献是提出了一种新颖的多模态交叉注意力音频检索（MMCAR）算法，使神经网络能够通过训练，在图像、文本和音频维度间学习一个共同的共享特征空间。我们利用这些学习到的特征，证明了所提方法优于现有最先进方法，并为研究社区在这一新数据集上提供了一个新的参考基准。

索引词— 音乐检索、多模态处理、交叉注意力。

1. 引言

随着按需使用的基于网页和移动端的图形设计软件日益普及[1, 2, 3]，，为各种场合设计创意文档变得十分容易，例如贺卡、活动邀请函/传单、社交媒体信息图等。在大多数情况下，这类设计文档往往具有多模态特性，即包含视觉图像和文本元素（见图1（右））。对于这类文档，添加相关音频/音乐文件这一额外模态，不仅能提升用户的使用体验，还能改善视障用户的文档可访问性。为此，我们的首项贡献是

提出并探讨一个新问题：给定一份多模态设计文档，检索相关音频文件。过去的许多研究 [1, 2]都聚焦于自然图像的音频检索，而在上述类型的设计文档领域，尚无相关研究。

本研究是在 Adobe Research 实习期间完成的。[1]

https://www.adobe.com/express[2]

https://www.canva.com[3]https://www.sketch.com

本文迈出了一步，旨在弥合文献中的这一空白（见图1）。



图1：自然图像（左）与设计文档（右）。

由于这一问题尚未得到探索，现有数据集 [3, 2] 用于音频检索的数据集仅包含自然图像，并不适用于我们提出的问题。为此，我们的第二项贡献是收集并整理了一个新的配对设计-音频数据集，其中包含从 Adobe Stock 公开数据中抓取的多模态设计文档[4]，以及从 MTG-Jamendo [4] 存储库中收集的相关音频文件。通过 $\approx 500k$ 将设计文档与超过 $\approx 7.5k$ 个音频文件配对，这是同类数据集中首个此类数据集，我们相信它将推动多模态设计理解领域的研究。

最后，我们的第三项贡献是一种新颖的多模态交叉注意力算法，使神经网络能够学习问题场景中图像、文本和音频模态之间的共享表示。具体而言，给定配对的设计-音频样本，我们提取各个模态的特征，并学习每对样本的权重以及整体权重，从而学习统一的设计-音频嵌入。通过在我们提出的新数据集上进行广泛实验，我们证明了我们的算法显著优于现有的最先进的音频检索方法。

2. 相关工作

如第 1 节所述，多模态音频检索领域的研究主要聚焦于自然图像。具体而言，在 Image2Song[3], 中，shuttersong数据集被用于将图像和歌曲歌词映射

[4]https://stock.adobe.com/

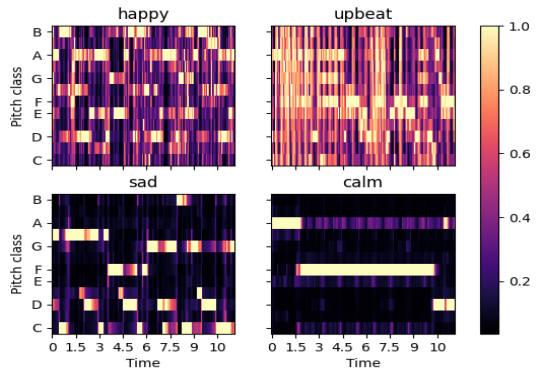


图 2：快乐、悲伤、欢快和平静等情绪下音频的时频色度特征对比。颜色（由深到浅）表示每个时频单元中的能量含量（由低到高）。

到同一特征空间中，随后用于检索等下游任务。Liang 等人使用同一数据集， [1]提出了一种通过融合视觉和声学特征来联合学习特征空间的方法。同样，诸如 VGGSound [2], 和 MUGEN [5], 这样的数据集虽然具有视频模态，但也主要聚焦于自然发生的人类动作或由游戏引擎生成的视频。相比之下，我们的工作通过提出一个全新的数据集，专门聚焦于贺卡、信息图等通常使用创意设计软件制作的多模态创意设计文档，具有独特性。我们的大规模数据集包含数十万份与音频文件配对的设计文档，为推进检索研究提供了具有挑战性的测试平台。

3. MELODIC DESIGN（MELON）- 一个新数据集

如第1节所述，对于图1中所示的多模态设计文档，我们要解决的问题是检索出一小组与输入中的各个元素相得益彰的音频文件。例如，图1 第一行第二列展示了一份以冒险为主题的设计文档，其中包含图像和文本。不同的元素包括背景图像/颜色、文本字段（例如“山脉”）、装饰元素和形状。请注意，这些元素各自构成设计文档中的一个图层（例如，背景图像是背景图层，而文字问候语是前景图层），从而形成多层的多模态设计文档。如第1节所述，并且从图1也可以看出，现有数据集仅关注自然图像，而我们要解决的问题涉及设计文档。为弥补文献中的这一明显空白，我们收集并整理了一个新数据集，其中包含多模态设计文档及其对应音频文件的配对样本，并将该数据集命名为“MELOdic desigN”（MELON）。

3.1. 收集原始数据集样本

我们使用公开可用的 MTG-Jamendo [4] 数据库作为音频文件来源，该数据库涵盖多种情绪／主题类别、乐器和流派。我们使用强度、音色、音高、速度和节奏等多种时频特征 [8] 来识别音频的情绪。例如，从“快乐”情绪转变为“悲伤”情绪时，音高会从很高变得很低。同样，“欢快”情绪的强度和速度都很高，而“平静”情绪的强度和速度则很低（见图 2）。为了将音频映射／关联到

数据集	视觉内容	模态	#图像	#音频	任务
VGGSound [2]	动作视频	I+A	199k	199k	C
Audio Set [6]	人类视频	I+ A	2.1m	2.1m	R + G
MUGEN [5]	游戏视频	I+A+T	233K	233K	R+G
Shuttersong [3]	图像	I+A+T	17k	17k	音乐检索 +C
IMEMNet [7]	图像	I+A	25k	1.8k	MR
MELON [提出]	多种	图像+音频+文本	488k	7.7k	音乐检索 + C

表 1：Melodic-Design 与其他数据集的对比。I、A 和 T 分别对应图像、音频和文本。Melodic-Design 中的“多种”涵盖插画、矢量图、模板和背景设计。C、R、G、MR 分别表示分类、检索、生成和音乐检索。

下文中的设计文档，我们使用 MTG-Jamendo 中对应 50 种情绪类别的音乐文件。由于 MTG-Jamendo 也包含标注了多个情绪类别的音频文件，为简化处理，我们仅保留只标注了一种情绪的数据样本。

我们使用来自 Adobe Stock 的公开可用数据作为来源，以收集由图像和文本内容组成的多模态设计文档。为抓取图像，我们构建了一个软件工具，可将任意情绪类别与数据类型一同作为输入来查询 Stock。例如，其中一个查询可以是获取冒险情绪的插画、矢量图、模板和背景图像。将搜索页面数量限制为 10 后，我们为每个情绪类别在上述所有文档类型中获得约 10,000 张图像。请注意，Adobe Stock 还提供图像元数据，其中包含对图像元素进行详细描述的人工生成说明文字。

3.2. 建立对应关系与数据集统计

在我们的数据集中，图像和文本已经配对，因为每个下载的图像都附带一个真实标注说明文字。我们使用 MTG 音乐数据集和我们提出的设计文档数据集中的共同情绪类别，构建图像-说明文字-音频配对。具体而言，给定一个情绪类别，我们首先提取图像-文本样本的 CLIP [9] 特征。对于每个标记有相同情绪的音频文件，我们提取 Wav2CLIP [10]嵌入，并计算图像-音频嵌入（记为 $s(i,a)$）与文本-音频嵌入（记为 $s(t,a)$）之间的余弦相似度。我们使用加权和 $\lambda_1 s(i,a) + \lambda_2 s(t,a)$，保留与相似度分数排名前 N 的结果相对应的音频文件。我们对每个情绪类别中的所有图像重复这一过程，以整理出我们的音频设计数据集。

我们提出的 MELON 数据集包含 488,510 张图像及其对应的说明文字，以及 7,737 首属于 50 种情绪/主题的音乐音频。每种情绪类别包含 $\approx 10k$ 张图像。补充材料中提供了各类别音频样本的分布图，以及按情绪分类的词云，以展示我们数据集的多样性和可变性[5]。在表 1中，我们将提出的 MELON 数据集与现有数据集进行了定量比较。可以看到，现有数据集主要关注自然图像和视频，而我们提出的数据集独具特色，包含创意插画、矢量图、模板和背景设计，并在说明文字中完整描述图像内容。凭借 $\approx 500k$ 张图像，我们的数据集将帮助该领域构建适用于音乐检索（MR）和分类（C）任务的稳健模型。

[5]补充文档链接：https://shorturl.at/hsDU5。我们计划在此链接公开相关材料。

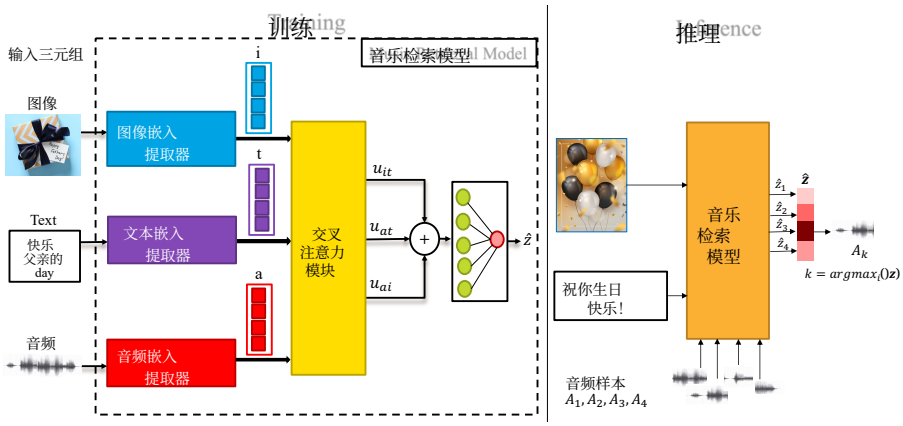


图 3：采用图像-文本对的所提音乐检索模型框图。

4. 多模态交叉注意力音频检索（MMCAR）

在此，我们介绍所提出的算法，该算法可根据输入的设计文档检索音频文件。我们在算法方面的主要创新，是一个多模态交叉注意力模块，它对来自全部三种输入模态（图像、文本和音频）的特征向量进行处理，并为下游检索任务学习一个共同的表示空间。

图3 直观地概述了我们提出的算法。在训练期间，给定来自 MELON 数据集的输入三元组，我们首先使用各模态的嵌入提取器计算特征向量**i**、**t**和**a**，分别对应图像、文本和音频模态。对于**i**和**t**，我们使用 CLIP [9] 模型分别获得 512 维嵌入。对于**a**，我们在公开可用的 VGGSound 数据集上训练 Resnet-18 模型（用于音频分类）。

给定**i**、**t**和**a** 嵌入，我们提出一种多模态交叉注意力操作，以学习统一的多模态设计-音频嵌入。给定三个特征向量，我们对任意两个模态$\mathbf{x}$ $\boldsymbol{y} \in \{\boldsymbol{i}, \boldsymbol{t}, \boldsymbol{a}\}$ 执行成对交叉注意力池化，使$\mathbf{x}$ $\in \mathcal{R}^d$ 作为查询，$\boldsymbol{y} \in \mathcal{R}^d$ 作为键，得到输出$\hat{x}$。类似地，通过交换$\mathbf{x}$ 和 $\boldsymbol{y}$，我们计算输出$\hat{y}$。随后使用这两个输出计算该特定模态对的公共嵌入$\mathbf{u}^{xy}$，其中 x 和 y分别表示这两个模态。我们对所有可能的模态对（$(x = i, y = t)$、$(x = i, y = a)$和$(x = t, y = a)$）重复此过程，并使用相应的输出得到所提出的统一嵌入$\mathbf{u}_{\text{all}}$，如下所示：

$$\begin{aligned} \boldsymbol{C}_{xy} &= \mathbf{x}\boldsymbol{y}^T \in \mathcal{R}^{dXd} \quad and \quad \boldsymbol{C}_{yx} = \boldsymbol{y}\mathbf{x}^T \\ \boldsymbol{S}_x &= \sigma(\boldsymbol{C}_{xy} * \boldsymbol{W} + \boldsymbol{B}) \in \mathcal{R}^{dXd} \\ \hat{\mathbf{x}} &= diag(\boldsymbol{S}_x \boldsymbol{C}_{xy}^T) \end{aligned} \tag{1}$$

$$\begin{aligned} \boldsymbol{S}_y &= \sigma(\boldsymbol{S}_{yx} * \boldsymbol{W} + \boldsymbol{B}) \in \mathcal{R}^{dXd} \\ \hat{\boldsymbol{y}} &= diag(\boldsymbol{S}_y \boldsymbol{C}_{yx}^T) \end{aligned} \tag{2}$$

$$\begin{aligned} \boldsymbol{u}_{xy} &= \hat{\mathbf{x}} \oplus \hat{\boldsymbol{y}} \in \mathcal{R}^{2dX1} \\ \boldsymbol{u}_{all} &= \boldsymbol{u}_{it} \oplus \boldsymbol{u}_{ia} \oplus \boldsymbol{u}_{ta} \in \mathcal{R}^{6dX1} \end{aligned} \tag{3}$$

随后将这一统一嵌入$\mathbf{u}_{\text{all}}$ 传入一个全连接神经网络单元，该单元通过 Sigmoid 操作生成一个范围为 $\in [0, 1]$ 的标量分数$\hat{z}$。我们将其与真实分数 $z = 1$（如果输入是正确的配对）进行比较，否则与 $z = 0$ 比较，从而得到一个二元交叉熵训练目标

冒险 广告 fun game 假日 恐怖 太空 运动 欢快	商业 企业 环境的 calm 放松 soft 醇和	戏剧 戏剧性的 电影 梦境 情感的 重 忧郁 sad	滑稽 时髦的 快乐 充满希望的 励志 旋律优美 儿童 圣诞节	love 浪漫的 自然 夏季 复古 背景

表 2：用于训练和评估的情绪/主题列表。

$$L = \frac{1}{B}\sum_{i=1}^{B} z\log(\hat{z}) + (1 - z)\log(1 - \hat{z}) \tag{4}$$

其中 B 为批量大小。在推理过程中，给定一个设计文档作为输入，以及一个用于检索相关文件的音频样本数据库，n 我们的模型会计算输入图文对与所有 $\hat{z}_1$、$\hat{z}_2$.....、$\hat{z}_n$ 个 n 音频样本之间的相似度分数。根据这些分数，我们选取分数最高的前 k 个音频文件作为检索结果（见图3（右））。

5. 实验与结果

由于我们提出的问题、数据和算法均围绕基于图文设计文档的音频检索，文献中最接近的基线方法包括 JTAV [1] 和 Wav2CLIP [10]。JTAV 基于图像、其标题说明以及音频的文本歌词来学习特征，而 Wav2CLIP 则使用 CLIP [9]图像编码器和音频自编码器，使音频和图像特征相互接近。

为评测这些算法的性能，并在我们的新数据集上将其与所提出的方法进行比较，我们采用了先前工作 [1] 提出的基于排名的评估指标。具体而言，我们使用 $Med\ r \in [1, M]$ 指标，它表示检索到的真实标注音频的中位排名，其中 M是所考虑类别的最大数量。较低的 $Med\ r$ 值表示更好的性能。我们还使用 Recall@k（$k = \{1, 5, 10, 15, 20\}$），即在所有测试样例中，排名前 k 的项目里检索到的真实标注音频所占的比例，较高的值表示更好的性能。

为训练和评估所有模型，我们根据全部 50 种情绪中的最大独特性，从数据集中选择了 38 种情绪（列表见表2）。我们使用 60% 的图像-描述对和音频样本构建训练集划分。每个三元组由图像、描述和音频组成，并附带一个 1/0 基于正确音频映射的标签。为构建正样本三元组，我们选择 10 个音频

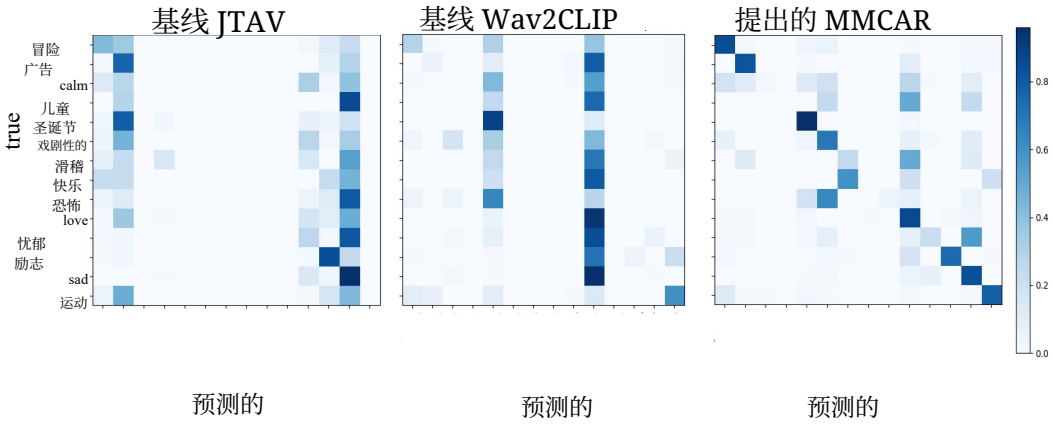


图 4：提出的 MMCAR 与基线 JTAV 和 Wav2CLIP 的混淆矩阵。

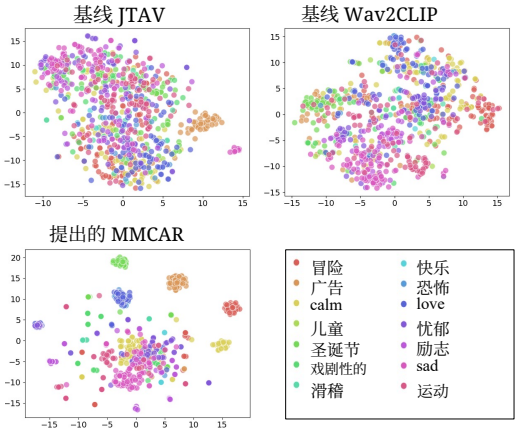


图 5：基线方法与所提出方法的 t-SNE 图。

样本，每个设计依据第3.2节所述的情绪选取。负样本三元组通过从图像情绪中随机选择5 种不同情绪，并从每种情绪中选取 2 个音频样本来构成。其余 40% 的数据被均等划分为验证集和测试集，其中每个集合包含20% 的样本。

接下来，我们展示评估结果。由于我们的目标是检索情绪最匹配的音频文件，因此首先从音频库中为每种情绪计算一个平均特征向量。然后，给定输入设计文档的特征向量，我们生成一个分数向量$\hat{}$ $\boldsymbol{z} \in \mathcal{R}^M$，其中 $M = 38$ 如上所述，对应于所有音频文件的平均特征。随后，我们使用参考情绪标签计算 recall@k 和 $Med\ r$ 指标。

在表 3中，我们比较了所提出的 MM-CAR 算法与 JTAV 和 Wav2CLIP 基线方法在验证集和测试集上的表现。可以看到，我们提出的 MMCAR 在所有排序中均取得了最低的 $Med\ r$ 3.9，且召回准确率最高。例如，在测试集中，79% 的准确率（k=5 ），而 JTAV 为 30% ，Wav2CLIP 为 39%，相对提升超过 100%。为进行更加公平的比较，我们还使用 Wav2CLIP 的特征重新实现了 MMCAR（表中记为 MMCAR*）。与原始的端到端学习 MMCAR 模型相比，MMCAR* 的性能有所下降，但仍显著优于基线 Wav2CLIP 方法。这一结果

方法	$Med\ r$	Recall@k				
		k=1	k=5	k=10	k=15	k=20
Val						
JTAV [1]	17.1	0.12	0.30	0.35	0.46	0.62
Wav2CLIP [10]	9.2	0.12	0.37	0.66	0.83	0.92
MMCAR [我们的方法]	**3.8**	**0.42**	**0.80**	**0.92**	**0.96**	**0.98**
MMCAR*	7.0	0.23	0.59	0.77	0.87	0.93
Test						
JTAV [1]	16.9	0.12	0.30	0.35	0.45	0.62
Wav2CLIP [10]	9.2	0.11	0.39	0.67	0.83	0.91
MMCAR [我们的方法]	**3.9**	**0.42**	**0.79**	**0.92**	**0.96**	**0.98**
MMCAR*	7.2	0.22	0.56	0.76	0.89	0.93

表3：提出的 MMCAR 与基线方法的比较。

为我们的多模态交叉注意力模块在共享图像-文本-音频特征空间中的判别能力提供了证据。

为提供更多证据，我们在图中展示了学习得到的设计文档嵌入的 t-SNE 可视化结果5 。可以看出，与基线方法相比，使用提出的 MMCAR 方法时，特征能够根据情绪更清晰地聚类。在图中4，我们将 MMCAR 的混淆矩阵与随机选取的 14 种情绪对应的基线方法的混淆矩阵进行了比较。可以发现，基线方法的预测偏向少数几种特定情绪，而提出的方法则如预期般呈现出接近对角线的矩阵。

6. 总结

我们探讨并提出了一个新问题：给定多模态设计文档作为输入，检索相关音频文件。由于文献中没有相关数据集，我们构建并介绍了首个此类多模态设计-音频数据集 MELON，其中包含数十万个设计文件及其对应的音频文件。随后，我们提出了一种多模态交叉注意力算法，使神经网络能够学习适用于设计文档的联合图像-文本-音频特征空间，并在测试时根据给定的设计输入检索相关音频。我们在新数据集上将该算法与现有最先进方法进行了基准测试，并希望这能推动该领域的进一步研究。

7. 致谢

作者谨向印度科学理工学院班加罗尔分校 LEAP 实验室的 Sriram Ganapathy 博士致以诚挚谢意，感谢他提供宝贵的意见和帮助，并提供运行实验所需的资源。

8. 参考文献

[1] Hongru Liang、Haozheng Wang、Jun Wang、Shaodi You、Zhe Sun、Jin-Mao Wei 和 Zhenglu Yang，“JTAV：通过融合文本、音频和视觉特征联合学习社交媒体内容表征”，载于第*27*届国际计算语言学会议论文集，2018年，第1269–1280页。[2] Honglie Chen、Weidi Xie、Andrea Vedaldi 和 Andrew Zis- serman，“Vggsound：大规模音视数据集”，载于*IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2020年，第721–725页。[3] Xuelong Li、Di Hu 和 Xiaoqiang Lu，“Image2Song：通过连接图像内容与歌词词语进行歌曲检索”，载于*IEEE*国际计算机视觉会议论文集，2017年，第5649–5658页。[4]Dmitry Bogdanov、Minz Won、Philip Tovstogan、Alastair Porter 和 Xavier Serra，“用于自动音乐标签生成的 MTG-Jamendo 数据集”，载于*ICML*，2019年。[5] Thomas Hayes、Songyang Zhang、Xi Yin、Guan Pang、Sasha Sheng、Harry Yang、Songwei Ge、Isabelle Hu 和 Devi Parikh，“MUGEN：用于视频、音频和文本多模态理解与生成的实验平台”，*arXiv* 预印本 *arXiv:2204.08058*，2022年。[6] Jort F Gemmeke、Daniel PW Ellis、Dylan Freedman、Aren Jansen、Wade Lawrence、R Channing Moore、Manoj Plakal 和 Marvin Ritter，“Audio Set：音频事件的本体与人工标注数据集”，载于*2017 IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2017年，第776–780页。[7] Sicheng Zhao、Yaxian Li、Xingxu Yao、Weizhi Nie、Pengfei Xu、Jufeng Yang 和 Kurt Keutzer，“效价-唤醒空间中基于情绪的图像与音乐端到端匹配”，载于第*28*届 *ACM* 国际多媒体会议论文集，2020年，第2945–2954页。[8] Aathreya S Bhat、VS Amith、Namrata S Prasad 和 D Murali Mohan，“一种通过音频特征提取对西方音乐和印地语音乐进行音乐情绪检测的高效分类算法”，载于第五届国际信号与图像处理会议。IEEE，2014年，第359–364页。[9] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等，“从自然语言监督中学习可迁移的视觉模型”，载于国际机器学习会议。PMLR，2021年，第8748–8763页。

[10]Ho-Hsiang Wu、Prem Seetharaman、Kundan Kumar 和 Juan Pablo Bello，“Wav2clip：从 CLIP 学习稳健的音频表征”，载于*IEEE* 国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2022年，第 4563–4567 页。